Automation of the Landing Error Scoring System using Inertial Measurement Units

A thesis

submitted in partial fulfilment

of the requirements for the degree

of

Master of Science (Research) in Computer Science

at

The University of Waikato

by

ZANE HAMILTON

THE UNIVERSITY OF
WAIKATO
Te Whare Wānanga o Waikato

2024

Abstract

The quantification of biomechanics is important for making informed decisions when implementing appropriate training interventions for improving performance and reducing injury risk. This study evaluated the viability of movement data provided by wearable inertial measurement units (IMUs) to automate the scoring of the Landing Error Scoring System (LESS). The LESS is an assessment tool used for identifying high-risk movement patterns in a double leg jump landing; however, the LESS is scored by experts using 2D video recordings, which limits large-scale screening. Movement data provided by three IMUs were used to train several out-of-the-box machine learning models, aiming to predict the result of the 17 LESS scoring items individually. Raw movement data was processed and segmented into the key phases of the movement, where additional features were derived from these segments. A supervised learning approach was taken, using a dataset containing 218 LESS scores derived from 40 participants as the desired output for each model. Comparisons were made between various subsets of features, and the features with the greatest importance on accuracy in the best performing models were extracted. Results showed limited improvements to a ZeroR approach, and features with the greatest importance on the best performing models had minimal relevance to the movement involved. Performance of LESS scoring automation using IMU data may be improved with further developments on this approach.

Acknowledgements

The completion of this Masters thesis was not possible without the support and compassion from a number of people. Firstly, I'd like to thank Dr. Ivana Hanzlíková for guiding me through the scoring process and using Kinovea for manual scoring. Additionally, I'd like to thank Markéta Zvonková for providing the LESS scores for the entire dataset. I'd also like to extend a massive thank you to Dr. Nick Lim for the open communication, guidance, and willingness to help support the implementation of the machine learning architecture and evaluation of relevant performance metrics.

I would like to acknowledge the University of Waikato for funding this Masters thesis through the generous Research Masters Scholarship, and inspiring my journey into research with the Summer Research Scholarship Programme. The support provided by the University has been integral to my academic experience.

A personal thank you to the friends who have supported me throughout the last two years. I have been endlessly supported by the people around me, which has allowed me to maintain a good life balance throughout my time of study.

I would like to express my utmost gratitude and appreciation to my three supervisors who have provided endless amounts of support and guidance throughout my academic study. Firstly, to Dr. Jemma König; thank you for allowing me to pursue this avenue of research over the last two years, with your endless support and unwavering positivity throughout every step of the journey. Next, I'd like to personally thank Dr. Kim Hébert-Losier, for taking the time to meet with me all the way back in 2020 in the early stages of my academic pathway.

Without your support, teachings, and guidance toward this inter-disciplinary area of research, none of this truly would have been possible. Your passion towards biomechanics and research is inspiring and has allowed me to apply myself in this area with confidence. Lastly, to Dr. Jessica Turner. I am forever indebted to you for putting your confidence in me over these last few years. Without you, the means to study in my hometown of Tauranga would not have been possible. You have been an anchor of support in all aspects of my career, and I am forever grateful for the pathway you have provided me to my future in computer science.

Lastly, I would like to take this moment to thank my family. To my brothers Rhys and Kyle, for their support and understanding in my pathway and providing the foundations for me to pursue my study. To my grandparents who passed away over these last few years of study; my Poppa and Nana Rose, the positivity and happiness you instilled in me for life and allowing me to appreciate the life around me, I am eternally grateful. To my Nana Barbara, thank you for our long in-depth chats about life, your solutions to the problems I have faced, and the confidence you have provided me as I have continued into my career.

To my parents, there are no words I can say that encapsulate the love and appreciation I have for you, especially in these last two years. To my dad Paul, thank you for your support, instilling the confidence in me to continue through my academic career and always being ready with a solution when things didn't go right. To my mum Raewyn, you were there around every corner, with a smile on your face and always ready to listen. Your contribution to my life will never be unnoticed, and I am eternally grateful to you both for making me the man I am today.

Contents

List of Figures

List of Tables

List of Code Listings

List of Abbreviations

3D

Three-Dimensional.

ACL

Anterior Cruciate Ligament.

ANN

Artificial Neural Network.

BMI

Body Mass Index.

CSV

Comma-Separated Value.

DFF

Deep Feedforward Neural Network.

DLJL

Double Leg Jump Landing.

GBC

Gradient Boosting Classifier.

GNB

Gaussian Naive Bayes Classifier.

IMU

Inertial Measurement Unit.

KNN

K-Nearest Neighbours.

LESS

Landing Error Scoring System.

OMC

Optical Motion Capture.

RFC

Random Forest Classifier.

S1

Segment of Data from Take-off to Initial Contact.

S1A2

Segment of Data from Take-off to Maximum Knee Flexion.

S2

Segment of Data from Initial Contact to Maximum Knee Flexion.

SGD

Stochastic Gradient Descent Classifier.

SHAP

Shapley Additive Explanation.

SVC

Support Vector Classifier.

XAI

Explainable Artificial Intelligence.

Chapter 1

Introduction

Human movement can be performed in many ways and under several environments, including at different speeds, on various surfaces, and under the influence of external factors, such as resistance. The performance of a movement is made up of many smaller-scale musculoskeletal interactions, referred to as biomechanics. Understanding and quantifying the biomechanics involved in a movement is helpful for evaluating performance and implementing appropriate training interventions to prevent injury, rehabilitate, or improve performance. The Landing Error Scoring System (LESS) is an example of a movement screening tool, aimed at highlighting high-risk biomechanics susceptible to cause injury in the lower extremity when performing a jump landing task.

1.1 Motivation and Scope

Lower extremity injuries are the most common injury site across most sports, with injury to the anterior cruciate ligament (ACL) being one of the most common knee-related injuries worldwide [3]. Addressing high-risk movement patterns that increase the likelihood of ACL injury in jump landings allows individuals to follow injury prevention programmes prior to vulnerable situations to mitigate injury risk.

Within the LESS, 17 different scoring items are scored manually by a trained clinician, performed by visually observing the jump landing task from 2D video

recordings and indicating whether an error is present or absent. With the need for manual intervention on every jump scored, the use of the LESS for large-scale testing is not applicable. The implementation of an automated scoring process may provide greater accessibility to the LESS, without the need for manual intervention.

To this point, wearable sensors, more specifically Inertial Measurement Units (IMUs), have been widely used in recent years for assessing the performance of movement. Three-dimensional (3D) optical motion capture (OMC) is widely considered as the "gold standard" of movement analysis [4, 5, 6], but are not practical for LESS assessment and scoring. Movement data provided by sensors has been integrated with 3D movement modelling to quantify movement, but these systems are highly technical, expensive, and require expert users. A more accessible option is to integrate the sensor data with machine learning models to automate the scoring process for the LESS.

1.2 Project Aims

We hypothesise that IMUs can be used to predict a LESS score using machine learning models. To evaluate this hypothesis, an investigation into the applicability of sensor data for scoring the high-risk movement patterns is required. In addition, the processing and implementation of sensor movement data and evaluating how it affects the performance of the machine learning models in scoring of the LESS are necessary. Therefore, we propose the following research questions:

1. How accurately can out-of-the-box machine learning models automate LESS scores from IMU data?
2. Which IMU-derived features are the most relevant to the automation of LESS scores?

1.3 Structure of Thesis

This thesis is structured as follows: Chapter 2 highlights how LESS scores are collected, and further motivates the potential use of wearable sensors to auto-

mate LESS scores. Chapter 3 covers related work using IMUs for movement quantification, and the further integration of IMU-derived data with machine learning models for classification. Chapter 4 outlines the solution design used for processing and automating the scoring process with the collected IMU data. Chapter 5 demonstrates the implementation of the solution design, specifically the algorithms created for processing sensor data into training data. Chapter 6 presents an evaluation on the accuracy of LESS score classification using the collected data, and how IMU-derived features influence this accuracy. Chapter 7 summarises the main findings in relation to the research questions posed. Finally, Chapter 8 concludes the thesis and highlights future work that can be completed based on the findings provided.

Chapter 2

Background

This thesis introduces an approach for automated quantification of the LESS using an integration of sensor data with machine learning. In this chapter, the motivations underpinning the use of the LESS are established. This chapter starts with a summary on the importance of biomechanics, and how human movement is quantified. Following this summary, the significance of lower extremity injury is highlighted, and the reasons why movement screening for vulnerable movement patterns is relevant. Subsequently, the LESS and related scoring protocol are discussed. Lastly, the use of wearable sensors for movement tracking is highlighted, including the types of sensors commonly used to capture biomechanics, and their specific applications in relation to the double leg jump landing (DLJL) task employed as part of the LESS.

2.1 Biomechanics of Movement

Human biomechanics refers to the evaluation of the structure and function of a living body performing a movement, including the interaction between musculoskeletal segments and the effect of forces on these segments [7]. This evaluation of movement typically occurs under two main contexts - clinic and sport.

The clinical assessment of movement focuses on guiding interventions for rehabilitating movement post-injury or preventing injury incidence. The data

provided from these assessments influences training interventions required to elicit changes for rehabilitation and reducing tendencies toward vulnerable movement patterns. Sports movement assessment typically seeks to optimise performance for competitive or high-performance needs. Sports-related interventions on movement patterns intend to maximise performance and optimise a given movement. There are overlaps between the clinical and sport applications of biomechanical assessments. For instance, screening for risk of injury is common practice in sports and applies to the clinical context.

The quantification of human movement is individualised depending on the specific data of interest and the movement sought to be improved. Using a 100-metre sprint as an example, a slower acceleration in the starting phase of the sprint motion could be a focus point for quantification in a sports-performance context. To quantify this starting phase, acceleration data could be captured from a participant and training programmes could be implemented that prioritise changes for improving strength and performance in this sprint phase and movement [8]. In a clinical context, the biomechanical interactions between muscles and joints could be monitored to identify potential injury risks during the acceleration phase, for instance. This monitoring could include using technology such as an optical motion capture system to monitor joint movement, force plates to measure magnitude and rate of force, or electromyography to quantify muscle activation patterns.

Understanding the biomechanics of a participant is important for implementing relevant training interventions. Whether it is for improving performance, reducing injury risk in vulnerable movement patterns, or both - movement data is useful for making informed decisions with regards to exercise prescription. Injuries occur not only due to high-intensity activities in sports, but also during low-to-moderate intensity activities [9]. Injuries do not only involve rehabilitation costs, but they can also lead to decreased physical activity levels [10] and long-term impairments like osteoarthritis [11, 12]. Additionally, injury may inhibit function and activities of daily living. Furthermore, when the lower extremities are involved, injuries can impact general human locomotion, such as gait cycle characteristics [13].

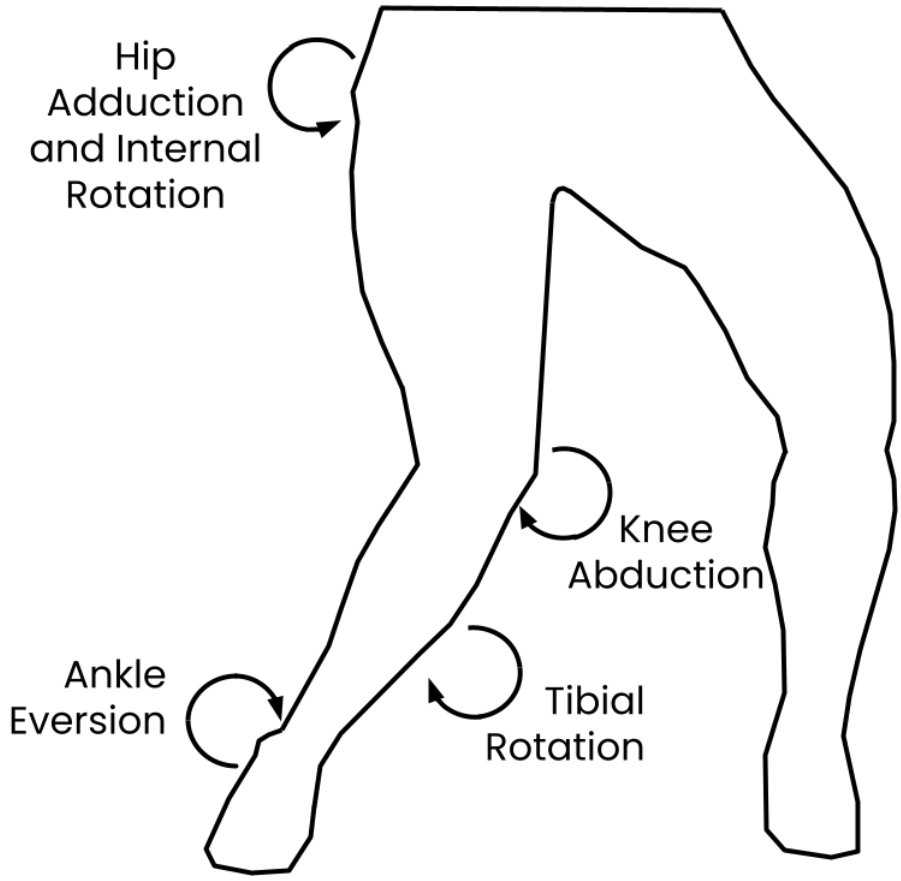


Figure 2.1: Example of the Dynamic Knee Valgus Position

2.2 Lower Extremity Injury Risk

Lower extremity injuries are the most common injuries observed across sports, making up over 50% of injuries [14]. A steady increase has been observed in sport-related injuries by the Accident Compensation Corporation (ACC) in New Zealand, with a total of 452,438 new injury claims in 2023 [15]. Since 2015, the cost of these active claims has increased to over $326 million dollars [15]. ACC claims relating to lower extremity injuries from 2012-2016 were most often moderate to serious injuries across five sports, with 19,026 of these lower extremity claims (53.7%) specifically for knee-related injuries [16]. ACL injuries are one of the most common knee-related injuries worldwide annually [3], with a reported 400 tears per 100,000 people in adolescents as recently as 2020 [17].

The ACL is essential in stabilising the knee, with primary function to restrain the anterior translation of the tibia on the femur [18]. ACL injuries occur most commonly in sporting environments, predominantly in non-contact conditions [19]. In previous studies, these injuries are most commonly associated with movements such as sudden changes in direction, reactive phases of play and quick or sudden deceleration [19, 20, 21]. Landing from a jump is considered one of the most vulnerable movement patterns for ACL injury that can occur in game situations, particularly when coupled with knee valgus [21, 22]. Dynamic knee valgus, as seen in Figure 2.1, describes a combination of motions in the lower extremity where the knee is displaced medially. In this position, excessive movement in the coronal and transverse plane are observed - combining hip adduction, hip internal rotation and knee abduction; often including tibial rotation and ankle eversion [23, 24]. This position places significant strain on the ACL [21] and is largely considered a primary mechanism of ACL injury [21, 25, 26].

There are numerous additional internal and external risk factors of ACL injuries, but addressing high-risk biomechanical movement patterns prior to vulnerable situations reduces the likelihood of injury [27, 28]. Specific interventions and feedback provide participants with knowledge on the correct and safest movement patterns, allowing greater neuromuscular control of a movement and further optimising a participant's understanding and ability to manoeuvre safely [29]. To understand the risks in different movement patterns, appropriate movement screening must be completed that can identify high-risk tendencies or weaker movement competency in participants, to then apply the relevant training interventions for improvement [25].

2.3 Landing Error Scoring System

As seen in Figure 2.2, the LESS is an assessment tool used for identifying high-risk movement patterns in a DLJL task. An examiner uses a scoring system where they observe various kinematic movement patterns that are linked to ACL injuries from 2D video recordings of a DLJL task [25]. The set up required for collecting the video for the LESS is presented in Figure 2.3. Participants are required to stand on a 30-cm high box and jump horizontally to 50% of their

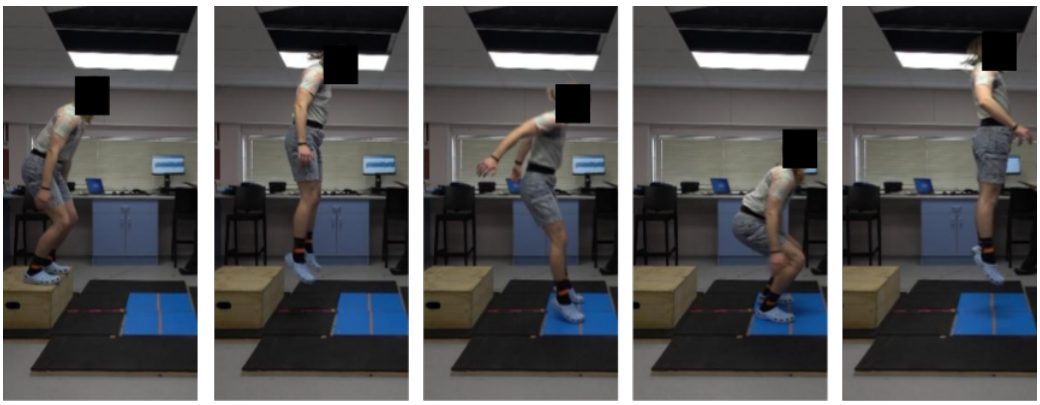


Figure 2.2: Standardised Double Leg Jump Landing Task

body height and upon landing, jump vertically for maximal height [1]. The task is repeated three times and the average score is recorded as the outcome.

Each jump is scored based on 17 different scoring items, as seen in Table 2.1. The first 15 items are classified as either a "1" when an error is present, or a "0" if an error is absent. Both items 16 and 17 are more subjective in nature, using a three-point scale where "0" indicates an excellent landing, "1" an average landing, and "2" a poor landing. The maximum possible score is 17 errors, as some items are directly linked and are not concurrently possible. Specifically, these items are items 7 and 8 where a participant may either have a wide or narrow stance, but not both; and items 9 and 10 where the foot position may be overly inward or outward, but not both. A lower overall LESS score indicates fewer movement errors and patterns linked to ACL injury; therefore, a lower risk of non-contact ACL injury and a better jump landing technique [1].

2.4 Scoring Protocol for the LESS

Each jump using the LESS protocol is performed in front of two standard video cameras, with one placed on the side to capture sagittal plane motion and one directly in front of the participant to record frontal plane motion [4], as seen in Figure 2.4. Movement in the sagittal plane refers to any forward and backward movement or flexion and extension movement of the joints. Frontal plane motion represents movement to the left and right, inward and outward,

Table 2.1: The Landing Error Scoring System Items, Adapted From [1]

	LESS Scoring Item	Definition of Error
1	Knee flexion at IC	Knee flexion less than 30°
2	Hip flexion at IC	Thigh is in line with the trunk (hips not flexed)
3	Trunk flexion at IC	Trunk is vertical or extended at the hips (i.e., not flexed)
4	Ankle plantar flexion at IC	Heel-to-toe or flat foot landing at IC
5	Knee valgus at IC	Centre of the patella is medial to the midfoot at IC
6	Lateral trunk flexion at IC	Midline of the trunk is flexed to the left or the right side of the body at IC
7	Stance width (wide)	Feet are positioned greater than shoulder width apart at IC
8	Stance width (narrow)	Feet are positioned less than shoulder width apart at IC
9	Foot position (toe-in)	Foot is externally rotated more than 30° between IC and MKF
10	Foot position (toe-out)	Foot is internally rotated more than 30° between IC and MKF
11	Symmetric foot contact at IC	One foot lands before the other foot or one foot lands heel to toe and the other foot lands toe to heel
12	Knee flexion displacement	Knee flexes less than 45° between IC and MKF
13	Hip flexion at MKF	Thigh does not flex more on the trunk between IC and MKF
14	Trunk flexion at MKF	Trunk does not flex more between IC and MKF
15	Knee valgus displacement	At the point of maximum medial knee position, the centre of the patella is medial to the midfoot
16	Joint displacement	Soft (0), Average (1), Stiff (2)
17	Overall impression	Excellent (0), Average (1), Poor (2)

* Note: IC: Initial Contact; MKF: Maximum Knee Flexion

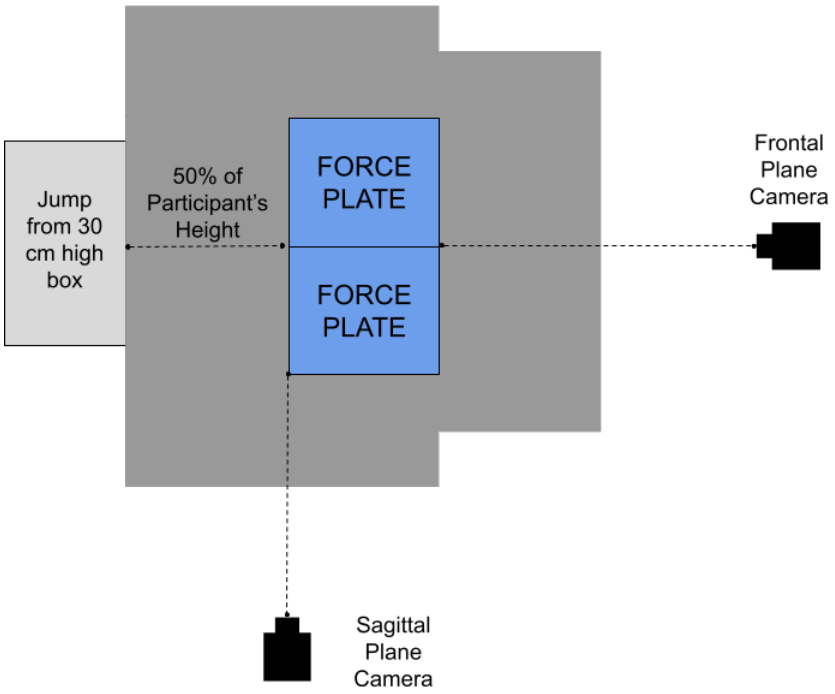


Figure 2.3: Double Leg Jump Landing Set Up

or abduction and adduction of the joints. Each respective item of the LESS is associated with movement in either the sagittal or frontal plane. For example, knee flexion occurs in the sagittal plane of motion; therefore, the side camera is used to visualise and evaluate the knee angle (between the thigh (femur) and shank (tibia) bones). Standard LESS evaluation can be performed on recorded motion as low as 30 Hz [30], although a refresh rate as high as 120 Hz has also been used [31].

The LESS has been shown to be a valid and reliable assessment for detecting poor jump landing biomechanics [1, 25]. Despite some LESS items being more subjective, good to excellent inter-rater and intra-rater reliabilities have been observed in LESS scoring [25], including LESS scoring performed by two different scorers with similar experience levels [32]. When comparing LESS results derived from 2D video to 3D motion capture, the validity of the LESS was concluded as item dependent [33]. Poor to moderate agreements between a number of items were mainly due to difficulty estimating segment angles and symmetric contact visually from 2D videos [25, 33]. A 1° difference in man-

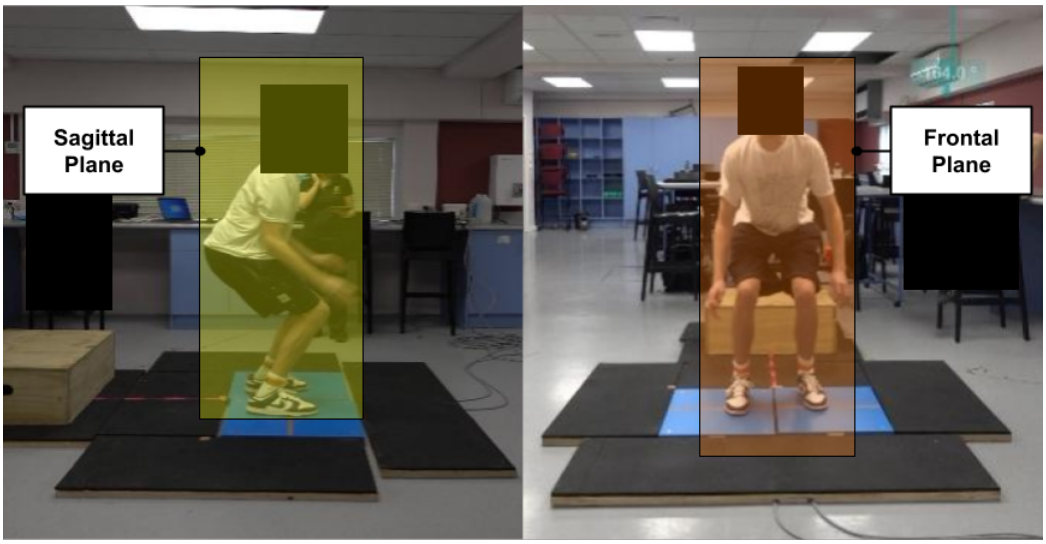


Figure 2.4: Planes of Motion Captured for the Landing Error Scoring System

ual video analysis software can be the borderline between an error present or absent [25].

In laboratory-based movement capture, 3D motion analysis systems using infra-red and retro-reflective markers have typically been considered the "gold standard" measurement for human motion [4, 5]. These systems use two or more cameras to record the position of physical markers in a 3D space [34]. However, these systems are expensive and require the correct preparation, a dedicated laboratory space, and an expert-user to operate the system and perform analysis [4]. The LESS is a "field-based" testing method, meaning it should be accessible. For example, 2D video recordings and cameras such as smartphones can be used to capture the jump landing task [4]. The LESS can also be completed on many landing surfaces, with no notable differences in LESS scores between court, grass, and laboratory surfaces [35].

There are additional factors that may affect the scores observed in the LESS, such as sex, where females demonstrate statistically significant higher overall LESS scores when compared to males [29, 36]. Injury intervention programmes, and previous injury history have also shown an influence on LESS performance [29, 37, 38]. Lastly, age has also been theorised to influence LESS scores [39], where younger individuals reported higher overall scores than older individuals.

Depending on the sample of individuals performing the DLJL, differences in LESS performance may be attributed to these factors. Additionally, some of the scoring items assessed in the LESS appear less commonly than others across datasets, or not at all - such as item 2, showing a frequency of 1% or less in large LESS datasets [1, 40].

In large-scale testing, the LESS has drawbacks. Scoring of the LESS takes time and an expert is required to manually review and score each video individually [4]. Automated quantification of LESS scores has been successful in research, including the use of markerless motion capture technology demonstrating similar levels of reliability as expert LESS scorers [41]. However, this approach was unable to automate the overall impression (LESS item 17). Additionally, access to these systems is limited, and the additional expense makes large-scale testing for the general population infeasible. Factors such as access to the correct laboratory and technology, and the need for an expert to operate the motion capture system correctly, need to be weighed against the positives of this automation [4].

Computer vision technology has also been applied to the LESS, using standard 2D video recordings in conjunction with OpenPose, an open-source system for detecting pose in 2D [42], to identify key frames of the movement and automate a participants score [4]. The use of this system facilitates large-scale testing, removes the need for an expert scorer to manually score each jump and introduces the possibility of using smartphone-based video footage [4]. Although this system allows far greater accessibility for the general population, two cameras are still required to automate a LESS score, meaning participants must ensure they have positioned the cameras in the correct position to allow the computer vision to track movement accurately. Furthermore, the accuracy of this automation method requires improvements to strengthen the agreement with the clinical method ($r = 0.63$). With the increasing popularity in accessible wearable technology for movement quantification, we propose the need for cameras to score the LESS may no longer be required.

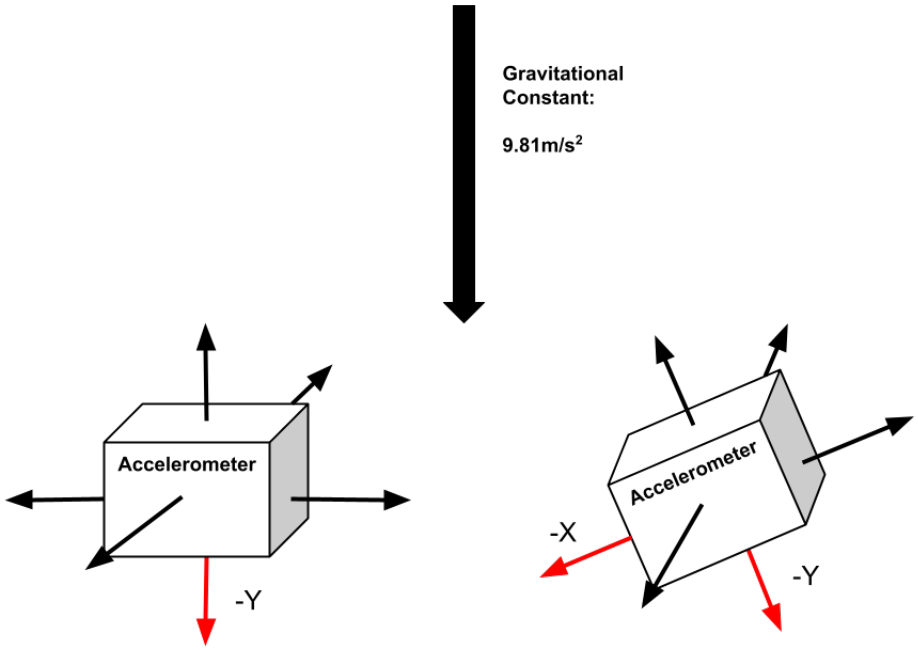


Figure 2.5: The Impact of the Gravitational Constant on an Accelerometer

2.5 Wearable Sensors for Movement Tracking

Wearable sensors have been shown to be a valid tool for tracking and assessing movement patterns [43, 44]. Wearables have been used for quantification in a range of contexts, including identifying key gait parameters [45], quantifying sports-based movements such as baseball pitching and batting [46], and even extracting different swimming styles from movement data [47]. Analysis of human movement allows practitioners to understand and describe the motor function of a participant in both a clinical and sports-based performance context [44]. There are three main types of wearable sensors that are used to describe and quantify human movement - an accelerometer, gyroscope and magnetometer [44, 48].

Accelerometers capture linear accelerations (m/s^2) of a body segment at the point where the sensor is attached [49]. There are two types of data recorded in the raw acceleration signal: gravity and physical movement. The influence of gravity acting on the body is included in accelerometer readings [50], with the gravitational acceleration constant being roughly 9.81 m/s^2 [51, 52] or 1 g.

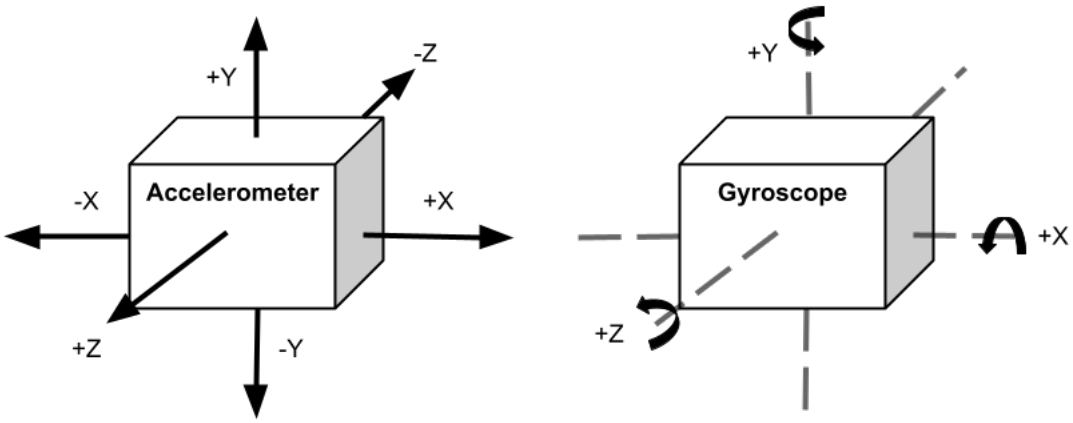


Figure 2.6: How Movement Data is Represented by Triaxial Accelerometer and Gyroscope

In addition to gravity, the acceleration generated from sudden movements or vibration of the sensor is also collected, referring to accelerations brought on by physical movement of the sensor by the wearer [50]. Triaxial accelerometers are used commonly in wearable sensor systems, measuring acceleration in three dimensions [44]. Each axis of a triaxial accelerometer provides detailed information of how the sensor is moving [53]. Changes in the raw accelerometer data can be caused by the sensor being accelerated, or the orientation of the sensor changing, meaning the influence of gravity is applied to other axes. As seen in Figure 2.5, the gravitational constant would only affect the y-axis (alone) in a perfectly aligned sensor; whereas a slight rotation of the IMU can impact two or all three axes unevenly. Accelerometers are typically applied to, and more easily interpreted from linear-based movement, such as detecting different phases of the gait cycle from accelerometer-derived data [54], or monitoring the acceleration of a barbell through a high-pull motion [55].

