The Identification of Foot-Strike Patterns and Prediction of Running Related Injuries

MSc Research Project
Data Analytics

Shane Gore
Student ID: x18174175

School of Computing
National College of Ireland

Supervisor: Dr Catherine Mulwa

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Shane Gore
Student ID:	x18174175
Programme:	Data Analytics
Year:	2020
Module:	MSc Research Project
Supervisor:	Dr Catherine Mulwa
Submission Due Date:	17/08/2020
Project Title:	The Identification of Foot-Strike Patterns and Prediction of Running Related Injuries
Word Count:	9260
Page Count:	27

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Shane Gore
Date:	17th August 2020

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

The Identification of Foot-Strike Patterns and Prediction of Running Related Injuries

Shane Gore
x18174175

Abstract

It is suggested that how the foot strikes the ground while running may be an important injury risk factor. Despite growing interest in foot-strike pattern and injury, only a limited number of studies have examined this relationship, with inconsistent findings. One source of this inconsistency may be how foot-strike pattern is defined. Indeed, current definitions are generally based on the arbitrary division of the foot into three equal parts. The aim of this project was twofold; 1) to identify foot-strike groupings using a clustering approach and to assess their relationship with injury and 2) to determine if any of the running biomechanics could predict injury classification. 3D biomechanical running data, collected prospectively, from 282 participants and 47,423 foot-strikes was explored along with injury occurrence. Six clustering algorithms were implemented and assessed with bootstrapped resamples of the Adjusted Rand Index (ARI). Mean ARI scores ranged from 0-0.007 indicating almost random assignment to the injury class. Six classification algorithms were then implemented and assessed with bootstrapped resamples of accuracy, sensitivity and specificity. The Random Forest model demonstrated the best performance (accuracy: 71%, specificity: 65%, sensitivity: 74%) and was significantly different than predicting the majority class ($p<0.01$). The final model contained ten features, of which, foot mechanics were not included. Collectively, these results suggest that foot-strike pattern is not important with respect to injury risk and the final features utilised by the Random Forest model likely represent the best targets for the prevention of running related injury.

1 Introduction

Running is a popular form of exercise with substantial physical, social and mental benefits (Rothschild, 2012). Indeed, it has been estimated that over 50 million people participate in running in Europe alone (Scheerder et al., 2015). Despite the popularity of this activity, up to 79% of recreational runners become injured every year (Van Gent et al., 2007), and the biomechanical risk factors for running related injuries (RRI) remain poorly understood (Bahr, 2016). While the cause of RRI is likely to be multifactorial, it has been suggested that foot-strike technique may represent an important modifiable risk factor (Pizzuto et al., 2016). This is because injuries are caused by excessive loading relative to tissue strength and modifying foot-strike technique has been shown to alter the characteristics of loading while running (Almeida et al., 2015). To date however, there have been a limited number of studies that have appropriately examined this concept and their findings have not been consistent. On the one hand some authors have identified an association between certain foot-strike patterns and injury (Goss and Gross, 2012, Daoud et al., 2012, Futrell et al., 2018) while others have found no relationship

(Warr et al., 2015, Paquette et al., 2017, Dudley et al., 2017, Donoghue et al., 2008, Kuhman et al., 2016, Messier et al., 2018). A methodological consideration that may, at least in part, explain these contrasting findings is the way foot-strike pattern is defined.

While running, it has been suggested that humans can utilise one of three foot-strike patterns; forefoot, midfoot and rearfoot, defined respectively based on which part of the foot contacts the ground first (Cavanagh and Lafortune, 1980) or using the angle of the foot at ground contact as a surrogate measure (Altman and Davis, 2012). However, there are at least two main challenges with this classification definition. Firstly, representing foot-strike pattern as a single instance in time may result in discarding potentially important information pertinent to injury risk. While it is common in biomechanics to represent a movement pattern using a single discrete point, this may misrepresent the movement pattern being examined (Pataky, 2012). A more informative approach would be to utilise the data contained in the whole movement pattern from the moment of initial contact to toe off at the end of the stance phase. Secondly, and perhaps most importantly, current classification approaches are based on the seminal work by Cavanagh and Lafortune (1980) who arbitrarily divided the foot into three equal parts. A more appropriate approach would leverage unsupervised clustering to identify groupings within the waveform data. To date however, to the best of the candidate's knowledge, no research has explictly investigated foot-strike patterns using unsupervised clustering methodologies.

Finally, while a number of studies have investigated the relationship between foot-strike pattern and injury, there remains a dearth of prospective research that have investigated foot-strike pattern and injury using a multivariate approach. This is important as the movement of the foot may only become relevant to injury risk, when considered in conjunction with the movement of other segments/joints of the body.

1.1 Research Question and Hypothesis

The primary aim of this project was to identify the presence of foot-strike groupings using unsupervised clustering approaches and determine their association with injury. The secondary research aim was to determine if any of the movement biomechanics from the lower limbs and trunk could predict those who would go on to be injured, and if the foot would be included in the final predictive models.

It was hypothesised that the association between injury class and the identified clusters would be significantly and substantially different than random assignment, and that the final classification models would include the foot in their prediction of injury class.

RQ: *"Can the use of unsupervised clustering (K-Means, Hierarchical, Mean Shift, OPTICS, HDBSCAN and Spectral clustering) enhance the identification of foot-strike patterns related to running injuries to inform best clinical practice?"*

Sub RQ: *"Can whole body movement biomechanics predict running injury classification using supervised learning (Naïve Bayes, Elastic Net Logistic Regression, Bagged SVM, Random Forest, AdaBoost and a weighted Stacked Ensemble) to identify targets for injury prevention and if so, will the movement of the foot be included in the best performing model?"* This is important, as the movement of the foot may only become informative with respect to injury risk when considered in conjunction with the biomechanics of the whole body.

1.2 Research Objectives and Contributions

Research Objectives The research questions were investigated through a series of objectives as outlined below.

1. Critically evaluate the literature pertaining to foot-strike and injury, and clustering in the biomechanics domain.
2. Implement event detection and dynamic time warping via landmark registration to enable feature extraction from the stance phase of a running cycle.
3. Extract relevant features using the concept of Analysis of Characterising Phases and using the TSFresh Python package.
4. Select relevant foot features for clustering using spectral feature selection.
5. Implement six clustering models (K-Means, Hierarchical, Mean Shift, OPTICS, HDBSCAN and Spectral clustering).
6. Evaluate and compare the six clustering models (K-Means, Hierarchical, Mean Shift, OPTICS, HDBSCAN and Spectral clustering).
7. Select relevant features for classification using a genetic search algorithm and tune the models using Bayesian optimisation.
8. Implement six predictive classification models (Naïve Bayes, Elastic Net Logistic Regression, Bagged SVM, Random Forest, AdaBoost and a weighted Stacked Ensemble).
9. Evaluate and compare the six predictive classification models (Naïve Bayes, Elastic Net Logistic Regression, Bagged SVM, Random Forest, AdaBoost and a weighted Stacked Ensemble).

Contributions The major contribution of this project was the identification of naturally occurring foot-strike patterns using unsupervised learning and determining their association with injury. To date, to the best of the candidate's knowledge, no research had utilised clustering to identify foot-strike patterns. Rather, previous research relied on the arbitrary division of the foot into three equal parts. As such, this current project can be seen as state of the art in this domain. Additionally, with respect to the secondary aim and sub research question, this current project extended upon the literature by utilising best analytical practices not utilised in the prospective research investigating running injuries. These included the use of continuous rather than discrete point analytics, the use of multiple machine learning models which had not been utilised in the literature, advanced feature engineering and feature selection techniques, and the use of out of sample testing to ensure generalizability of the findings. Finally, the overall findings from this research project would suggest that the movement of the foot in the sagittal plane is not important with respect to running related injury. This finding will influence best clinical practice for the prevention and rehabilitation of running injuries. Rather than focusing on foot-strike pattern, clinicians can be advised to target the final features utilised in the random forest model to predict injury classification. Furthermore, the results of this project will have implications for manufacturers who produce products based on the concept of foot-strike pattern.

The remaining document is structured as follows: Section 2 presents a literature review on foot-strike classification as well as the use of clustering within the biomechanics domain. Section 3 will present the scientific methodology implemented in this project. Section 4 will illustrate the three tier design specification used in this project. Section 5 will detail the implementation undertaken in this project including the clustering and classification approaches. Section 6 will present the evaluations findings from the experiments in this project along with a detailed discussion of the results. Finally, section 7 will summarize the take home messages from this project and propose future research.

2 Related Work

2.1 Foot-strike Pattern and Injury

Recently there has been a growing interest in the association between RRI and foot-strike pattern as a potentially modifiable risk factor. Despite this, a quasi-systematic review of the literature between January 1960 and January 2020 identified only nine studies that have explicitly explored the relationship between foot-strike pattern and injury (Goss and Gross, 2012, Daoud et al., 2012, Warr et al., 2015, Futrell et al., 2018, Paquette et al., 2017, Dudley et al., 2017, Donoghue et al., 2008, Kuhman et al., 2016, Messier et al., 2018) with conflicting findings. Full results and study characteristics are presented in table 1. For further details on the systematic review approach taken, please see section 6.1.1 of the configuration manual.

Table 1: Relationship between Foot-strike pattern and injury

Author	Design	Subjects	Survalience Time	Footstrikes	Test Surface	Relationship to injury.
(Goss and Gross, 2012)	Retro	n = 881	12 months	N/A	N/A	RFS >MFS >FFS
(Daoud et al. , 2012)	Retro	n = 52	5 seasons *	N/R	Treadmill	RFS >FFS
(Warr et al. , 2015)	Retro	n = 341	60 months	2	Overground	-
(Futrell et al. , 2018)	Retro	n = 125	N/R	10	Treadmill	RFS >FFS **
(Paquette et al. , 2017)	Retro	n = 44	12 months	10	Treadmill	-
(Dudley et al. , 2017)	Prosp	n = 31	14 weeks	5	Overground	-
(Donoghue et al. , 2008)	Retro	n = 22	12 months	5	Treadmill	-
(Kuhman et al. , 2006)	Prosp	n = 19	3 months	5	Overground	-
(Messier et al. , 2018)	Prosp	n = 300	24 months	3	Overground	-

Retro = retrospective, Prop = prospective, RFS = Rearfoot-strike, MFS = Midfoot-strike, FFS = Forefoot-strike, N/R = not reported,* = 5 collegiate cross-country seasons, ** = not statistically examined, - = No significant difference

The majority of studies found no statistically significant difference in the number of injuries sustained between runners utilising different foot-strike patterns (Warr et al., 2015, Paquette et al., 2017, Dudley et al., 2017, Donoghue et al., 2008, Kuhman et al., 2016, Messier et al., 2018). However, the most consistent association was that runners utilising a RFS pattern have an increased prevalence of running related injuries (Goss and Gross, 2012, Daoud et al., 2012, Futrell et al., 2018). Interestingly, all three studies which identified an association between RFS and injury utilised a categorical definition of foot-strike pattern. While the reason for this trend is unclear, it may be related to non-linear relationship between foot-strike angle and loading (Stiffler-Joachim et al., 2019). Only three of the nine studies were prospective in nature (Messier et al., 2018, Dudley et al., 2017, Kuhman et al., 2016) and all three found no significant relationship between foot-strike pattern and injury. While generally speaking, prospective research offers a higher level of evidence in comparison to retrospective research, two of these studies had small sample sizes with relatively short surveillance periods (Dudley et al., 2017, Kuhman et al., 2016) which somewhat limits the weight of this evidence.