Gyroscopes measure the angular velocity occurring around the point where a sensor is attached [56]. Angular velocity is typically captured in degrees per second (°/s), but some devices produce revolutions per second. Triaxial gyroscopes are also used commonly in human movement analysis [44], but data given by a gyroscope represents the rotation of the sensor [56], as seen in Figure 2.6. Gyroscopes are beneficial in capturing data in rotational based movements, such as assessing heel-strike and toe-off based on changes in the

shank during gait movement [57] or recognising basic turns and motions in snow sports-related movements from gyroscopes placed on a snowboard [58].

Lastly, a magnetometer provides information on the magnetic field in relation to the placement of a sensor [44]. Similar to both an accelerometer and a gyroscope, magnetometers in wearable sensors typically use a three-axis system to capture this information [44]. Changes in the data provided by a magnetometer relates to the orientation of the sensor in respect to Earth's magnetic field [59]. Wearable sensors can be subject to "drift," where the accuracy of data provided by sensor systems may be affected due to the integration of noise in the sensor signals [1, 60, 61]. With the inclusion of magnetometers in modern sensor units, the data provided by a triaxial magnetometer can be used to correct sensor drift [60]. Despite this correction, magnetometers require consistent recalibration as the data may be disrupted by magnetically different environments [62], such as collecting data indoors where magnetic wall and floor structures may disrupt the Earth's magnetic field [60].

Each sensor type has their unique uses for movement tracking, and may either be used individually or in combination with each other for a more comprehensive movement dataset [63]. For example, placing an accelerometer and gyroscope on the knee could monitor both the acceleration of segments around the knee as well as angular velocities of these segments during jump landing. Using these sensors in combination is typically known as an Inertial Measurement Unit (IMU). An IMU is a sensor-embedded system with an accelerometer and gyroscope, and sometimes a magnetometer [44, 63]. IMUs are the most commonly used sensors for quantifying movement in both clinical and sports-based movement contexts [43, 44]. With the movement data provided by an IMU, further kinematic variables can be calculated, such as velocity or displacement from acceleration [64]. There are many applications for sensors in a range of applications for quantifying biomechanics. Additionally, there are a number of possible sensor brands and models that exist. The use of an IMU or multiple IMUs in the DLJL for the LESS application requires a logical implementation that provides relevant movement data to derive the LESS items and represent high-risk movement patterns.

Chapter 3

Related Work

Wearable technology provides an exciting avenue for capturing and tracking movement in field environments and clinical settings. While automated video quantification of the LESS has had success [4, 41], applying sensors to the DLJL may provide even greater accessibility to movement data and remove the need for camera systems and manual scoring. Understanding how sensors have been used in related research is important for establishing how IMU-derived movement data can be processed best and applied for machine learning automation during the DLJL task.

In this chapter, we focus on the usage of wearable sensors in research and the integration of sensor data with machine learning models. This chapter begins with an introduction to how movement is quantified using wearables, including the specific use of IMUs and how these sensors compare to current benchmark technology. Next, the methods used for automating movement quantification are highlighted, with focus on the common automation methods used and the classification of jump-related and landing-related movements. Finally, a summary of the relevant work is discussed, and how these key ideas are relevant to automation of the LESS scoring process.

3.1 Movement Quantification with Wearables

As highlighted in previous sections, a single movement can be of relevance in both a clinical and sports context. For example, a standing calf raise motion involves lifting the heel up from the ground and back down. This motion can be used as a resistance training exercise to develop and build lower limb muscles [65], or used in clinics to assess lower limb muscle strength and endurance and detect lower limb deficit [66]. Wearable sensors have enabled greater accessibility to movement data as they are easy to access, transport, set up, and do not require an expert to collect and monitor movement data [63]. Having greater access to movement data allows people to better understand their own capabilities and provide a monitoring tool. Among the wearable sensors available, IMUs are commonly applied for capturing and quantifying the biomechanics of a range of different movements.

3.1.1 Quantification Using Inertial Measurement Units

Example applications of IMUs include the quantification of basic human locomotion [67]. Gait analysis is defined as the biomechanical science behind human locomotion [68, 69] - a fundamental aspect of any human movement. The gait cycle can be separated into seven different key phases of the movement, from initial contact with a heel strike to terminal swing [70]. Many gait biomechanical metrics have been quantified using IMUs, such as stride time, cadence, walking speed, and lower extremity joint angles (e.g., hip, knee, and ankle) [45, 71].

Additionally, IMUs have also been applied to a range of sporting-related movement patterns, including low velocity movements like dancing motion analysis [72] or higher speed movements involved in tennis [73, 74], and overhead baseball pitching movement [75]. The DLJL used in the LESS involves relatively high impacts and accelerations during ground contact. Therefore, it is important to establish whether IMUs can provide meaningful data from high impact movements.

Related research focused on quantifying high-impact movements - specifically, jumps and landings - has had success with IMU application. Jump-related

movement patterns are commonly used in research due to its multifunctional relevance in many sporting activities [76]. IMUs to date have been most commonly applied to measuring jump height when applied to jump testing [77, 78, 79]. Jump height allows for an assessment of lower extremity functional performance and can be used as a marker of explosive strength, useful for monitoring the benefits from training interventions in sports players [80]. Jump height is also useful in injury rehabilitation, such as identifying interlimb differences following ACL reconstruction [81]. Different calculation methods were used to derive jump height from IMU data, such as focusing on the peaks in the IMU data [77], implementing algorithms to measure take-off and landing velocity based on accelerometer data [78], or registering relevant movement data when calculated displacement reaches a threshold value [79]. In the context of the LESS used to assess athletes, one may seek to reduce injury-risk biomechanics without compromising the subsequent maximal vertical jump.

Similarly, angles of different body segments have been quantified and monitored using IMUs, primarily during the landing phases of jumps [82, 83, 84]. As discussed in Section 2.2, landings are considered one of the most vulnerable movement patterns for ACL injury. When there is not enough flexion occurring throughout the joints when landing, the landing is considered "stiff," with greater ground reaction forces present [85]. Assessment of the relevant segment angles involved in the LESS using an IMU may improve accuracy in classification of a scoring item, as visual estimation of a segment angle is difficult [25, 33].

In the case of calculating specific biomechanical metrics such as angles in jumps and landings, specific positioning of the sensor is important. Placement of a sensor must avoid inhibiting the participant from performing a movement. If movement data is not captured in the participants "normal" motion, the data is not representative of the actual movement pattern tendencies of the participant. The anatomical landmark at which a sensor is attached must also provide movement data that is relevant to the biomechanics being quantified. Using the detection of a jump landing as an example, a sensor placed at the wrist is not going to provide movement data as insightful as an ankle, knee or pelvis sensor could. Furthermore, the sensor should not be placed where there is considerable soft tissue artifact unless movement of the soft tissue is

of primary issue [86].

Soft tissue artifact refers to the biological tissues between a skin-based marker and bone, where the separation between the bone and marker lead to inaccuracies in monitoring [87]. For segment angles, two sensors are typically used where the axes of the sensors coincide with the axis of the joint [71, 88, 89, 90]. Placement of a sensor that aligns with the joint centre can be complex due to the considerations raised above (i.e., movement obstruction and soft tissue artifact) [71, 86]. One study looking at the difference between sensor data captured for gait analysis from a prosthetic and human leg found greater accuracy from the prosthetic leg, concluding that the prosthetic more closely represented a "perfect hinge joint" [71].

With so many differing options for sensor placement and calculation methods, the accuracy of IMU-derived biomechanics must be evaluated against current benchmark movement capture technology. Although wearable sensors are accessible, the movement data provided by the sensors is not useful unless the data is representative of a participants movement tendencies. Despite success quantifying high-impact movements using IMUs, inaccurate movement data may lead to missed injury risk factors or an overestimation of the performance of a movement, leading to the incorrect interpretation of metrics and associated prescribed training interventions.

3.1.2 Accuracy Comparison to Benchmark Technology

An OMC is commonly listed as the "gold standard" of movement capture and quantification [4, 5]. These systems involve multiple high-resolution cameras capturing a participant performing a movement from multiple angles, with each view used to reconstruct a movement onto a 3D model [91]. 3D motion capture can be done in many ways, however these systems typically use infra-red light cameras that record the light reflections from retro-reflective markers placed on a participant in strategic locations [6, 92]. IMU movement data can be captured simultaneously with OMC data to compare the same outcomes from both datasets. IMU-derived data have been shown accurate when compared to OMCs for quantifying various movement types, including detection and analysis of the various gait phases and angular velocity in tennis based on

gyroscope data [69, 93], accelerations in a variety of straight-line sports-specific movements [94], and extraction of exercise performance measures in resistance training [95].

In jump and landing-related research, IMU movement data have been compared to both OMC and force plates, where the latter assesses ground reaction forces important in jump and landing tasks [96]. Strong correlations have been observed between IMU-derived movement data and force plate data, including quantifying step times from IMU data during gait [97], reactive strength index and temporal-related measures in a countermovement jump [98], and peak ground reaction of five different jump and rise tests [99]. Despite different calculation methods for jump height, as discussed in Section 3.1.1, result comparisons to respective "gold standards" of movement capture all produced similar results to IMU-derived data [77, 78, 79].

When comparing IMU movement data to current benchmarks of movement capture technology, a clear application is shown for quantifying a number of biomechanical metrics. Different approaches exist for calculating the same biomechanical metrics using IMU movement data, with similar levels of accuracy. As touched on in Section 2.4, high quality OMCs and force plates can be expensive technology, which does not allow usage for the general population, in addition to the need for an expert to operate the technology and analyse the data for performance evaluation.

In the case of the LESS, 2D video recordings are used as the benchmark for manual analysis of the different scoring items. As previously discussed, the two last scoring items involve subjective scoring such as visually assessing whether a landing is soft or stiff. Other scoring items are difficult to visually estimate accurately, including flexion angles and simultaneous contact time between both feet [25, 33]. In consideration of the success in quantifying related biomechanics, IMU application may provide more accurate and objective quantification of the LESS.

3.2 Automation of Movement Quantification

As discussed in Section 2.4, the LESS is inaccessible for large-scale testing due to the need for an expert to manually review and score each video individually [4]. The ability to perform this test with an expert is limited by the availability of the expert, and time required to score each individual jump. As discussed in Section 2.4, automation of the LESS has been relatively successful previously using other means than IMUs [4, 41]; but access to these approaches for the general population is still limited. With the growing popularity of IMU applications for human movement quantification, implementation of a system that is able to score the LESS items using IMU data would remove the need for a "hands-on" expert and allow greater access to LESS quantification.

3.2.1 Common Automation Methods

To assess movement from sensor data, two approaches are quite common: 3D modelling and the use of machine learning methods. With 3D modelling, sensor data are used as input to replicate a participant's biomechanics based on an existing 3D model, allowing researchers to visualise and reconstruct a movement [100]. Modelling can be completed using different software, such as Visual 3D[1], that includes techniques for signal processing and filtering [101]. Open-source libraries have also had success with automating movement from IMU data, such as OpenPose, where multiple key points in various body parts are identified and tracked through images and videos [74]. Although these tools have been commonly used for movement assessment, OMC data are required to calibrate and create the 3D model. As stated previously, an OMC system is unobtainable for regular use and still requires expert intervention to evaluate the LESS performance.

3.2.2 Machine Learning for Jumps and Landings

There are many machine learning models that have been used in assessing biomechanics. For example, neural networks have been used for quantifying movements within gait cycle research [102] and a soccer ball kick [103], while

[1] https://www.has-motion.ca/biomechanics-software/#visual3d

Table 3.1: Machine Learning Models in Jump and Landing Research

Model Used	Literature	Use Case
Support Vector Machine	[104]	Classification on flight or landing phase of jump
	[107]	Prediction of peak power
Neural Network	[104]	Estimation of ground reaction force
	[108]	Estimation of ground reaction force
	[106]	Quantification of jump height
	[67]	Predicting knee joint forces
Linear Regression	[109]	Estimation of flight time, force and impulse
	[107]	Prediction of peak power
	[105]	Estimation of ground reaction and knee forces

a binary classifier has been used for tennis-related stroke style predictions [73]. In research focused on the integration of machine learning and the quantification of jump and landing performance, seven relevant articles were found. These articles typically used IMU data to calculate relevant jump landing biomechanics, such as ground reaction forces [104], knee joint forces [105], and jump height based on accelerometer values [106].

It has been proposed that the most commonly used algorithms for quantifying biomechanics are support vector machines, neural networks and generalised linear models [110]. From the related research covered, as seen in Table 3.1, support vector machines were typically used to distinguish between key phases of the movement, rather than the biomechanics involved in the jumps and landings [104, 107]. Neural networks were used for the estimations of ground reaction forces [104, 108], jump height [106], and knee joint forces [67]. These studies also indicated their reason for using a neural network was related to success in other impact-related studies.

Neural networks were used individually to examine movement in some studies,

while others compared the performance of multiple models against each other. Generalised linear regressions were used for predicting various biomechanical metrics, such as flight time or net impulse [109], peak power [107, 109] and ground reaction forces [105]. The LESS items cannot be considered linear, rather each item needs to be classified into one class. The majority of items are either scored a "0" or a "1," as discussed in Section 2.3. Therefore, the implementation of binary classification models is likely the best option and more appropriate than generalised linear regression models applied to LESS scoring based on the sensor data, with multi-label classification for scoring items with three targets.

Additionally, the LESS is scored between two main key events of the movement - initial contact and maximum knee flexion. These events occur in the first landing phase of the movement. Therefore, the segmentation of the IMU data into key phases of the first DLJL may improve the accuracy of classification, rather than just using a full raw dataset of the entire movement. Segmentation was frequently used when processing the IMU data to highlight key points in various other movements, such as the breakdown of a two-handed ball throw into throwing and follow-through phases [103], or extracting the heel strike/toe-off phases of the gait cycle [111].

Segmentation of jump-related IMU data has also been applied elsewhere. For example, some authors [106] have split the accelerometer data from an IMU into five key stages of the jump, from weightbearing to landing. There were 26 kinematic-related features taken from all five stages and used to train a neural network. Elsewhere [109], critical points in the accelerometer data of a jump were highlighted to calculate flight time and peak force. With the appropriate algorithms and data processing, critical points in a jump can be extracted based on the accelerometer data signals. Machine learning can then be applied to classify these phases in the raw IMU data. For example [104], some authors trained a support vector machine using the calculated magnitude of all three axes of accelerometer data to classify whether a segment of data was in the flight phase or landing phase of a ballet jump. Depending on the classification, the data were then either passed to a flight-based artificial neural network (ANN) or a ground-based ANN to compute relevant jump metrics.

In a similar study focused on the DLJL, researchers [105] identified the relevant events from the LESS using IMU data - initial contact and maximum knee flexion. These phases were extracted in a step-by-step process, firstly finding the relative region of interest, and using maxima and minima points in the different sensors to extract the full movement window. Positioning of the sensors at the shank and thigh provided the movement data required for predicting knee biomechanics, so the segmentation of the DLJL may differ depending on the movement data available.

In addition to various processing techniques applied in the related work, a filter was also commonly applied to the raw movement data prior to feature extraction and training of multiple models. When applying a filter, the aim is to remove sensor noise while also maintaining the significance within the magnitude of the signal [112]. Most commonly, a low-pass Butterworth signal filter was applied to the accelerometer and gyroscope data [67, 105, 106, 108, 109]. The cut-off frequencies ranged from 10 Hz [106] to 32 Hz [108], with authors [105] comparing filter applications from 15 Hz up to unfiltered. A fourth-order Butterworth filter was applied most commonly [67, 108, 109], with some using a second and sixth-order filter [105, 106]. Assessing the orders and cut-off frequency for the LESS movement data will be important for ensuring signal magnitude is preserved, while noise is removed.

In the review of relevant work, the use of IMU movement data as input for machine learning models was not limited to one sensor alone. For example, some authors [105] compared the use of accelerometer features against the combination of both gyroscope and accelerometer features, allowing a greater amount of features to be integrated for training a model. These authors [105] also extracted statistic-based features from both accelerometer and gyroscope data - mainly centred around the maxima and minima points in all axes of data. This approach was common among the other related studies, with many using features that detail maxima and minima points in all axes [105, 106, 109]. Additionally, the mean and standard deviation of these axes were extracted to detail the variability between jumps and landings [2, 109].

In addition to statistical descriptions of the dataset, temporal measures were also used in conjunction with segmentation. Temporal measures such as the

Table 3.2: The Different Models and Their Accuracy, Data Provided by [2]

Model	**Accuracy (%)**
Naive Classifier	7.0
K-Nearest Neighbours	95.9
Gaussian Naive Bayes	88.9
Support-Vector Classification	94.2
Gradient Boosting Classifier	93.0
Stochastic Gradient Descent	91.5
Deep Feedforward Neural Network	96.4
Convolutional Neural Network	96.1

time to a maxima or minima point, the start and end time of a region of interest, or the time taken to complete a movement were extracted and used to train multiple machine learning models [105, 106, 109]. Additionally, the times between key events or phases of movement may correlate to biomechanic measures, such as flight time [109].

In the case of the LESS, appropriate models need to be selected to classify whether IMU-derived data are capable of accurately predicting a LESS score. One relevant study involved the comparison of 10 different supervised learning machine learning models in assessing the difficulty of a jump in trampoline gymnastics [2]. Supervised learning involves training a model with both the input, and desired output - where the model tries to map the connection between the two sets and apply this same mapping to further input data [113].

In one study [2], the extracted features from the IMU data acted as the input, while the jump type acted as the labelled output. There were 45 features extracted from accelerometer and gyroscope data used to train each model, including segmented data for each phase and statistical features such as mean and standard deviation values for each phase. In relation to the application of machine learning models to the LESS data, it is important to distinguish between accelerometer and gyroscope readings, and acknowledge that the time course of the curve will differ from other jumps not performed in a similar manner.

As seen in Table 3.2, the models selected mostly produced accuracies greater than 90% for identifying jump type [2]. Notably, the two neural networks produced the greatest accuracy when classifying jump types. Shapley additive explanation (SHAP) values were used to assess which of the phases and features had the greatest importance on the training of each model. With a similar approach with LESS data, the phase between initial contact and maximum knee flexion should produce the greatest importance in consideration that all the scoring items of the LESS are scored in this phase [1]. Considering the success of models used in trampolining [2] with complex jumps and landings, applying LESS movement data to these models and extracting similar statistical features based on the phases of the movement will help to establish the applicability of IMU sensor data to scoring the LESS. Additionally, a supervised learning approach using manually scored data as the labels may provide greater clarity on the application of IMU sensors for classifying the score of each scoring item.

3.3 Summary of Findings

The usage of wearable sensors has demonstrated clear applicability for the quantification of human biomechanics, with machine learning models being helpful in the process. Wearable sensors provide a greater accessibility to movement data for the general population, allowing assessment of movement without the need for expensive technological systems and an expert to capture the data. IMU sensors have been commonly used to assess jumps and landings, with various calculation methods present for deriving biomechanical metrics from the raw sensor data. IMUs have also demonstrated excellent results when compared to "gold standard" movement assessment technologies.

The machine learning models commonly used for quantifying biomechanics were both ANN and support vector machine models. Both these models are used in supervised learning approaches. Model selection for the LESS will benefit from a supervised learning approach as manually scored LESS scores can be input in conjunction with the processed IMU data to help inform the model on the expected output. Using multiple models allows comparison between different machine learning architectures for prediction, while also showing which

extracted features from the IMU data have the greatest importance for accurately classifying the LESS.

From the raw data provided by IMUs, there are a number of approaches to processing the data prior to feature extraction. From the movement data provided by a DLJL, the jump can be split into the primary phase where the LESS items are scored, i.e. from initial contact to maximal knee flexion. The features in these phases are expected to have the most importance to the performance of the machine learning models for automating the LESS. Segmenting the IMU data into phases and comparing the performance of the models will determine whether patterns in the IMU data are relevant to LESS scoring, or whether patterns in the other phases of the data are also worth considering.

Various features can be extracted from IMU movement data, including statistical features surrounding the key phases of the movement and the maximum and minimum points. Temporal features can also be extracted from IMU data in the DLJL, such as flight time or the times between phases of the task. Extracting both statistical and temporal features from the IMU data may have a greater importance to the accuracy of the models in comparison to the raw data, allowing for an evaluation on whether raw IMU data from each jump demonstrates any patterns relating to each respective scoring item.

Chapter 4

Design

There are many approaches to applying sensor data to the calculation of biomechanics in a jump and landing movement, including different data processing techniques, possible features to extract, and machine learning algorithms that can be used for predicting movement performance. The implementation of customised algorithms for processing sensor data must provide relevant training data to allow a fair evaluation on the application of machine learning algorithms for the automation of the LESS. A design based on the literature must be formalised to determine how LESS data can be integrated with machine learning to automate the scoring process.

In this chapter, the decisions made based on the literature are highlighted and a formal design is established for evaluating the applicability of IMU sensors to the LESS. This chapter begins with discussing the research questions and hypothesis. Next, the design requirements are outlined in relation to how movement data will be processed for integration with machine learning for automation. Finally, the design of the proposed solution is covered in relation to the collection of LESS movement data, and how these data will be processed and used to train and test machine learning models to predict the scoring items.

4.1 Problem Discussion

As discussed in Section 2.4, the scoring of the LESS requires the use of two video cameras correctly placed to monitor movement in the sagittal and frontal plane of a DLJL. Some of the scoring items involved in a LESS score are hard to accurately measure using visual estimates [25, 33]. From the work highlighted in Section 3.1.1, IMUs have shown to be applicable for quantifying various movement patterns. Additionally, the related work covering automation of movement quantification in Section 3.2 showed IMU-derived data can also be used in conjunction with machine learning algorithms with high levels of accuracy. Based on these findings, the main hypothesis posed for this thesis is:

1. IMU sensors can be used to predict the LESS score using out-of-the-box machine learning models.

There are many applications of machine learning in various contexts such as healthcare and criminal justice that leverage an out-of-the-box machine learning approach when making predictions [114]. Interpretation of the decisions made by these types of machine learning models cannot be made, as decisions are independent. The accuracy of selected machine learning models will therefore be based on the input data, and not the tuning of various hyperparameters. Additionally, the accuracy of the models using movement data can be used as a proof-of-concept to assess whether any patterns in the raw IMU data are identified that relate to each scoring item of the LESS. To evaluate whether IMU sensors can predict a LESS score, the following research questions were formulated:

1. How accurately can out-of-the-box machine learning models automate LESS scores from IMU data?
2. Which IMU-derived features are the most relevant to the automation of LESS scores?

Excellent accuracy has been shown between IMU data and current benchmark technology of movement quantification (OMC systems and force plates), as discussed in Section 3.1.2. However, to the best of our knowledge, no studies were found comparing the application of IMU sensors to manual quantification

of jump and landing movements from 2D videos. The automation of the LESS using markerless tracking and computer vision has shown successful results, but the need for adequate laboratory space and technological requirements does not allow accessibility to the general population [4]. With the use of wearable sensors, we are interested in evaluating whether scoring using the LESS can be automated and thereby implemented with no manual intervention other than sensor placement. Additionally, the work covered in Section 3.2.2 showed many potential features that can be derived from raw IMU movement data for jump and landing movement types. Assessing different IMU-derived measures is important for not only improving the overall accuracy of the out-of-the-box models, but also for evaluating whether additional features exist that impact LESS scoring.

4.2 Design Requirements

To evaluate the applicability of IMU sensor data to LESS automation, appropriate design requirements need to be established. As such, the design requirements are split into two main subheadings: processing requirements and training requirements. Processing requirements relate to the algorithms that will need to be implemented for processing LESS movement data for integration with machine learning models. Additionally, training requirements outline the methodology of how the processed data will be used to train the out-of-the-box models and evaluate the performance of these models when automating LESS scores.

4.2.1 Processing Requirements

The data output from IMU sensors will require a certain level of processing to be used as training data. Raw data from motion capture can include large time windows of irrelevant data (e.g., before jumping and after landing for the DLJL movement). Therefore, the following requirements have been outlined for the processing of the data:

- **Organising:** movement data are likely to be provided in various files and formats by end users. The files required for automation will need to

be organised into a logical order to ensure relevant movement data are correctly collated in one location.

- **Cropping:** in consideration of the data collection process, raw movement data are likely to have periods of noise before and after the DLJL task is performed. An algorithm will need to be implemented to remove this noise and remove the need for additional processing on large quantities of irrelevant data.
- **Alignment:** with the application of more than one sensor, it is possible that sensors may be worn in differing orientations. Depending on the number of sensors included, all axes of movement data will have to be aligned so movement at any point of the body occurs in the same axis.
- **Segmentation:** as discussed in Section 3.2.2, peaks in the accelerometer data can be used to identify key phases of a landing movement. Segmenting the output from each sensor into respective phases of the movement may help to understand which IMU-derived features influence the accuracy of automation the most.
- **Feature Extraction:** using the key phases of the DLJL, temporal and statistical features can be extracted for use in training the machine learning models.
- **Merging:** the LESS is made up of 17 different scoring items. These scores will need to be merged with the IMU-derived data into one dataset for training the machine learning models. This process will involve implementing an algorithm that merges the score of a jump correctly with its corresponding IMU data.

4.2.2 Training Requirements

To automate the LESS scoring process, the selection and structuring of appropriate machine learning models must be implemented correctly to ensure a black-box approach is maintained. Additionally, the training of a model using the processed data must be structured correctly to allow the selected models to take advantage of any identifiable patterns in the data. Therefore, the requirements listed below are outlined for evaluating the performance of the

automation:

- **Structuring:** the corresponding architecture for the machine learning models must be implemented correctly to ensure training of the models for each respective scoring item and data subset is done correctly.
- **Metric Extraction:** ensuring each of the selected performance metrics used for evaluating model performance is correctly extracted, and that these are outputted for further comparison between models.
- **Comparison:** the performance of each model and feature subsets are compared against one another to evaluate the performance of different models and IMU-derived features for each respective LESS scoring item.

4.3 Solution Design

To meet the processing and training requirements as specified in Section 4.2, the design for a solution that is capable of processing and training machine learning models must be formalised. As shown in Figure 4.1, there are three main stages outlined for using movement data collected by an IMU to train and evaluate the prediction of LESS scores using machine learning models. These stages include (1) data collection, (2) data processing, and (3) prediction using processed data.

4.3.1 Data Collection

Sensor data in relation to the DLJL task as part of the LESS were collected as part of a larger study that was previously completed. The data collection protocol was approved by the Health, Engineering, Computing, and Science Division Human Research Ethics Sub-committee of the University of Waikato (HREC(HECS)2023#4) (Appendix A).

The sensor selected for capturing movement data throughout the DLJL task was the IMeasureU Blue Trident IMU sensor[1], as seen in Figure 4.2a. The IMeasureU sensor uses both a low-g and high-g three-axis accelerometer simultaneously, capable of tracking high-value peaks in movement data. The

[1]https://imeasureu.com/imu-sensor/

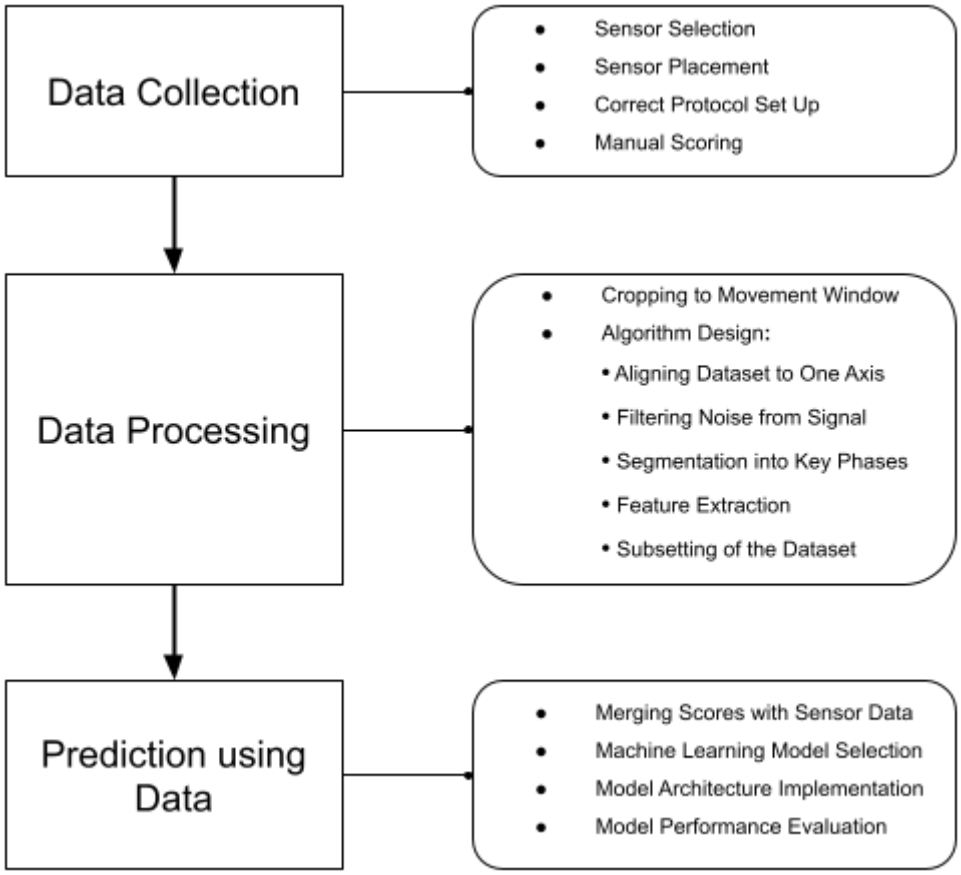
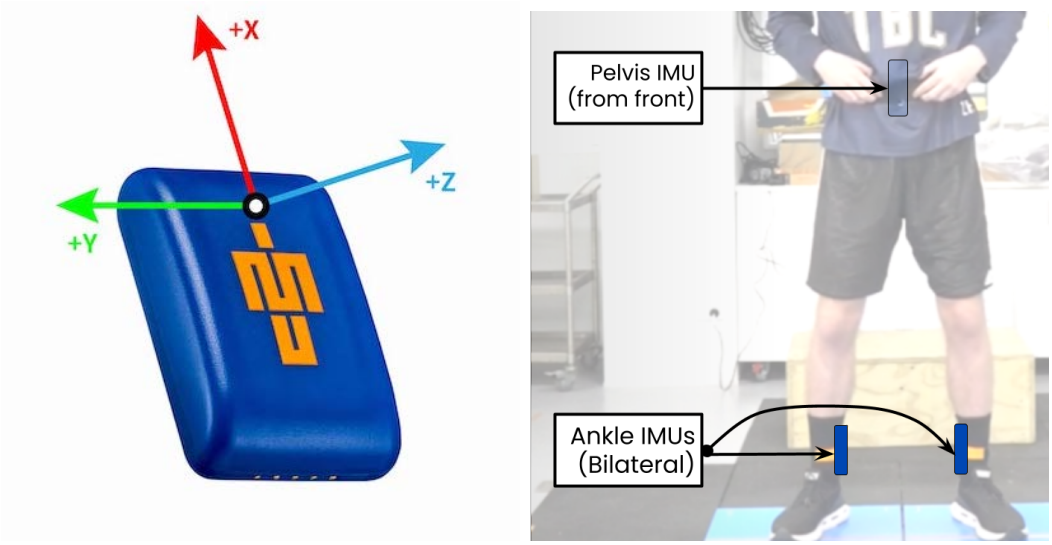


Figure 4.1: Flowchart for Solution Design

low-g accelerometer has a frequency of 1125 Hz, in comparison to a 1600 Hz frequency for the high-g. Additionally, a three-axis gyroscope with a 1125 Hz frequency with a range of ± 2000 °/sec is used, along with a three-axis magnetometer at a frequency of 100 Hz - all capturing movement data simultaneously. The data were captured using Bluetooth connection to an iOS device, through the Vicon Capture mobile application. Three IMeasureU Blue Trident IMU sensors were placed on the participants prior to their jump performance. Two sensors were placed bilaterally at the ankles, proximal to the medial malleolus, with the final sensor placed on the posterior aspect of the pelvis/sacrum, as seen in Figure 4.2b

The scoring protocol implemented is a slightly modified example of the typical LESS scoring protocol [1]. This protocol required each participant to jump



(a) Blue Trident IMeasureU Sensor Used (b) Placement of the Sensors

Figure 4.2: Blue Trident IMeasureU Sensor and Related Placement

horizontally from a 30 cm high box, to 50% of their body height, and then immediately jump vertically for maximal height. Each participant was asked to complete the LESS in two conditions - three jump landings with no intervention and three jump landings while performing a dual task. The dual task included either turning the head to the right or counting multiple high and low tones while performing the jump movement. As the LESS is typically performed and scored without the head being in a rotated position, LESS data from these jumps were disregarded for the purpose of this thesis. The dual task data with auditory cues were maintained as the head was in the usual position.

Two digital video cameras (Sony RX10 II, Sony Corporation, Tokyo, Japan) were placed in front and to the right of the participants to capture the frontal and sagittal plane motion at 120 frames per second. Each video was started prior to the first jump of a given participant, and recording was stopped following the sixth and final jump. Each participant performed three jumps without intervention and three jumps with the dual task, in a randomised order.

Manual scoring of the LESS was performed by the author of this thesis (ZH) to better under the scoring process, the movement patterns that constitute an error, and how these movements may be identified within sensor data.

This scoring was performed on a subset of the data (30 total jumps). Scoring of the full dataset was completed by an expert LESS scorer, with validation by another experienced scorer. To evaluate the performance of the machine learning models, the full dataset will be used.

4.3.2 Data Processing

Customised algorithms that are capable of processing a large window of movement data will need to be implemented. The approach to processing movement data will need to consider the additional noise, differing axes of movement involved, and movement data that does not correspond to the LESS scoring items. The Python language[2] will be used to implement the processing algorithms.

The protocol followed for collecting data in relation to the DLJL involves a potential issue where there is many rows of time series data, but only a small subset of this data relates to the jump performance. Due to the high-impact movements, using prominent peaks in the movement data will be used to recognise the two key landings (first landing from the 30 cm box, and second landing from the subsequent maximal vertical jump). With these landings, an algorithm will be required to crop out large windows of irrelevant data to speed up further processing algorithms. Additionally, due to the irrelevance of magnetometer data to the scoring process, magnetometer data will not be included and removed.

With three sensors involved in data collection at different anatomical locations, there are a number of axes involved. Simultaneous movement in one direction could occur in three separate axes. Based on the movement data provided and the video recordings, an algorithm should be implemented to ensure all movement is represented correctly in one direction. The consistent axis that will be implemented is shown in Figure 4.3. Directional movement will need to be recognised in the movement data and manipulated into the new axis for all three sensors. The x-axis refers to right/left horizontal movement, y-axis movement refers to up/down vertical movement, and z-axis corresponds to forward/backward movement. Positive acceleration values represent movement to

[2]version 3.10.0, https://www.python.org/

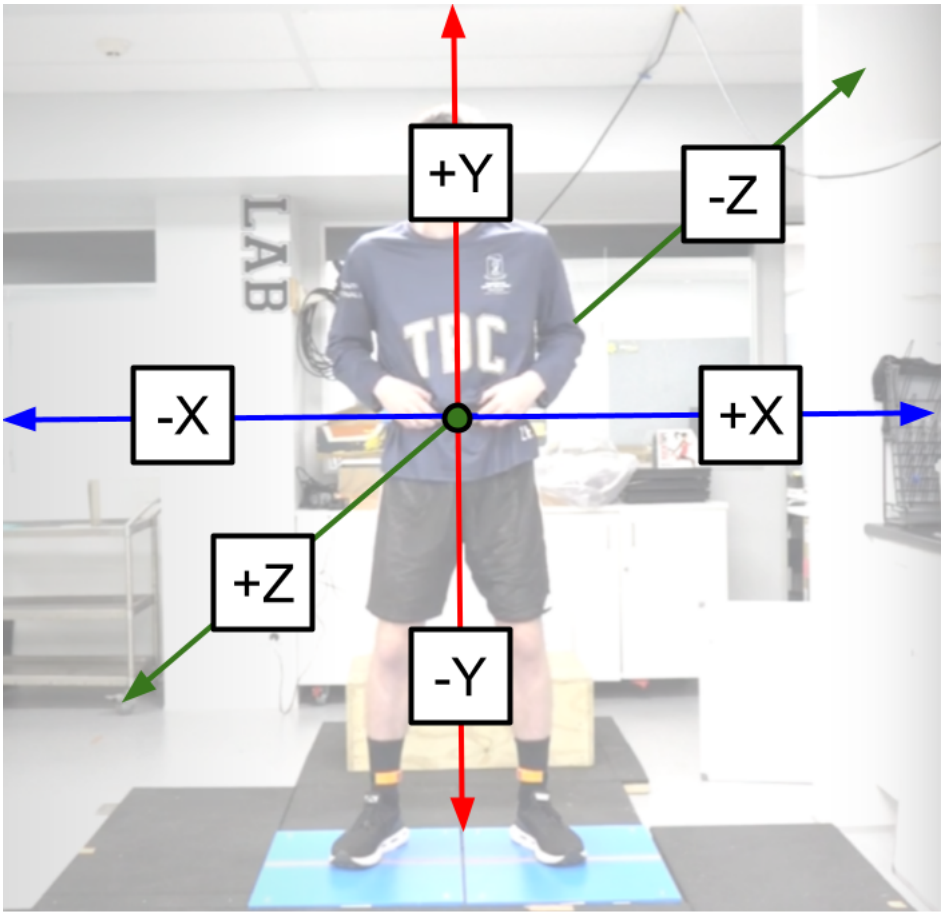


Figure 4.3: Selected Axis for Consistent Alignment

the right in the x-axis, upward in the y-axis, and forward in the z-axis. Negative acceleration values represent movement to the left in the x-axis, downward in the y-axis, and backward in the z-axis.

The Butterworth filter was applied commonly in machine learning research in jump and landing-related movements, as seen in Section 3.2.2. Most commonly, the filter implemented was a fourth-order Butterworth filter at a low-pass cut-off rate ranging from 10 Hz to unfiltered. Due to this filter appearing in similar movements, a fourth-order Butterworth filter will be applied to the LESS movement data. Due to the impact expected from the DLJL, a range of low-pass rates from 10-100 Hz will be evaluated, and the most appropriate will be selected based on visual inspection. A further evaluation from a first-order up to a tenth-order Butterworth filter will validate whether a fourth-order is an

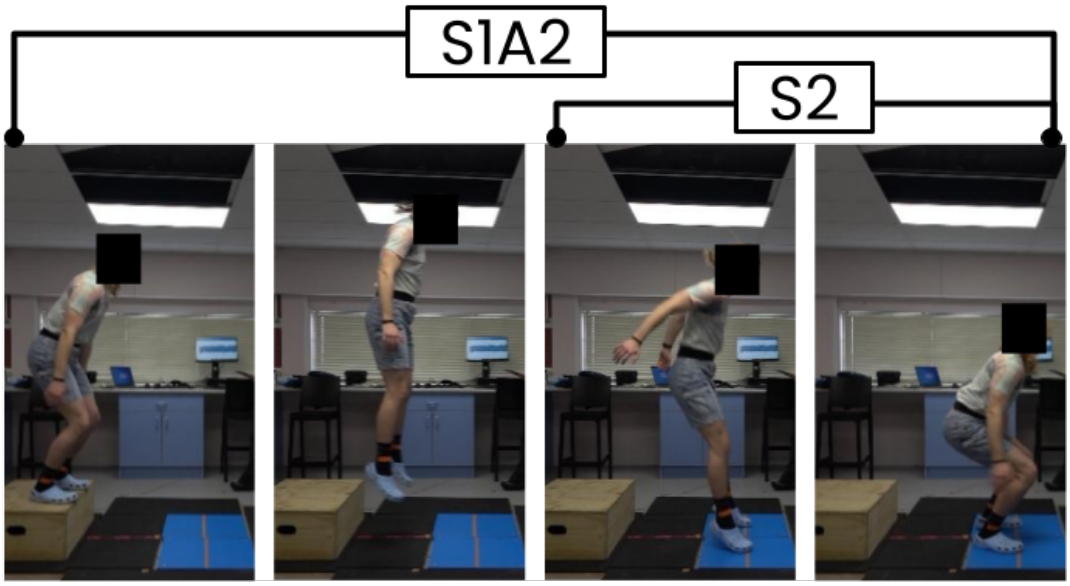


Figure 4.4: The Selected Key Phases of the Double Leg Jump Landing Task

applicable solution, and the movement data does not demonstrate a dramatic time shift in comparison to other orders.

Segmentation was a common data processing technique used in related work covering machine learning and movement data integration, as discussed in Section 3.2.2. In the DLJL, the key events initial contact and maximum knee flexion can be identified [105]. This segment of movement data will be labelled as segment two (S2). For the LESS, in consideration of all the scoring items taking place between initial contact and maximum knee flexion, algorithm implementation will be focused on identifying the start and end of this phase within accelerometer data. Additionally, extracting the phase between take-off from the box and initial contact allows an evaluation of the flight time and the movement of the participant before they come into contact with the ground. This segment of data will be labelled as segment one (S1). Movement data from S1 alone is not likely to provide meaningful insight for classification considering the LESS items are scored in S2. Therefore, S1 and S2 will be merged together (S1A2) to provide features from the full movement window. These key phases of the LESS can be seen in Figure 4.4.

Using the key phases identified in segmentation, a variety of features can be

Table 4.1: Selected Features to be Derived from Sensor Data

Temporal	Statistical-Based	Characteristic
Flight Time to Initial Contact (S1)	Maximum	Age
	Mean	BMI
Time from Initial Contact to Maximum Knee Flexion (S2)	Minimum	Height
	Root Mean Square	Mass
Total Movement Time	Standard Deviation	Sex
	Variance	Past Injury

extracted and further used in machine learning training and evaluation. From the related work covered in Section 3.2.2, a number of temporal and statistical features have been used with successful results. In relation to the movement data provided in the DLJL and the features provided in Section 3.2.2, six temporal measures, six statistical measures, and six participant-related characteristics are the selected features, as seen in Table 4.1.

Since asymmetric contact with the ground is a scoring item of the LESS [1], initial contact extracted from the left ankle sensor may be different to initial contact at the right ankle sensor. Flight time from take-off to initial contact details the time taken to complete S1. This feature will be two different features in itself - flight time from take-off to the left ankle initial contact, and time from take-off to the right ankle initial contact. Additionally, the time taken from initial contact to maximum knee flexion details the time taken for a participant to complete S2. This feature will also be two different features - time from left ankle initial contact to maximum knee flexion, and the time from the right ankle initial contact to maximum knee flexion. Finally, the full movement will refer to the time taken to complete S1A2.

The selected statistical features help to highlight individual participant performance characteristics. Statistical feature extraction will be performed on each segment, for each axis of the sensors used, and for each participant, as performed in similar related work discussed in Section 3.2.2. The mean and standard deviation have both shown to be an influential measure in the accuracy of classification for several indicators of countermovement jump performance [109] and the detection of trampoline gymnastic jumps [2]. The maximum and

Table 4.2: Selected Subsets to Evaluate Classification Performance

Segments	Sensors Involved	Feature Grouping
Take-off to Maximum Knee Flexion (S1A2)	All Sensors	All Features
	Ankles	Raw Movement Data
Initial Contact to Maximum Knee Flexion (S2)	Pelvis	Extracted Features

minimum points also show the differences in acceleration and rotational velocities between participants, implemented in related work [105, 108, 109]. Other than the mean as an input feature, two additional features will be included. The root mean square provides another interpretation of the mean, focused on measuring the mean magnitude of numbers. Variance will also be included to detail how all the values in an axis of movement data differ from the mean value.

In addition to the features extracted from the movement data, participant-related characteristics will be included as features. Considering participant movement data is individualised, the characteristics of the participant may assist the machine learning models to recognise features. Age, sex, and past injury were factors shown or speculated to influence LESS scores, as discussed in Section 2.4; therefore, these characteristics were included. The height, mass, and body mass index (BMI) are participant-specific, and can influence movement. For example, a participant with greater mass will accelerate toward the ground faster, which may influence accelerometer signals upon landings.

The final step of data processing involves dividing the movement data into smaller subsets to assess the influence of data processing on the classification performance for the LESS. The selected subsets to be evaluated are highlighted in Table 4.2. As previously discussed, all the scoring items take place between initial contact and maximum knee flexion (S2) - therefore, it is the main segment of interest. A comparison can also be made to the combination of the take-off segment with the landing segment (S1A2) to see whether this phase provides meaningful data that relates to the scoring items. The scoring items

Table 4.3: Selected Supervised Learning Classifiers

Model
ZeroR or Dummy
K-Nearest Neighbours
Gaussian Naive Bayes
Support-Vector Classification
Gradient Boosting Classifier
Stochastic Gradient Descent
RandomForest Classifier
Deep Feedforward Neural Network

of the LESS all occur within S2, so additional segments by themselves do not provide any relevant movement data.

Sensor position is an important parameter for collecting movement data from a sensor, as previously discussed in Section 3.1.1. The anatomical placement of the sensors was already decided when the data was collected. The positioning of the sensors was not selected to calculate movement of the joints, but rather positioned as a proxy of ground reaction force values via the ankle sensors and centre of mass displacement for the pelvis sensor. Nonetheless, an evaluation can be made on whether these sensor locations provide movement data that can accurately quantify the 17 LESS items.

As observed in Section 3.2.2, some related work using IMU-derived data had a high classification accuracy in their machine learning models when only using statistical or temporal features. Therefore, the data will be split into three groups: (1) all features together, (2) the raw movement data, and (3) the extracted features (temporal, statistical, and characteristic-based features). By using the different subsets of the data, an assessment on whether derived measures influence the classification can be made.

4.3.3 Automation of Scoring Using Sensor Data

Classification and automation of the scoring process will leverage a supervised learning approach. In consideration that the manual scoring process has al-

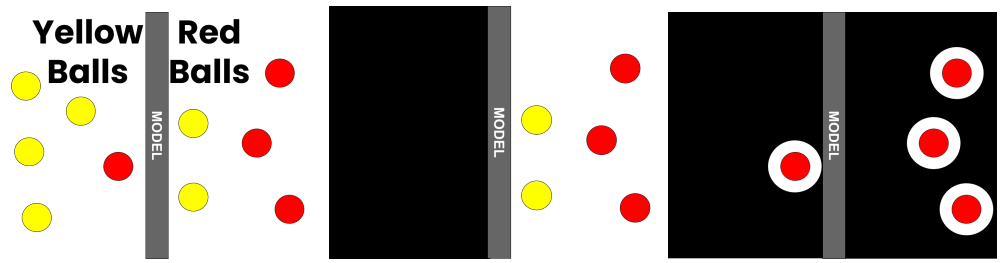
ready been completed for this dataset by an expert scorer, these scores will act as the desired output for the classifiers to try to predict and recognise any patterns in the dataset. From the related work covered in Section 3.2.2, seven supervised learning classifiers are selected to be implemented. These selected classifiers can be seen in Table 4.3.

The selected classifiers are similar to the selection elsewhere [2], as they cover a good range of supervised learning machine learning algorithms. Due to the accuracy demonstrated by the K-nearest neighbours (KNN), Gaussian Naive Bayes (GNB), support-vector (SVC), gradient boosting (GBC), and stochastic gradient descent (SGD) classifiers, they were implemented for evaluation in the case of the DLJL. Additionally, the deep feedforward neural network (DFF) used by the authors in [2] provided the greatest accuracy, therefore a similar model architecture will be implemented. The naive classifier was removed due to the poor performance observed. Due to the inconsistent sizes of the dataset in the selected subsets, a convolutional neural network is not applicable as 2D resizing of the data would be different for every subset iteration. To replace this model, a random forest classifier (RFC) will be included.

ZeroR will be implemented and act as a comparison measure to ensure any performance metrics derived from the other classifiers is not due to a skewed or imbalanced dataset. This classifier will classify all of the classes in the dataset as the most frequent class in the dataset. A high accuracy value may be observed in an imbalanced dataset, but classifying an entire dataset as one label does not provide a robust confidence for further applications to larger datasets of LESS data.

To ensure an out-of-the-box approach, all classifier architecture will be implemented using Scikit-learn[3]. The models implemented in Scikit-learn are constructed and maintained by a large group of machine learning experts, with extensive documentation and easy integration with the Python language. For ZeroR classification, the dummy classifier will be used. In the case of the DFF specifically, the multi-layer perception classifier will be implemented, a type of feedforward neural network and considered as the foundational architecture of deep learning [115]. Each model will be imported using the Scikit-learn library

[3]version 1.5, https://scikit-learn.org/stable/



(a) Example of a Model (b) Precision Calculation (c) Recall Calculation

Figure 4.5: Calculation Process of Precision and Recall

and fitted with the divided subsets of movement data.

Binary classification will be performed for 15 of the scoring items where the output labels were either a "0" for the error being absent, or "1" for the error being present. Multi-class classification will be performed for scoring items 16 and 17. Some classifiers typically do not natively support multi-class classification, such as a SVC. However, Scikit-learn provides a default behaviour to handle and perform multi-class classification.

As discussed in Section 2.3, the frequency of some scoring items in the LESS differ, where some items may be present for 50% of a sample, while other scoring items have a rare chance of being observed. Therefore, if we have an imbalanced dataset for a scoring item, the classifiers may just label every input as the most frequent class. Accuracy would therefore be a representation of the imbalance in the dataset, rather than the actual performance of the classifier.

Due to this potential imbalance, the metrics to be implemented for assessing the performance of the sensor data will be precision, recall, and the F1-score. All three metrics will be implemented using the Scikit-learn metrics library. Values range from 0-1, where 0 is the worst-case, and 1 is the best-case model performance. The calculation process of these metrics can be seen in Figure 4.5.

Using the example in Figure 4.5b, precision refers to the amount of true positives (TP) (red ball identified as red, so 3) divided by the total number of positive observations (TP + false positives (FP)) (5 total balls classified as red). Precision would be equal to 60%. The equation for this calculation is:

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

Table 4.4: Example of Weighted F1-Score Calculation

Label	F1-Score	Instances	Proportion	Weighted F1-Score
Soft	0.75	3	3/40 = 0.075	(0.75 * 0.075) +
Average	0.96	33	33/40 = 0.825	(0.96 * 0.825) +
Stiff	0.66	4	4/40 = 0.1	(0.66 * 0.1) =
Total	-	40	1.0	0.91

Furthermore, using the example shown in Figure 4.5c, recall refers to the number of total true positives (red ball identified as red, 3) divided by the total number of classes input into the model (TP + false negatives (FN)) (all the red balls, 4). Therefore, recall would be equal to 75%. The equation for calculating recall using the Scikit-learn library is:

$$Recall = \frac{TP}{TP + FN}$$

When an imbalanced dataset is used to train a model, a classifier may default to labelling the entire dataset as the dominant class. Say for example there were 95% for one class, and 5% for another - the accuracy of the model would be equal to 95%. The precision and recall would be equal to zero for the secondary class, which shows that the model is not capable of classifying both classes. Therefore, for evaluating overall accuracy, the F1-score will be used as the primary metric. The F1-score is calculated using both precision and recall to provide a harmonic mean of both metrics. The F1-score is calculated as:

$$F1 = \frac{2 * TP}{2 * TP + FP + FN}$$

Precision, recall and the F1-score are calculated individually for each class in a model. To calculate the overall performance of a metric, the metrics are ave raged across all the class labels. In consideration of the possible imbalance for different scoring items, as discussed in Section 2.4, the weighted average is going to be used for comparing model performance. The weighted average calculation uses the F1-score for each class, and multiplies it by the percentage of instances the class has in the dataset. All values are added together and divided by the total number of class labels in the model. An example calcu-

lation can be seen in Table 4.4. A weighted F1-score of 1 indicates a perfect model, where all the predictions are correct.

To evaluate the features that have the greatest importance on classifying each scoring item, the SHAP[4] library will be used to calculate the features with the greatest importance on prediction. A SHAP value is an explainable artificial intelligence (XAI) approach that assigns a weight to each feature of a trained model, depending on how a prediction is changed by the feature [116]. Other libraries exist for extracting feature importance; however, SHAP provides global feature importance, meaning the SHAP values of each prediction in a model are aggregated together to demonstrate the impact of features on the model as a whole [117]. Furthermore, SHAP demonstrated success in explaining model performance in trampoline-based jump classification [2].

[4]version 0.45.1, https://shap.readthedocs.io/en/latest/

Chapter 5

Implementation

With a research design constructed based on the related work covered, the implementation of customised algorithms must be capable of processing multiple jump occurrences accurately. With multiple key phases of the movement, features to extract, and subsets of the data to derive, multiple algorithms must work together to provide input data that can be integrated correctly with all of the selected machine learning models.

In this chapter, the implementation of algorithms for the data processing and integration of machine learning models are discussed. This chapter begins with a look into the collection of IMU data, and how all the movement data files are structured for processing. Next, the required processing pipeline of all the sensor data is discussed, including the steps taken at each stage of processing and how the data are manipulated throughout. Finally, the integration of this processed data with the selected machine learning models is covered, including the architecture of the models and how performance metrics were extracted from each model.

5.1 Collection of IMU Data

From the LESS data collected, each sensor produced three different comma-separated value (CSV) files - a high-g, low-g, and a mag file. The "g" in this case refers to a unit of acceleration, in relation to standard gravity (1 g

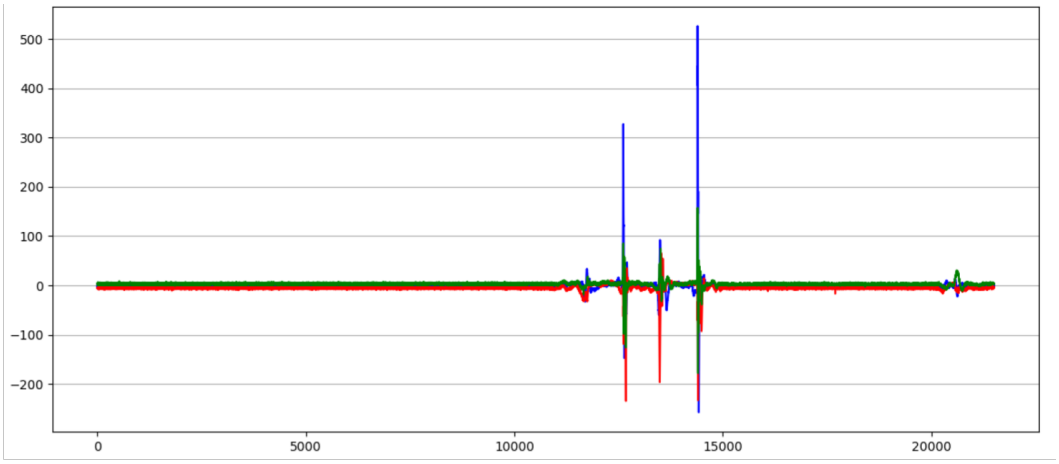


Figure 5.1: Example Accelerometer Output from the Double Leg Jump Landing

= 9.81 m/s^2). As discussed in Section 4.3.1, the high-g file corresponds to the accelerometer with a range of ± 200 g, while the low-g file refers to the movement data from both the ± 16 g accelerometer, and the ±2000 °/sec gyroscope capturing data at 1125 Hz. Lastly, the mag file corresponds to the magnetometer readings. Each of the files were labelled with the ID of the sensor and the date and time the data were collected. Each output file included the Unix timestamp in microseconds, recording the time series data in timestamp order. Collection of the movement data was manually started and stopped within a couple of seconds before and after the jump landing was performed. The windows of data collection differed between jumps, meaning the relevant data related to the jump landing could appear anywhere within the time series. An example of this time window can be seen in Figure 5.1. The x-axis is represented by blue, the y-axis represented by red, and the z-axis represented by green.

The video files and CSV files were supplied in bulk, meaning the first step was to organise the files to ensure all relevant files were processed together. At the beginning of each video, a whiteboard was held up with the ID of the participant. These IDs were manually extracted and used to pair the sagittal and frontal plane videos with the CSV files that corresponded with all six jumps.

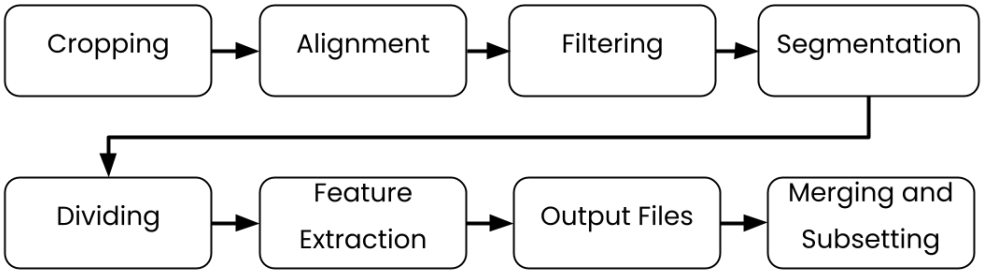


Figure 5.2: Processing Pipeline using Sensor Data

5.1.1 Manual Scoring

From the ordered directories of LESS movement data, 10 participants were randomly selected for manually scoring the LESS scoring items. Scoring was performed only on the non-intervention group of jumps, meaning only 50% of the data were used to implement the processing algorithms. Videos were uploaded and analysed using Kinovea[1]. Each jump was scored based on the 17-item LESS sheet [1]. The Kinovea coordinate system and angle drawing tool were used to evaluate each scoring item. The two key phases of the LESS were manually located by navigating through each frame of the video and selecting the point in which the participant first makes contact with the ground, and the point of maximum knee flexion upon the first landing. Scoring of each jump video took approximately 10 minutes.

5.2 Processing Sensor Data

Based on the output provided by each CSV file, an eight-step pipeline was implemented for processing the sensor data accordingly, as seen in Figure 5.2. These steps involved: (1) cropping, (2) alignment, (3) filtering, (4) segmentation, (5) dividing, (6) feature extraction, (7) output file processing, and (8) merging and subsetting. All movement data were put through each step consecutively. At each step, the CSV data files were manipulated accordingly in separate directories to keep track of how the data were processed throughout each stage. The algorithms required for implementing this pipeline were per-

[1]version 0.9.5, www.kinovea.org

formed on a subset of 30 jumps (ten participants performing three jumps). The customised algorithms created as part of this thesis are publicly available[2].

5.2.1 Cropping

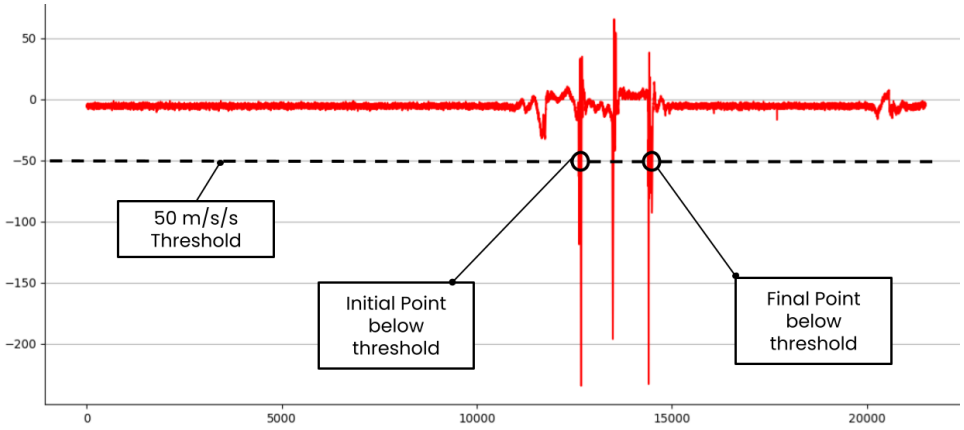
As seen in Figure 5.1, there is a large potential window of irrelevant data where a participant is standing still before and after performing the DLJL. Processing many CSV files where a majority of the file contains irrelevant data is not optimal. The time taken to process through these files will be longer, in comparison to cropping the original data to a smaller window around the relevant movement data.

Upon investigating the output of each CSV file, the key phases of the DLJL occur at random points in the data, as data collection was not consistently started and stopped at exact times. To crop the data in a consistent manner, all the files were grouped into one directory. Two main points were selected for an easy cropping process - the two main peaks observed in the first and second landing. To identify the two landing points, the high-g accelerometer data was selected. With a 1600 Hz frequency, movement data are captured every 625 microseconds, giving the greatest resolution on changes in the movement data throughout the entire movement.

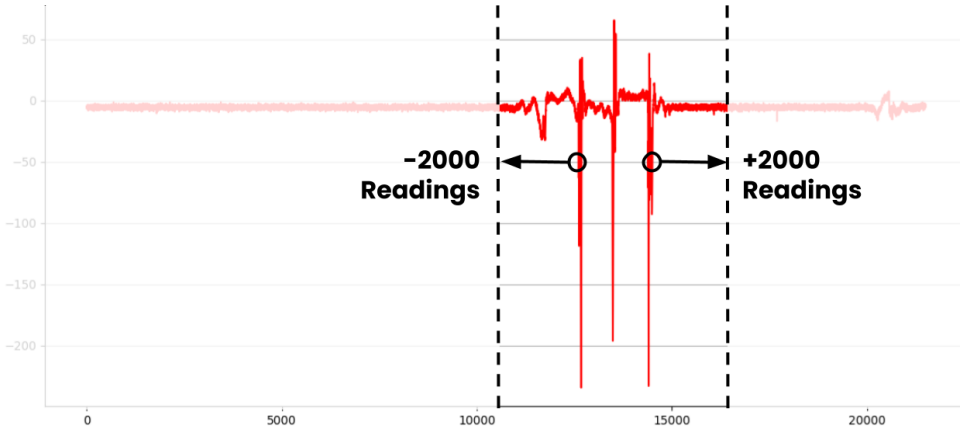
All CSV files involved in the movement were cropped consecutively. As stated in Section 4.3.1, the IMeasureU sensors capture data simultaneously between all three sensors using the Vicon iOS application. Therefore, only one of the three sensors will need to be used for identifying these landings. Impact upon landing is likely to have greater raw acceleration in the ankle-placed sensors, therefore either of the left and right ankle sensors can be used for extracting both landing timestamps. The initial cropping of the files does not need to be precise, as proper alignment of the data will be performed when segmenting the data into the key phases. A window of 2000 sensor readings (an average of 1.25 seconds of data) before and after the two landing peaks was selected to delineate cropping of the files.