When exploring the remaining characteristics of the studies, the large heterogeneity in study design makes identifying clear trends challenging. Study sample sizes ranged from 19 participants to 881 participants [(median: 52 participants (interquartile range: 26.5 – 320. 5 participants)] and included an analysis of 2-10 foot-strikes per participant [median: 5 foot- strikes (interquartile range: 3 – 7.5 foot-strikes)]. With respect to injury surveillance, the timeframe for participant tracking ranged between 3 months and 5 years [(median: 12.3 months (interquartile range: 8.3 - 20. 5 months)]. Foot-strike technique was explored via over ground running (Messier et al., 2018, Dudley et al., 2017, Kuhman et al., 2016, Warr et al., 2015), on a treadmill (Futrell et al., 2018, Paquette et al., 2017, Donoghue et al., 2008) and even via a self-reported questionnaire (Goss and Gross, 2012). However, it is perhaps worth noting, that one study explored foot-strike pattern on both a treadmill and over ground and obtained identical classification in both surfaces (Daoud et al., 2012), suggesting that for experimental assessments of foot-strike, the surface does not matter. While it is unclear why inconsistences exist in the literature, it may be related to methodological variations in the studies. In particular, considerable variation exists in how foot-strike pattern is defined. The following section will review foot-strike definitions utilised within the literature.

2.2 Definitions of Foot-strike Technique and Identified Gaps

While both continuous and categorical representations of foot-strike technique have been used in the literature, the general consensus is that while running, humans tend to use one of three foot-strike techniques. These can be broadly defined as a forefoot-strike when the anterior aspect of the foot contacts the ground first, a rearfoot-strike when the heel contacts the ground first and a midfoot-strike when the foot contacts the ground in a flat position (Forrester and Townend, 2015). Foot-strike technique was first assessed using the concept of strike index (Cavanagh and Lafortune, 1980). This approach utilises the centre of pressure location along the foot at initial contact, as a percentage of total foot length, to group foot-strike patterns. Using strike index, a forefoot-strike would be defined when the strike index was >69%, a rearfoot-strike would be defined with a strike index <33% and a midfoot-strike would be defined if the strike index was between these two thresholds (Cavanagh and Lafortune, 1980). However, a limitation of this approach is the need for a force plate which may not be feasible as in the case of running on a conventional treadmill or in an outdoor environment. As a result, authors have proposed kinematic based methods as a surrogate measure for strike index.

The first proposed surrogate measure was strike angle (Altman and Davis, 2012), which utilises the sagittal plane angle of the foot with respect to the ground at initial contact. In comparison to the strike index method, the authors identified a strong and significant correlation between the two approaches ($r = 0.86$, $p < 0.01$). The authors subsequently devised a classification by firstly isolating foot-strikes identified as midfoot by the strike index (Cavanagh and Lafortune, 1980), and creating upper and lower thresholds for midfoot-strike based on the mean angle of those in the midfoot-strike group +/- 3 * standard errors of the mean. Using this threshold, a forefoot-strike would be defined when the foot angle was <- 1.6°, a rearfoot-strike would be defined with a foot angle of >8°and a mid-foot-strike is defined when the foot angle is between - 1.6°and 8°(Altman and Davis, 2012). While these thresholds defined by Altman and Davis (2012) are most commonly utilised in the literature, others have suggested that midfoot-strike should be represented by a foot-angle of 0°, with forefoot-strikes and rearfoot-strikes represented by positive and negative angles respectively (Lieberman et al., 2010). Similarly, others have utilised ankle angle rather than foot angle and proposed that a neutral ankle angle should reflect a midfoot-strike, while a forefoot and rearfoot-strike are represented by a positive and negative ankle angle respectively (Donoghue et al., 2008).

While the above quantitative approaches of defining foot-strike technique are an improvement on qualitative methods, they are subject to two main challenges. Firstly, current definitions of foot-strike patterns (continuous or categorical) are defined based on a single instance in time, that is, initial contact. However, it is plausible that foot movement following initial contact is also important for injury risk and/or the classification of foot-strike pattern. Indeed previous research has highlighted the value in considering the action of distal joints/segments which can influence joint stiffness (Farley and Morgenroth, 1999), energy absorption (Yeow et al., 2011) and general running mechanics (Almeida et al., 2015). While representing a movement pattern using a discrete time point is typical within the biomechanics literature, this approach has been criticized as it may not accurately characterise the movement pattern being analysed (Marshall et al., 2015, Pataky, 2010, Richter et al., 2014c). A potential solution to limitation, is to examine the data contained in whole foot time series from initial contact to toe off. In comparison to examining a discrete point in time, the benefit of exploring the information contained in whole movement pattern has been illustrated in several applications including distinguishing between healthy and injured subjects (Donoghue et al., 2008), improved explanation of jump height (Richter et al., 2014a) and enhanced ability to detect movement asymmetries in a variety of exercises (Marshall et al., 2015). To date however, to the best of the candidate's knowledge, no research has used the information from the full stance phase to identify foot-strike patterns or determine its relationship with injury.

The second limitation that is applicable to methods involving the classification of foot-strike technique, is that there is no justification for the three groupings reported in the literature (rearfoot, midfoot, forefoot). Rather, these three classifications are based on the seminal work by Cavanagh and Lafortune (1980), who arbitrarily divided the foot into three equal parts. A more robust approach would be to leverage unsupervised methods to cluster the data and identify foot-strike patterns in foot movement during the stance phase of running. To date however, to the best of the candidate's knowledge, no studies have explicitly utilised unsupervised clustering to identify foot-strike patterns. The following section will review the use of clustering within the biomechanics domain.

2.3 Cluster Analysis in Biomechanics and Identified Gaps

Within biomechanics it is common to investigate the risk factors for an injury using a single group design. However, this approach risks masking potentially important injury risk factors should the single group's movement not be suitably homogeneous. While often authors will delimit a study based on known cofounding factors, this is not always possible and/or the cofounding factors may be unknown. An alternative to relying on a-priori knowledge of cofounding factors, is to utilise an unsupervised machine learning technique known as clustering. Clustering is a prominent methodology which is used to group unlabelled data into clusters sharing similar qualities, typically based on a distance or similarity metric (Xu and Tian, 2015). Within biomechanics, unsupervised clustering has been utilised successfully for a number of applications, from recognising pathological walking gaits (Chau, 2001, Roche et al., 2014) to identifying performance determining factors during jumping (Richter et al., 2014b).

For example, in the clinical setting, Franklyn-Miller et al. (2017) examined over 300 patients with Athletic Groin Pain during a side stepping task and identified three subgroups that were independent of anatomical pain location. The authors concluded that rehabilitation should target the possible propagative biomechanics identified in the form of the movement clusters. While in contrast, Richter et al. (2014a), compared the use of Hierarchical clustering, K-means and the Expectation–Maximization algorithms to help explain jump height. The authors determined that Hierarchical clustering performed best and extended the ability of a stepwise regression analysis to predict jump height in comparison to a single group design by 7%.

Within the literature exploring running biomechanics, cluster analysis has been reported for non-injured runners (Phinyomark et al., 2015) and injured runners alike (Dingenen et al., 2020, Jauhiainen et al., 2020, Watari et al., 2018). For example, Phinyomark et al. (2015) applied Hierarchical cluster analysis to 121 healthy runners and identified two running pattern clusters which were independent of participant demographics or running velocity. When exploring the risk factors for patellofemoral pain syndrome, the comparison with the healthy clusters identified two independent biomechanical risk factors which could be subsequently targeted with rehabilitation. Later, Watari et al. (2018) also explored patellofemoral pain syndrome using Hierarchical clustering applied to pelvic acceleration profiles. The authors identified two subgroups of runners, but later determined that the variability observed in the running biomechanics occurred mainly due to known sex-related factors. Similarly, Jauhiainen et al. (2020) identified a five clusters solution using Hierarchical clustering in injured runners. Despite identifying well defined clusters, the authors concluded that the homogeneous biomechanical patterns existed independent of injury location suggesting the clusters were of little relevance to the injuries explored.

With respect to foot biomechanics and clustering, the primary application of clustering has been with regard to diabetic feet and pressure measures (Sawacha et al., 2010, Deschamps et al., 2013, Bennetts et al., 2013). Using K-means clustering all three studies attempted to identify plantar pressure distribution clusters during walking as a means to determine mechanical interventions for the prevention and/or treatment of the diabetic foot. To date, to the best of the candidate's knowledge, only one study has explored clustering with respect to foot-strike pattern (Forrester and Townend, 2015). However, rather than exploring the presence of foot-strike patterns per se, the authors investigated how the traditional foot-strike patterns (as defined by foot angle at initial contact) changed with increasing velocity. Using a regression mixture model, the authors explored the influence of increasing running velocities on foot angle in 102 runners. The authors identified three clusters describing the foot-strike angle vs running velocity behaviour of the participants and suggested that the clusters could represent a novel and relevant means of grouping athletes for further biomechanical running assessment. To date however, to the best of the candidate's knowledge, no research has explored if naturally occurring clusters exist in terms of foot-strike pattern.