To extract the timestamp of the first peak, the selected ankle accelerometer

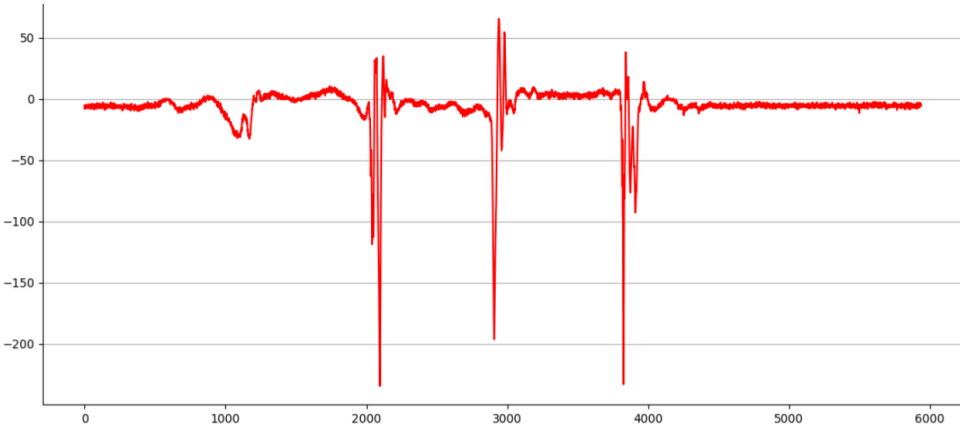
[2]See github.com/zanephamilton/Landing-Error-Scoring-System-Automation-Algorithms



(a) Extraction of Points Below Threshold Using the Y-Axis



(b) Cropping to New File Windows



(c) Example of the New Cropped File

Figure 5.3: Process for Cropping Irrelevant Movement Data

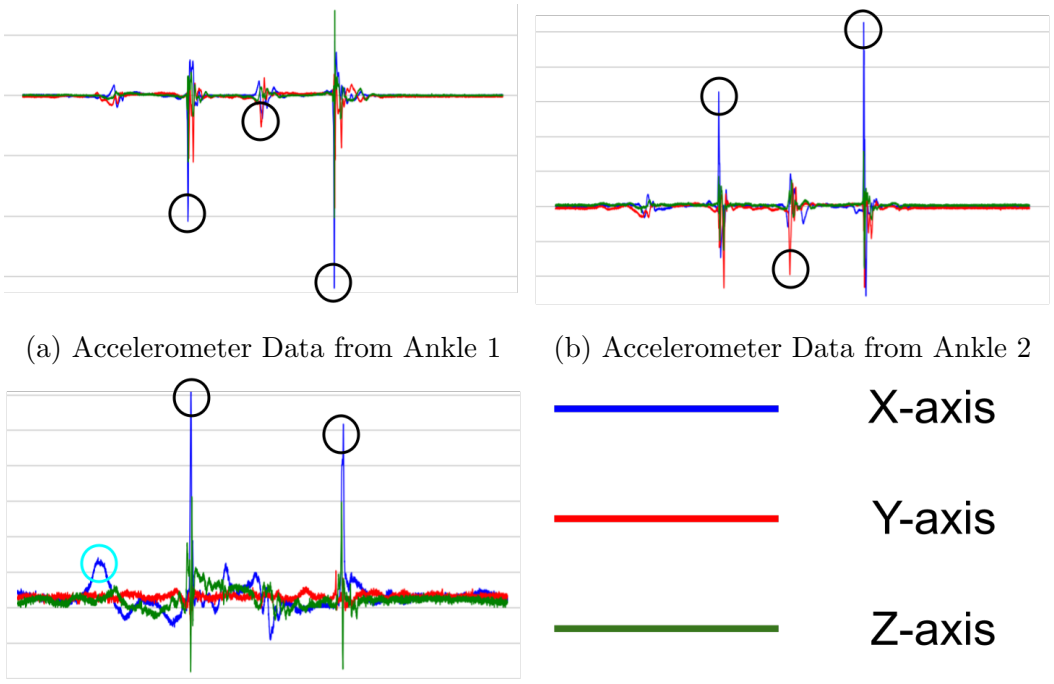
data was processed left to right to find a point below a selected threshold value. Within the current ankle accelerometer file, the timestamp and current acceleration vector in the y-axis were extracted, as the y-axis was already aligned in both ankles, meaning the same threshold can be used between both sensors. In this case, any y-axis value below -50 m/s^2 was flagged as the start of the first landing peak, as shown in Figure 5.3a. This value was low enough to avoid flagging any additional movement that may occur outside of the DLJL such as nervous movement or walking, while also being just enough to capture all of the landing peaks correctly.

Using the timestamp extracted for the initial contact, the 2000 sensor readings prior to the timestamp were stored for output. From this point, the accelerometer data were processed into the output until another x-axis value below -50 m/s^2 was observed. Following the final contact extraction, another 2000 additional sensor readings were appended to the other values. This window can be seen in Figure 5.3b. Using the two timestamps at the start and end of the cropped ankle accelerometer data, the remaining movement files were processed to find the closest absolute timestamp to these points, and cropped to the same timestamp points. An example of a cropped file can be seen in Figure 5.3c.

Following the cropping of each file, it was observed that corresponding datasets were cropped to differing lengths across the three different sensors. Despite being advertised as simultaneous data capture from the three IMUs, timestamp differences between each row of data differed between 600-670 microseconds in the accelerometer data, and 800-840 microseconds in the gyroscope data. Input data for machine learning models typically must be the same length, therefore this observation was considered for further data processing in Section 5.2.7.

5.2.2 Alignment

Each of the sensors used for data collection had an ID that related to the anatomical positioning of the sensor (left ankle, right ankle or pelvis). The sensors were placed consistently at the same location for every participant, but this location was not recorded. Therefore to align each sensor to its respective anatomical positioning, visual comparisons between three accelerometer out-



(a) Accelerometer Data from Ankle 1

(b) Accelerometer Data from Ankle 2

(c) Accelerometer Data from Pelvis

(d) Key

Figure 5.4: Visualisation for Matching Sensor Data to Position

put files were used to determine the relative location of each sensor. As seen in Figure 5.4, a clear distinction between pelvis and ankle-related accelerometer data can be observed. Large ground reaction acceleration can be observed at the ankle sensor upon initial contact, and final contact due to the proximity of the ankle to the contact points of a landing. Additionally, the accelerometer data demonstrates a third prominent peak between both contact points, whereas the pelvis accelerometer data does not. Additionally, the pelvis data shows a small peak prior to the first landing, an indication of the take-off point. Therefore, pelvis data was labelled to the data shown in Figure 5.4c.

Distinction between the left and right ankle sensor IDs using accelerometer data was not as clear as there were no indicators of how the acceleration data differed between ankles. Therefore, comparisons were made between both the accelerometer output and video for each participant. At the starting point of the movement, some participants did not take off using both legs - one leg was accelerated forwards, and the posterior leg followed a short time afterwards.

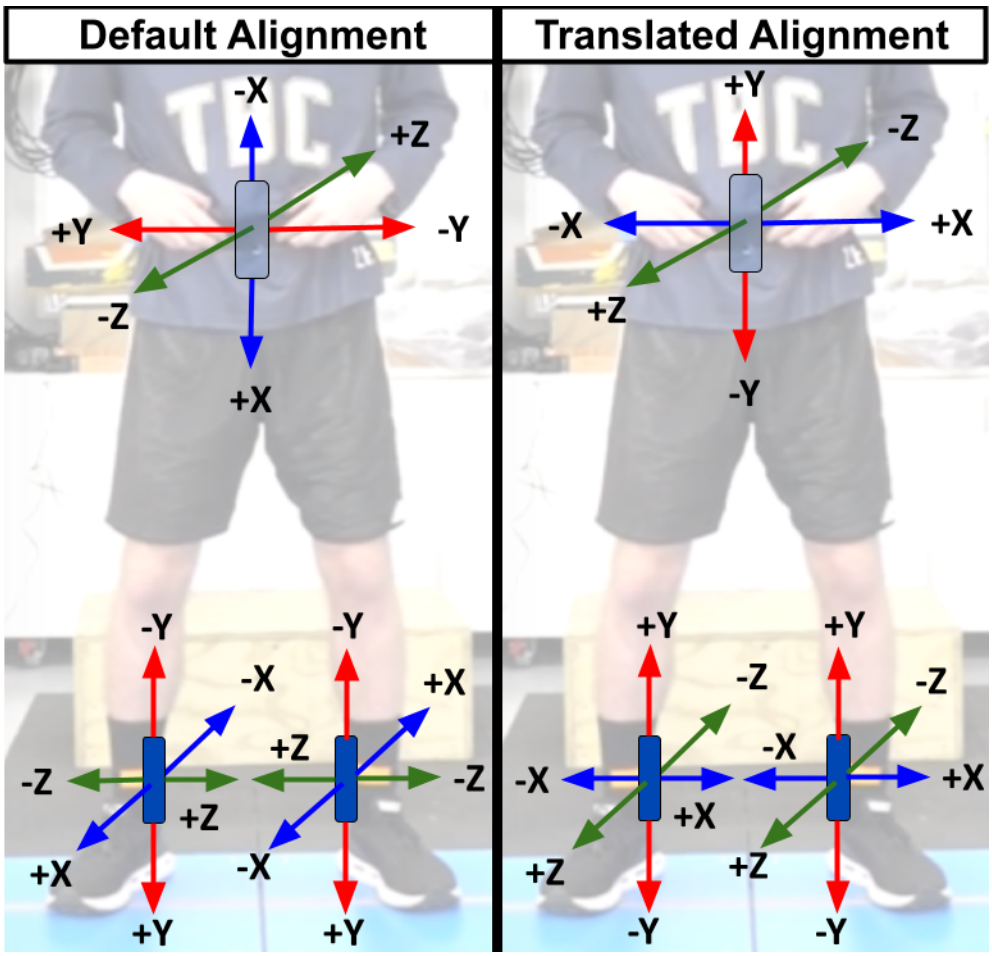


Figure 5.5: Alignment Process for Sensor Axes

Using this subset of the participants, the movement data prior to the first prominent peak was observed for any significant noise indicating a take-off point. Based on the timestamps at these points for each participant in Figure 5.4, left ankle data was labelled to Figure 5.4a, and right ankle data was labelled to Figure 5.4b.

With all three datasets labelled to their respective anatomical location, the axes of movement were evaluated. As seen in Figure 4.2a, the BlueTrident IMU uses the x-axis to determine up/down vertical movement, y-axis as right/left horizontal movement, and z-axis as forward/backward movement. In consideration of each sensor being attached in different orientations to the participant, re-alignment of the dataset was performed to ensure movement occurring at

all three sensors was performed in the same respective axes. To correctly align the datasets to this consistent axis, as set out in Section 4.3.2, the axes of each sensor must be understood. Visual observation of both the accelerometer and video data was performed to determine how each axis needed to be manipulated into the new axis.

All three sensors were attached to the participant in three orientations, as seen in Figure 5.5. Using these orientations, the axes of each sensor were drawn onto the participant to understand how each sensor was moving. Next, the axes were shifted to match the standard anatomical position of a participant. Each of these axes were then re-labelled into the consistent axis. To implement the changes of each axis label onto the actual raw data, an algorithm was created to process each column of data, and convert the data axes to the new translated alignment.

The ID of each sensor was used to perform the correct manipulation for each sensor, and whether the current file was an accelerometer or gyroscope file. The manipulation process can be seen in Figure 5.5. In this case, the left ankle sensor was aligned by swapping the X and Z-axes and inversing the values, so left and right movement at the ankle were represented correctly. The right ankle required similar alignment, with the x-axis being swapped with the z-axis. Finally, the x-axis was swapped with the y-axis in the pelvis data, and then all three axes were inversed to correctly represent each direction of movement. With the algorithm changes implemented, the data could now be consistently evaluated in the same axis of movement and therefore, segmentation at key points can be performed more consistently than using individualised axes.

5.2.3 Filtering

As discussed in Section 3.2.2 and Section 4.3.2, the use of a fourth-order Butterworth filter on jump and landing-related data was prominent, but a range of critical frequencies were used to filter the respective data. To select a critical frequency for this dataset and evaluate the application of different orders, data was filtered using a range of frequencies (10-100 Hz) and orders (1 to 10). Selection of a frequency and order was based on the smoothing of the data, the elimination of the noise without significant loss of magnitude and extent

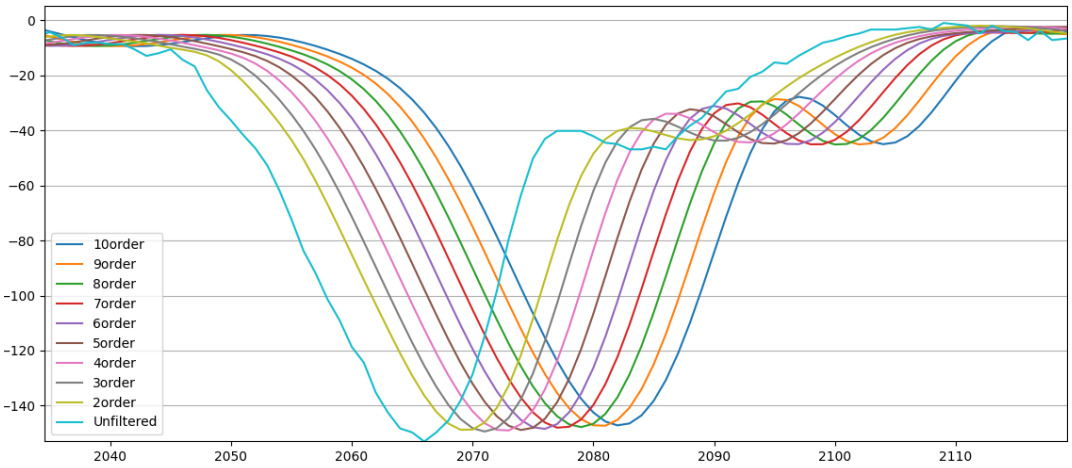


Figure 5.6: Difference in Butterworth Filter Orders

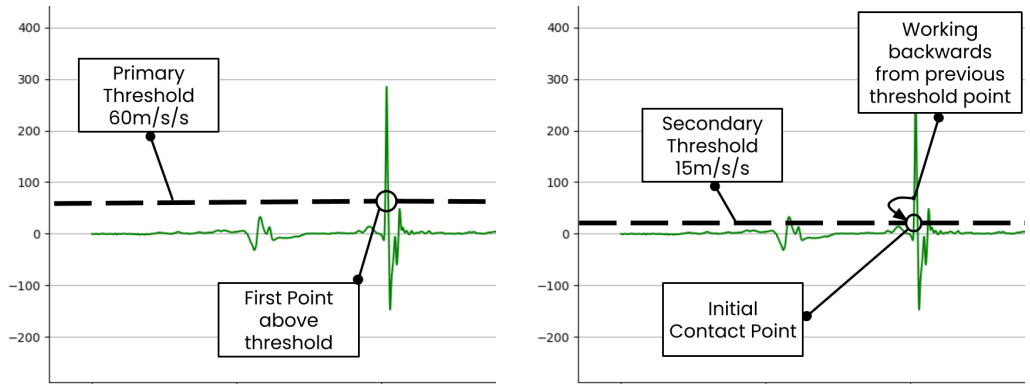
of a time shift. As the filtering frequency was decreased from 100 Hz, there was a loss of magnitude in the main peaks of the movement data. At 100 Hz, the majority of the acceleration magnitude was maintained and the noise present in the raw signals smoothed. To evaluate the application of a fourth order filter to this dataset, a comparison between the orders can be seen in Figure 5.6. As the filter order increases, a larger time-shift is observed as the data and peaks move. Based on the success in related work and the minimal shift in the data, a fourth order was applied to the entire dataset.

To filter the IMU data using a Butterworth filter, the SciPy signal processing library[3] was used. To implement this filter, a simple algorithm was implemented to loop through the new directory of aligned IMU data. Accelerometer and gyroscope data was processed separately to ensure the correct labels were used. The filter is set up using a selected order, and a critical frequency. The filter is applied to each axis individually, and the output of these filtered values are used.

5.2.4 Segmentation of Key Events

As discussed in Section 4.3.2, there are three key events of interest in the DLJL. This consists of the take-off from the box, initial contact with the

[3]version 1.13.1, https://docs.scipy.org/doc/scipy/reference/generated/scipy.signal.butter.html



(a) Locating Initial Landing (b) Locating Value Below Threshold

Figure 5.7: Process for Extracting Initial Contact Point

ground, and the point of maximum flexion in the knee. Identification of these events within the accelerometer data needs to be correctly implemented to ensure the algorithm is accurate in extracting each phase of the movement for a wide range of different jump and landing conditions. The algorithm was implemented to find all six files corresponding to one jump occurrence (accelerometer and gyroscope files from all three sensors) and process them at the same time.

Initial Contact

The first step for segmenting the data involved the evaluation of an initial contact point. In consideration of possible asymmetric contact - the timestamp of initial contact observed in the left ankle accelerometer may be different to the right ankle accelerometer. Therefore, an algorithm was implemented to perform the same processing on both the left and right ankle data.

The extraction of the initial contact event can be seen in Figure 5.7. The initial contact event is easily identifiable in the data by the first prominent peak, as seen in Figure 5.7a, due to the quick deceleration experienced when coming into contact with the ground. However, this peak is not representative of the initial contact point as the participant has likely been in contact with the ground for long enough to further load the body. Therefore, the point at which initial contact occurs is toward the start of the ascending portion of this peak.

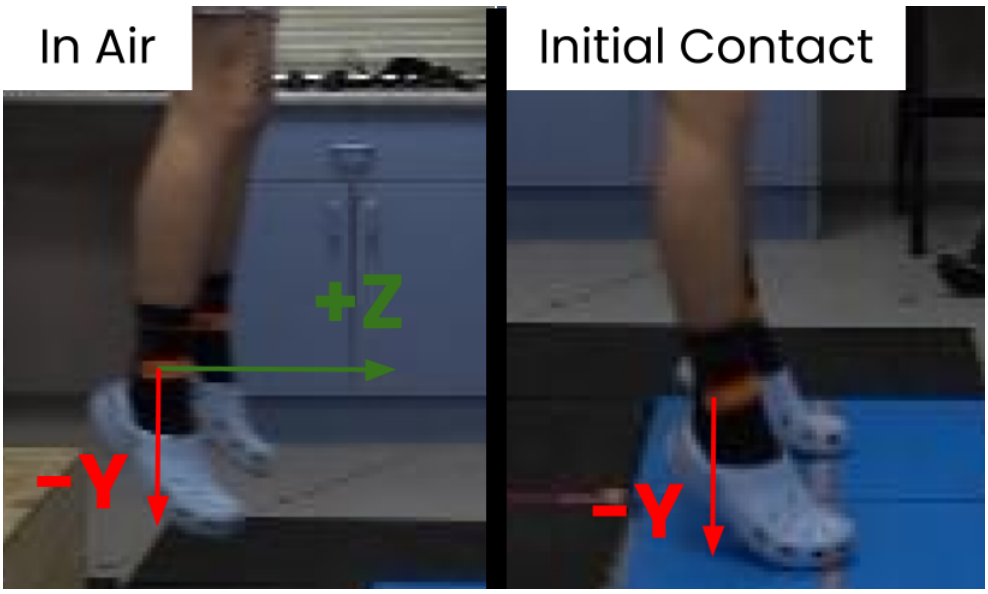
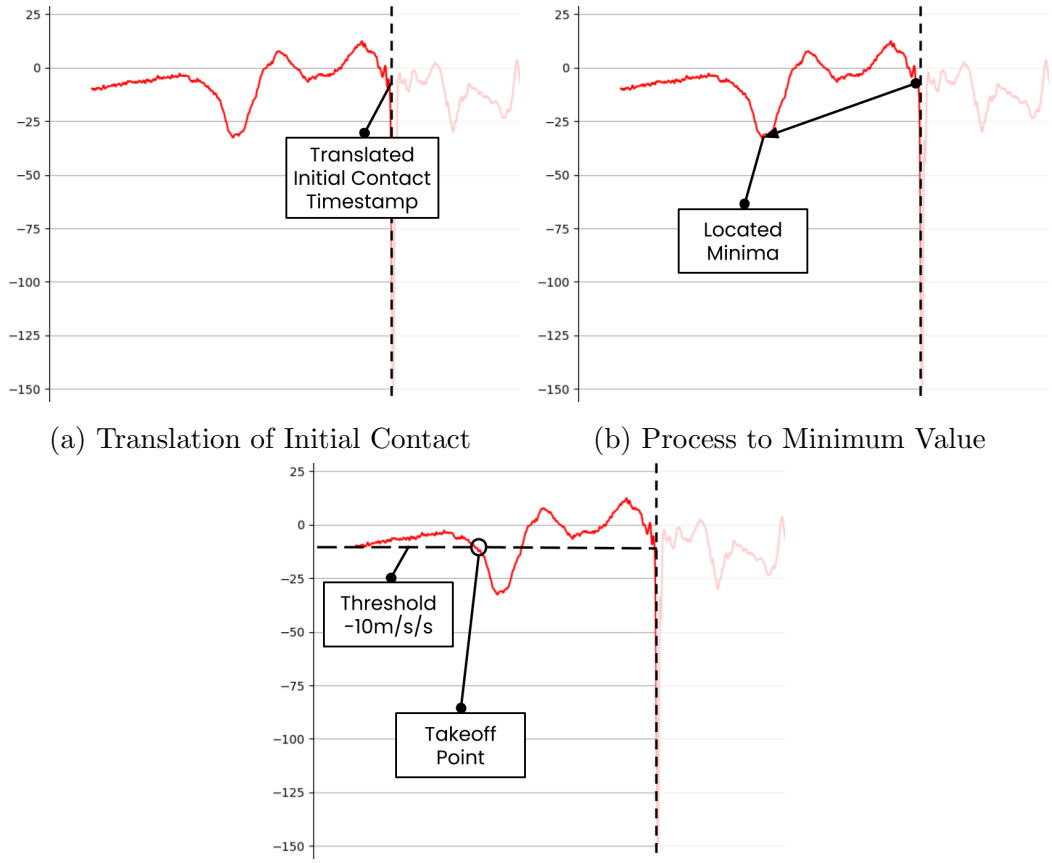


Figure 5.8: Movement of Sensor Upon Initial Contact

To get the point of initial contact from a sensor, relevant information about the raw data is extracted. This process involves using the full dataset, a starting index point, and a threshold that acts as a cut-off point at which a peak is identified, without the risk of being mistaken for additional noise or irrelevant data.

Firstly, the algorithm evaluates the values within the z-axis. As seen in Figure 5.8, the z-axis will observe a peak prior to the y-axis as the top of the foot is in contact with the ground before the heel. This position means the forward motion of the sensor will be in contact first and decelerate before the downward vertical motion of the sensor.

Once a value above 60 m/s^2 is observed, the peak has been located, but the starting point needs to be located instead. Processing continues until a secondary threshold close to 0 m/s^2 is reached, as seen in Figure 5.7b. A secondary threshold value of 15 m/s^2 was selected to consider any noise and gravitational influence occurring in the z-axis. At this point, the timestamp is extracted and used for further processing.



(a) Translation of Initial Contact

(b) Process to Minimum Value

(c) Located Threshold Point

Figure 5.9: Process for Extracting the Take-off Event

Box Take-off

When taking off for a jump, a small peak is observed in the pelvis due to the biomechanics in the lower extremities. With flexion occurring at the knee and hip for a take-off phase, the participant moves downward in the y-axis. The downward movement of the pelvis is likely to reflect the beginning of a take-off as a proxy to the displacement of the centre of mass of participants. Muscle mass of the legs is also concentrated more heavily toward the hip, with much less near the ankle [118]. Both the left and right ankle will demonstrate some degree of noise prior to the initial contact, but this does not represent the true beginning of the take-off. Therefore, a take-off point should be identified using the pelvis movement data. The retrieval process of the take-off point can be seen in Figure 5.9.

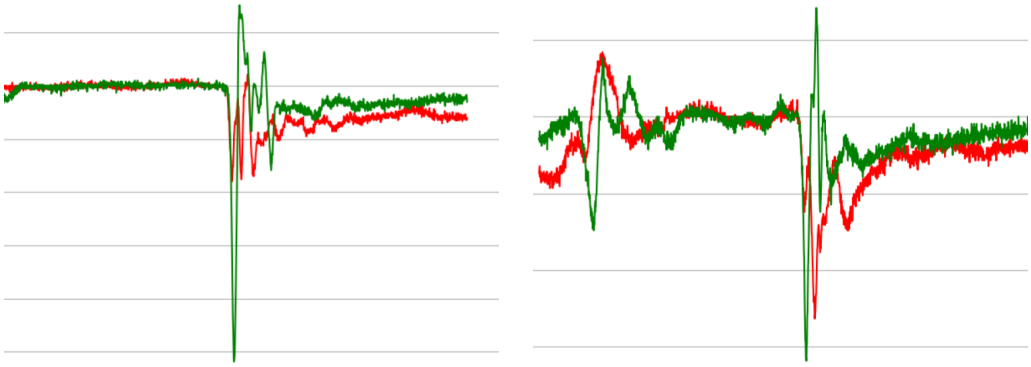
For retrieving the take-off point, either of the initial contact timestamps (left ankle or right ankle) can be used to find the initial contact timestamp in the pelvis accelerometer data, as seen in Figure 5.9a. As discussed in Section 5.2.1, the timestamps between all three sensors in the accelerometer files differed in a range of 600-670 microseconds. This difference meant that finding the same timestamp within another accelerometer file was not possible, as data could not be captured in all three sensors at the exact same time. Therefore, an additional function was implemented to find the closest absolute timestamp in a file and find the index value of that timestamp for further processing of the pelvis data.

Using the closest timestamp to initial contact from the pelvis data, the pelvis file is processed backward from this point. To jump from the box, downward movement of the pelvis is observed in the accelerometer data. Therefore, the minimum value in the y-axis between the beginning of the file, and the initial contact event timestamp was used for identification of the take-off event, as shown in Figure 5.9b.

Similar to initial contact, the peak of the signal for the pelvis does not represent the start of take-off, rather it represents a portion of the movement. Therefore, a threshold of -10 m/s^2 was used in consideration of gravitational influence and noise in the y-axis. Once -10 m/s^2 was reached, the index of the take-off point and the equivalent timestamp was extracted, as shown in Figure 5.9c.

Maximal Knee Flexion

Detection of maximal knee flexion in the three accelerometer datasets was not a clear-cut process. Without the addition of a knee-based sensor, exact calculation of the knee angle throughout the DLJL is limited. In both the ankles and their respective accelerometer data, a peak in data from any axis is an indication of the second take-off point as the ankle does not move until the participant has left the ground. However, a large incline of acceleration in the y-axis is a potential representation of maximal knee flexion. As the participant reaches their point of maximal knee flexion, movement at the pelvis will observe a turning point - where the pelvis begins to accelerate vertically upwards to take off again.



(a) Prior to Maximum Knee Flexion (b) Following Maximum Knee Flexion

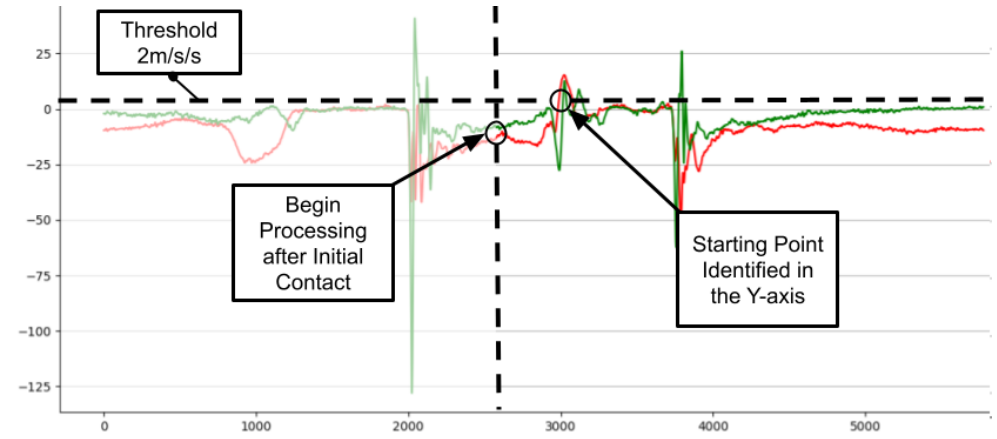
Figure 5.10: Visualisation of Maximal Knee Flexion in the Pelvis Accelerometer using Video Timestamps

To verify this theory, the timestamps from the manual scoring process were used. The accelerometer data was visualised using the Matplotlib[4] Python library. This data was visualised in two halves: the accelerometer readings before and after the timestamp where the participant was considered to be in maximal knee flexion. An example of this visualisation can be seen in Figure 5.10.

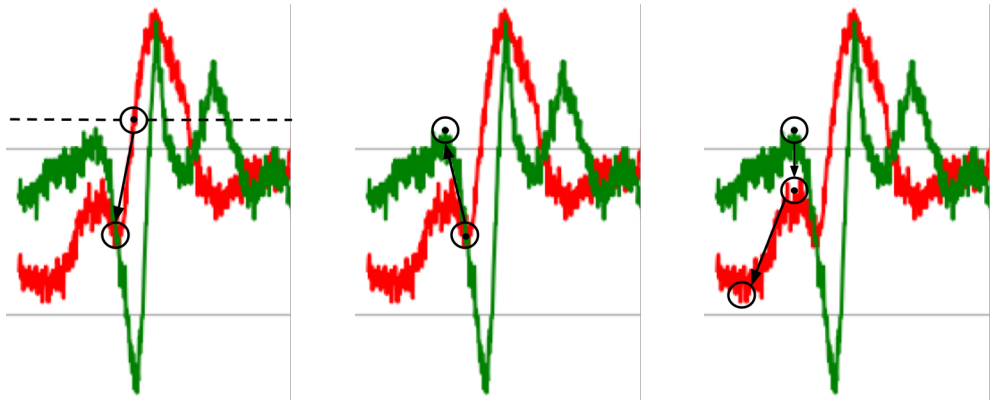
In some cases, the incline was not one peak - rather it was split into two or more quick and progressive peaks. An example can be seen in Figure 5.10b. Therefore, the algorithm also processed the movement using the z-axis. As seen from Figure 5.10a to Figure 5.10b, the z-axis also observes a sudden peak in acceleration. Due to the orientation of the trunk in a jump landing (as indicated by scoring item 3 and 14), trunk flexion will affect the orientation of the pelvis sensor (where the z-axis of the sensor may now represent half horizontal and half vertical motion). Forward and backward motion of the pelvis (occurring in the z-axis) will reflect some of the gravitational acceleration and the extension of the trunk as the second take-off point approaches. Therefore, both axes can be used together to identify a point relative to the timestamp of maximal knee flexion.

A visualisation of the maximal knee flexion segmentation can be seen in Fig-

[4]version 3.8.3, https://matplotlib.org/



(a) Initial Identification of Threshold Point in the Y-Axis



(b) Threshold to Minimum (c) Minimum to Maximum (d) Maximum to Minimum

Figure 5.11: Visualisation of Maximal Knee Flexion Segmentation

ure 5.11. After using the smaller initial contact timestamp between the left and right ankle data for the take-off point, the larger timestamp is used to process forwards to detect maximal knee flexion. A small buffer was used to begin the processing away from the initial peak, as there is additional sensor movement as the participant begins descending towards maximal knee flexion.

Once a participant begins accelerating upward for the second take-off, acceleration will begin to head toward 0 m/s^2 as positive y-axis acceleration represents upward vertical motion, as established in Section 4.3.2. Therefore, the y-axis is processed first. A threshold level of 2 m/s^2 was used for identifying the first point on the maximal knee flexion incline.

In the first stage of detecting the minimum, the first value above a 2 m/s^2 threshold value was considered as the starting point for processing a peak. Using this value, the pelvis data is processed backward until acceleration was no longer decreasing. A timeout of 50 sensor readings was selected as it produced the closest accuracy to the video-based maximal knee flexion points in the jump subset used. Once the timeout was reached, this point was established as the first minima of the y-axis.

The z-axis is then used for the second stage of processing, seeking to detect for a maximum. Forward movement in the pelvis is represented by positive acceleration values, as established in Section 4.3.2. This function uses the same logic as the minimum detection, but any point greater than the previous value is stored and returned after a timeout counter of 50 sensor readings. This maximum is then used in the final detection process for extracting the final minima value in the y-axis. This final minima is used as the point at which maximum knee flexion occurs.

5.2.5 Dividing the Data

The key events derived from the movement data were used to divide the data files into their respective segments. The calculated index values were used to make slices of the movement data into smaller-length segments. The key events corresponded to two segments: (1) take-off to initial contact, and (2) initial contact to maximum knee flexion

Firstly, the key events as calculated using the pelvis-related IMU data was translated into index values for the left and right ankle files respectively. These key events included the take-off and maximal knee flexion timestamps. Each individual ankle accelerometer file was used to find the closest absolute timestamp related to the key events, and the corresponding index value at that event.

This process was also replicated for the key events as calculated by the ankle accelerometer data: the initial contact and final contact points. In this case, one initial contact timestamp may occur later than the other, as asymmetric contact is possible, as stated in Section 4.3.2. Initial contact in the pelvis relates to the point at which at least one foot is in contact with the ground.

This distinction was selected as the pelvis must consider both legs coming into contact with the ground as potential points of initial contact, whereas a single ankle sensor only collects movement data in relation to one point. In this case, the initial contact points for both the left and right ankle are compared. Whichever foot first makes contact with the ground first was used as the initial contact point in the pelvis movement data.

In consideration that the gyroscope operates at a lower sampling rate, the index values calculated for the accelerometer data does not correlate with the gyroscope data, as data are collected less often (difference of roughly 260 ms). Therefore, the timestamps calculated from the accelerometer data are used for the three gyroscope files of each sensor, and the closest absolute timestamp is used to derive index values that correlate to the same relative timestamp point in the gyroscope data.

When creating each segment slice, overlapping movement data between segments was avoided using the derived key event as the first value of a segment. For example, the first timestamp of a S1 slice would be the take-off timestamp, and the final timestamp would be the closest timestamp of data prior to the initial contact timestamp.

5.2.6 Feature Extraction

As discussed in Section 3.2.2, there are a number of potential temporal and statistical features that can be derived from IMU-related movement data. To implement the selected features outlined in Section 4.3.2, temporal features can be derived using the derived index values of each segment. Statistical features present a different challenge, where features can be derived from the two key phases of the DLJL, at three sensor locations, within three different movement axes. Feature importance can be evaluated which allows assessment on the implemented features that produce greater or weaker accuracy in the selected models. All the derived features were stored so they could be later merged with the LESS scores, participant characteristic data, and movement data.

To derive the selected temporal features from the segmented phases, the timestamps as calculated from segmentation are used. In consideration that these

timestamps are all stored as Unix timestamps in microseconds, two timestamps can be subtracted from each other to get a microsecond representation of the time taken in each phase of the movement.

In the case of deriving statistical features, each axis of all the sensors was used, meaning a total of six features were calculated for every axis (nine potential axes). For each segment, this total came out to 54 potential statistical features for every axis within the movement data. All the accelerometer and gyroscope data within one axis were extracted to calculate these features.

Firstly, the root mean square value was extracted using a function that calculated the square of all movement data that appeared in that axis, calculated the mean value of that square value, and returned the square root. For extracting variance, the NumPy var function[5] was used. The respective functions from the Python statistics library[6] were used to derive the mean and standard deviation. Finally, the maximum and minimum values were extracted using the Python max and min functions.

5.2.7 Processing Output Files

With the features extracted and movement data divided into smaller segments of the movement, new files were used to keep track of each participants movement data and further integration with machine learning models as input data. Output was produced in two main stages: (1) divided segment output, and (2) feature output.

Accelerometer and Gyroscope Segment Output

Following segmentation and dividing the data into segments, each jump now had nine total segment slices (three segments per three sensor locations). Considering all the segments were consistent, it was decided to merge the movement data of the left ankle, right ankle, and the pelvis into one combined CSV file under the participant ID and jump number. An example of an output file can be seen in Table 5.1. Each output file corresponded to one segment, with nine different columns corresponding to each axis of data captured.

[5]version 2.0, https://numpy.org/doc/stable/reference/generated/numpy.var.html
[6]version 3.12.4, https://docs.python.org/3/library/statistics.html

Table 5.1: Format of an Accelerometer File with all Axes of Movement

lax	lay	laz	rax	ray	raz	pax	pay	paz
1.044	-1.510	6.059	4.953	-3.293	2.444	-8.561	-10.106	-6.462
1.017	-1.584	6.127	4.670	-3.204	2.596	-8.551	-10.333	-6.467
1.018	-1.612	6.205	4.424	-3.158	2.665	-8.548	-10.549	-6.357
1.047	-1.588	6.282	4.237	-3.135	2.636	-8.559	-10.721	-6.178
1.113	-1.510	6.347	4.125	-3.118	2.525	-8.575	-10.834	-5.995

The main problem to consider in this case, was the difference in segment lengths due to the movement data not being collected consistently every 625 microseconds, as discussed in Section 5.2.1. It is also important to note, that using the raw data for machine learning input requires a consistently-sized dataset. To combat this problem, the lengths of the three segment files were compared for every segment. Using the maximum length out of the three sensor files, the decision was made to pad empty columns with a consistent value of "-1." A consistent value acts as an identifier of the end of the data stream, which allows for the data to be the same length, without losing any additional data during processing, especially around key points of the movement changing between each segment.

List of Derived Features

Considering that the derived features from each participant were extracted at the same time as segmentation, the features were output following the processing of all participants. Following segmentation and feature derivation, all the feature data was output to a separate CSV file of all the features for the entire dataset. The CSV file was formatted simply using two main "groups" of columns, one for all the temporal features, and the rest for the statistical features. Each row was identified with two columns for the participant ID and jump number, for further use in merging and subsetting the three main datasets together.

5.2.8 Merging and Subsetting Multiple Datasets

With a file for each segment of the DLJL, a file with all the derived features and a file of the LESS scores, each dataset had to be merged correctly. Additionally, the subsets to be tested, as highlighted in Table 4.2, involved the sensor positions (both ankles and pelvis individually), and the differing segment groups (S1A2 and S2).

The first process involved creating new merged files of S1A2. As highlighted previously in Section 5.2.7, the consistent value "-1" was used to make up for missing data where the three sensors capture rate did not precisely align. Since S1A2 is one consistent window, the "-1" values at the end of the S1 dataset must be removed, and the beginning values of S2 must be added.

To merge the two segments, an algorithm was created to work backwards through the two segment files and remove the filled values until a valid row of acceleration or rotational velocity has been reached. This algorithm also avoids replacing a cell that actually has an acceleration or rotational velocity value of -1 m/s^2 as all three cells must be filled to be replaced. With S1A2 now joined, the size of each sensor dataset must be consistent, as previously discussed in Section 5.2.7. Using the same process as Section 5.2.7, the maximal length of the three sensors was taken and the remaining sensors were filled with "-1" to ensure the lengths were matched.

For merging the two segment subsets (S1A2, or S2 alone), an algorithm was created for merging and subsetting the data. For an easier indexing process, the desired output for the machine learning models (the manually scored LESS scores) were selected as the first column group of the merged file. The second group were the feature-related columns, with temporal measures first followed by the statistical features. Lastly, the raw sensor data of the jump occurrence was the last group of columns. To match all three files together, the participant and jump IDs were used to match each one to a row. Each row was written at a time, with all three files being merged in the same process.

As highlighted in Section 5.2.8, the combined accelerometer and gyroscope files have nine columns made up of a large number of rows per segment. Therefore, the decision was made to flatten the movement data into one array. This

decision meant for every nine columns of movement data, three groups of three made up the sensors and their respective three axes of movement data. Considering there are both accelerometer and gyroscope data to integrate, accelerometer data was flattened first. Once the end of the file had been reached, the gyroscope data was flattened second, and inserted onto the end of the accelerometer data.

For subsetting into three positional-related subsets (all three sensors, just the ankles, or just the pelvis), as covered in Section 4.3.2, the merging algorithm had to consider the columns we wanted to implement before the dataset was flattened. Therefore, depending on the current subset selected, the merging algorithm only flattened values that were relevant using the headers set for each column. For the full dataset of all three sensors, this merging involved nine columns, while an ankle-only subset involved six columns, and a pelvis-only subset involved just three columns of movement data per sensor reading. The same process was implemented for the feature-related data, where the column headers contained a relevant ID for the segment the statistical feature corresponded to. All temporal measures were included for every subset to evaluate whether the time taken to get to a key phase of the movement influenced further phases.

With the nature of time-series data, timing of participants to complete different segments of the movement varied. To ensure all the columns were the same length, a further algorithm was implemented to find the participant with the largest time window for both the accelerometer and gyroscope respectively. The length for both the maximum accelerometer length and gyroscope length were used to fill all remaining rows with "-1" up to the required lengths to ensure the length of columns were consistent for all participants.

5.3 Machine Learning Integration

Following the use of the merging algorithms, there were six main dataset files to be used - including three positional-related datasets for both S2 data, and the merged dataset of S1A2 respectively. Model architecture was implemented, and an appropriate algorithm was made to extract the performance measures

on each iteration of classification.

5.3.1 Evaluation Architecture

As discussed in Section 4.3.3, Scikit-learn was selected for implementation of the machine learning models. To preserve an out-of-the-box approach using the Scikit-learn classifiers, each classifier was initialised with no additional parameters, leaving the set up of these classifiers to the default parameters as set out by the documentation[7].

Each of the subsetted data files were used as input for classification. As shown in Table 4.2, the input data can be split into further subsets to evaluate the performance of the IMU-derived measures, the raw movement data, or both combined. These subsets include using just the raw movement data, the temporal and statistical features, or both. To get the correct set of input data, the full dataset was sliced into smaller subsets. Depending on the subset, the algorithm returned the appropriate slice of the data being evaluated. The manually scored LESS scores and participant characteristics were all in individual columns, therefore these values were retrieved using their index values. With input data and desired output now split appropriately, the training of the selected machine learning models was completed.

5.3.2 Extracting Performance Measures

Each scoring item was evaluated using a score-by-score process. All eight selected classifiers were initialised and passed to a separate algorithm to perform five-fold cross validation. This function can be seen in Code Listing 5.1. Cross-fold validation allowed sampling of a limited data sample into five different subsets. Five folds provided a 1/5 test split, meaning 80% of the data was used to train each model, while 20% was used as a test set to evaluate the performance of the model for classifying a scoring item. On each fold, the model was re-initialised again to ensure the same model was not being trained multiple times on the full dataset. This process was performed five times, and the average performance metrics were returned.

[7] version 1.4.1, https://scikit-learn.org/stable/supervised_learning.html

Code Listing 5.1: Five-Fold Cross Validation and Performance Extraction

```
...
raw_classifier = sklearn.base.clone(classifier)

for train_index, test_index in cv.split(X_set, y_set):
    #Resetting model to default on every fold.
    classifier = sklearn.base.clone(raw_classifier)

    X_train, X_test = X_set.iloc[train_index], X_set.iloc[
        test_index]
    y_train, y_test = y_set.iloc[train_index], y_set.iloc[
        test_index]

    classifier.fit(X_train, y_train)
    predictions = classifier.predict(X_test)

    full_report = classification_report(y_test, predictions,
        target_names = class_names, output_dict = True)

    #Storing all metrics for averaging, and output.
    accuracy_store.append(full_report['accuracy'])
    ...
```

As stated in Section 4.3.3, the metrics selected for evaluating the performance of the different models were precision, recall, and the F1-score. The metrics stored included the individual performance of label classification (precision, recall and F1-score per class), and a weighted average of performance across all the classes. For each cross-fold, these metrics were stored, and once five folds were completed, the performance metrics were averaged. The selected metrics were output into two separate CSV files; one for the weighted average values, and one for the individual label performance. These files were then used for observing performance results and evaluating the viability of the different subsets in predicting LESS scores.

Chapter 6

Results and Evaluation

Using the full dataset collected for the DLJL, an evaluation can be completed on the performance of the selected machine learning models for automating the LESS scoring process. Using the data processing and integration as set out in Chapter 5, the performance metrics within each data subset can be integrated into the processing pipeline, and output can be analysed to determine the best performers across the entire scoring table. Using the best performing models, a further evaluation on the features that have the greatest importance in classification of a score can be extracted and discussed.

In this chapter, we cover the evaluation process for assessing the performance of the trained models in the LESS scoring items, and how these results relate to the performance of movement. This chapter begins with a breakdown of the methodology used for evaluation, including the characteristics of participants, the spread of available data, and the metrics used for evaluating each model. Following the methodology, the results are highlighted and shown for each scoring item. Lastly, these results are summarised and the top performing features are discussed in relation to the real performance of a movement pattern.

Table 6.1: Characteristics of Participants

Characteristics	Males (n= 21)	Females (n = 17)	Other (n = 2)	All (n = 40)
Age (y)	20.0 ± 3.9	20.8 ± 4.1	18.4 ± 0.03	20.3 ± 3.9
Height (cm)	183.1 ± 5.8	168.6 ± 6.0	166.5 ± 0.7	176.1 ± 9.3
Mass (kg)	81.6 ± 12.7	70.0 ± 10.1	64.1 ± 2.1	75.8 ± 12.8
Previous Injury	14	10	2	26

6.1 Data Collection

To evaluate the performance of the machine learning models in classifying a scoring item, the entire dataset was fed through the data processing pipeline, as discussed in Section 5.2. A total of 218 different jump occurrences were successfully processed (i.e., five to six jumps per participant). The subsets as generated from this step were individually integrated with the machine learning models to return the overall performance metrics of each model.

6.1.1 Participants

Forty participants (21 males, 17 females, 2 other) were involved in this study and provided movement data and LESS scores for this thesis. Participants were university students who volunteered to participate in data collection. Demographic characteristics of the participants can be seen in Table 6.1. On average, participants were involved in moderate physical activity three times per week. Twenty-six participants had history of injury prior to data collection, but these were resolved and they could participate in data collection safely. All participants were involved with sport (20% Gym/Weightlifting, 15% Netball, 12.5% Rugby, 7.5% Basketball, and 55% Other Sports).

Table 6.2: Distribution of Classes in the Landing Error Scoring System From 218 Jumps

(a) Overall Landing Error Scoring System Score Counts

Total Score	1	2	3	4	5	6	7	8	9	10
Counts	2	2	4	22	52	54	47	17	14	4
Distribution (%)	0.9	0.9	1.8	10.1	23.9	24.8	21.6	7.8	6.4	1.8

(b) Binary Landing Error Scoring System Item Counts

	Scoring Item	Absent (n)	Present (n)	Distribution (%)
1	Knee flexion at initial contact	39	179	82
2	Hip flexion at initial contact	218	0	100
3	Trunk flexion at initial contact	218	0	100
4	Ankle plantar flexion at initial contact	180	38	83
5	Knee valgus at initial contact	62	156	72
6	Lateral trunk flexion at initial contact	186	32	85
7	Stance width (wide)	205	13	94
8	Stance width (narrow)	106	112	51
9	Foot position (toe-in)	94	124	57
10	Foot position (toe-out)	217	1	100
11	Symmetric foot contact at initial contact	136	82	62
12	Knee flexion displacement	202	16	93
13	Hip flexion at maximal knee flexion	217	1	100
14	Trunk flexion at maximal knee flexion	194	24	89
15	Knee valgus displacement	56	162	74

(c) Multi-Class Landing Error Scoring System Item Counts

	Scoring Item	Absent (n)	Present (n)	Penalty (n)	Distribution (%)
16	Joint displacement	69	144	5	66
17	Overall impression	6	197	15	90

6.1.2 Data Spread of Landing Error Scoring Items

The value counts provided by the manual scoring of the LESS can be seen in Table 6.2. For classifying the overall total score of the DLJL, a total of ten scores were available for classification, as seen in Table 6.2a. The shape of this data follows a bell-curve, with majority of the class instances being overall scores between 5-7. No scores above 11 or scores of 0 were available from this dataset.

For the individual scoring items in Table 6.2b, 11 out of the total 15 binary scoring items were able to be used for training the selected machine learning models. Scoring items 2, 3, 10, and 13 did not provide enough data to classify whether an error was absent or present. Collecting more data does not guarantee these errors will be present and would likely require hundreds of additional participants to identify these errors - therefore, the decision was made to omit these scoring items from classification. Both multi-class scoring items in Table 6.2c were included for classification. Therefore, the classification of 14 different scores (overall score, and 13 individual scoring items) was available for performance evaluation.

6.1.3 Model Performance Evaluation

All of the seven selected machine learning models were trained on the 18 different subsets of data. This training was completed on a score-by-score basis, which meant there were 126 total models created and trained to classify the same score. All of the classifiers were initialised with a consistent random state value to ensure the best performing model could be replicated for extracting the top performing features. The best performing model was selected by comparing the weighted F1-score of all 126 models, as this metric gives a harmonic mean that also considers imbalanced datasets, as discussed in Section 4.3.3. Using the best performing model, the precision and recall values for each of the predicted classes were compared and discussed in relation to the performance of ZeroR.

Additionally, the best performing model for all scoring items were used to extract the features that had the greatest impact on prediction. The Kernel-

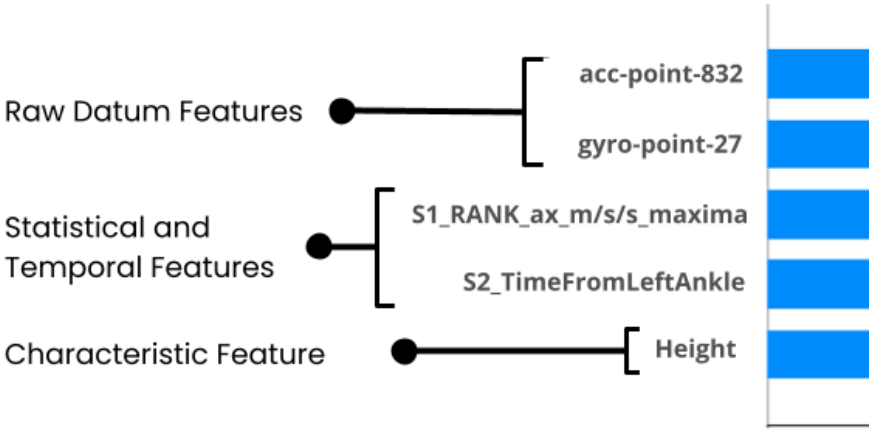


Figure 6.1: Formatting of the Extracted Top Performing Features

Explainer function, provided by the SHAP library[1], was used to extract the top five features from each model, which assigns a weight to every feature within a model, dependent on how much each feature influences the probability of a class. These features were evaluated in relation to the performance of the scoring item, and whether they provide confidence in accurate classification with future LESS datasets. In scoring items where the best performing subset of data included the raw movement data, the SelectPercentile function from Scikit-learn was implemented to extract 1% of the dataset. The 1% was selected based on the features that provided the greatest relationship to the target LESS score.

Each of the top five features were extracted and formatted depending on the type of the feature, as shown in Figure 6.1. For a raw movement data feature, both the sensor type and the dataset index of the data point was returned. For statistical features, the output string was formatted with five main values: (1) the segment of data the feature was extracted from, (2) a four letter code for the sensor position ("LANK" for the left ankle, "RANK" for the right ankle, and "PELV" for the pelvis sensors), (3) the sensor type and axis the feature was extracted from, (4) the unit of data returned (m/s^2 for an accelerometer, deg/s for a gyroscope), and (5) the statistical feature. Additionally, temporal features detail the phase of movement, and whether specific phase times are

[1] version 0.45.1, https://shap.readthedocs.io/en/latest/generated/shap.KernelExplainer.html

Table 6.3: Model Performance for Classification of Total Overall Score

Model	Processing Subset			Accuracy (%)	F1-Score (%)
	Data	Segment	Sensors		
ZeroR	-	-	-	25.0	10.0
SVC	Raw	S2	Ankles	25.7	20.7
KNN	Raw	S2	All	25.7	22.4
DFF	Raw	S2	Ankles	23.8	23.0
RFC	Stats	S2	Pelvis	28.4	24.0
GNB	Both	S1A2	Ankles	27.0	24.9
SGD	Raw	S1A2	All	27.0	25.6
GBC	Raw	S1A2	Ankles	27.9	26.3

to/from the left ankle initial contact, or the right ankle initial contact. Lastly, all the characteristic features are output as normal.

6.2 Results

Using the results provided, an evaluation of the metrics of the best performing machine learning models can be made across the total score, and all scoring items. This evaluation involves assessing the best model for each respective scoring item, making comparisons in performance to the ZeroR classifier, and the five features that demonstrated the largest importance for each particular scoring item.

6.2.1 Total Score

The best performing models from all seven selected models for classifying the total score of the LESS are shown in Table 6.3. The best performing subset of data differed between models. These differences are likely due to the different algorithms employed by each model. Some features from one subset may have greater impact in one model than another. The KNN produced the highest F1-score when using features from the raw movement data from all three sensors in S2. The GNB used both raw and statistical features in S1A2 from the ankle sensors. The SVC used the features from the raw movement data provided by

the ankle sensors in S2. The GBC also used the features extracted from the raw movement data provided by the ankles in S1 and S2. Additionally, the SGD also used the raw data features from all three sensors in S1 and S2. The RFC used only the statistical features from the pelvis sensor in S2. Lastly, the DFF used all features from the raw movement data provided from the ankle sensors in S2.

Classification of the total score of a jump demonstrated the weakest performance of all respective scoring items predicted in this dataset. Six was the most common total score observed in the dataset, with a percentage of 25%. By picking the most frequent label, the difference between the model with the highest accuracy (RFC) and ZeroR classifier accuracy is 3.4%. In a real-world application, this improvement in accuracy is not a viable choice for automating further LESS score datasets. Furthermore, due to the limited performance across all seven models, the top performing features for each model are not likely to provide accurate insight into the most relevant IMU-derived features.

Across all seven models implemented, accuracy was relatively similar, with a range of 4.6% between the best model (RFC) and worst model (DFF) for predicting the overall LESS score. The DFF providing the lowest total accuracy in a complex classification example is an interesting finding due to the success of neural networks in related research, further suggesting that the data available for classifying total score is not a viable option for LESS automation.

Using the averaged F1-scores across all classifiers, the ZeroR classifier is outperformed by a minimum of 10.7%. The GBC produced the greatest F1-score (26.3%) - suggesting the GBC performed better predicting multiple scoring labels than the other models. Additionally, five out of seven models performed best when using the raw movement data as the processing subset (71.4%). The statistical and temporal-based features only outperformed the other subsets for the RFC model, while the GNB used raw and statistical features together. Models also leveraged data from either S1A2 or S2 for classification, with minimal changes between models. Subset data from S2 performed well in terms of classification accuracy with the RFC, while S1A2 produced the top F1-scores across all models (24.9% for the GNB, 26.3% for the GBC, and 25.6% for the SGD). In terms of positional-related data, the ankle sensor data

was used more commonly in six out of seven models (85.7%).

The low performance of the machine learning models does not provide confidence in applying this approach to further LESS datasets. There are two main factors to consider. Firstly, total score is a value that is largely individualised between participants with differing errors contributing to one overall value. There are 17 different items that contribute to the LESS total score. Recognising all these items simultaneously, without an indication of the individual items making up this total score value may be too complex.

Additionally, the scores in this dataset range from 1 to 10, with majority of scores occurring within the middle of this range (between scores 5 and 7), as seen in Section 6.1.2. The models are being trained in stratified folds, meaning the distribution percentage is going to be the same across the training and test set. With such a heavy imbalance towards the middle scores, other total scores from this dataset are not being trained as regularly by the model - meaning the subtle differences between individual scoring items are not able to be recognised well.

6.2.2 Individual Scoring Items

By applying machine learning to the scoring items individually with a binary and multi-class classification approach, the high-risk movement patterns that make up an overall LESS score can be better recognised in stages. Instead of multiple different movement patterns making up one value, individual scoring classification can focus on one movement item and aim to find the patterns linking the input data to whether an error is absent, or present. As discussed in Section 6.1.2, scoring items 2, 3, 10 and 13 did not provide enough data to be used for classification. Classification was performed for all other LESS items. The best performing models for each scoring item can be seen in Table 6.4

Score 1 - Knee Flexion at Initial Contact

Knee flexion greater than 30° at initial contact was an error that was presented commonly in the dataset (82.1%). When comparing performance, the ZeroR classifier outperformed the DFF model in both accuracy, and the weighted F1-score, as seen in table 6.5. In other words, this finding means none of the

Table 6.4: Summary of the Best Performing Models for all Scoring Items

Score	Model	Processing Subset			Accuracy (%)	F1-Score (%)
		Data	Segment	Sensors		
1	DFF	Raw	S2	All	80.7	77.9
4	GNB	Stats	S1A2	All	89.0	87.8
5	RFC	Stats	S2	All	77.1	73.3
6	RFC	Both	S2	All	84.4	78.9
7	RFC	Stats	S1A2	All	95.0	92.9
8	KNN	Both	S1A2	Ankles	58.3	57.4
9	GNB	Stats	S2	Pelvis	60.0	57.0
11	SGD	Raw	S1A2	All	64.3	63.7
12	SGD	Stats	S2	Pelvis	92.7	90.2
14	RFC	Stats	S2	Pelvis	91.7	89.8
15	RFC	Stats	S2	All	78.4	75.4
16	GBC	Raw	S2	Ankles	70.2	68.8
17	GBC	Stats	S2	Pelvis	93.2	94.9

selected models were able to recognise enough of a distinction between absent and present, and that classifying all labels as present by default was a better approach for overall performance. Similar to the total score classification, the application of a neural network did not perform highly.

Focusing more on the individual performance of the DFF model, as seen in Table 6.5, the DFF model demonstrates a higher precision rate than the ZeroR classifier (82% and 85% respectively) in the dominant class (error present). Using the recall rate for the DFF, 93% of the error present classes are being classified correctly, while 7% are being classified as error absent. On the other hand, a recall of 26% for the error absent class suggests 74% of the error absent classes are being misclassified as error present.

The subset of data used to produce the best model was the raw movement data from all three sensors in S2. As seen in Figure 6.2, the top performing features involved various accelerometer and gyroscope points. All five features are within similar ranges of overall average impact, meaning the importance of each feature is relatively similar. As the dataset used was provided from

Table 6.5: Summary Performance of Scoring Item 1

Metric (%)	ZeroR	DFF
Accuracy	82.1	80.7
Weighted F1	74.1	77.9

(a) Summary Comparison

Class Label	Metric (%)	ZeroR	DFF
Absent (0)	Precision	0	42
	Recall	0	26
Present (1)	Precision	82	85
	Recall	100	93

(b) Performance of Individual Class Labels

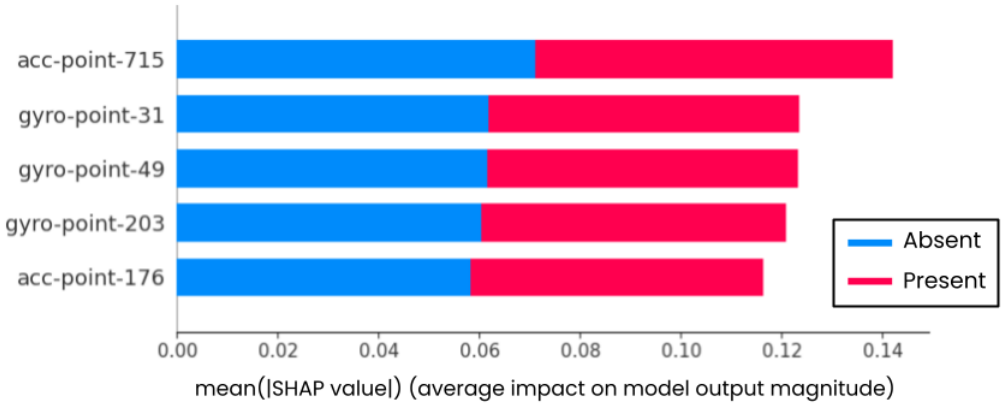


Figure 6.2: Features with the Greatest Influence in Classifying Score 1

S2, the knee flexion is assessed at the point of initial contact. Therefore, the features with the highest importance are expected to be toward the beginning of the raw movement data (lower point numbers).

Score 4 - Ankle Plantarflexion at Initial Contact

Unlike knee flexion in scoring item 1, ankle plantarflexion was not as commonly observed in this LESS dataset. With a percentage of 83% of the examples, the error absent label was the dominant class. When investigating the accuracy values between the ZeroR and GNB classifier, seen in Table 6.6, the GNB outperforms the ZeroR overall accuracy by 6.4%. Similarly in the weighted F1-score, the GNB outperforms the ZeroR by 13.1%. Therefore, this finding indicates that the GNB model is able to distinguish between at least some of the absent and present labels.

Table 6.6: Summary Performance of Scoring Item 4

Metric (%)	ZeroR	GNB
Accuracy	82.6	89.0
Weighted F1	74.7	87.8

(a) Summary Comparison

Class Label	Metric (%)	ZeroR	GNB
Absent (0)	Precision	83	91
	Recall	100	96
Present (1)	Precision	0	82
	Recall	0	54

(b) Performance of Individual Class Labels

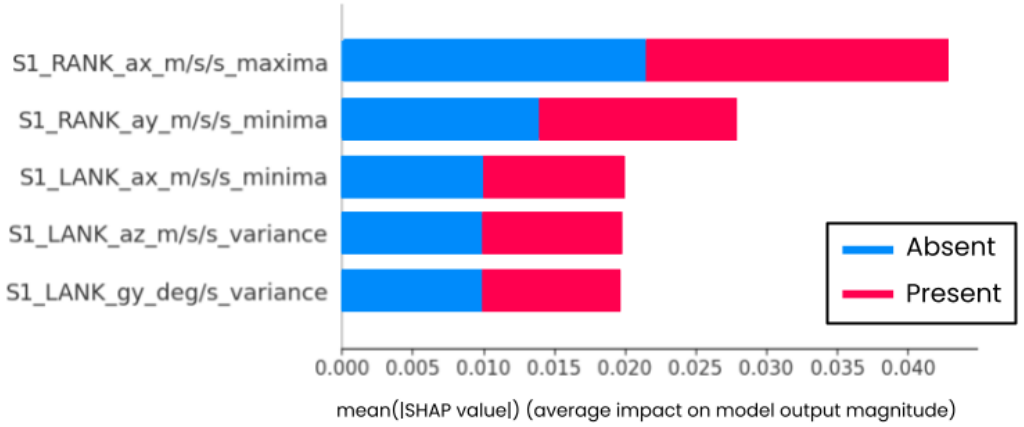


Figure 6.3: Features with the Greatest Influence in Classifying Score 4

Diving into the finer performance of each class label in the GNB, as seen in Table 6.6, the GNB demonstrates a higher precision rate (91%) than the ZeroR classifier (83%) in the dominant class. Using the recall rate for this classification, we can calculate 96% absent examples being correctly classified. Additionally, the recall of the present label class is 54% - meaning 54% of the error present classes are misclassified as error absent.

With this scoring item being an ankle-based movement, the expectation is the ankle sensors would provide the greatest clarity out of all sensor data available, to indicate whether plantarflexion is occurring upon landing. Although the best performing subset uses movement data from all three sensors, the features with the highest importance, as shown in Figure 6.3, also support this theory - with three left ankle features and two right ankle features. These features also correspond to S1, indicating that the movement of these ankle sensors prior

Table 6.7: Summary Performance of Scoring Item 5

Metric (%)	ZeroR	RFC
Accuracy	71.6	77.1
Weighted F1	59.7	73.3

(a) Summary Comparison

Class Label	Metric (%)	ZeroR	RFC
Absent (0)	Precision	0	73
	Recall	0	31
Present (1)	Precision	72	78
	Recall	100	95

(b) Performance of Individual Class Labels

to initial contact are giving some insight on the position of the foot at initial contact.

As discussed in Section 5.2.4, forward motion of the ankle sensors (z-axis) is expected to decelerate quickly prior to the downward descent of the ankle to the ground (y-axis) upon initial contact. In the case of plantarflexion, this pattern is true; but when the error is present, the heel makes contact with the ground prior to the front of the foot, causing movement in the z-axis and y-axis to both decelerate at the same time. From this, we would expect features extracted from the y-axis and z-axis of the ankle sensors to have a degree of influence in the scoring of this item. However, this influence was not demonstrated strongly in the GNB model.

The overall performance for this scoring item provides average evidence that machine learning is able to recognise patterns within the available movement data to classify this error in the jump landing. Notably with the error present being the less-dominant class, the model trained on less instances of this error - but was still able to correctly classify 54% of these examples. Further investigation would be required to improve this performance, possibly with the inclusion of a sensor closer to the front of the foot, where the fusion of data may allow additional relevant features to be extracted.