There are two major limitations in the current use of clustering within the biomechanics domain. Firstly, a challenge with clustering algorithms is that they may identify well defined clusters which have little practical implication with respect to the research question being examined. This was well illustrated in the study by Jauhiainen et al. (2020) and who identified that the clusters identified had no relevance to the injuries explored and by Watari et al. (2018), who observed that the identified clusters simply reflected known sex differences in running biomechanics. This is an important challenge with respect to unsupervised learning, that is too often overlooked in biomechanics research

(Dingenen et al., 2020, De Cock et al., 2006). Future research that utilises clustering, should explore means of assigning value to the identified clusters in addition to metrics of separability and compactness. A second challenge with the literature reviewed in the biomechanics domain, is that all but one study (Richter et al., 2014b) reported the use of a single clustering algorithm. As per the no free lunch theorem (Wolpert, 1996), there does not exist a single universally best performing machine learning algorithm, therefore clustering research should explore a wider range of algorithms to determine the optimal solution for the data being examined. The following section will review the prospective studies that have investigated running related injuries.

2.4 Prospective Biomechanical Risks for Running Injuries

Within the biomechanics literature there has been considerable interest in the biomechanical risk factors for running related injuries (Pohl et al., 2008, Taunton et al., 2002). However, the majority of this research has been retrospective in nature. An alternative and more robust research design is to investigate the risk factors for running related injuries prospectively. A recent systematic review of the biomechanical risk factors for running related injuries identified 16 prospective studies (Ceyssens et al., 2019). Overall, in terms of kinematics (movement), the movement of the foot and ankle were most commonly identified as a risk factor for injury (Dudley et al., 2017, Kuhman et al., 2016, Hein et al., 2014) followed by the knee (Hein et al., 2014, Messier et al., 2018). Despite this, there was considerable inconsistencies in the research findings. While, it is unclear what the source of these inconsistencies are, a primary review conducted of the methods within the studies, revealed that there were considerable statistical limitations associated with the research. These included not controlling for multiple comparisons, not conducting out of sample testing and only statistically examining discrete time points within biomechanical waveforms. For a more detailed review of this literature, please see section 6.1.2 of the configuration manual.

To conclude, this review of related work identified several studies which have investigated the association between foot-strike technique and injury. Despite the growing interest in this area, there has been a lack of prospective research that has examined the relationship between foot-strike technique and injury appropriately, and in general, the findings from the literature have not been consistent. A methodological consideration which may in part help explain this conflicting evidence, is how foot-strike technique is defined. Current definitions of foot-strike technique utilised in the literature are based on a single instance in time (initial contact) and are often based on the arbitrary division of the foot into three equal parts. It was suggested that a more objective approach would leverage unsupervised clustering to identify appropriate foot-strike patterns. While a review of the literature would indicate that the use of clustering in the biomechanics domain is not uncommon, to date no research has explored the use clustering to identify foot-strike patterns. When exploring the prospective biomechanical risk factors for running related injuries, it would appear that while the foot and ankle were most commonly identified as risk factors, there was conflicting evidence in the literature. While the source of this conflicting evidence is unclear, it may be related to the specific injuries being studied or the statistical limitations identified. The following section will outline the methodology employed in this project to explore the presence of foot-strike patterns and the prediction of running related injuries.

3 Foot-strike Methodology Approach

The methodology employed in this project, will follow a modified Knowledge Discovery in Databases (KDD) approach (Fayyad et al., 1996) as illustrated in Figure 1.

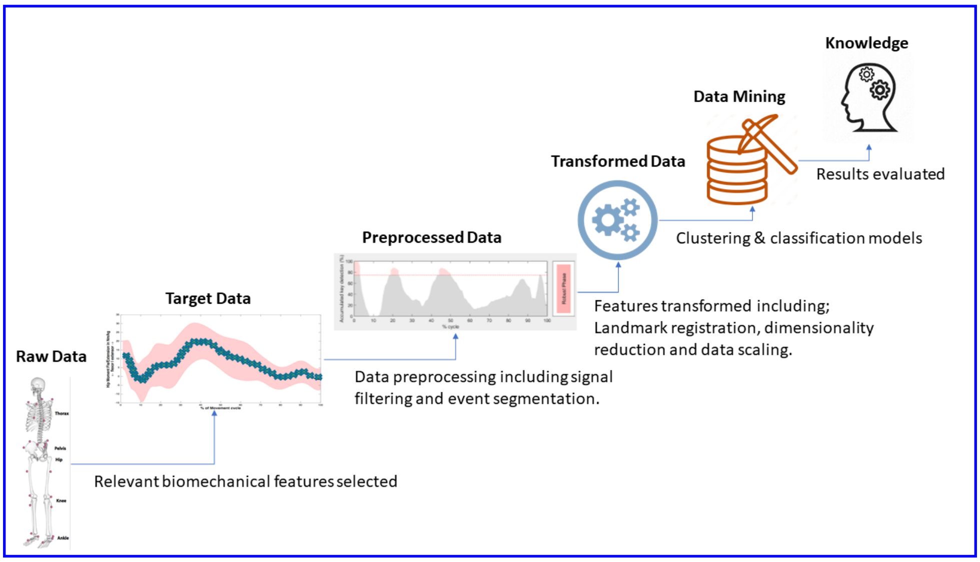


Figure 1: Foot-strike methodology approach

3.1 Data collection and Selection

The 3D motion of 47,423 foot-strikes, collected prospectively from 282 participants, was explored along with 18 months of injury surveillance (see config manual section 6.2.2). This data was captured in Dublin City University as part of the Running Injury Surveillance Centre (RISC) study and ethical approval was granted (Ref: DCUREC/2017/186).

3.1.1 Data Capture Description

Prior to data capture, 32 reflective markers were placed on known anatomical landmarks and participants where asked to run on a motorised treadmill (Flow Fitness, DTM3500i, Netherlands) at 9 km/h for one minute. The 3D motion data was captured with sixteen Vicon cameras (Vicon, UK) recording 200 HZ (Figure 2) and saved as C3d file format.

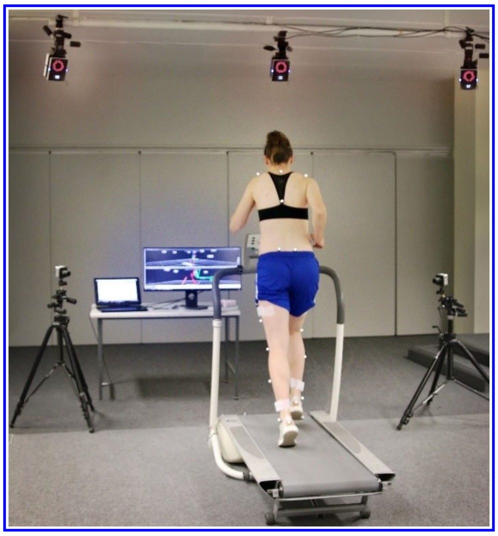


Figure 2: Data capture set up

3.2 Pre-processing

Pre-processing included filtering of the data and screening for inappropriate waveforms which were corrected in an iterative process using a custom written application and motion capture software.

3.3 Transformation

Transformation included time warping of the biomechanical signals using landmark registration (Moudy et al., 2018) and generating features using the concept of Analysis of Characterizing Phases (Richter et al., 2014a) and also the TSFresh python package. The generated features were screened for outliers and imputed when appropriate (Christ et al., 2018).

3.4 Data Mining

Data mining involved implementing the clustering and classification approaches utilized within this project. For clustering this inlcuded; K-means, Hierarchical, Mean Shift, OPTICS, HDBSCAN and Spectural clustering. For the classification, this included Elastic Net Logistic Regression, Naive Bayes, Random Forest, Bagged SVM, AdaBoost and a weighted Stacked Ensemble.

3.5 Knowledge

Finally, the results of the data mining process are evaluated and visually inspected to bring about knowledge. For the clustering approaches the primary evaluation metric was the adjusted rand index. For the classification approaches it was a combination of accuracy, specificity and sensitivity.

4 Design Specification

The following three-tier architecture diagram depicts the design process used in this project for the identification of foot-strike patterns and prediction of injury classification (Figure 3).

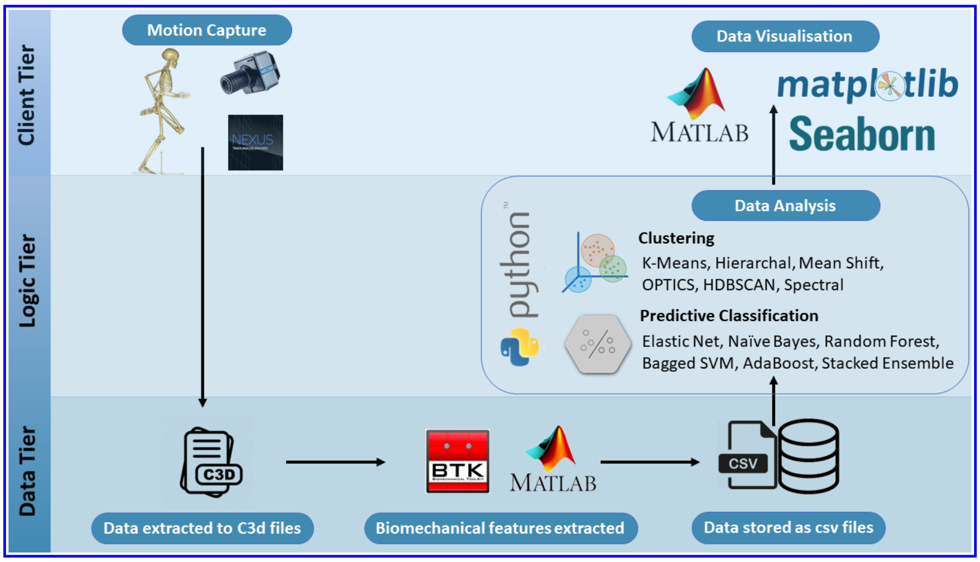


Figure 3: Three tier design specification for the foot-strike and running injury analysis

The process flow begins in the client tier where motion capture and visualisation are conducted using Vicon Nexus software. Data is passed from the motion capture process to the data (persistent) tier where the biomechanical data is extracted from the C3d files, pre-processed and stored as csv files. This pre-processed data is then passed to the logic tier, where it is modelled using both unsupervised clustering and predictive classification models to answer the primary and secondary research question of the project respectively. The results are then passed to the client tier again where the findings are visualised using MATLAB, and the Python packages; Seaborn and Matplotlib. The following section will detail the implementation of this research design.

5 Implementation

This project is split into two main aims. Firstly, to identify the presence of foot-strike groupings using unsupervised clustering approaches and determine their association with injury and secondly, to identify the relationship between all the biomechanical features and injury and determine if any foot was included in the final classification models. As such, the implementation pipeline splits following feature engineering and will be detailed separately (Figure 4).