Score 5 - Knee Valgus at Initial Contact

Similarly to scoring item 1, movement at the knee is complex to calculate based on the movement data available from this study. As discussed in Section 2.2, knee valgus (i.e., abduction) is one of the highest risk movement patterns

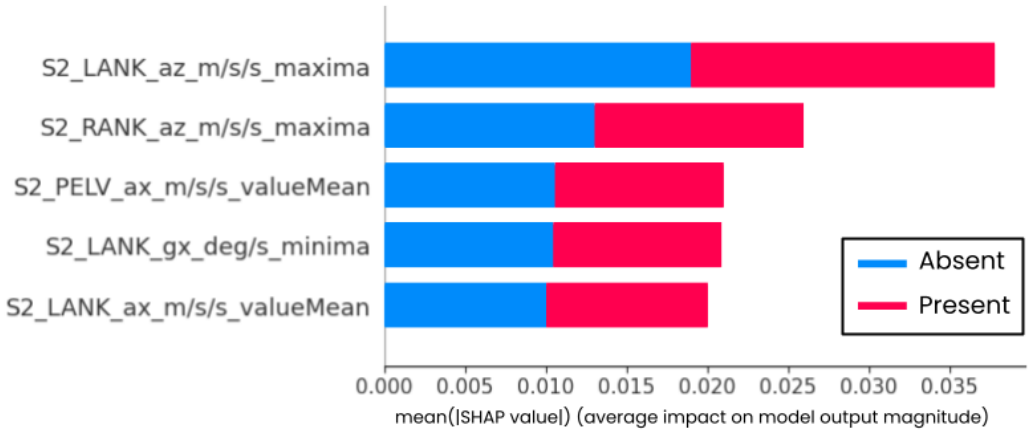


Figure 6.4: Features with the Greatest Influence in Classifying Score 5

for ACL injury. Knee valgus at initial contact appeared as the dominant class, with a total of 72% of the dataset presenting this error. The accuracy from the RFC model shows a 5.5% increase (77.1%) from ZeroR classification (71.6%), as seen in Table 6.7. Furthermore, the weighted F1 score of the RFC model (73.3%) outperforms the ZeroR score (59.7%) by 13.6%. From these comparisons, there is an indication that the RFC model may be able to recognise differences between both labels.

Individual class performance, as seen in Table 6.7, of the RFC model on the error present class demonstrated a 6% improvement in precision (78% for RFC, 72% for ZeroR). From the recall rate given for this class, 95% of the error present classes were correctly classified - meaning 5% were incorrectly classified as error present. Furthermore, the RFC is able to recognise a small subset of the absent class labels, with a recall rate of 31% - indicating 69% of the error absent classes were misclassified as error present. There are some patterns being recognised for some of the present examples, but absent example performance does not provide enough strong enough evidence that both classes are distinct enough.

Since the performance of the DFF in scoring item 1 did not outperform ZeroR, the subset used for establishing knee flexion may not be the same as identifying knee valgus. To this point, the dataset of statistical features from all three sensors in S2 provided the greatest performing model. Scoring item 1 trained

Table 6.8: Summary Performance of Scoring Item 6

Metric (%)	ZeroR	RFC
Accuracy	85.3	84.4
Weighted F1	78.6	78.9

(a) Summary Comparison

Class Label	Metric (%)	ZeroR	RFC
Absent (0)	Precision	85	86
	Recall	100	98
Present (1)	Precision	0	20
	Recall	0	3

(b) Performance of Individual Class Labels

using only the raw movement data, while this scoring item benefited from the statistical features instead. As seen in Figure 6.4, the maximum value in the z-axis of both ankles contributes the highest impact on the RFC model. As discussed previously, the maximum value in the z-axis represents the full deceleration of the ankle upon initial contact. This finding indicates that the amount of deceleration observed in the horizontal motion of the ankle is a key indicator on how the knee is positioned upon landing. Additionally, as discussed in Section 2.2, knee valgus can cause ankle eversion. None of the top features observed in this case were focused on the z-axis of the ankle gyroscopes.

The application of machine learning to this scoring item shows an average level of evidence that knee valgus can be quantified using this approach. Further investigation, similar to scoring item 1, with sensor positioning more focused around the knee may improve results in this case, and provide greater confidence in the application of IMUs to this classification.

Score 6 - Lateral Trunk Flexion at Initial Contact

The lateral flexion of the trunk at initial contact was another error that was not present commonly in this dataset, with a dominance of 85% in the absent class label. Comparing the accuracy between ZeroR (85.3%) and RFC (84.4%), as seen in Table 6.8, shows that baseline prediction of the dominant class label outperformed all other models trained for scoring item 6. However, the weighted F1-score of the RFC (78.9%) marginally outperforms the ZeroR (78.6%) by 0.3%. Therefore, it is likely that this dataset and approach does

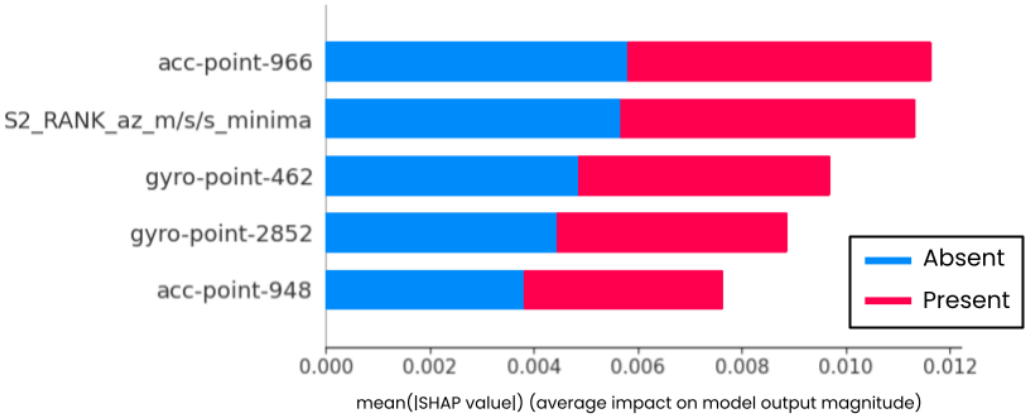


Figure 6.5: Features with the Greatest Influence in Classifying Score 6

not provide enough data to separate the two scoring conditions.

Using the performance metrics for each class label of the RFC, as seen in Table 6.8, this claim is strengthened. In the dominant class label, the RFC only slightly outperforms the ZeroR by 1% of precision (85% vs 86%). A 98% recall for the RFC model means 2% of the error absent classes are being misclassified. For the error present classes in the RFC, only 3% of the error present classes are correctly predicted. This result is the weakest performance across all scoring items.

In addition to this performance, the subset of data with the greatest performance uses both the raw and statistical features from all three sensors in S2 to classify trunk flexion, as seen in Figure 6.5. Lateral trunk flexion occurs when the body is inclined to the left or right of the participant in the x-axis. Movement in this scoring item would likely be best represented by lateral movement of the pelvis in the x-axis, or rotation in the z-axis of the pelvis gyroscope. As discussed in scoring item 1, the earlier value points in the raw movement data of S2 indicate movement upon initial contact. These points were not recognisable, making it difficult to understand where in the movement window these accelerometer and gyroscope positions occur. Furthermore, the only statistical feature to have a high degree of influence was the backward motion of the pelvis, as indicated by the minimum value of the z-axis (maximum acceleration backwards). The performance metrics provided in scoring item 6 indicate that

Table 6.9: Summary Performance of Scoring Item 7

Metric (%)	ZeroR	RFC
Accuracy	94.0	95.0
Weighted F1	91.2	92.9

(a) Summary Comparison

Class Label	Metric (%)	ZeroR	RFC
Absent (0)	Precision	94	95
	Recall	100	100
Present (1)	Precision	0	40
	Recall	0	17

(b) Performance of Individual Class Labels

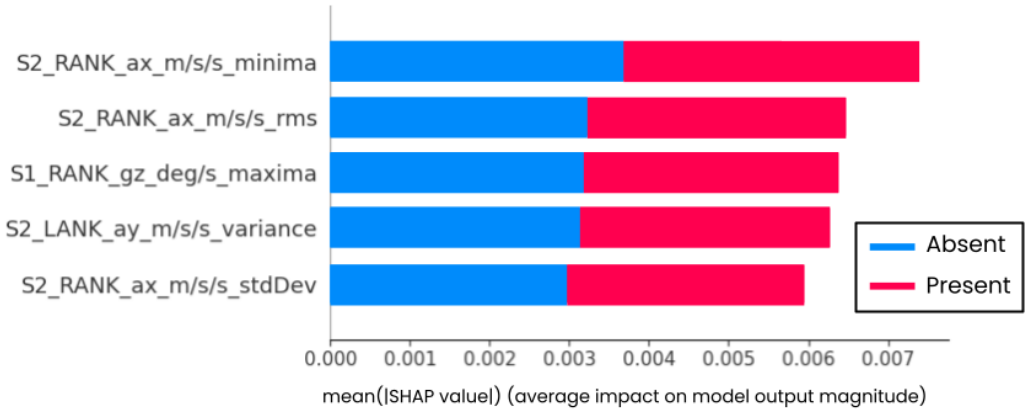


Figure 6.6: Features with the Greatest Influence in Classifying Score 7

this approach for classifying lateral trunk flexion is not applicable for further investigation.

Score 7 - Wide Stance Width

With two possible stance width scoring items, a wide stance width was not commonly observed within this dataset, with a strong dominance of 94% in the absent class for this dataset. As seen in Table 6.9, the accuracy of the RFC model (95.0%) shows a 1% improvement over the ZeroR classifier (94.0%). Weighted F1-score also shows a smaller margin of improvement, with the RFC F1-score (92.9%) showing a 1.7% improvement over the ZeroR (91.2%). Similar to scoring item 6, these smaller margins indicate classification between the two labels is not strong.

In terms of precision for the dominant absent class, as seen in Table 6.9, the

RFC provides a 1% improvement on ZeroR precision (95% vs 94%). Recall for RFC is also 100%, meaning all 205 absent labels are classified correctly. Performance of the present class indicates only a 17% recall rate, equating to 83% of the error present classes being correctly identified. Although two examples have demonstrated some level of distinction between present and absent, the dataset is limited for classifying this scoring item.

At initial contact, the ankle sensors are already positioned close to the ground, without any indication of how the ankle sensors are positioned prior to landing. To assess the stance width at initial contact, the subset of data with the greatest relevance would likely be the movement data captured in S1A2. With S1 included, movement in the x-axis of the ankle sensors represents the left and right movements of the ankle as a participant descends to the ground from take-off. To this point, the best performing subset was the statistical features between S1A2 with all three sensors. Predominantly, the right ankle accelerometer provided the most features with high impact on the model. As seen in Figure 6.6, these features included the movement to the right (minimum acceleration value), root mean square in the x-axis (the mean magnitude of the acceleration values), and the standard deviation in this axis. However, these features came from S2 - indicating the movement of the ankle after already making contact with the ground influences the stance width at initial contact, which is a contradiction.

This scoring item is limited in terms of the available data that can be used to train each model. With the stratified cross fold approach, each training set will only ever have 6% of error present classes to train a model with, meaning the models are not seeing enough of the class to make distinctions between labels. Therefore, this dataset does not provide enough evidence to use this approach for classifying a wide stance width, but narrow stance width may provide stronger results to evaluate given the greater data spread.

Score 8 - Narrow Stance Width

Following on from scoring item 7, narrow stance width has a far greater spread of data to evaluate. An error present was the dominant class with 51% of the dataset, and 49% of the dataset being error absent classes. As seen in

Table 6.10: Summary Performance of Scoring Item 8

Metric (%)	ZeroR	KNN
Accuracy	51.4	58.3
Weighted F1	34.9	57.4

(a) Summary Comparison

Class Label	Metric (%)	ZeroR	KNN
Absent (0)	Precision	0	57
	Recall	0	63
Present (1)	Precision	51	61
	Recall	100	54

(b) Performance of Individual Class Labels

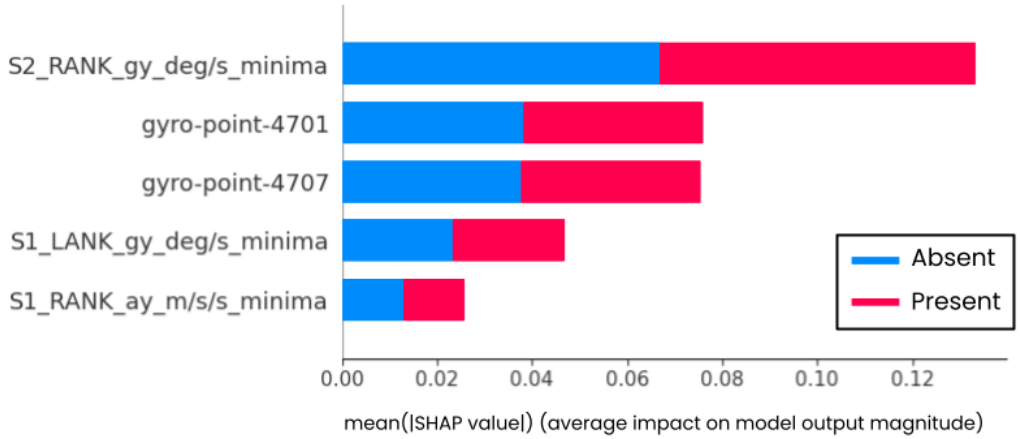


Figure 6.7: Features with the Greatest Influence in Classifying Score 8

Table 6.10, the accuracy of the KNN model (58.3%) outperforms the ZeroR accuracy (51.4%) by 6.9%. Furthermore, the weighted F1-score of the KNN (57.4%) outperforms the ZeroR by 22.5% (34.9%). With a more even spread of the dataset, the margins between ZeroR and KNN may appear higher, but the F1-score of the ZeroR is also affected due to a more balanced dataset.

To evaluate the performance of scoring item 8, comparisons are made between the precision and recall of both classes, seen in table 6.10. For the dominant class (error present), the precision of the KNN (61%) outperforms the ZeroR classifier by 10% (51%). A total of 54% of the error present classes are correctly classified by the KNN. Furthermore, 63% of the error absent classes were correctly classified. Notably, the dominant class (error present) was outperformed by the error absent class, meaning some level of distinction between the two error conditions has been recognised, more so than many other scoring

Table 6.11: Summary Performance of Scoring Item 9

Metric (%)	ZeroR	GNB
Accuracy	56.9	60.0
Weighted F1	41.3	57.0

(a) Summary Comparison

Class Label	Metric (%)	ZeroR	GNB
Absent (0)	Precision	0	55
	Recall	0	34
Present (1)	Precision	57	62
	Recall	100	80

(b) Performance of Individual Class Labels

examples where the dominant class typically performed best.

As discussed in scoring item 7, x-axis movement within the ankle sensors for the full S1A2 window is likely to best represent the positioning of the feet for stance width. To this point, the subset of data that produced the best performing model used a combination of both raw and statistical features from just the ankle sensors, in the full S1A2 window. However, the features with the greatest importance extracted from this model, seen in Figure 6.7, do not reflect x-axis movement, rather internal rotation of the right ankle (minimum value in the y-axis of the gyroscope) throughout S2 produced the greatest impact. Additionally, the internal rotation of the left ankle on descent to initial contact, and the downward acceleration of the right ankle were also important features. Notably from the raw dataset, two gyroscope points in close proximity (4701 and 4707) provided similar influence on the model, but it is difficult to recognise where these points occur in the overall window.

Classification of a narrow stance width showed more potential than classifying a wide stance width, as the higher percentages of recall for both labels in this dataset indicate that the KNN model was recognising some patterns between the two. However, the features with the greatest importance on the model do not have a great relation to the movements that contribute to stance width in a DLJL, therefore further investigation would be required into how to improve the classification performance of both wide and narrow stances.

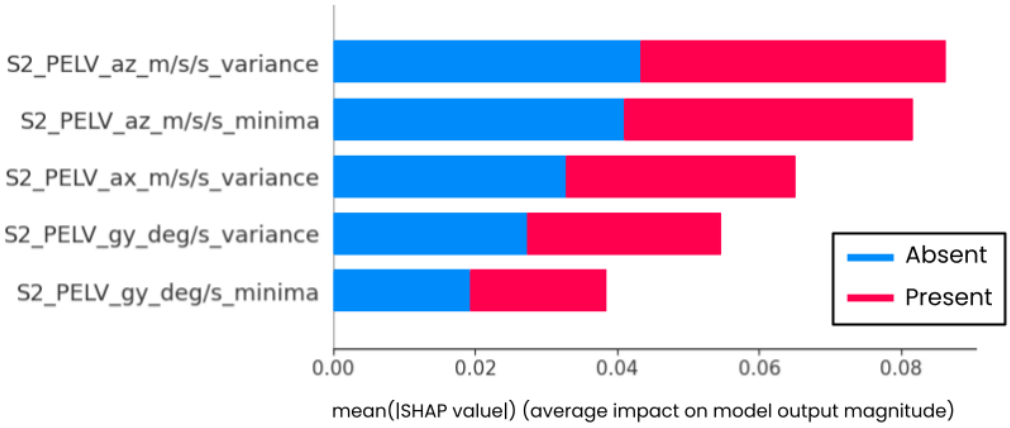


Figure 6.8: Features with the Greatest Influence in Classifying Score 9

Score 9 - Toe-in Foot Position

Scoring item 9 was another error that was dominated by the present class label, with 57% of the dataset demonstrating this error. A 3.1% difference is observed when comparing the accuracy of the GNB (60.0%) to the ZeroR (56.9%), as seen in Table 6.11. Additionally, a 15.7% difference between the weighted F1-scores of the GNB (57.0%) and the ZeroR classifier (41.3%) is observed. Similar to scoring item 8, the margin difference between F1-scores is due to the weight influence that a more evenly-spread dataset has on the output of the ZeroR metric.

Using the summary results provided in Table 6.11, individual performance in the dominant class labels (error present) shows a 5% improvement in precision between the ZeroR (57%) and GNB (62%). Recall in the GNB model indicated 80% of the error present classes were correctly classified. For the error absent classes, only 34% of the classes were correctly classified.

Toe-in foot position refers to the amount of internal rotation seen at the foot between the initial contact and maximum knee flexion (S2). However, statistical features from the pelvis provided the best performing model in this scoring item, with no inclusion of features from the ankle sensors. The top performing features, shown in Figure 6.8, included the variance and minima values in the accelerometer and gyroscope axes at the pelvis. These features suggest movement at the pelvis influences the rotation of the foot, more so than the

Table 6.12: Summary Performance of Scoring Item 11

Metric (%)	ZeroR	SGD
Accuracy	62.4	64.3
Weighted F1	47.9	63.7

(a) Summary Comparison

Class Label	Metric (%)	ZeroR	SGD
Absent (0)	Precision	62	70
	Recall	100	73
Present (1)	Precision	0	54
	Recall	0	50

(b) Performance of Individual Class Labels

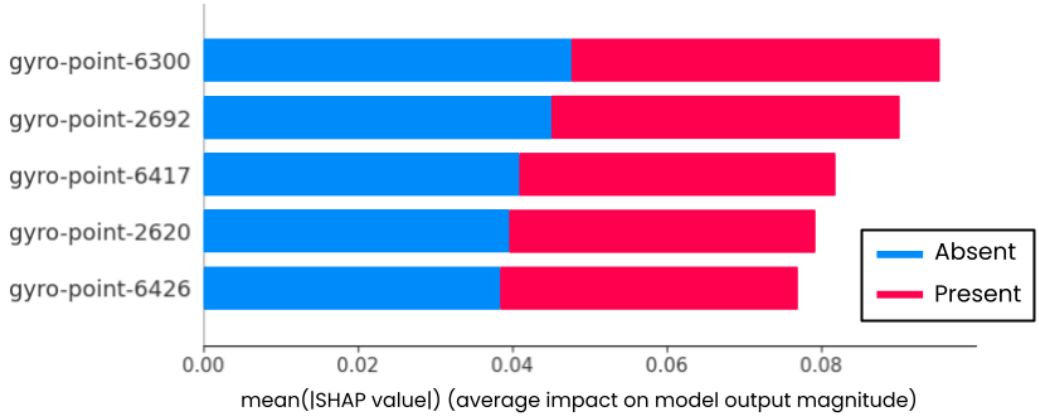


Figure 6.9: Features with the Greatest Influence in Classifying Score 11

movement of the ankles do - which is a contradiction, as the ankle bone is rotating when the position of the foot is rotated inwards. This difference would suggest the gyroscope within both ankle sensors should provide greater quality movement data for foot rotation, meaning the patterns being recognised in the data are not consistent with real performance of the movement. Therefore, the application of this model to the automation of scoring item 9 is limited. The accuracy might have been better with a sensor positioned on the foot itself.

Score 11 - Symmetric Foot Contact at Initial Contact

With 62% of the dataset, symmetric foot contact at initial contact was dominated by the error absent class. As seen in Table 6.12, the accuracy of the SGD model (64.3%) outperforms ZeroR accuracy (62.4%) by 1.9%. Weighted F1-score performance of the SGD (63.7%) outperformed the ZeroR model (47.9%)

by 15.8%.

Using the individual performance metrics, as seen in Table 6.12, the classification precision of the SGD in the dominant class (absent) (70%) outperformed the ZeroR model by 8% (62%). The recall rate for the error absent classes was 73%. For the present class, a recall rate of 50% means 50% of the dataset was also incorrectly classified. Once again, there is an indication a pattern has been observed for distinguishing between symmetric and asymmetric contact, but the quality of these patterns is dependent on the features with the greatest importance.

As discussed in Section 4.3.2, temporal features in relation to initial contact were extracted from the IMU data. These included time from take-off to both left and right initial contact timestamps, and the time taken from left and right initial contact to the maximum knee flexion timestamp. These features provide clear comparison between foot performance and would allow manual assessment of which foot made contact first without the need for video. However, this scoring item used the raw movement data from all three sensors in S1A2. From these features with the greatest impact on the model, five different gyroscope points are given, as seen in Figure 6.9. These points may correlate to any sensor, and any axis of movement from these sensors - therefore, we cannot establish how viable these sensors are for predicting these classes. However, based on the exclusion of a clear metric for establishing symmetric initial contact - the best performing model has not been able to recognise this pattern. Therefore, the application of this model for further examples is not applicable for automating scoring item 11.

Score 12 - Knee Flexion Displacement

In comparison to knee flexion at initial contact in scoring item 1, knee flexion displacement was not as commonly observed, with a dominance in the absent class of 93%. Comparison between the accuracies of the ZeroR (92.7%) and SGD model (92.7%) show both models had the same overall classification performance, as seen in Table 6.13. However, the SGD model outperforms the weighted-F1 score of the ZeroR model (89.1%) by a slight margin of 1.1% (90.2%), indicating there is some classification of the error present

Table 6.13: Summary Performance of Scoring Item 12

Metric (%)	ZeroR	SGD
Accuracy	92.7	92.7
Weighted F1	89.1	90.2

(a) Summary Comparison

Class Label	Metric (%)	ZeroR	SGD
Absent (0)	Precision	93	93
	Recall	100	99
Present (1)	Precision	0	27
	Recall	0	13

(b) Performance of Individual Class Labels

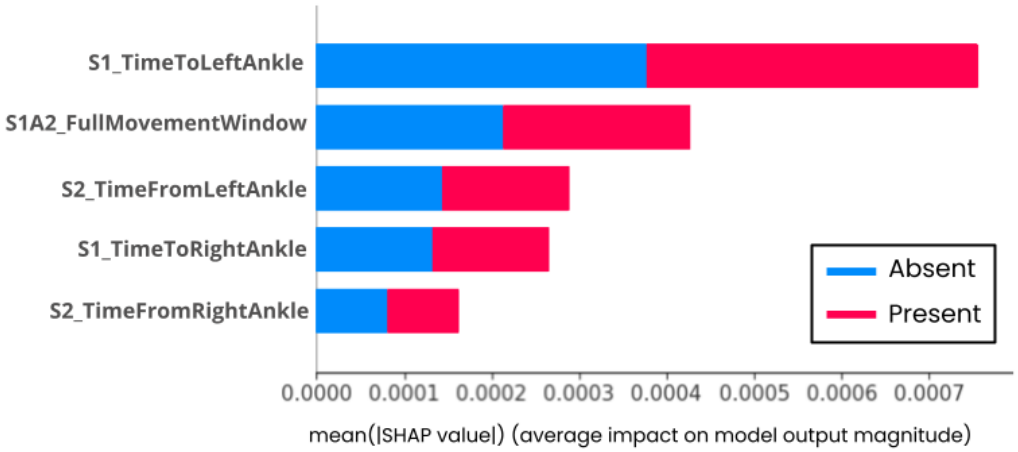


Figure 6.10: Features with the Greatest Influence in Classifying Score 12

classes occurring using the SGD, but the performance of this classification is not outperforming picking the dominant class.

We can further observe this pattern by using the individual performance metrics of scoring item 12, seen in Table 6.13. Precision rate between both models is identical in the dominant class at 93%. However, the recall rate of the SGD shows 99% of the absent classes are correctly classified, indicating 1% of the error absent classes are classified as error present. Within the present class labels, performance does not outperform the ZeroR model by a large margin. Only 13% of the error present classes are correctly classified, with the remaining 87% classified as absent.

As discussed in scoring item 1, exact knee flexion calculation is not represented well using pelvis and ankle movement data. However, the more a participants'

Table 6.14: Summary Performance of Scoring Item 14

Metric (%)	ZeroR	RFC
Accuracy	89.0	91.7
Weighted F1	83.8	89.8

(a) Summary Comparison

Class Label	Metric (%)	ZeroR	RFC
Absent (0)	Precision	89	92
	Recall	100	99
Present (1)	Precision	0	63
	Recall	0	32

(b) Performance of Individual Class Labels

pelvis displaces downwards in the DLJL reflects vertical displacement of the centre of mass, which is linked with the amount of knee flexion that occurs during ground contact. The subset of data used to train the SGD model was the statistical features from the pelvis sensor in the S2 segment. Despite the use of the pelvis sensor subset, the top performing features, provided in Figure 6.10, showed a variety of temporal measures influenced the model the most of any other features. Notably, the mean SHAP values for all five features are incredibly low, indicating the use of these features is likely not outperforming all other statistical features by a large margin.

These features suggest the time taken for a participant to complete different segments of the movement influence the classification of a present or error absent condition. As discussed in Section 6.2.1, movement data is largely individualised. In a real-world example of the LESS, when a participant is in contact with the ground for longer, this is typically associated with greater flexion of the knee. Therefore, these top performing features are relevant to knee flexion. However, with a dominance of 93% in the error absent class, the error present class is not represented well in the training and testing of the models. This low distribution is likely the cause of low performance in the error present recall and the overall impact of the top performing features on the SGD model. An even distribution of classes would provide greater confidence on the influence of these features to classification of scoring item 12.

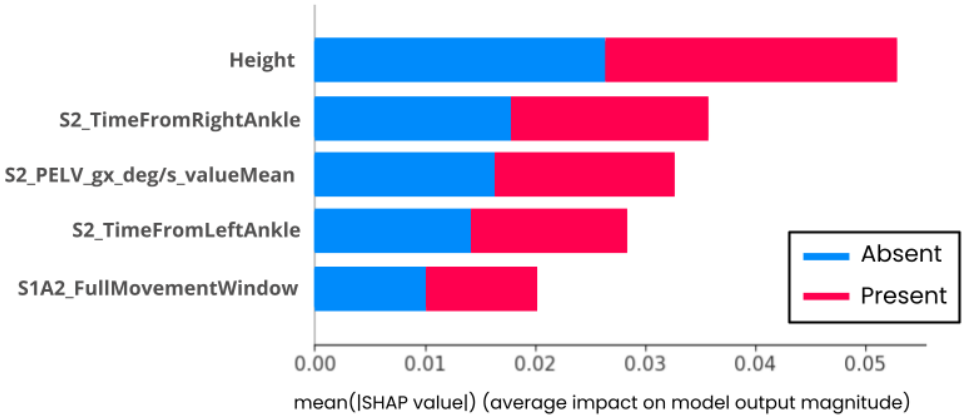


Figure 6.11: Features with the Greatest Influence in Classifying Score 14

Score 14 - Trunk Flexion at Maximal Knee Flexion

Hip flexion at the point of maximal knee flexion was another error that was not well represented in this dataset, with a 89% dominance in the absent class. As seen in Table 6.14, the accuracy of the RFC model (91.7%) outperformed the ZeroR model (89.0%) by 2.7%. The weighted F1-score of the RFC (89.8%) also outperformed ZeroR (83.8%) by 6%.

With the individual performance metrics, as seen in Table 6.14, precision was improved by 3% between ZeroR (89%) and RFC (92%) performance in the absent class. Furthermore, a recall rate of 99% for the error absent class in the RFC means only 1% were misclassified. However, a recall rate of 32% for the error present class indicates 68% is misclassified as error absent.

In scoring item 14, trunk flexion is likely to be best represented by movement in the x-axis of the pelvis gyroscope, as these movements represent the forward and backward rotation at the pelvis. In contrast to scoring item 6, lateral trunk flexion was the movement to the left and right in the x-axis of a participant, while typical trunk flexion occurs when the pelvis is tilted forward in the z-axis. To this point, the statistical features from the pelvis provided the greatest importance on the RFC model performance, as seen in Figure 6.11. The mean value in the x-axis of the pelvis produced the third highest impact on the model, which aligns with this type of movement - but all other features include a variety of temporal measures and the height of the participant. This finding

Table 6.15: Summary Performance of Scoring Item 15

Metric (%)	ZeroR	RFC
Accuracy	74.3	78.4
Weighted F1	63.4	75.4

(a) Summary Comparison

Class Label	Metric (%)	ZeroR	RFC
Absent (0)	Precision	0	62
	Recall	0	35
Present (1)	Precision	82	81
	Recall	100	93

(b) Performance of Individual Class Labels

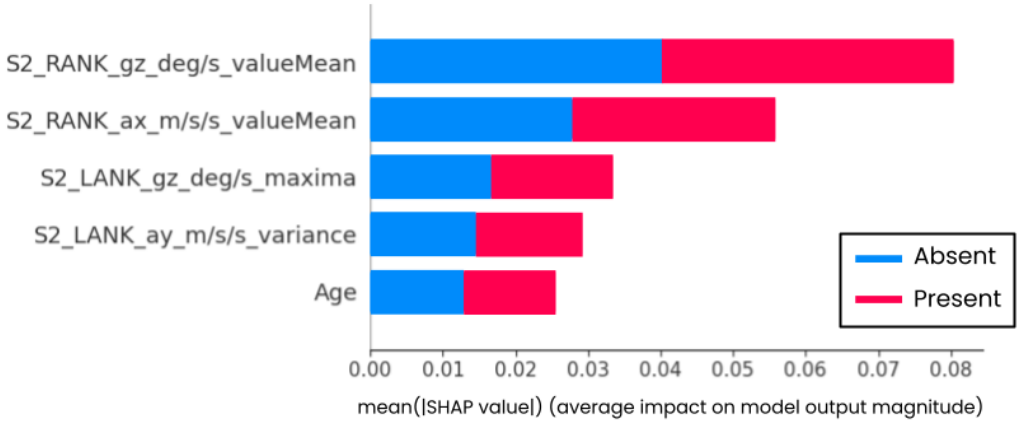


Figure 6.12: Features with the Greatest Influence in Classifying Score 15

suggests both the height of the participant, and the time taken from initial contact of both the left and right ankles influenced whether additional trunk flexion took place from the beginning to the end of S2. Further investigation would be required, with more representation of the error present to determine whether these features influence real movement performance in all participants.

Score 15 - Knee Valgus Displacement

Scoring item 15 is another example of knee valgus, similar to scoring item 5, but this time an error is represented toward the end of S2. This error condition appeared often, with 74% of the dataset representing the error present class. Accuracy comparison between ZeroR (74.3%) and RFC (78.4%) showed a 4.1% increase in total classification performance, as seen in Table 6.15. Furthermore, weighted F1-score of the RFC (75.4%) outperformed ZeroR (63.4%) by 12%.

The precision rate of the RFC in the dominant class is only 81%, a 1% increase on ZeroR performance (82%), as seen in Table 6.15. Furthermore, recall rate of the RFC for the dominant class was 93%, meaning only 7% of the classes were misclassified. For the error absent classes, recall was 35% - equating to a misclassification of 65%. From these values, there is some level of recognition occurring between the classes, but the applicability of this recognition needs to be evaluated.

The best performing model (RFC) for knee valgus in scoring item 5 used the statistical features from all three sensors in S2. For scoring item 15, RFC was also the best performing model, using the same subset of data as scoring item 5 - showing some consistency between two scoring items looking at the same type of movement pattern. As discussed in Section 2.2 and scoring item 5, knee valgus may include ankle eversion. To this point, the rotation of the ankles in the z-axis would represent this pattern. From the features shown in Figure 6.12, the mean value of the z-axis in the right ankle, and the maximum value in the z-axis of the left ankle lean towards the rotation of the ankles as indicators of knee valgus. The mean value in this case represents the average amount of rotational velocity observed in this axis over the entire movement, while the maximum value in the left ankle indicates how fast the peak rate at which the ankle was rotated towards eversion throughout the entire S2 window. Age of the participant was also seen as an indicator of knee valgus, which would require further investigation to assess the correlation of this participant characteristic to knee valgus.

Additionally, the mean value in the x-axis of the right ankle accelerometer represents left and right movement of the ankle, which could also be a sign of horizontal movement caused by the action of ankle eversion combined, which is a logical pattern to assess for knee valgus. Comparing these top performing features to knee valgus in scoring item 5, the features may appear more relevant - but performance between the two models recognising the same pattern do not differ by much in terms of overall performance. Despite having different features that had the highest impact, both produced similar accuracy and performance in classification of the same movement pattern, in different parts of the movement. Therefore, a further investigation with a combination of knee-placed sensors in conjunction with the ankles and pelvis may provide

Table 6.16: Summary Performance of Scoring Item 16

Metric (%)	ZeroR	GBC
Accuracy	66.1	70.2
Weighted F1	52.6	68.8

(a) Summary Comparison

Class Label	Metric (%)	ZeroR	GBC
Soft (0)	Precision	0	59
	Recall	0	43
Average (1)	Precision	66	75
	Recall	100	84
Stiff (2)	Precision	0	30
	Recall	0	40

(b) Performance of Individual Class Labels

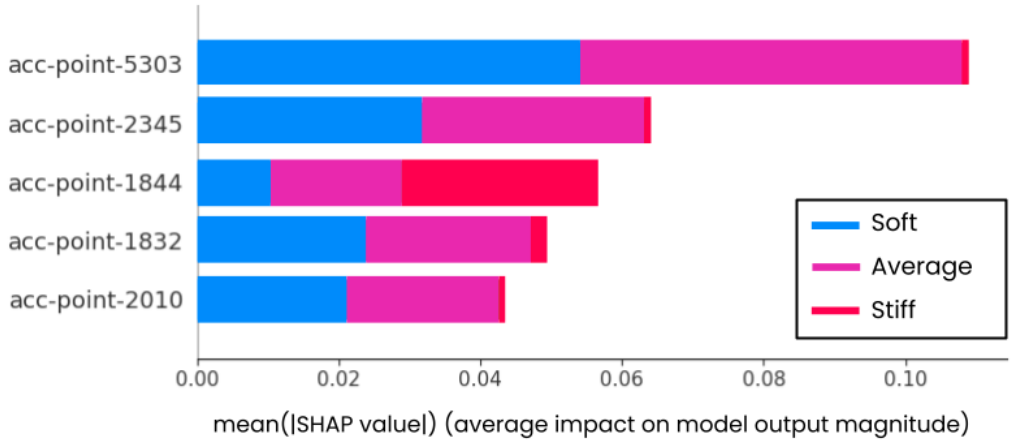


Figure 6.13: Features with the Greatest Influence in Classifying Score 16

further insights into the features that truly contribute to this scoring item.

Score 16 - Joint Displacement

Scoring item 16 is made up of multiple joints to consider for displacement, with the conditions of this scoring item predominately being more subjective depending on the scorer. Additionally, with a multi-class approach, pattern recognition of the machine learning models must now distinguish three different potential values. For this scoring item, 66% of the dataset included "average" joint displacement, making this class the dominant class. To this point, accuracy of ZeroR is 66.1%, compared to an accuracy of 70.2% for the GBC, as seen in Table 6.16. Furthermore, weighted F1-score of the GBC (68.8%) shows

an improvement in performance over the ZeroR (52.6%) by 16.2%.

As seen in Table 6.16, classification of the dominant class of an average landing condition showed a 9% increase in precision between the GBC (75%) and ZeroR (66%). A recall of 84% equates to a total of 16% of the classes being misclassified as either soft or stiff classes. Moving onto a stiff landing condition, the recall rate produced was 40% - meaning 60% of the stiff landing classes were misclassified. Lastly, the recall rate of a soft landing was equal to 43% - equating to a 57% misclassification. In a general sense, there are some distinguishing patterns observed between the two lesser-count classes with a relatively similar recall between both, but total overall classification accuracy could be further improved for further application.

Additionally, there are many potential features we could consider to be an optimum subset for the classification of joint displacement. We can also look at the previous scoring items related to these joints and what subsets produced the best performing models. Firstly, trunk flexion (item 14) used statistical features from the pelvis in S2, while knee-related scoring items (items 1, 5, 12, and 15) have most commonly benefitted from statistical features from all three sensors in S2. The best performing models also differed, including RFC, DFF, and SGD models. However, overall joint displacement provided the best results using a GBC, trained with the raw movement data from the ankles in S2. Because of this, the features shown with the highest importance, shown in Figure 6.13, are difficult to evaluate as these can appear anywhere in the performance of the DLJL.

Most notably, all features were accelerometer related, with overall importance ranging between classes. Each of these points allow visualisation of which points in the accelerometer data were used to distinguish between two or more of the classes. Accelerometer-point 1844 had greater impact on the classification between stiff and average, than stiff and soft. Additionally, classification between soft and average classes were impacted by the remaining features, while stiff landing classes were mostly impacted by one accelerometer point in comparison to the other important features.

Table 6.17: Summary Performance of Scoring Item 17

Metric (%)	ZeroR	GBC
Accuracy	90.4	93.2
Weighted F1	85.8	94.9

(a) Summary Comparison

Class Label	Metric (%)	ZeroR	GBC
Excellent (0)	Precision	0	83
	Recall	0	100
Average (1)	Precision	90	100
	Recall	100	92
Poor (2)	Precision	0	80
	Recall	0	100

(b) Performance of Individual Class Labels

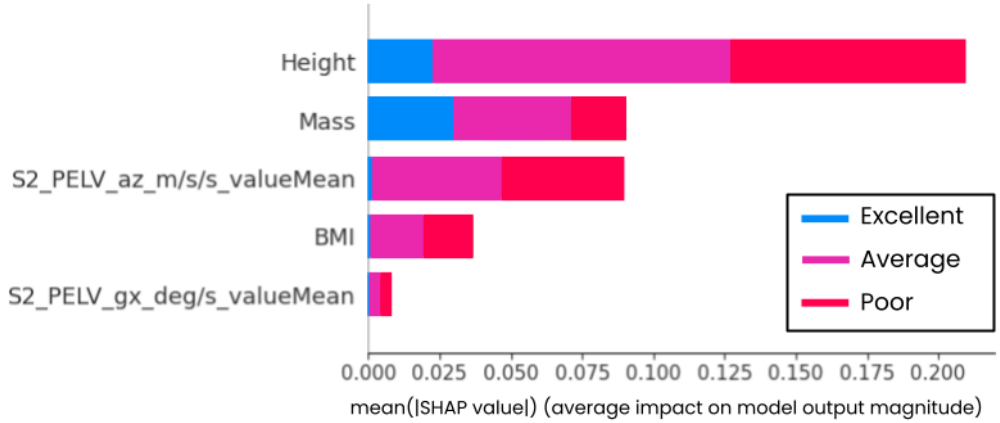


Figure 6.14: Features with the Greatest Influence in Classifying Score 17

Score 17 - Overall Impression

Overall impression is another relatively subjective scoring item, up to the decision of the scorer. An overall impression is classified as either excellent, average or poor landing condition. Similarly to scoring item 16, an "average" overall impression was the dominant class at 90% of the dataset. The overall accuracy of the GBC (93.2%) outperformed ZeroR accuracy (90.4%) by 2.8%, as seen in Table 6.17. Additionally, the weighted F1-score in the GBC (94.9%) outperformed ZeroR (85.8%) by 9.1%. With three different classes and a low distribution between excellent, and poor impressions - this F1-score suggests a few recognisable patterns have been identified between the three classes.

Performance of the GBC in the classification of this scoring item, as seen in Table 6.17, had the greatest results of all other scoring items. In the precision of the dominant class, a 100% precision indicates the scores that were classified as "average" were all average landing classes. Overall, 92% of the average classes were correctly classified, while the other 8% of the classes were misclassified as either excellent or poor landing classes. The recall rate for both poor and excellent classification was 100%, which therefore means the six excellent, and 15 poor classes were all correctly classified. The precision rates of 83% in the excellent class, and 80% in the poor class are made up of the 8% of average classes that were misclassified. Therefore, patterns were identified that clearly classify both less-dominant classes with no overlap between the two.

There are a number of smaller movement patterns that may make up this scoring item, as the pattern involves a number of side-to-side and rotational motions (frontal and transverse plane motion). Additionally, a stiff landing in scoring item 16 can influence the overall impression of a jump. Unlike total score, the classification performance in this scoring item was able to distinguish between three classes. The subset of data with the greatest relevance was the statistical features from the pelvis in S2. The movement data provided from a pelvis sensor represents the centre of mass movement pattern. For an overall impression of a jump landing, a smoother signal in the accelerometer and gyroscope data may be being used to represent an excellent landing, whereas a rough signal may represent a poor landing condition.

To this point, the top five features from Figure 6.14 show three out of five of the main features are characteristic-based features. Height shows an impact on all three classes, while mass seems to have more of a role in distinguishing between average and excellent examples. BMI values have greater use in distinguishing between average and poor classes. The mean value in the pelvis gyroscope does not have a substantial impact on the model, while the mean value in the forward motion of the pelvis is used as further support for distinguishing between average and poor classes. With the current performance of the GBC model in this case, it is reasonable to suspect the height, mass and BMI do have influence on the overall jump impression. The application of this model for further LESS scores may improve results further with a greater distribution of data; more precisely, to improve recognition of the average class.

Chapter 7

Discussion

With all the information collected, the performance of this integration of machine learning and IMU data can be used to evaluate the applicability of IMU sensor data to the scoring of the LESS. The subsets and features extracted from the movement data had varied amounts of impact on the quality of classification in the models selected. The comparisons of performance between ZeroR and the best performing models for each respective scoring item can be used to establish whether this application of machine learning classification is capable of accurately scoring a LESS score, without the need for manual intervention.

In this chapter, the main findings of this thesis are discussed in relation to the research questions driving this implementation. This chapter begins with a summary of the performance in automation of LESS scores, including overall accuracies and the factors that influenced performance. Next, the relevance of the subsets and features derived from IMU data are highlighted, including the subset contribution to different scoring items, and the applicability of the top performing features to every scoring item. Lastly, the applicability of this solution is outlined, including factors contributing to the results provided from this thesis.

7.1 Accuracy of Automated Scores

The application of machine learning models for the automation of the LESS had a mixed performance across both the individual scoring items, and the overall score. In some scoring items, accuracy and the weighted F1-score provided a strong performance for identifying patterns within the movement data, and being able to distinguish whether an error is present or absent with the data available. In other cases, classification of all the labels using just the dominant class provided greater results than over 126 individually trained models.

7.1.1 Classification Accuracy

Classification performance in the total score provided a weak performance across all seven different models and the respective 18 subsets of movement data. The highest performing model in terms of overall accuracy only provided a 28.4% accuracy rating. Additionally, the greatest weighted F1-score was 26.3% for the GBC. In terms of real-world application, this level of accuracy and performance is not fit for automation. In a real-world application, the inaccurate classification of nearly 75% of the dataset prompts inaccurate training interventions to accommodate this error.

SVC and neural networks were proposed as the most common supervised learning models used in biomechanics [110], as discussed in Section 3.2.2. Additionally, the DFF model that produced the greatest accuracy in a similar trampoline-based jumping study [2]. However, the best performing model in accuracy from this thesis was a RFC, outperforming the SVC by 2.7%, and the DFF by 4.6%. Furthermore, the best performing weighted F1-score was 26.3% using a GBC, outperforming ZeroR by 16.3%. The GBC also outperformed the SVC by 5.6% and the DFF by 3.3%.

The best performing models differ to the models discussed in Section 3.2.2, however the low performance across all models in total score classification indicates there are limitations in this approach to automation in comparison to related work. There are a number of high-risk movement patterns encapsulated within the one total score - which does not provide individualised feedback for a participant to implement the training interventions required for reducing this

risk.

There were a mix of different models that provided the greatest performance in each individual item of the LESS. A RFC was the best performing model for five scoring items. GBC, SGD, and GNB were the best performing models for two scoring items each. Finally, a KNN and DFF gave the best performance in one scoring item respectively. In relation to the SVC and neural network popularity in biomechanics mentioned previously, both the SVC and DFF did not appear regularly as best performing models. Similarly to total score, there are limitations in the results provided.

The classification of each scoring item individually provided strong, but specious results in accuracy and performance. The best performing models in scoring items 7, 12, and 14 provided accuracy and weighted F1-scores of 90% and above. On first observation, this appears as promising in automation performance, but heavily dominated class counts ultimately played the biggest influence on all scoring items. In scoring items 8, 9, and 11, the weakest overall performances are observed with accuracy and weighted F1-scores both ranging between 55-65%. There is one key link between these items - the distribution split of absent and present classes ranges between 51-62%. Furthermore, the highest performing models of 7, 12, and 14 demonstrated distribution splits between 89-94%. Due to these misleading results, many of these models do not provide a strong performance in pattern recognition for real-world application.

Additionally, this distribution plays a factor in recommendations of automation for many scoring items. Although some error absent and error present classes are distinguishable in some scoring items, there are not enough classes to confidently establish whether these patterns are applicable in further datasets. Using scoring item 7 as an example, it provides the second greatest performance in terms of weighted F1-scores across all scoring items, with a score of 92.9%; however, recall of the less-dominant class indicated that only 17% of the error present classes were correctly classified. If we consider the application of this same model with a much larger dataset of individualised LESS scores, the misclassification of many more participants indicates this automation approach is not a viable option for assessing injury risk.

One scoring item out of the 17 scoring items provided a very strong perfor-

mance in classification of the different error conditions. Scoring item 17, the overall impression, produced a high rate of recall for all three scoring labels, with the misclassification of only 8% of the classes. In addition to having three different classes and being a more subjective score, three of five features with the greatest impact on this classification did not come from IMU-derived features, but rather correlated to the participant characteristics.

7.2 Relevant Derived Features and Subsets

The importance of features in each model provided an insight into the decisions made by out-of-the-box machine learning models. As discussed in Section 4.1, interpretation of the decisions made from out-of-the-box models is limited, with the architecture of the models being pre-set without user intervention. Therefore, in addition to the overall performance of the models applied to the LESS data, subsets of the full dataset were used. This approach allowed for an assessment on the quality of the input data for automating the LESS scores, rather than the tuning of the hyperparameters that influence prediction in each respective model.

7.2.1 Processing of Movement Data Into Subsets

To assess which IMU-derived measures are the most relevant for automation, the movement data had to be processed using a number of processing steps. Alignment of the movement data provided a better look into the movement of a participant, while segmentation allowed for the relevant movement window to be used for accuracy evaluation. As discussed in Section 4.3.2, all scoring items occurred between initial contact and maximum knee flexion (S2). Segmentation of movement data is a common processing approach for IMU quantification [103, 111], with specific uses in jump landing tasks [104, 106, 109] and the DLJL task specifically [105].

Segmentation of IMU data involved processing for critical points in the accelerometer data that acted as an indication of a phase [105, 109]. In this thesis, observable peaks in the IMU data were used as indicators for extracting key phases and compared to video reference. This approach to processing

the movement data allowed for greater consistency with the LESS and the scoring items involved, ensuring the movement data relating to these movement patterns were included in all subsets. Identifying and isolating S2 was also a focus for feature extraction, as the top performing features in each model should relate to the scoring item being assessed.

7.2.2 Subset Differences

There were 18 different subsets implemented, including two different time windows for comparison (S1A2 and S2), three positional-related subsets including the ankle movement data, the pelvis data, or both in combination, and either the raw movement data, the statistical features or all features together. These IMU-derived subsets had varying influences across all scoring items.

From the related work discussed in Section 3.2.2, statistical and temporal features derived from IMU data were frequently used as input datasets for training models [2, 105, 106, 109]. For total score classification, raw movement data produced the best performing models in 5/7 of the models selected. Statistic-based features were only used in 2/7 models. However, the differences across all models show limited improvement across subsets. The statistical features produced the model with the highest accuracy (RFC with 28.4%), however raw data produced the model with the greatest F1-score (GBC with 26.3%). Across all models in total score classification, subset difference played minimal influence on the accuracy and performance of LESS score automation.

However, for classification of the LESS items individually, the statistical features had the greatest influence on best performing models with 10/13 models using this subset of data to automate the score. The raw movement data was used by 5/13 models, while a combination of both was seen in 2/13 models. Statistical datasets recorded an average accuracy of 84.6%, with an average F1-score of 82.7%, while the use of raw data provided an average accuracy of 71.7% and a F1 score of 70.1%. The combination of both datasets produced an average accuracy of 71.4%, and an average weighted F1-score of 68.2%.

With S2 as the segment of interest for all scoring items, the expectation was that the majority of the subsets would provide greater results with movement data provided from this segment. However, some scoring items benefited from

the inclusion of the take-off to initial contact segment (S1) with S2. S1A2 was used by 4/13 models, mainly in scoring items that involve the movement of a joint moments prior to initial contact. These items included scoring item 4, where movement data is focused on how the ankle is positioned prior to initial contact. Scoring items 7 and 8 referred to the stance width at initial contact, therefore S1A2 was likely a better option to understand how the feet are moving prior to initial contact, and whether they are moving in opposite directions or closer together. Finally, scoring symmetric contact on initial contact in item 11 likely benefits from the raw movement data to recognise the point at which each ankle sensor identifies initial contact with the ground.

Subsets that produced the greatest models were not consistent for some similar scoring items. Four LESS scoring items pertain to knee-related movement. Within the two knee flexion-related scoring items, scoring item 1 benefited from the raw movement data from all three sensors in S2. In contrast, knee flexion displacement (scoring item 12) produced the best model using the statistical features from the pelvis in S2. However, the two scoring items relating to knee valgus (scoring item 5 and 15) both produced the respective best models using the statistical features from all sensors in S2. Furthermore, the two stance width scoring items (items 7 and 8) used two different sets of data, including the statistical features from all sensors in S1A2 for a narrow width, and the combination of both raw data and statistical features from the ankles in S1A2. Despite the similar movement patterns occurring, different subsets and features with greater importance differed in scoring items where it would be expected to have a similar application. Due to these inconsistencies, the patterns being recognised by each model are likely not applicable for larger LESS datasets as these patterns indicate an overfitting to the dataset, rather than the best performing models training on features that relate to the performance of the movement.

7.2.3 Relevance of Extracted Features

In one relevant study [2], the SHAP values extracted allowed the authors to evaluate relevant features and the most expressive phase in various types of trampoline jumps. The resulting top performing features from the best

performing model (DFF) indicated multiple features that were relevant to the real-world performance of the jump. However, from the results provided in this thesis, the features extracted from the best performing models had limited relevance to the real-world movement patterns of each scoring item.

Scoring item 1 used a variety of values in the accelerometer and gyroscope raw data, which does not allow for any meaningful insight into the associated best-performing model. However, five different temporal measures produced the best model for scoring item 12. In the classification of both of these knee flexion LESS scoring items, no features overlapped between the two best performing models. Furthermore for knee valgus, a mix of statistical features such as maximum and minimum values made up the top feature list for scoring item 5. In contrast, the top performing features for scoring item 15 involved mean values, and the age of the participant. Lastly, using stance width scoring items 7 and 8, none of the statistical features overlapped between the two models.

Across the classified scoring items, the top performing features in many scoring items had little relevance to the real-world application of the LESS scoring item. Some examples of this are the features in scoring item 9 - where the rotation of the foot denotes whether an error is present or absent. In the GNB model, pelvis-based data provided greater relevance to the performance of the model than any ankle-based features did. In the motion of rotating the foot inward or outward, the ankle bone is rotating as well - therefore, the ankle sensor is also rotating at the same time. However, the best performing model did not use this data in any top performing features, which is a contradiction.

In many cases, some obvious patterns were not identified that would have benefited the performance of a model. Using scoring item 11 as an example, symmetric initial contact involved scoring an error when both feet did not come into contact in the same frame of a 2D video. Temporal features included the time from take-off to left and right initial contact separately, or time from left and right initial contact to maximum knee flexion were primary features for detection of this error. In a manual scoring scenario, the timestamps of each value would indicate symmetric contact even without the use of a video. It could be that the higher sampling rate of the IMU sensors lead to higher resolution of symmetry and asymmetry that do not agree with the visual observations from

the videos collected at 120 Hz. Feature relevance varied in majority of the scoring items, with top performing features having no real-world relevance to the movement pattern involved. Once again, this observation is an indication that further application of this approach to additional datasets is likely to be inaccurate.

7.3 Solution Applicability

From the results shown, this approach using sensor data for automation of the LESS is not viable currently. The overall accuracy of many of the models were misleading, with distribution of the classes imbalanced in many cases. Furthermore, the top performing features across many of the scoring items demonstrated limited relevance to the performance of the real movement patterns in manual assessment.

There are a number of factors that may influence or improve performance of sensor data to the LESS scoring items. Primarily, the positioning of the sensors is likely a large contribution to this overall result. As discussed in Section 3.1, positioning of the sensors was essential in providing accurate results in calculating joint angles. Many of the scoring items are in relation to the foot, ankle, knee, hip, and trunk. To this point, the sensor data available from the pelvis and ankles are only in proximity to these joints. Therefore, the features and data available are not going to be exact representations for a large majority of these scoring items. The positioning of the sensors was not selected to align with the movement involved in the 17 LESS items. Nonetheless, this automation approach has demonstrated that despite this positional limitation, some distinguishable patterns were recognisable. Additional or different positioning of the sensors may yield different results.

With an overall LESS score being derived from 17 LESS items, a better approach may have been to combine the classification of individual scoring items into the classification of a total score. Using each of the best performing models to score an individual item for one participant, and summing all these individual scores may have provided a greater accuracy in total score classification than using the raw values. Additionally, this approach may have been benefi-

cial for scoring item 16 and 17, where the impression and quality of the jump is made up of previously-scored items of the LESS. However, this approach was not possible because there was insufficient examples of errors or absences of errors for certain of the items for automation.

The individualisation of movement data created an interesting problem for integrating raw movement data as input data. Each participant moves differently in three axes of sensor data. Some key features or processing techniques may not have had 100% accuracy for all 218 jump occurrences. This inconsistency included the setting of threshold values for capturing initial contact or take-off values, and segmentation of the maximum knee flexion. Threshold values may have picked up noise around the segmented phases, such as detecting the take-off point too early, or detecting maximum knee flexion prior to actual maximum knee flexion. Features were then extracted from these segments, which may have led to inconsistency between participants in some metrics, such as flight time or shorter/longer segment widths than the actual segment.

Additionally, participants perform the DLJL task in many ways, which is reflected in sensor data. The time taken to complete different phases of the landing is different among all participants, therefore aligning raw movement data into one dataset is a complex approach. Features cannot be simply removed from participants that took a longer time to complete the DLJL, as the data provided in these phases of movement is still relevant to the movement patterns of the participant. However, all the values within a feature are compared to one another with the same weight. Replacing the empty values with a filler value ("-1") means the models are finding patterns based on the comparisons between the filler values and other raw data values.

With the use of subsets including a large number of features (where the entire raw movement data is used), feature extraction was not possible on the entire subset. Some of the libraries implemented were limited in their application with large feature counts. The main example in this thesis is the use of the SHAP library, as each feature is evaluated individually. Therefore, SelectPercentile was required to select 1% of the best performing features, rather than the full dataset. In contrast, feature selection techniques could have been im-

plemented to improve the accuracy of models - however, this approach leads to inconsistent feature selection among the models, and therefore model performance comparison would be on different input datasets. Additionally, this thesis focused on the use of raw movement data specifically, which involves the input of the entire movement window.

7.4 Summary of Findings

The application of IMU-derived data and machine learning models for the automation of LESS scores was a more complex problem than initially expected. The weakest performance was observed in total score classification. Weak classification performance was observed across all the other scoring items, despite the use of best performing models and data subsets. In addition to weak performances, the top performing features differed across all scoring items with minimal crossover between scoring items, and limited relevance to the performance of the movement. Additionally, there was limited consistency between the top performing features in scoring items that contained the same movement patterns, indicating the patterns recognised from this evaluation is likely not compatible with further implementations of LESS data. Based on these findings, it is not recommended to apply this approach for the automation of LESS scores using IMU-derived data.

Chapter 8

Conclusion

In this thesis, we investigated the integration of IMU movement data, gathered from the DLJL, with out-of-the-box machine learning models. Customised algorithms were used to process the data into various subsets to assess the performance of the models with differing input features. An evaluation on the performance of these models with the different subsets, and the five features that provide the greatest impact on the models were extracted to assess whether the processed data provides accurate automation of a LESS score.

8.1 Research Questions

To evaluate the applicability of automating the LESS scoring process, this thesis set out to answer two research questions:

1. How accurately can out-of-the-box machine learning models automate LESS scores from IMU data?
2. Which IMU-derived features are the most relevant to the automation of LESS scores?

We address research question one starting in Chapter 4, by discussing the proposed solution design, aimed at manipulating the sensor data into a dataset that can be integrated with machine learning models. Chapter 5 details the implementation of custom algorithms to process the sensor data into a series

of subsets, using a combination of different features extracted from segmented points in the IMU data. Lastly, the evaluation of performance in Chapter 6 effectively addresses the accuracy of the out-of-the-box models, demonstrating a weak performance in the automation of LESS scoring conditions.

To address research question two, we extracted the feature importance of the models used for classifying the results in each scoring item of the LESS, as discussed in Chapter 6. The process used for extracting the top five features with the greatest SHAP values (greatest impact on the prediction of a class) in each scoring item is detailed in Chapter 5. Using the features provided from the results of the evaluation, Chapter 7 details the applicability of these derived measures to future application of the best performing models, and whether these derived measures are both accurate and relevant to real-world prediction of LESS scores. The top performing features provided by the models differed between scoring items, showing minimal relevance to the movement involved in each scoring item.

8.2 Future Work

There are many potential future research points that can be covered, based on the conclusions provided in this thesis. The primary influence on the results of this classification is likely due to the positioning of the sensors. Considering the movement data was provided for this thesis, the positioning of the sensors was not a controllable variable. Movements at the pelvis do not relate to any of the scoring items employed by the LESS explicitly, and the ankle sensors have limited relation to the actual movement patterns. Therefore, results may improve with sensor placement around critical areas of the LESS. Placements could include tracking of the knee joint, hip joint, and trunk. These placements alone would cover a total of nine of the scoring items (53%). Additionally, the placement of a sensor near the top of the foot may provide a measure for tracking stance width, the symmetry in foot contact, and foot position. With different placements, the movement data provided may better relate to the LESS scores. However, including more sensors will introduce additional expenditure, and still requires user intervention for placement of sensors. Therefore, this application would have to be weighed against the accessibility to additional

sensors, and how multiple data sources are used together for automation.

New sensor placements may also provide greater accuracy in the segmentation of the key phases of the DLJL. The algorithm created for identification of maximum knee flexion does not find the exact moment due to the limited movement data available close to the knee. Placement of sensors proximal to the knee joint can provide a more accurate knee flexion angle, and therefore the timestamp at which the maximum flexion occurs.

Due to the varying distribution of classes across the scoring items, many of the best performing models do not get an opportunity to train on the class with less instances and learn the patterns that constitute an error class. Some of the existing LESS scores can be fed back into the training dataset, to improve the balance of the class counts, and therefore allowing the models to learn from more examples of the less-dominant class.

In consideration of the weak performance of the total score automation, using the unique models for each scoring item may provide greater accuracy. In a different approach, each row of LESS data could be used to predict the 17 scoring items, and the classes provided by each model can be summed together to provide a predicted overall score. Comparisons could then be made by the summed overall score, with the actual score provided in the dataset, and whether or not the performance of individual scoring item models outperforms one model for the total score classes.

The subsets using just the IMU-derived features often performed the greatest across the scoring items. This finding is an indication that using the raw movement data is overfitting the model, and may be impacting the accuracy and performance of the models of the test set classification negatively. An improved merging algorithm could be implemented, spreading each jump occurrence in the dataset to the same length and filling the empty values with statistical values (mean, or median between two points for example). Furthermore, feature selection may be a better approach with the IMU-derived features to provide an optimised dataset for the models to work with. Considering these models are also out-of-the-box, customised models may perform better with the data available. Additional approaches could include the combination of both feature selection and hyperparameter tuning to optimise both the dataset and

the model.

8.3 Concluding Remarks

The quantification and understanding of biomechanics involved in a movement is essential for evaluating human movement performance and allowing the appropriate training interventions to be implemented for rehabilitating, mitigating injury risk, or improving performance. The LESS provides a movement screening tool which allows participants to understand their tendency to perform high-risk movement patterns upon a DLJL. Accessibility to performing and scoring the LESS is limited, therefore the automation of the LESS scoring process is proposed.

The solution provided in this thesis was implemented to provide an approach to using collected sensor data from both ankles, and the pelvis to try and accurately classify all the scoring items within the LESS using multiple supervised learning models. An eight-step processing pipeline was outlined, and customised algorithms implemented to process the collected movement data. These stages included the cropping, aligning, filtering, and segmentation of the data, in addition to the dividing, extraction of features, outputting, and merging of the LESS scores with various subsets of features. Subsets included the raw movement data, extracted features, and the combination of both in one dataset.

The best performing models for 12 different scoring items were extracted using a weighted F1-score, and evaluated against ZeroR to determine whether model performance showed a high rate of accuracy. From the results provided, the performance in majority of the scoring items were misleading, indicating high accuracy rates due to the distribution of the classes in the dataset. Only one scoring item model showed a potential in distinguishing between three different classes. Therefore, it was concluded that out-of-the-box machine learning models do not provide a performance great enough to apply this automation approach to further LESS datasets when sensors are located as they were in this thesis.

In addition to the accuracy, the IMU-derived features with the greatest im-

portance on the 12 models did not provide confidence for further application of this automation approach. The features with the greatest impact on these models did not indicate any relevance to real-world performance of a scoring item. Examples of these features included random points in the raw movement data, and extracted features from different sensor axes that did not correlate to the expected axes involved in a movement pattern, such as pelvis sensor features in an ankle-based scoring item. Additionally, top performing features were not consistent between scoring items involving similar movement patterns, indicating the models were identifying patterns from the dataset itself, rather than the relevant features of the DLJL. From these sensor placements, the collection of more data is likely to provide an even weaker performance in automation, as a different sample of participants and scoring classes may not have the same patterns as the dataset used to train these models.

From the insights provided by the evaluation of automated LESS scores, the outcomes of this thesis can be concluded. The out-of-the-box machine learning models did not provide significant improvements in the accuracy of classification in comparison to the ZeroR approach. Additionally, the top performing features were not consistent among scoring item models, showing a large variety in the IMU-derived features that were the most relevant to the LESS scores. Despite a limited confidence in the application of this approach to automating the LESS scoring, this thesis has highlighted areas of development that may prompt a greater performance from the integration of machine learning and IMU-derived data. Although results were misleading, some scoring items provided recognisable patterns in the movement data from anatomical positions that had limited significance to the scoring items of the LESS. By including additional sensors at joint segments relevant to more of the LESS scoring items, better results are likely to be observed. Additionally, the distribution of classes influenced the results of all the models for the scoring items. With the collection of a larger dataset of data from relevant segments, there is potential to automate the LESS scoring process and provide greater accessibility to LESS data for both large-scale and general population testing.

Bibliography

[1] D. A. Padua, S. W. Marshall, M. C. Boling, C. A. Thigpen, W. E. Garrett Jr, and A. I. Beutler, "The Landing Error Scoring System (LESS) Is a Valid and Reliable Clinical Assessment Tool of Jump-Landing Biomechanics: The JUMP-ACL Study," *The American Journal of Sports Medicine*, vol. 37, no. 10, pp. 1996–2002, 2009.

[2] L. Woltmann, C. Hartmann, W. Lehner, P. Rausch, and K. Ferger, "Sensor-Based Jump Detection and Classification With Machine Learning in Trampoline Gymnastics," *German Journal of Exercise and Sport Research*, vol. 53, no. 2, pp. 187–195, 2023.

[3] B. Moses, J. Orchard, and J. Orchard, "Systematic Review: Annual Incidence of ACL Injury and Surgery in Various Populations," *Research in Sports Medicine*, vol. 20, no. 3-4, pp. 157–179, 2012.

[4] K. Hebert-Losier, I. Hanzlikova, C. Zheng, L. Streeter, and M. Mayo, "The 'DEEP' Landing Error Scoring System," *Applied Sciences*, vol. 10, no. 3, 2020.