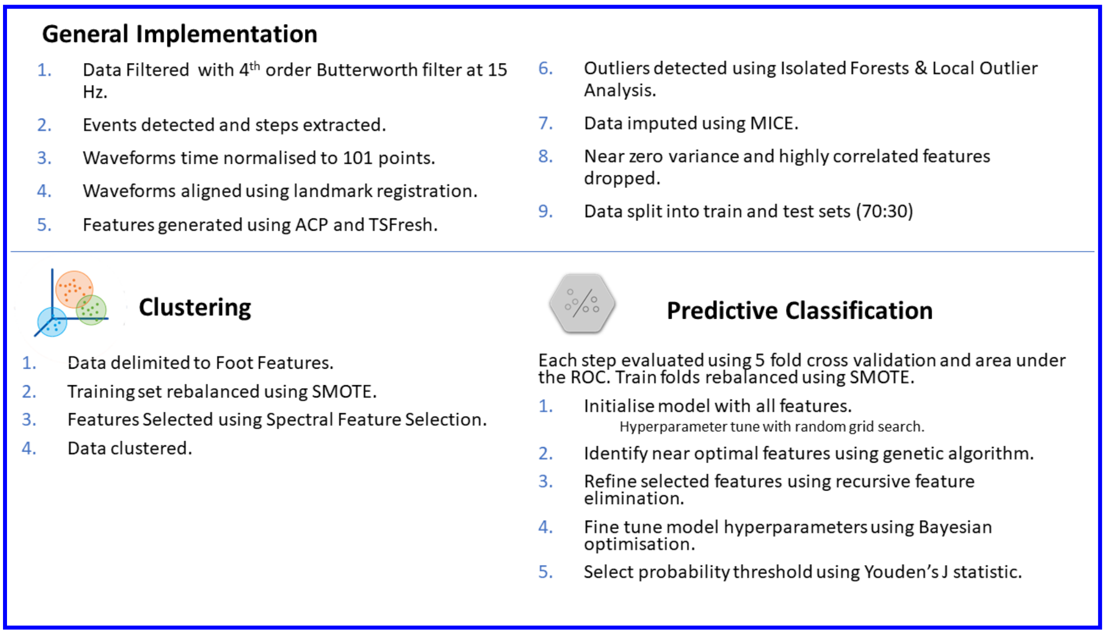


Figure 4: The key steps in implementation

ACP = Analysis of Characterizing Phases, MICE = Multivariate imputation by chained equations, SMOTE = Synthetic minority oversampling technique. For further details, please see the text below.

5.1 General Implementation

This section involves the general pre-processing and feature engineering steps. It directly relates to the implementation and achievement objectives 2 and 3 detailed in section 1.2.

5.1.1 Pre-processing

3D marker trajectories were initially tracked, cleaned and screened in Vicon Nexus software. Data was then filtered using a 4th zero order Butterworth filter at 15Hz to remove impact artefacts and joint/segment angles were modelled for the trunk and lower limbs as per the Vicon Plug in Gait model (Vicon, UK). 15Hz was selected as the cut off frequency following a residual analysis (Winter, 2009).

5.1.2 Data Transformation

Joint angular data as defined by the Plug in Gait model (Vicon, UK) was extracted from the C3d files using the biomechanical toolkit and MATLAB 2018b (MathWorks, USA) for each foot-strike, from initial contact to toe off. Initial foot contact was defined by firstly identifying a window in which the ankle marker was within 10cm of its local minima and subsequently using the first peak negative horizontal velocity of the toe or heel marker as the foot contact event. Toe off was then identified in the same window of time, using the toe jerk maxima (3rd derivative of toe marker position) following peak knee extension as a combination of two previously published algorithms (Handsaker et al., 2016, Dingwell et al., 2001). The segmented data was then time normalised to 101 data points using a cubic spline to represent 0 - 100% of stance.

Foot angle was manually calculated as the angle between the heel and toe markers as previously proposed (Altman and Davis, 2012). Foot angular velocity and angular acceleration were then calculated as the 2nd and 3rd derivative of the foot angle data respectively. The biomechanical waveform data from the whole body was then screened using a custom written application developed with MATLAB 2018b (see configuration manual section 3.3). In an iterative process, inappropriate waveforms were then re-screened and corrected as appropriate. After screening for outliers using both statistical methods and manual inspection of the biomechanical waveforms, and removing those individuals who dropped out from the prospective arm of this project, the data set contained 43,184 foot-strikes and a matrix size of 1,813,728 x 101.

In order to remove unwanted temporal variation in the normalised kinematic signals, a landmark registration algorithm as previously described (Moudy et al., 2018), was employed using custom python code (Figure 5). This approach was taken, as it has been shown to improve the predictive power of classification algorithms (Moudy et al., 2018). In contrast to the algorithm described by Moudy et al. (2018), an Akima spline (Akima, 1970) was used rather than a cubic spline to reduce fitting errors and a divide and conquer binary search algorithm was utilised to speed up convergence. In comparison to the more traditional approach of dynamic time warping, using a global landmark approach (Ramsay and Silverman, 2005) as employed in this current project, retains the relationship between segments and joints which is important for the interpretation of biomechanical data.

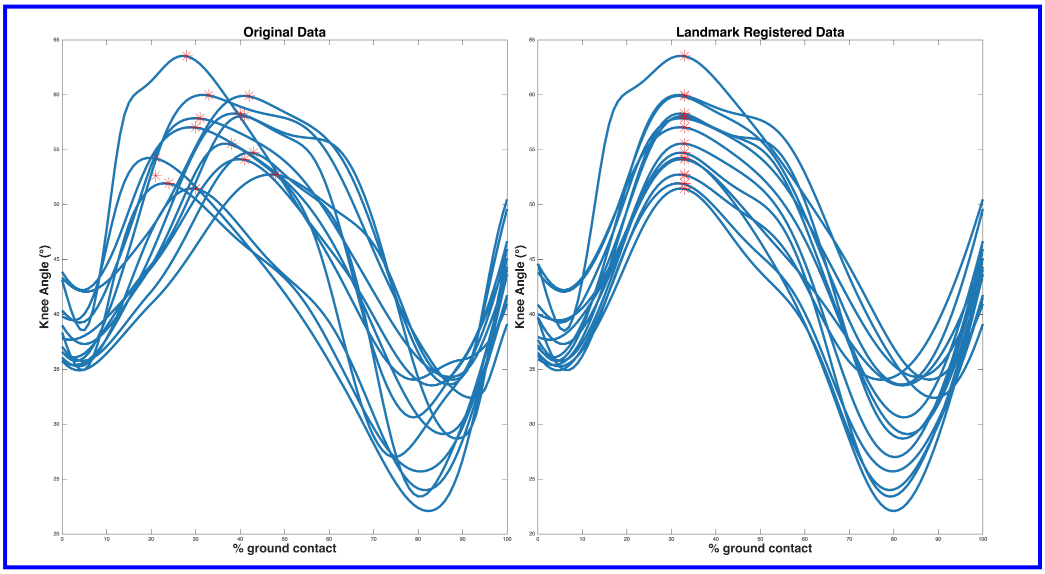


Figure 5: The effect of landmark registration

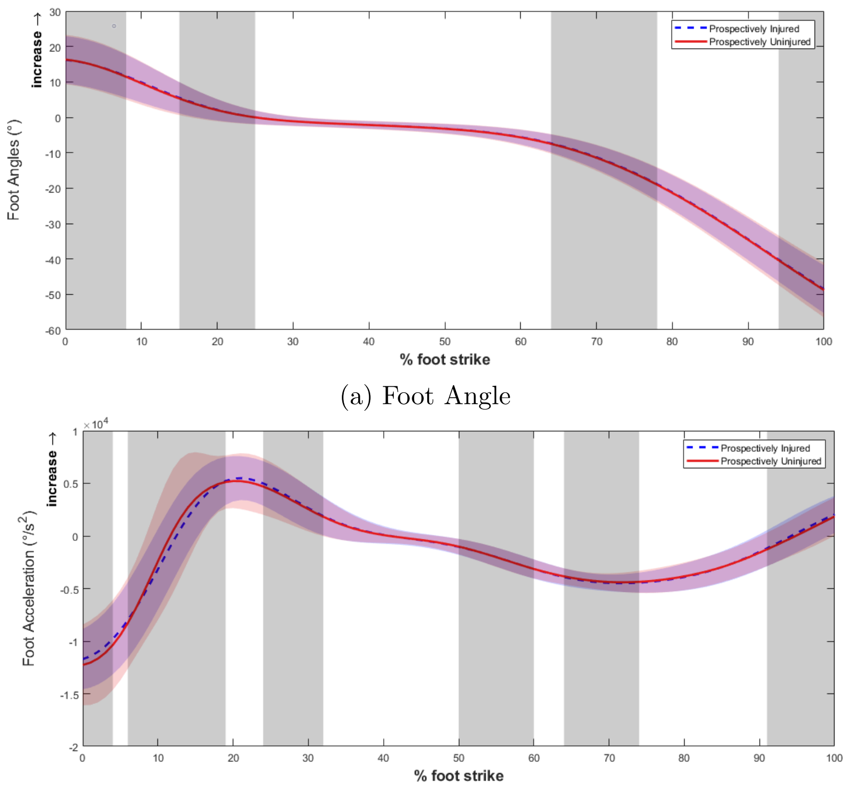
The left graph depicts the original data with considerable temporal variation as highlighted by the red star marking peak knee flexion while the right graph depicts the same data post landmark registration

5.1.3 Feature Engineering

In order to reduce the dimensionality of the data and extract key features, the concept of 'Analysis of Characterizing Phases' (ACP) was used to generate participant scores that represent the movement of each participant within key phases of variation using VARIMAX rotated principal components (Richter et al., 2014a). This approach was utilised as it has been demonstrated to outperform other popular continuous waveform techniques such as functional principal component analysis (Richter et al., 2014c). Using ACP, each score captures the samples movement for each identified phase (k) as the summed difference between a participant' s waveform (p) and the mean waveform (q) for each time point (i) between the start (n) and end (m) of a phase. This was completed for each biomechanical waveform (j) (Equation 1):

$$feature_{j,k} = \sum_{i=n}^{m} p(i) - q(i) \tag{1}$$

To enhance the generalisability of this method, the above approach was conducted 100 times on a random 70% subsample. Only robust phases were then retained, defined as being identified more than 80% of the time (Richter et al., 2019) (Figure 6). All biomechanical waveforms depicting the ACP phases are presented in section 6.2.1 of the configuration manual.



(a) Foot Angle

(b) Foot Acceleration

Figure 6: Example biomechanical waveforms for foot angle and foot acceleration

Data segmented by injury status, identified robust ACP phases are shaded in grey

In addition to the features generated by ACP, basic statistical features were generated from the time series data using the Python package, TSfresh (Christ et al., 2018). Outliers in the generated features were detected using isolated forests and local outlier factor.