[5] Y. Liao, A. Vakanski, M. Xian, D. Paul, and R. Baker, "A Review of Computational Approaches for Evaluation of Rehabilitation Exercises," *Computers in Biology and Medicine*, vol. 119, 2020.

[6] E. Van der Kruk and M. M. Reijne, "Accuracy of Human Motion Capture Systems for Sport Applications; State-of-the-Art Review," *European Journal of Sport Science*, vol. 18, no. 6, pp. 806–819, 2018.

[7] J. Hamill and K. M. Knutzen, *Biomechanical Basis of Human Movement*. Lippincott Williams & Wilkins, 2006.

[8] M. Čoh, K. Tomažin, and S. Štuhec, "The Biomechanical Model of the Sprint Start and Block Acceleration," *Facta Universitatis: Series Physical Education & Sport*, vol. 4, no. 2, 2006.

[9] J. Parkkari, P. Kannus, A. Natri, I. Lapinleimu, M. Palvanen, M. Heiskanen, I. Vuori, and M. Järvinen, "Active Living and Injury Risk," *International Journal of Sports Medicine*, vol. 25, no. 3, pp. 209–216, 2004.

[10] D. Murphy, D. Connolly, and B. Beynnon, "Risk Factors for Lower Extremity Injury: A Review of the Literature," *British Journal of Sports Medicine*, vol. 37, no. 1, pp. 13–29, 2003.

[11] E. M. Roos, "Joint Injury Causes Knee Osteoarthritis in Young Adults," *Current Opinion in Rheumatology*, vol. 17, no. 2, pp. 195–200, 2005.

[12] D. T. Felson, "Osteoarthritis as a Disease of Mechanics," *Osteoarthritis and Cartilage*, vol. 21, no. 1, pp. 10–15, 2013.

[13] K. Shah, M. Solan, and E. Dawe, "The Gait Cycle and Its Variations With Disease and Injury," *Orthopaedics and Trauma*, vol. 34, no. 3, pp. 153–160, 2020.

[14] J. M. Hootman, R. Dick, and J. Agel, "Epidemiology of Collegiate Injuries for 15 Sports: Summary and Recommendations for Injury Prevention Initiatives," *Journal of Athletic Training*, vol. 42, no. 2, pp. 311–319, 2007.

[15] Accident Compensation Corporation New Zealand, "Sport and Recreation Injury Statistics," 2023. Retrieved from: `https://www.acc.co.nz/newsroom/media-resources/sport-and-recreation-injury-statistics`.

[16] D. King, P. A. Hume, N. Hardaker, C. Cummins, C. Gissane, and T. Clark, "Sports-Related Injuries in New Zealand: National Insurance (Accident Compensation Corporation) Claims for Five Sporting Codes From 2012 to 2016," *British Journal of Sports Medicine*, vol. 53, no. 16, pp. 1026–1033, 2019.

[17] J. T. Bram, L. C. Magee, N. N. Mehta, N. M. Patel, and T. J. Ganley, "Anterior Cruciate Ligament Injury Incidence in Adolescent Athletes: A

Systematic Review and Meta-Analysis," *The American Journal of Sports Medicine*, vol. 49, no. 7, pp. 1962–1972, 2021.

[18] J. Richards, C. Docherty, B. Arnold, K. Hébert-Losier, C. Häger, B. Mazuquin, and P. Monga, "Biomechanics of Sports Injuries, Their Management and Clinical Considerations," *A Comprehensive Guide to Sports Physiology and Injury Management: An Interdisciplinary Approach*, pp. 47–61, 2020.

[19] B. P. Boden, F. T. Sheehan, J. S. Torg, and T. E. Hewett, "Noncontact Anterior Cruciate Ligament Injuries: Mechanisms and Risk Factors," *American Academy of Orthopaedic Surgeons*, vol. 18, no. 9, pp. 520–527, 2010.

[20] A. S. Fox, "Change-of-Direction Biomechanics: Is What's Best for Anterior Cruciate Ligament Injury Prevention Also Best for Performance?," *Sports Medicine*, vol. 48, no. 8, pp. 1799–1807, 2018.

[21] F. Della Villa, M. Buckthorpe, A. Grassi, A. Nabiuzzi, F. Tosarelli, S. Zaffagnini, and S. Della Villa, "Systematic Video Analysis of ACL Injuries in Professional Male Football (Soccer): Injury Mechanisms, Situational Patterns and Biomechanics Study on 134 Consecutive Cases," *British Journal of Sports Medicine*, vol. 54, no. 23, pp. 1423–1432, 2020.

[22] R. Fagenbaum and W. G. Darling, "Jump Landing Strategies in Male and Female College Athletes and the Implications of Such Strategies for Anterior Cruciate Ligament Injury," *The American Journal of Sports Medicine*, vol. 31, no. 2, pp. 233–240, 2003.

[23] T. E. Hewett, G. D. Myer, K. R. Ford, R. S. Heidt Jr, A. J. Colosimo, S. G. McLean, A. J. Van den Bogert, M. V. Paterno, and P. Succop, "Biomechanical Measures of Neuromuscular Control and Valgus Loading of the Knee Predict Anterior Cruciate Ligament Injury Risk in Female Athletes: A Prospective Study," *The American Journal of Sports Medicine*, vol. 33, no. 4, pp. 492–501, 2005.

[24] L. C. Mitchell, K. R. Ford, S. Minning, G. D. Myer, R. E. Mangine, and T. E. Hewett, "Medial Foot Loading on Ankle and Knee Biomechanics,"

North American Journal of Sports Physical Therapy: NAJSPT, vol. 3, no. 3, pp. 133–140, 2008.

[25] I. Hanzlíková and K. Hébert-Losier, "Is the Landing Error Scoring System Reliable and Valid? A Systematic Review," *Sports Health*, vol. 12, no. 2, pp. 181–188, 2020.

[26] J. Larwa, C. Stoy, R. S. Chafetz, M. Boniello, and C. Franklin, "Stiff Landings, Core Stability, and Dynamic Knee Valgus: A Systematic Review on Documented Anterior Cruciate Ligament Ruptures in Male and Female Athletes," *International Journal of Environmental Research and Public Health*, vol. 18, no. 7, 2021.

[27] D. Sugimoto, G. D. Myer, K. D. B. Foss, and T. E. Hewett, "Specific Exercise Effects of Preventive Neuromuscular Training Intervention on Anterior Cruciate Ligament Injury Risk Reduction in Young Females: Meta-Analysis and Subgroup Analysis," *British Journal of Sports Medicine*, vol. 49, no. 5, pp. 282–289, 2015.

[28] D. Sugimoto, G. D. Myer, J. M. McKeon, and T. E. Hewett, "Evaluation of the Effectiveness of Neuromuscular Training to Reduce Anterior Cruciate Ligament Injury in Female Athletes: A Critical Review of Relative Risk Reduction and Numbers-Needed-to-Treat Analyses," *British Journal of Sports Medicine*, vol. 46, no. 14, pp. 979–988, 2012.

[29] I. Hanzlíková, J. Athens, and K. Hébert-Losier, "Factors Influencing the Landing Error Scoring System: Systematic Review with Meta-Analysis," *Journal of Science and Medicine in Sport*, vol. 24, no. 3, pp. 269–280, 2021.

[30] D. R. Bell, M. D. Smith, A. P. Pennuto, M. R. Stiffler, and M. E. Olson, "Jump-Landing Mechanics After Anterior Cruciate Ligament Reconstruction: A Landing Error Scoring System Study," *Journal of Athletic Training*, vol. 49, no. 4, pp. 435–441, 2014.

[31] I. Hanzlíková and K. Hébert-Losier, "Clinical Implications of Landing Distance on Landing Error Scoring System Scores," *Journal of Athletic Training*, vol. 56, no. 6, pp. 572–577, 2021.

[32] K. Hébert-Losier, C. Boswell-Smith, and I. Hanzlíková, "Effect of Footwear Versus Barefoot on Double-Leg Jump-Landing and Jump Height Measures: A Randomized Cross-Over Study," *International Journal of Sports Physical Therapy*, vol. 18, no. 4, pp. 845–855, 2023.

[33] J. Onate, N. Cortes, C. Welch, and B. Van Lunen, "Expert Versus Novice Interrater Reliability and Criterion Validity of the Landing Error Scoring System," *Journal of Sport Rehabilitation*, vol. 19, no. 1, pp. 41–56, 2010.

[34] A. M. Valevicius, P. Y. Jun, J. S. Hebert, and A. H. Vette, "Use of Optical Motion Capture for the Analysis of Normative Upper Body Kinematics During Functional Upper Limb Tasks: A Systematic Review," *Journal of Electromyography and Kinesiology*, vol. 40, pp. 1–15, 2018.

[35] K. Jacobs, D. Riveros, H. K. Vincent, and D. C. Herman, "The Effect of Landing Surface on Landing Error Scoring System Grades," *Sports Biomechanics*, 2018.

[36] C. M. Kuenze, S. Trigsted, C. Lisee, E. Post, and D. R. Bell, "Sex Differences on the Landing Error Scoring System Among Individuals With Anterior Cruciate Ligament Reconstruction," *Journal of Athletic Training*, vol. 53, no. 9, pp. 837–843, 2018.

[37] C. John, A. Stotz, J. Gmachowski, A. L. Rahlf, D. Hamacher, K. Hollander, and A. Zech, "Is an Elastic Ankle Support Effective in Improving Jump Landing Performance, and Static and Dynamic Balance in Young Adults With and Without Chronic Ankle Instability?," *Journal of Sport Rehabilitation*, vol. 29, no. 6, pp. 789–794, 2019.

[38] E. O'Malley, J. C. Murphy, U. M. Persson, C. Gissane, and C. Blake, "The Effects of the Gaelic Athletic Association 15 Training Program on Neuromuscular Outcomes in Gaelic Football and Hurling Players: A Randomized Cluster Trial," *The Journal of Strength & Conditioning Research*, vol. 31, no. 8, pp. 2119–2130, 2017.

[39] H. C. Smith, R. J. Johnson, S. J. Shultz, T. Tourville, L. A. Holterman, J. Slauterbeck, P. M. Vacek, and B. D. Beynnon, "A Prospective Evaluation of the Landing Error Scoring System (LESS) as a Screening Tool

for Anterior Cruciate Ligament Injury Risk," *The American Journal of Sports Medicine*, vol. 40, no. 3, pp. 521–526, 2012.

[40] P. Lisman, J. N. Wilder, J. Berenbach, E. Jiao, and B. Hansberger, "The Relationship Between Landing Error Scoring System Performance and Injury in Female Collegiate Athletes," *International Journal of Sports Physical Therapy*, vol. 16, no. 6, pp. 1415–1425, 2021.

[41] T. C. Mauntel, D. A. Padua, L. E. Stanley, B. S. Frank, L. J. DiStefano, K. Y. Peck, K. L. Cameron, and S. W. Marshall, "Automated Quantification of the Landing Error Scoring System With a Markerless Motion-Capture System," *Journal of Athletic Training*, vol. 52, no. 11, pp. 1002–1009, 2017.

[42] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7291–7299, 2017.

[43] F. Porciuncula, A. V. Roto, D. Kumar, I. Davis, S. Roy, C. J. Walsh, and L. N. Awad, "Wearable Movement Sensors for Rehabilitation: A Focused Review of Technological and Clinical Advances," *PM&R*, vol. 10, no. 9, pp. S220–S232, 2018.

[44] M. Iosa, P. Picerno, S. Paolucci, and G. Morone, "Wearable Inertial Sensors for Human Movement Analysis," *Expert Review of Medical Devices*, vol. 13, no. 7, pp. 641–659, 2016.

[45] T. Gujarathi and K. Bhole, "Gait Analysis Using IMU Sensor," in *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pp. 1–5, IEEE, 2019.

[46] M. Lapinski, E. Berkson, T. Gill, M. Reinold, and J. A. Paradiso, "A Distributed Wearable, Wireless Sensor System for Evaluating Professional Baseball Pitchers and Batters," in *2009 International Symposium on Wearable Computers*, pp. 131–138, IEEE, 2009.

[47] Z. Wang, X. Shi, J. Wang, F. Gao, J. Li, M. Guo, H. Zhao, and S. Qiu, "Swimming Motion Analysis and Posture Recognition Based on Wear-

able Inertial Sensors," in *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*, pp. 3371–3376, IEEE, 2019.

[48] S. Z. Homayounfar and T. L. Andrew, "Wearable Sensors for Monitoring Human Motion: A Review on Mechanisms, Materials, and Challenges," *SLAS TECHNOLOGY: Translating Life Sciences Innovation*, vol. 25, no. 1, pp. 9–24, 2020.

[49] J. H. Migueles, C. Cadenas-Sanchez, U. Ekelund, C. Delisle Nyström, J. Mora-Gonzalez, M. Löf, I. Labayen, J. R. Ruiz, and F. B. Ortega, "Accelerometer Data Collection and Processing Criteria to Assess Physical Activity and Other Outcomes: A Systematic Review and Practical Considerations," *Sports Medicine*, vol. 47, pp. 1821–1845, 2017.

[50] D. Mizell, "Using Gravity to Estimate Accelerometer Orientation," in *Seventh IEEE International Symposium on Wearable Computers, 2003. Proceedings*, pp. 252–253, Citeseer, 2003.

[51] B. Akinnuli, O. Dahunsi, S. Ayodeji, and O. Bodunde, "Investigation and Validation of Consumer Device Accelerometers for the Assessment of Whole-Body Vibration," *Cogent Engineering*, vol. 4, no. 1, 2017.

[52] A. Nadeau, K. Dinesh, G. Sharma, and M. Xiong, "In-Situ Calibration of Accelerometers in Body-Worn Sensors Using Quiescent Gravity," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2192–2196, IEEE, 2017.

[53] A. R. Javed, M. U. Sarwar, S. Khan, C. Iwendi, M. Mittal, and N. Kumar, "Analyzing the Effectiveness and Contribution of Each Axis of Tri-axial Accelerometer Sensor for Accurate Activity Recognition," *Sensors*, vol. 20, no. 8, 2020.

[54] M. Boutaayamou, C. Schwartz, J. Stamatakis, V. Denoël, D. Maquet, B. Forthomme, J.-L. Croisier, B. Macq, J. G. Verly, G. Garraux, *et al.*, "Development and Validation of an Accelerometer-Based Method for Quantifying Gait Events," *Medical Engineering & Physics*, vol. 37, no. 2, pp. 226–232, 2015.

[55] K. Sato, S. L. Smith, and W. A. Sands, "Validation of an Accelerometer for Measuring Sport Performance," *The Journal of Strength & Conditioning Research*, vol. 23, no. 1, pp. 341–347, 2009.

[56] V. M. Passaro, A. Cuccovillo, L. Vaiani, M. De Carlo, and C. E. Campanella, "Gyroscope Technology and Applications: A Review in the Industrial Perspective," *Sensors*, vol. 17, no. 10, 2017.

[57] D. McGrath, B. R. Greene, K. J. O'Donovan, and B. Caulfield, "Gyroscope-Based Assessment of Temporal Gait Parameters During Treadmill Walking and Running," *Sports Engineering*, vol. 15, pp. 207–213, 2012.

[58] T. Holleczek, J. Schoch, B. Arnrich, and G. Tröster, "Recognizing Turns and Other Snowboarding Activities With a Gyroscope," in *International Symposium on Wearable Computers (ISWC) 2010*, pp. 1–8, IEEE, 2010.

[59] H. J. Williams, M. D. Holton, E. L. Shepard, N. Largey, B. Norman, P. G. Ryan, O. Duriez, M. Scantlebury, F. Quintana, E. A. Magowan, *et al.*, "Identification of Animal Movement Patterns Using Tri-axial Magnetometry," *Movement Ecology*, vol. 5, pp. 1–14, 2017.

[60] F. Wittmann, O. Lambercy, and R. Gassert, "Magnetometer-Based Drift Correction During Rest in IMU Arm Motion Tracking," *Sensors*, vol. 19, no. 6, 2019.

[61] R. Takeda, G. Lisco, T. Fujisawa, L. Gastaldi, H. Tohyama, and S. Tadano, "Drift Removal for Improving the Accuracy of Gait Parameters Using Wearable Sensor Systems," *Sensors*, vol. 14, no. 12, pp. 23230–23247, 2014.

[62] M. Kok and T. B. Schön, "Magnetometer Calibration Using Inertial Sensors," *IEEE Sensors Journal*, vol. 16, no. 14, pp. 5679–5689, 2016.

[63] I. Poitras, F. Dupuis, M. Bielmann, A. Campeau-Lecours, C. Mercier, L. J. Bouyer, and J.-S. Roy, "Validity and Reliability of Wearable Sensors for Joint Angle Estimation: A Systematic Review," *Sensors*, vol. 19, no. 7, p. 1555, 2019.

[64] A. Blake, G. Winstanley, and W. Wilkinson, "Deriving Displacement From 3-Axis Accelerometers," in *Computer Games, Multimedia & Allied Technology*, 2009.

[65] W. Kassiano, B. Costa, G. Kunevaliki, D. Soares, G. Zacarias, I. Manske, Y. Takaki, M. F. Ruggiero, N. Stavinski, J. Francsuel, *et al.*, "Greater Gastrocnemius Muscle Hypertrophy After Partial Range of Motion Training Performed at Long Muscle Lengths," *The Journal of Strength & Conditioning Research*, pp. 10–1519, 2022.

[66] K. Hébert-Losier, A. G. Schneiders, R. J. Newsham-West, and S. J. Sullivan, "Scientific Bases and Clinical Utilisation of the Calf-Raise Test," *Physical Therapy in Sport*, vol. 10, no. 4, pp. 142–149, 2009.

[67] B. J. Stetter, S. Ringhof, F. C. Krafft, S. Sell, and T. Stein, "Estimation of Knee Joint Forces in Sport Movements Using Wearable Sensors and Machine Learning," *Sensors*, vol. 19, no. 17, 2019.

[68] Y. Luo, S. M. Coppola, P. C. Dixon, S. Li, J. T. Dennerlein, and B. Hu, "A Database of Human Gait performance on irregular and uneven surfaces collected by wearable sensors," *Scientific Data*, vol. 7, no. 1, 2020.

[69] T. Liu, Y. Inoue, and K. Shibata, "Development of a Wearable Sensor System for Quantitative Gait Analysis," *Measurement*, vol. 42, no. 7, pp. 978–988, 2009.

[70] S. Mohammed, A. Same, L. Oukhellou, K. Kong, W. Huo, and Y. Amirat, "Recognition of Gait Cycle Phases Using Wearable Sensors," *Robotics and Autonomous Systems*, vol. 75, pp. 50–59, 2016.

[71] T. Seel, J. Raisch, and T. Schauer, "IMU-Based Joint Angle Measurement for Gait Analysis," *Sensors*, vol. 14, no. 4, pp. 6891–6909, 2014.

[72] R. Aylward and J. A. Paradiso, "A Compact, High-Speed, Wearable Sensor Network for Biomotion Capture and Interactive Media," in *Proceedings of the 6th International Conference on Information Processing in Sensor Networks*, pp. 380–389, 2007.

[73] L. Büthe, U. Blanke, H. Capkevics, and G. Tröster, "A Wearable Sensing System for Timing Analysis in Tennis," in *2016 IEEE 13th International*

Conference on Wearable and Implantable Body Sensor Networks (BSN), pp. 43–48, IEEE, 2016.

[74] N. P. Brouwer, T. Yeung, M. F. Bobbert, and T. F. Besier, "3D Trunk Orientation Measured Using Inertial Measurement Units During Anatomical and Dynamic Sports Motions," *Scandinavian Journal of Medicine & Science in Sports*, vol. 31, no. 2, pp. 358–370, 2021.

[75] M. Lapinski, C. Brum Medeiros, D. Moxley Scarborough, E. Berkson, T. J. Gill, T. Kepple, and J. A. Paradiso, "A Wide-Range, Wireless Wearable Inertial Motion Sensing System for Capturing Fast Athletic Biomechanics in Overhead Pitching," *Sensors*, vol. 19, no. 17, 2019.

[76] K. Karatrantou, V. Gerodimos, V. Voutselas, N. Manouras, K. Famisis, and P. Ioakimidis, "Can Sport-Specific Training Affect Vertical Jumping Ability During Puberty?," *Biology of Sport*, vol. 36, no. 3, pp. 217–224, 2019.

[77] S. Marković, M. Dopsaj, S. Tomažič, A. Kos, A. Nedeljković, and A. Umek, "Can IMU Provide an Accurate Vertical Jump Height Estimate?," *Applied Sciences*, vol. 11, no. 24, 2021.

[78] P. Picerno, V. Camomilla, and L. Capranica, "Countermovement Jump Performance Assessment Using a Wearable 3D Inertial Measurement Unit," *Journal of Sports Sciences*, vol. 29, no. 2, pp. 139–146, 2011.

[79] K. MacDonald, R. Bahr, J. Baltich, J. L. Whittaker, and W. H. Meeuwisse, "Validation of an Inertial Measurement Unit for the Measurement of Jump Count and Height," *Physical Therapy in Sport*, vol. 25, pp. 15–19, 2017.

[80] J. M. Sheppard, A. A. Dingley, I. Janssen, W. Spratford, D. W. Chapman, and R. U. Newton, "The Effect of Assisted Jumping on Vertical Jump Height in High-Performance Volleyball Players," *Journal of Science and Medicine in Sport*, vol. 14, no. 1, pp. 85–89, 2011.

[81] A. Kotsifaki, S. Van Rossom, R. Whiteley, V. Korakakis, R. Bahr, V. Sideris, and I. Jonkers, "Single Leg Vertical Jump Performance Identifies Knee Function Deficits at Return to Sport After ACL Reconstruc-

tion in Male Athletes," *British Journal of Sports Medicine*, vol. 56, no. 9, pp. 490–498, 2022.

[82] B. Fan, H. Xia, J. Xu, Q. Li, and P. B. Shull, "IMU-Based Knee Flexion, Abduction and Internal Rotation Estimation During Drop Landing and Cutting Tasks," *Journal of Biomechanics*, vol. 124, 2021.

[83] E. M. Nijmeijer, P. Heuvelmans, R. Bolt, A. Gokeler, E. Otten, and A. Benjaminse, "Concurrent Validation of the Xsens IMU System of Lower-Body Kinematics in Jump-Landing and Change-of-Direction Tasks," *Journal of Biomechanics*, vol. 154, 2023.

[84] L. Chia, J. T. Andersen, M. J. McKay, J. Sullivan, T. Megalaa, and E. Pappas, "Evaluating the Validity and Reliability of Inertial Measurement Units for Determining Knee and Trunk Kinematics During Athletic Landing and Cutting Movements," *Journal of Electromyography and Kinesiology*, vol. 60, 2021.

[85] C.-M. Fong, J. T. Blackburn, M. F. Norcross, M. McGrath, and D. A. Padua, "Ankle-Dorsiflexion Range of Motion and Landing Biomechanics," *Journal of Athletic Training*, vol. 46, no. 1, pp. 5–10, 2011.

[86] V. Camomilla, R. Dumas, and A. Cappozzo, "Human Movement Analysis: The Soft Tissue Artefact Issue," *Journal of Biomechanics*, vol. 62, pp. 1–4, 2017.

[87] A. Ancillao, E. Aertbeliën, and J. De Schutter, "Effect of the Soft Tissue Artifact on Marker Measurements and on the Calculation of the Helical Axis of the Knee During a Gait Cycle: A Study on the CAMS-Knee Data Set," *Human Movement Science*, vol. 80, 2021.

[88] N. Oliveira, J. Park, and P. Barrance, "Using Inertial Measurement Unit Sensor Single Axis Rotation Angles for Knee and Hip Flexion Angle Calculations During Gait," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 11, pp. 80–86, 2022.

[89] S. Bakhshi, M. H. Mahoor, and B. S. Davidson, "Development of a Body Joint Angle Measurement System Using IMU Sensors," in *2011*

Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 6923–6926, IEEE, 2011.

[90] T. H. Jeon and J. K. Lee, "IMU-Based Joint Angle Estimation Under Various Walking and Running Conditions," *Journal of the Korean Society for Precision Engineering*, vol. 35, no. 12, pp. 1199–1204, 2018.

[91] G. Guerra-Filho, "Optical Motion Capture: Theory and Implementation," *RITA*, vol. 12, no. 2, pp. 61–90, 2005.

[92] M. Menolotto, D.-S. Komaris, S. Tedesco, B. O'Flynn, and M. Walsh, "Motion Capture Technology in Industrial Applications: A Systematic Review," *Sensors*, vol. 20, no. 19, 2020.

[93] G. Delgado-García, J. Vanrenterghem, E. J. Ruiz-Malagón, P. Molina-García, J. Courel-Ibáñez, and V. M. Soto-Hermoso, "IMU Gyroscopes Are a Valid Alternative to 3D Optical Motion Capture System for Angular Kinematics Analysis in Tennis," *Proceedings of the Institution of Mechanical Engineers, Part P: Journal of Sports Engineering and Technology*, vol. 235, no. 1, pp. 3–12, 2021.

[94] M. Roell, H. Mahler, J. Lienhard, D. Gehring, A. Gollhofer, and K. Roecker, "Validation of wearable sensors during team sport-specific movements in indoor environments," *Sensors*, vol. 19, no. 16, 2019.

[95] S. Milanko and S. Jain, "LiftRight: Quantifying Strength Training Performance Using a Wearable Sensor," *Smart Health*, vol. 16, 2020.

[96] M. S. Walsh, K. R. Ford, K. J. Bangen, G. D. Myer, and T. E. Hewett, "The Validation of a Portable Force Plate for Measuring Force-Time Data During Jumping and Landing Tasks," *The Journal of Strength & Conditioning Research*, vol. 20, no. 4, pp. 730–734, 2006.

[97] M. R. Patterson, W. Johnston, N. O'Mahony, S. O'Mahony, E. Nolan, and B. Caulfield, "Validation of Temporal Gait Metrics From Three IMU Locations to the Gold Standard Force Plate," in *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 667–671, IEEE, 2016.

[98] P. Miranda-Oliveira, M. Branco, and O. Fernandes, "Accuracy of Inertial Measurement Units When Applied to the Countermovement Jump of Track and Field Athletes," *Sensors*, vol. 22, no. 19, 2022.

[99] A. Pouliot-Laforte, L. Veilleux, F. Rauch, and M. Lemay, "Validity of an Accelerometer as a Vertical Ground Reaction Force Measuring Device in Healthy Children and Adolescents and in Children and Adolescents With Osteogenesis Imperfecta Type I," *Journal of Musculoskeletal and Neuronal Interactactions*, vol. 14, no. 2, pp. 155–161, 2014.

[100] L. Kakadiaris and D. Metaxas, "Model-Based Estimation of 3D Human Motion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, pp. 1453–1459, 2000.

[101] M. P. Mavor, G. B. Ross, A. L. Clouthier, T. Karakolis, and R. B. Graham, "Validation of an IMU Suit for Military-Based Tasks," *Sensors*, vol. 20, no. 15, 2020.

[102] O. Dehzangi, M. Taherisadr, and R. ChangalVala, "IMU-Based Gait Recognition Using Convolutional Neural Networks and Multi-sensor Fusion," *Sensors*, vol. 17, no. 12, 2017.

[103] W. Kim and M. Kim, "On-line detection and segmentation of sports motions using a wearable sensor," *Sensors*, vol. 18, no. 3, 2018.

[104] D. Hendry, R. Leadbetter, K. McKee, L. Hopper, C. Wild, P. O'sullivan, L. Straker, and A. Campbell, "An Exploration of Machine-Learning Estimation of Ground Reaction Force From Wearable Sensor Data," *Sensors*, vol. 20, no. 3, 2020.

[105] C. R. Chaaban, N. T. Berry, C. Armitano-Lago, A. W. Kiefer, M. J. Mazzoleni, and D. A. Padua, "Combining Inertial Sensors and Machine Learning to Predict vGRF and Knee Biomechanics During a Double Limb Jump Landing Task," *Sensors*, vol. 21, no. 13, 2021.

[106] G. Mascia, B. De Lazzari, and V. Camomilla, "Machine Learning Aided Jump Height Estimate Democratization Through Smartphone Measures," *Frontiers in Sports and Active Living*, vol. 5, 2023.

[107] M. G. White, N. E. Bezodis, J. Neville, H. Summers, and P. Rees, "Determining Jumping Performance From a Single Body-Worn Accelerometer Using Machine Learning," *PLOS ONE*, vol. 17, no. 2, pp. 1–25, 2022.

[108] S. Cerfoglio, M. Galli, M. Tarabini, F. Bertozzi, C. Sforza, and M. Zago, "Machine Learning-Based Estimation of Ground Reaction Forces and Knee Joint Kinetics From Inertial Sensors While Performing a Vertical Drop Jump," *Sensors*, vol. 21, no. 22, 2021.

[109] Y.-Y. Chiang, W.-Y. Shih, W.-H. Chen, J.-L. Huang, and T.-Y. Shiang, "A Machine Learning-Based Countermovement Performance Measurement Method Using a Wearable IMU," in *2020 International Conference on Pervasive Artificial Intelligence (ICPAI)*, pp. 79–85, IEEE, 2020.

[110] E. Halilaj, A. Rajagopal, M. Fiterau, J. L. Hicks, T. J. Hastie, and S. L. Delp, "Machine Learning in Human Movement Biomechanics: Best Practices, Common Pitfalls, and New Opportunities," *Journal of Biomechanics*, vol. 81, pp. 1–11, 2018.

[111] H. Lim, B. Kim, and S. Park, "Prediction of Lower Limb Kinetics and Kinematics During Walking by a Single IMU on the Lower Back Using Machine Learning," *Sensors*, vol. 20, no. 1, 2019.

[112] A. de Cheveigné and I. Nelken, "Filters: When, Why, and How (Not) to Use Them," *Neuron*, vol. 102, no. 2, pp. 280–293, 2019.

[113] P. Cunningham, M. Cord, and S. J. Delany, "Supervised Learning," in *Machine Learning Techniques for Multimedia: Case Studies on Organization and Retrieval*, pp. 21–49, Springer, 2008.

[114] C. Rudin, "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

[115] I. H. Sarker, "Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions," *SN Computer Science*, vol. 2, no. 6, 2021.

[116] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[117] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, "From Local Explanations to Global Understanding With Explainable AI for Trees," *Nature Machine Intelligence*, vol. 2, no. 1, pp. 56–67, 2020.

[118] B. R. Umberger, "Mechanics of the Vertical Jump and Two-Joint Muscles: Implications for Training," *Strength & Conditioning Journal*, vol. 20, no. 5, pp. 70–74, 1998.

Appendix A: Ethics Approval

The University of Waikato
Private Bag 3105
Hamilton, New Zealand, 3240
0800 WAIKATO (924 528)

HECS Human Ethics Committee
Brett Langley
Telephone +64 77 838 4060
Hecs-ethics@waikato.ac.nz

THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

4 August 2023

Kim Hebert-Losier
Steve Finlayson
Merel Hoskens
Liis Uiga
Ivana Hanzlikova

Re: HECS Ethics Approval of Application HREC(HECS)2023#43 "Landing error scoring system: Effect of head position and dual tasking"

Dear Kim:

Thank you for submitting your amended application HREC(HECS)2023#43 for ethical approval.

We are pleased to provide formal approval for your project, including the following activities:

- Recruitment of approximately 48 physically active and healthy male and female (>16 years of age) participants for a study that aims to examine the influence of head position and dual tasking on the Landing Error Scoring System (LESS) scores.
- Have participants complete a baseline data collection form.
- Have participants complete a cognitive test.
- Have participants jump from a 30-cm box, and immediately jump upwards for maximal height. This will be done three times for each of three conditions (control, head turned to the right, or while performing a dual task -counting backwards- that enhances cognitive load) and recorded using 2D video capture and force plates.
- In total, the entire session should take approximately 35 minutes to complete.

Please contact the committee by email (hecs-ethics@waikato.ac.nz) if you wish to make changes to your project as it unfolds, quoting your application number with your future correspondence. Any minor changes or additions to the approved research activities can be handled outside the monthly application cycle.

We wish you all the best with your research.

Kind regards,

Brett Langley, PhD
Chairperson
HECS Human Ethics Committee
University of Waikato