These outliers were explored for patterns, and when absent, were removed. Missing data was then imputed using multivariate imputation by chained equations and a Bayesian ridge regression approach, based on the twenty nearest features. After feature engineering and dimensionality reduction, the final data set was of size 43,184 x 362. The data was partitioned into two class labels: prospectively injured and prospectively uninjured. The data was imbalanced with 66% becoming injured. A full list of the features and their description is provided in section 6.3.1 of the config manual.

5.2 Clustering Implementation

This section relates to the implementation of the clustering models. It addresses the implementation and achievement of objectives 4 and 5 as outlined in section 1.2.

5.2.1 Feature Selection

Given the aim of this section to identify naturally occurring foot-strike groups, the full feature space was first delimited to foot related features. After removing features with zero, or near zero variance (<98% variance) and randomly removing highly correlated features (r >0.90), feature selection was conducted in an unsupervised manner using the concept of Spectral Feature selection (SPEC) (Zhao and Liu, 2007). This approach involves constructing a Laplacian matrix and evaluating the relevance of each feature by its consistency with the structure of the graph induced from the similarities among objects (Tang et al., 2014). While this approach is often used to select a predefined number of features, within this current project, the feature scores were sorted and plotted (Figure 7). The number of features to retain were then selected by identifying an elbow point in the graph. The retained five features (maximum foot acceleration, mean foot acceleration from 6-20% of the foot-strike, foot acceleration variation, maximum foot velocity and median foot velocity) had a Hopkin's statistic of 0.96 suggesting the dataset had high clusterability.

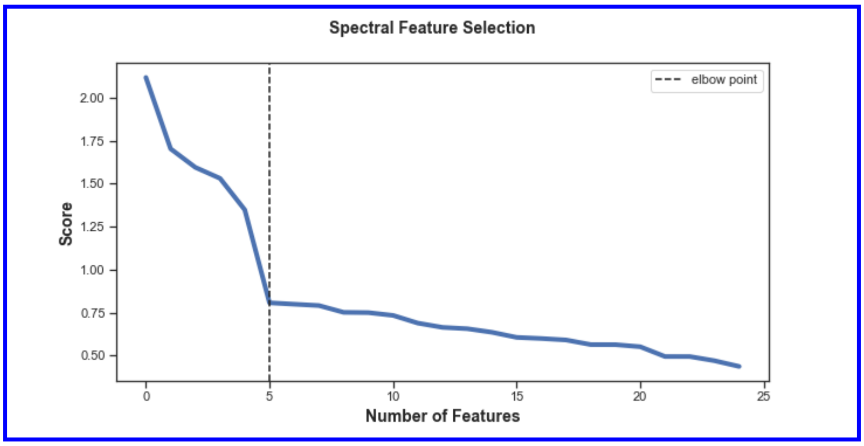


Figure 7: The SPEC scores plotted in sorted order

5.2.2 Clustering Models

In line with the no free lunch theorem (Wolpert, 1996), a wide range of clustering models were explored in this project in order to identify a suitable clustering solution in the foot-strike patterns. In addition to the most commonly implemented models in the biomechanics literature (K-means and Hierarchical clustering), four other cluster models were explored (Mean Shift, OPTICS, HDBSCAN and Spectral clustering). All clustering models were implemented using scikit-learn (Pedregosa et al., 2011), with the exception of HDBSCAN which was implemented using the HDBSCAN Python package.

Each clustering model was assessed with a variety of hyperparameters using a greedy grid search. In total 51 cluster variations were assessed (see config manual section 6.4). The evaluation and results of the clustering models will be detailed in section 6.1.

5.3 Classification Implementation

This section relates to the implementation of the predictive classification models. It addresses the implementation and achievement of objectives 7 and 8 in section 1.2.

5.3.1 Feature Selection and Model Tuning

Given the aim of this section to identify if any of the biomechanical features could classify those who became injured, all features were included in this analysis. After removing features with zero, or near zero variance (<98% variance) and randomly removing highly correlated features (r >0.90), feature selection and model tuning was conducted over several steps (Figure 4). Data was firstly split into training and testing datasets using a 70:30 ratio in a stratified manner. All training and tuning was conducted on the training dataset. After identifying a suitable hyperparameter solution for each machine learning model using a random search of the hyperparameter grid with all features, feature selection was conducted in two stages. Firstly, a genetic search algorithm was implemented with five-fold cross validation to identify an optimal or near optimal solution to feature selection. This was completed using the sklearn-genetic package (Calzolari, 2019). The parameters chosen included a population of 50, cross over probability of 0.8, mutation probability of 0.2 and a tournament size of 3. These values were chosen similar to the defaults proposed by Kuhn and Johnson (2019), who suggest these work well in practice. Finally 200 generations were tested to find the best solution, with a constraint of retaining a maximum number of features less than the total number of features * 0.25. The second step of feature selection was to implement five-fold cross validated recursive feature elimination to identify the best subset of features from the delimited feature space following the genetic search approach. This was implemented as genetic algorithms often tend to select larger feature subsets than other feature selection methods since there is little penalty for keeping a feature that has no impact on predictive performance (Kuhn and Johnson, 2019). In order to enhance generalisability, a sparse model was encouraged by choosing the smallest number of features within one standard error of the feature set which maximised the area under the receiver operating characteristic curve (James et al., 2013). The model was then fine-tuned using a 3 x 2-fold nested cross validated Bayesian optimisation approach via the scikit-optimize package which uses Bayes theorem to explore and exploit the hyperparameter space. Finally, the probability classification threshold was selected to balance sensitivity and specificity by using Youden's J statistic (Figure 8).

5.3.2 Predictive Classification Models

Given that there is no such thing as a universally best machine learning model (Wolpert, 1996), several models were assessed in this project. These included the Naïve Bayes model, Elastic Net Logistic Regression, Bootstrapped Aggregated (Bagged) SVMs, Random Forest, AdaBoost and a weighted Stacked Ensemble of the aforementioned approaches. All classification models with the exception of the Stacked Ensemble were implemented using scikit-learn (Pedregosa et al., 2011). Given that scikit-learn does not directly support weighted stacking, a pragmatic equation was proposed which took into account the base classifier's class vote, it's average performance and the confidence of it's class vote. Each model was tuned over an extensive range of hyperparameters using Bayesian optimisation. For further details on the chosen models, the hyperparameter grid space for each model and the equation for the weighted stacked ensemble, please

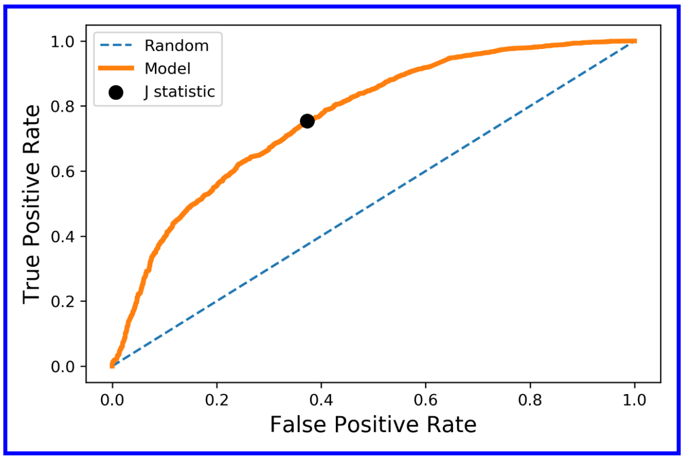


Figure 8: Example ROC curve for the random forest model

The probability threshold was set using Youden's J statistic indicated by the black dot

see the configuration manual (section 6.5). The evaluation and results of the predictive classification models will be detailed in the following section (section 6.2).

6 Evaluation

6.1 Cluster Evaluation and Results

This section details the evaluation, comparison and results of the clustering models. It addresses the implementation and achievement of objective 6 as outlined in section 1.2.

6.1.1 Evaluation Methods

Each clustering solution was initially assessed using three internal evaluation metrics. Firstly, silhouette coefficient, which is a measurement of within to between cluster compactness based on a distance measure. Secondly, to appropriately evaluate clustering solutions based on density, the density based validity index, which is a measurement of within to between cluster density was implemented (Moulavi et al., 2014). Finally, a custom function based on the cluster validity index was implemented, that has been shown to be a generic cluster validation score, that outperforms other commonly utilised metrics (Rodríguez et al., 2018). This cluster validity index validates the clustering solutions via an ensemble of supervised classifiers. The idea of this approach, is that good clustering partition should induce the construction of a good classifier. Within this project, a Logistic Regression, Random Forest, Gaussian Naïve Bayes, Support Vector Machine (SVM), K-Nearest Neighbour and a Linear Discrimination Analysis was implemented in a stacked voting ensemble similar to the method described by Rodríguez et al. (2018). Finally, once the best performing cluster solutions were identified using the above criteria, the clusters were evaluated using the adjusted rand index (ARI) across 100 random bootstrapped samples of the hold-out test set. Levene's test for homogeneity of variance, suggested unequal variances across the groups ($p < 0.01$), so Welch's ANOVA was conducted, which is robust to both unequal variances and non-normality (Delacre et al., 2019). Similarly, post hoc comparisons when required, were conducted using Games-Howell tests which like the Tukey HSD test, uses Tukey's studentized range distribution but is based on Welch's degrees of freedom correction. Given the non-parametric nature of this test, it is also robust to both unequal variances and non-normality (Ruxton and Beauchamp, 2008). For pair wise comparisons, standardised effect sizes were reported using Cohen's D as small (<0.5), medium (0.5 - 0.8) and large (>0.8) (Cohen, 1988). Alpha level was set at 0.05 for the statistical tests.

6.1.2 Results

Of the six clustering models assessed, both Mean Shift and HDBSCAN failed to converge to a suitable solution and were not considered any further. As a metric, the recently proposed cluster validity index (VIC) (Rodríguez et al., 2018) provided little discriminative information (results were either excellent or undefined). As such, the VIC was not considered when delimiting the best clustering models. The best hyper parameter solution for the four retained clustering models are presented below (Figure 9). All of the clustering models assessed are presented in section 6.6 of the configuration manual.

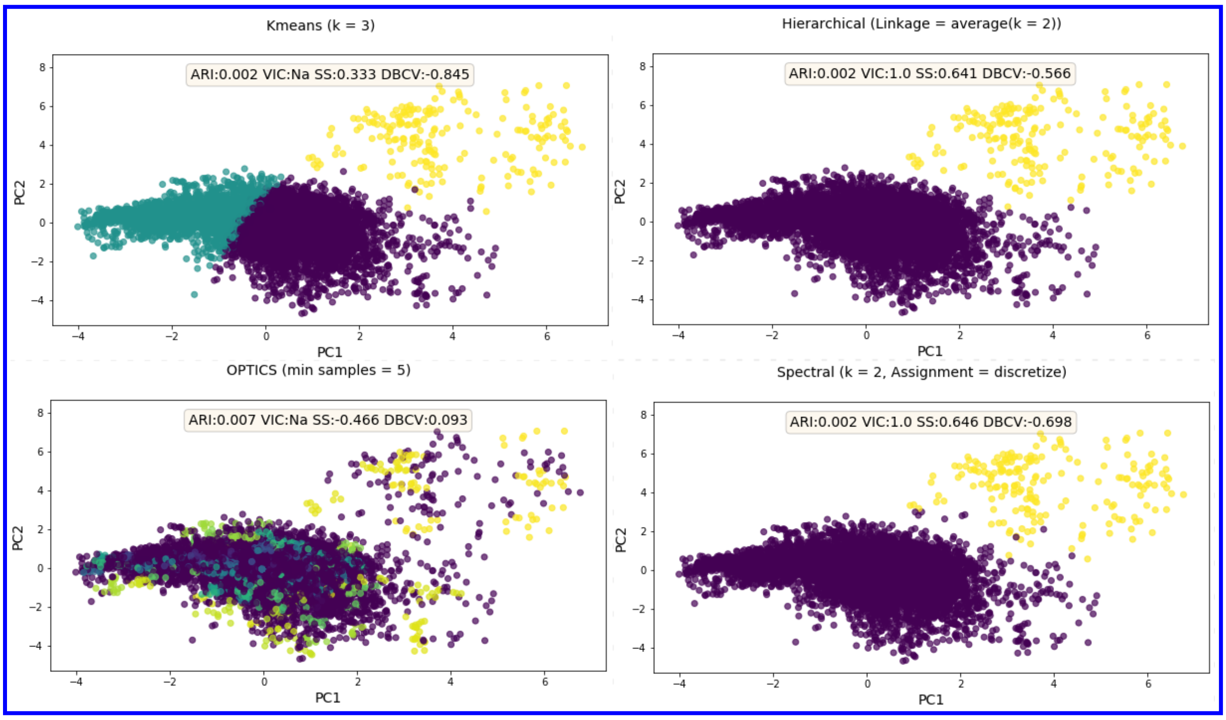


Figure 9: Visualisation of the cluster labels on a plot of the first two principle components

Each plot represents the best hyperparameter solution for the four retained clustering models; K-means, Hierarchical, OPTICS and Spectral.

The one-way Welch's ANOVA demonstrated a significant difference between the classification models for ARI scores, F (4,208.4) = 330, p <0.01 (Figure 10). A Games-Howell post hoc test indicated that the only approachs that were not statistically significantly different from another were the Spectral and Hierarchical clustering approaches (p = 0.35, D = 0.26). All other pairwise comparisons were significantly different with Cohen's effect size ranging from medium to large (p <0.05, D = 0.42 − 4.17). Furthermore, multiple one-sample welch t-tests with holm's correction, indicated that all approaches, with the expectation of the traditional approach, were significantly different than zero (p <0.05). The best performing clustering approach was the OPTICS model. Despite this, the mean ARI scores across all approaches were low (0 - 0.007) indicating almost random grouping relative to the injury label. Footstrikes were classified as RFS, MFS and FFS 87%, 11% and 2% of the time respectivly using the traditional classification approach. Full post hoc comparisons are presented in the configuration manual (section 6.6.1).

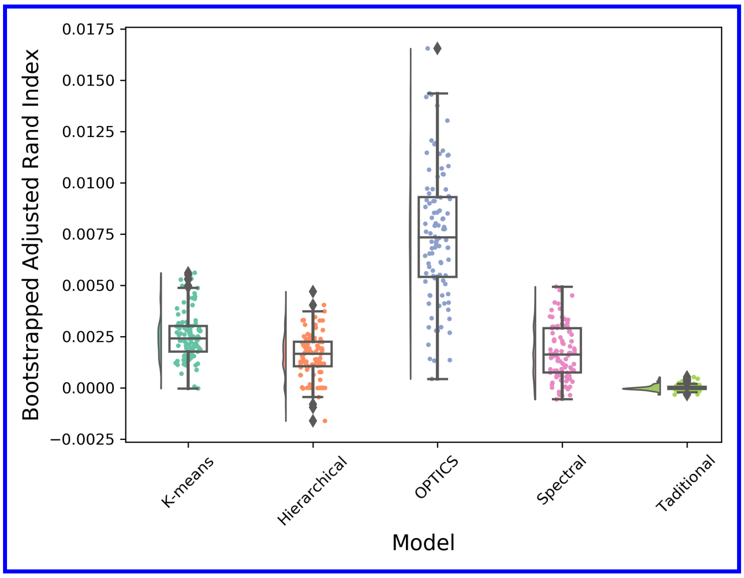


Figure 10: Bootstrappred Adjusted Rand Index scores

6.2 Classification Evaluation and Results

This section details the evaluation, comparison and results of the predictive classification models. It addresses the implementation and achievement of objective 9 as outlined in section 1.2.

6.2.1 Evaluation Methods

The final tuned classification models were evaluated using 100 random bootstrapped permutations. Given the imbalanced nature of the class labels, the classification models were assessed using a combination of accuracy, specificity and sensitivity. For all three assessments, Levene's test for homogeneity of variance suggested unequal variances across the groups ($p < 0.01$), so Welch's ANOVA was conducted, which is robust to both unequal variances and non-normality (Delacre et al., 2019). Similarly, post hoc comparisons when required, were conducted using Games-Howell tests which uses Tukey's studentized range distribution with Welch's degrees of freedom correction. Given the non-parametric nature of this test, it is also robust to both unequal variances and non-normality (Ruxton and Beauchamp, 2008). For pair wise comparisons, standardised effect sizes were reported using Cohen's D (Cohen, 1988). For the accuracy findings, multiple one sample welch t-tests were also conducted with holm's correction between the classification approaches and 50% (random assignment). Finally, once the best classification model was identified, feature importance and relationship with the class label was determined using a permutation approach and dependency plots respectively. Alpha level was set at 0.05 for the statistical tests.

6.2.2 Results

In terms of accuracy, the one-way Welch's ANOVA demonstrated a significant difference between the classification models, $F (6,299.3) = 7924$, $p < 0.01$ (Figure 11). A Games-Howell post hoc test indicated the only approaches that were not statistically significantly different from another were AdaBoost vs Naïve Majority comparison ($p = 0.29$, $D = 0.31$) along with Random Forest vs Stacked Ensemble comparison ($p = 0.89$, $D = 0.17$). All other pairwise comparisons were statistically different with large effect sizes ($p < 0.01$, $D = 1.4 – 28$) (see config manual section 6.7.1). All approaches were also statistically significantly different from 50 (random assignment) as determined using holm corrected one sample welch t-tests ($p < 0.01$), however only the Random Forest ($p < 0.01$, $D = 3.86$) and the Stacked Ensemble ($p < 0.01$, $D = 3.88$) models were statistically significantly greater than the Naïve Majority classifier. Overall, the best performing model in terms of accuracy was the Random Forest model (mean accuracy = 0.71), while the worse performing model was the Bagged SVM (mean accuracy = 0.51).

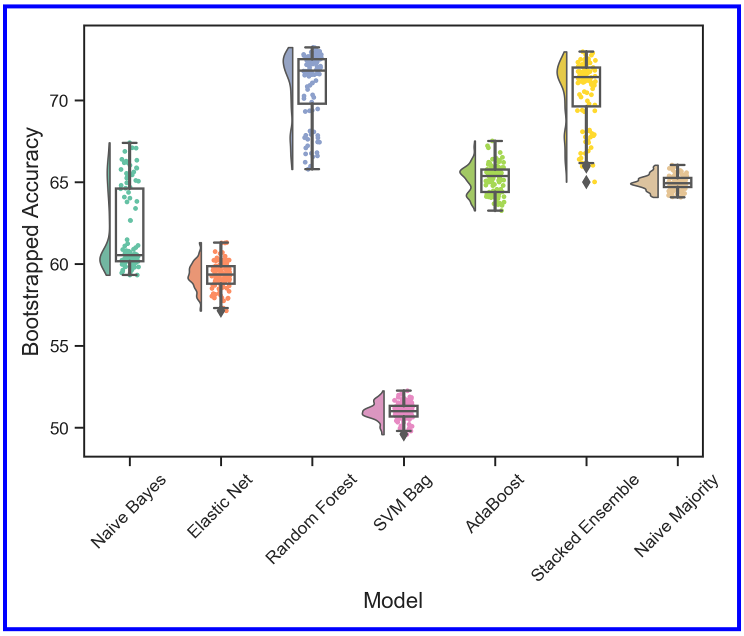
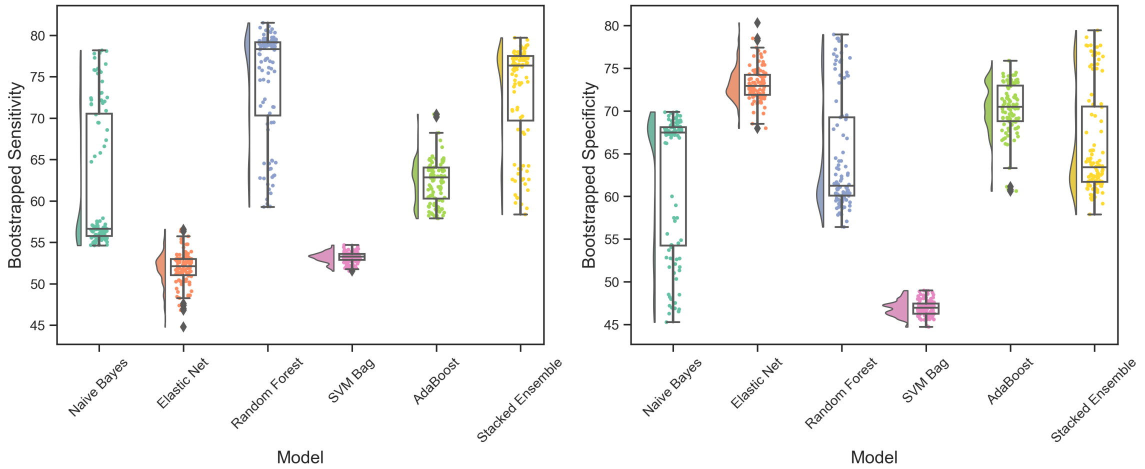


Figure 11: Bootstrappred accuracry for the six models and the Naive Majority classifier

In terms specificity, the one-way Welch's ANOVA demonstrated a significant difference between the classification models $F(5,252.6) = 3703$, $p < 0.01$ (Figure 12 [a]). Games-Howell post hoc tests indicated that all methods were statistically significantly different from one another ($p < 0.01$), with the exception of the Random Forest model vs Stacked Ensemble comparison ($p = 0.71$, $D = 0.2$) and the Naive Bayes vs Random Forest comparison ($p = 0.32$, $D = 0.30$). In terms of specificity, the best performing model was the Elastic Net Logistic regression (mean specificity = 0.73), while the worst performing model was the Bagged SVM (mean specificity = 0.47). Full post hoc comparisons from the specificity bootstrapped comparisons, are presented in section 6.7.2 of the configuration manual.

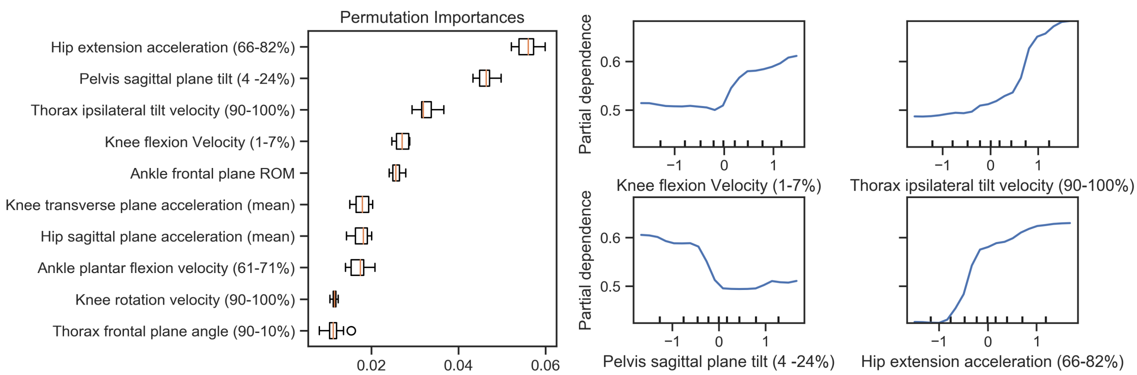
In terms sensitivity, the one-way Welch's ANOVA demonstrated a significant difference between the classification models $F(5,244.2) = 618$, $p < 0.01$ (Figure 12 [b]). Games-Howell post hoc tests indicated that all methods were statistically significantly different from one another ($p < 0.01$), with the exception of the Random Forest model vs Stacked Ensemble comparison ($p = 0.78$, $D = 0.20$) and the Naive Bayes vs Adaboost comparison ($p = 0.9$, $D = 0.12$). In terms of sensitivity, the best performing model was the Random Forest Model (mean sensitivity = 0.74) while the worst performing model was the Elastic Net regression model (mean specificity = 0.52). Full post hoc comparisons from the sensitivity bootstrapped comparisons, are presented in section 6.7.3 of the configuration manual.



(a) Bootstrapped sensitivity (b) Bootstrapped specificity

Figure 12: Bootstrappred sensitvity (a) and specificity (b) for the six classificaiton approachs

In terms of feature importance, while both the Stacked Ensemble and the Random Forest models were deemed the best performing models in this project, given the non-significant difference between the two models in terms of accuracy, specificity and sensitivity, only the Random Forest model was considered in terms of feature importance. The final Random Forest model contained ten features [Knee flexion Velocity (1-7%), Thorax frontal plane angle (90-10%), Thorax ipsilateral tilt velocity (90-100%), Knee rotation velocity (90-100%), Pelvis sagittal plane tilt (4 -24%), Hip extension acceleration (66-82%), Ankle plantar flexion velocity (61-71%), Ankle frontal plane ROM, mean Hip sagittal plane acceleration for the full stance phase, and mean Knee transverse plane acceleration for the full stance phase] (Figure 13 [a]). When considering the partial dependency of the four top features (Figure 13 [b]), with the exception of Pelvis sagittal plane tilt (4 -24%), all feature demonstrated a generally positive partial dependency relationship with injury classification.



(a) Permutation Importance Test (b) Partial Dependency Plots

Figure 13: Feature Importance (a) Permutation Importance Test (b) Partial Dependency Plots

6.3 Discussion

The primary aim and research question of this project was to identify the presence of naturally occurring foot-strike groupings, using unsupervised clustering and to determine their association with injury. To date, to the best of the candidate's knowledge, this was first application of clustering to identify foot-strike groupings. Rather, the traditional approaches of foot-strike classification are commonly based on the arbitrary division of the foot into three equal parts (Cavanagh and Lafortune, 1980). Within this current project, well defined clustering solutions were identified (e.g. Spectral clustering with a silhouette score of 0.65), that compare favorably with other studies that have explored clustering in general running biomechanics studies [e.g. silhouette score of 0.53 (Dingenen et al., 2020)]. Despite these promising findings, and in contrast to the hypothesis of this project, the bootstrapped analysis of the Adjusted Rand Index scores demonstrated almost random assignment to injury class for the six clustering models explored in this project. This is similar to the findings by Jauhiainen et al. (2020) who identified a five cluster solution in general running biomechanics, but concluded the clusters had no relevance to the injuries explored. Furthermore, and in agreement with the other prospective analysis of foot-strike angle and injury (Messier et al., 2018, Dudley et al., 2017, Kuhman et al., 2016), this project found no association between the traditional approach of classification and injury. These results suggest that the movement of the foot by itself, contains little information relevant to injury risk in runners. Unlike previous prospective research however, this current project explored a larger number of foot-strikes (>47,000 foot-strikes) compared to what has been reported (155 – 900), and as such adds substantial weight to the body of evidence.

The secondary aim and sub research question of this project was to determine if any of the biomechanical features captured could predict risk of injury, and to determine if the best classification models contained any foot features. The justification for this approach was that the movement of the foot may only become important with respect to injury when considered in conjunction with other biomechanical features. To answer this second aim, six classification models were trained, and the performance of the models were evaluated using bootstrapped resamples of accuracy, sensitivity and specificity. The results from this experiment illustrated that the both the Random Forest model and the Stacked Ensemble were useful classifiers, reporting a mean accuracy of 71% and 70% respectively, both of which were statistically significantly different than classifying the majority class ($p < 0.01$). While direct comparison to the literature is difficult given the different evaluation metrics utilised by previous research (e.g. Hazard Ratio, Pseudo R^2), in comparison to the 16 studies included in a recent systematic review of prospective risk factors for running injury (Ceyssens et al., 2019), it would appear that this current project is state of the art (see config manual, section 6.1.2 for a detailed review). For example, all the studies in the systematic review which utilised multivariate models, failed to use any form of out of sample testing and only explored a single model. This can lead to poor generalisability of the studies' findings, and as per the no free lunch theorem (Wolpert, 1996), runs the risk of utilising a non-optimal model for the data being examined.

Within this current project, the bagged SVM was the worst performing model with accuracies little better than random guessing. While the reason for this poor performance is unclear, it may be related to suboptimal bootstrap parameters. Indeed, the hyperparameter grid space for this model focused on the SVM base estimator rather than the bootstrapping approach. Future research should further evaluate this process. Interestingly, the weighted Stacked Ensemble did not outperform the Random Forest in this project in terms of accuracy. This may be due to two reasons. Firstly, as scikit learn does not directly support weighted stacking, a pragmatic equation was proposed (see config manual section 6.5) which may not have been optimal. Secondly, the Stacked Ensemble was tuned on Youden's J statistic and as such, rather than optimising for accuracy, it was a slightly more balanced classifier in comparison to the Random Forest model. Despite this, the differences between the two models were not significant and as such, the Random Forest was considered in further detail.

In terms of sensitivity and specificity, the Random Forest model was able to correctly classify 74.3% of those who went on to become injured and 65% of those who did not. These are compelling findings when considering the simple binary classification of injury and the fact that those who go on to become injured may do so due reasons other than movement biomechanics. For example, age, genetics and previous injury are all known risk factors for injury (Ceyssens et al., 2019), that were not considered within the scope of this current project. In a similar light, it is possible that those who were misclassified as injured have not yet encountered the cumulative threshold of load required to cause an injury and/or due to genetic reasons have not become injured despite presenting with injurious movement biomechanics. It is worth nothing however, that despite the high performance reported by the Random Forest model on average, there was considerable variation in the findings. Future research should therefore explicitly explore the characteristic of the subsamples that cause this variation. When considering the final features included in the Random Forest model, it is of note, that the movement of the foot was not included. This finding again suggests that foot-strike pattern is not amongst the most important biomechanical features when trying to predict risk of injury. Rather, clinical practitioners may be advised to target the final features utilised by the Random Forest model with particular emphasis on the features most related to an increased probability of being classified as injured, as determined by the feature importance test and subsequent partial dependency analysis. For example, just before toe-off, an increase in thorax ipsilateral velocity is related to increased probability of being classified as injured. This can be explained as during running, it is important to have a neutral trunk at the time take off to avoid lateral projection of the body's centre of mass. The injured group must readjust their trunk position, as for the majority of stance, they are in a more contralaterally tilted position in comparison to the prospectively uninjured group (see config manual section 6.2.1). This, trunk positioning is a compensatory pattern for weak hip abductors known as Trendelenburg's sign and is one of the most commonly cited risk factors for knee injury (Ferber et al., 2011).

6.3.1 Limitations

Within this current project, the participants ran at a standardized speed on the treadmill. While, this allowed direct comparison between participants, it is possible that a self-selected training pace would provide a more ecologically valid representation of the participant's biomechanics. Furthermore, while at the time of writing this project, only a binary classification of injury was available, it is likely that a breakdown of injury by location and type would substantially improve the predictive performance of the models explored. Similarly, within this project, for the research question related to classification, the participants were assumed to be suitably homogeneous. Future research should determine if there are any sub clusters in the full movement biomechanics, that could improve the classification performance of the models explored. Finally, while the size of the dataset prohibited the use of greedy search algorithm, the use of the genetic search algorithm and the Bayesian optimization may not have resulted in the optimal solution.

7 Conclusion and Future Work

As a result of this project, all objectives listed in section 1.2 have been implemented and the research questions have been answered. The primary aim and research question of this project was to identify the presence of naturally occurring foot-strike groupings, using unsupervised clustering approaches and to determine their association with injury. To date, to the best of the candidate's knowledge, no research had utilised clustering models to explicitly explore foot-strike patterns. Within this present project, six clustering approaches were explored (K-means, Hierarchical, OPTICS, Mean Shift, Spectral and HDBSCAN), of which, the four best were explored in further detail. In contrast to the hypothesis of this project (see section 1.1), the distribution of the injury class across clusters demonstrated almost random assignment. This finding suggests that neither the identified clusters or the traditional approach to foot-strike classification were able to distinguish between prospectively injured and uninjured subjects. The secondary aim and sub research question of this project was to determine if any of the biomechanical features captured could predict risk of injury and to determine if the best classification model contained any foot features. After an extensive pipeline of feature selection (genetic algorithms, recursive feature elimination) and hyperparameter optimisation, six models (Elastic Net Logistic Regression, Naïve Bayes, Bagged SVM, Random Forest, AdaBoost and a weighted Stacked Ensemble model) were evaluated using bootstrapped resampling on a hold out test set and the best performing model was chosen for further investigation. A Random Forest model containing ten features achieved an accuracy of 71% which was statistically different than a majority classifier and was deemed the best performing model. Interestingly, and again in contrast to the hypothesis of this project (see section 1.1), the foot was not included in final Random Forest model, suggesting that the movement of the foot is not amongst the most important features related to injury classification. Clinicians should therefore be advised to target the final features utilised by the Random Forest model rather than foot-strike pattern. These results also have considerable implications for manufactures, who have developed products (e.g. footwear, wearable sensors) around the concept of foot-strike pattern to reduce risk of injury and may have to re-evaluate current designs.

In order to enhance the ability to predict running related injuries, future research should provide a more detailed breakdown of injury type beyond a binary classification and consider testing participants at a more ecologically valid, self-selected running pace. Furthermore, research should also explore the use of unsupervised clustering to determine if there are any subgroups in the whole-body biomechanics that could enhance the predictive ability of the classification models. Finally, future research should implement a clinical intervention study aimed at targeting the most important biomechanical features identified in this project. It is anticipated that the results from this project will lead to more targeted injury prevention interventions which will reduce the incidence of running related injury and ultimately improve the health of the population.

8 Acknowledgement

I would like to start by thanking my supervisor, Dr. Catherine Mulwa for her guidance and support throughout this research project. I would also like to thank the RISC research team in DCU for allowing me to use the running data. On a personal level, I would like to thank my family for their continued support in life. In particular, I would like to thank my financé Róisín for all her love and for supporting me once again through another degree.

References

Akima, H. (1970). A New Method of Interpolation and Smooth Curve Fitting Based on Local Procedures, *J. ACM* **17**(4): 589–602.

Almeida, M. O., Davis, I. S. and Lopes, A. D. (2015). Biomechanical differences of foot-strike patterns during running: A systematic review with meta-analysis, *J. Orthop. Sports Phys. Ther.* **45**(10): 738–755.

Altman, A. R. and Davis, I. S. (2012). A kinematic method for footstrike pattern detection in barefoot and shod runners, *Gait Posture* **35**(2): 298–300.

Bahr, R. (2016). Why screening tests to predict injury do not work-and probably never will.: A critical review, *Br. J. Sports Med.* **50**(13): 776–780.

Bennetts, C. J., Owings, T. M., Erdemir, A., Botek, G. and Cavanagh, P. R. (2013). Clustering and classification of regional peak plantar pressures of diabetic feet, *J. Biomech.* **46**(1): 19–25.

Calzolari, M. (2019). manuel-calzolari/sklearn-genetic: sklearn-genetic 0.2 (Version 0.2).

Cavanagh, P. R. and Lafortune, M. A. (1980). Ground reaction forces in distance running, *J. Biomech.* **13**(5): 397–406.

Ceyssens, L., Vanelderen, R., Barton, C., Malliaras, P. and Dingenen, B. (2019). Biomechanical Risk Factors Associated with Running-Related Injuries: A Systematic Review, *Sport. Med.* **49**(7): 1095–1115.

Chau, T. (2001). A review of analytical techniques for gait data. Part 1: Fuzzy, statistical and fractal methods, *Gait Posture* **13**(1): 49–66.

Christ, M., Braun, N., Neuffer, J. and Kempa-Liehr, A. W. (2018). Time Series FeatuRe Extraction on basis of Scalable Hypothesis tests (tsfresh – A Python package), *Neurocomputing* **307**: 72–77.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences, *Stat. Power Anal. Behav. Sci.* **2nd**: 567.

Daoud, A. I., Geissler, G. J., Wang, F., Saretsky, J., Daoud, Y. A. and Lieberman, D. E. (2012). Foot strike and injury rates in endurance runners: A retrospective study, *Med. Sci. Sports Exerc.* **44**(7): 1325–1334.

De Cock, A., Willems, T., Witvrouw, E., Vanrenterghem, J. and De Clercq, D. (2006). A functional foot type classification with cluster analysis based on plantar pressure distribution during jogging, *Gait Posture* **23**(3): 339–347.

Delacre, M., Leys, C., Mora, Y. L. and Lakens, D. (2019). Taking Parametric Assumptions Seriously: Arguments for the Use of Welch's F test instead of the Classical F test in One-Way ANOVA, *Int. Rev. Soc. Psychol.* **32**(1): 13.

Deschamps, K., Matricali, G. A., Roosen, P., Desloovere, K., Bruyninckx, H., Spaepen, P., Nobels, F., Tits, J., Flour, M. and Staes, F. (2013). Classification of forefoot plantar pressure distribution in persons with diabetes: A novel perspective for the mechanical management of diabetic foot?, *PLoS One* **8**(11).

Dingenen, B., Staes, F., Vanelderen, R., Ceyssens, L., Malliaras, P., Barton, C. J. and Deschamps, K. (2020). Subclassification of recreational runners with a running-related injury based on running kinematics evaluated with marker-based two-dimensional video analysis, *Phys. Ther. Sport* **44**: 99–106.

Dingwell, J. B., Cusumano, J. P., Cavanagh, P. R. and Sternad, D. (2001). Local dynamic stability versus kinematic variability of continuous overground and treadmill walking, *J. Biomech. Eng.* **123**(1): 27–32.

Donoghue, O. A., Harrison, A. J., Laxton, P. and Jones, R. K. (2008). Lower limb kinematics of subjects with chronic Achilles tendon injury during running, *Res. Sport. Med.* **16**(1): 23–38.

Dudley, R. I., Pamukoff, D. N., Lynn, S. K., Kersey, R. D. and Noffal, G. J. (2017). A prospective comparison of lower extremity kinematics and kinetics between injured and non-injured collegiate cross country runners, *Hum. Mov. Sci.* **52**: 197–202.

Farley, C. T. and Morgenroth, D. C. (1999). Leg stiffness primarily depends on ankle stiffness during human hopping, *J. Biomech.* **32**(3): 267–273.

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996). The KDD Process for Extracting Useful Knowledge from Volumes of Data, *Commun. ACM* **39**(11): 27–34.

Ferber, R., Kendall, K. D. and Farr, L. (2011). Changes in knee biomechanics after a hip-abductor strengthening protocol for runners with patellofemoral pain syndrome, *J. Athl. Train.* **46**(2): 142–149.

Forrester, S. E. and Townend, J. (2015). The effect of running velocity on footstrike angle - A curve-clustering approach, *Gait Posture* **41**(1): 26–32.

Franklyn-Miller, A., Richter, C., King, E., Gore, S., Moran, K., Strike, S. and Falvey, E. C. (2017). Athletic groin pain (part 2): A prospective cohort study on the biomechanical evaluation of change of direction identifies three clusters of movement patterns, *Br. J. Sports Med.* **51**(5): 460–468.

Futrell, E. E., Jamison, S. T., Tenforde, A. S. and Davis, I. S. (2018). Relationships between Habitual Cadence, Footstrike, and Vertical Load Rates in Runners, *Med. Sci. Sports Exerc.* **50**(9): 1837–1841.

Goss, D. L. and Gross, M. T. (2012). Relationships among self-reported shoe type, footstrike pattern, and injury incidence., *US. Army Med. Dep. J.* pp. 25–30.

Handsaker, J. C., Forrester, S. E., Folland, J. P., Black, M. I. and Allen, S. J. (2016). A kinematic algorithm to identify gait events during running at different speeds and with different footstrike types, *J. Biomech.* **49**(16): 4128–4133.

Hein, T., Janssen, P., Wagner-Fritz, U., Haupt, G. and Grau, S. (2014). Prospective analysis of intrinsic and extrinsic risk factors on the development of Achilles tendon pain in runners, *Scand. J. Med. Sci. Sport.* **24**(3).

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013). *An introduction to Statistical Learning*, Springer Texts in Statistics, Springer New York, New York, NY.

Jauhiainen, S., Pohl, A. J., Äyrämö, S., Kauppi, J. P. and Ferber, R. (2020). A hierarchical cluster analysis to determine whether injured runners exhibit similar kinematic gait patterns, *Scand. J. Med. Sci. Sport.* **30**(4): 732–740.

Kuhman, D. J., Paquette, M. R., Peel, S. A. and Melcher, D. A. (2016). Comparison of ankle kinematics and ground reaction forces between prospectively injured and uninjured collegiate cross country runners, *Hum. Mov. Sci.* **47**: 9–15.

Kuhn, M. and Johnson, K. (2019). *Feature Engineering and Selection : a Practical Approach for Predictive Models.*, CRC Press LLC.

Lieberman, D. E., Venkadesan, M., Werbel, W. A., Daoud, A. I., Dandrea, S., Davis, I. S., Mangeni, R. O. and Pitsiladis, Y. (2010). Foot strike patterns and collision forces in habitually barefoot versus shod runners, *Nature* **463**(7280): 531–535.

Marshall, B., Franklyn-Miller, A., Moran, K., King, E., Richter, C., Gore, S., Strike, S. and Falvey, É. (2015). Biomechanical symmetry in elite rugby union players during dynamic tasks: an investigation using discrete and continuous data analysis techniques, *BMC Sports Sci. Med. Rehabil.* **7**(1): 13.

Messier, S. P., Martin, D. F., Mihalko, S. L., Ip, E., DeVita, P., Cannon, D. W., Love, M., Beringer, D., Saldana, S., Fellin, R. E. and Seay, J. F. (2018). A 2-Year Prospective

Cohort Study of Overuse Running Injuries: The Runners and Injury Longitudinal Study (TRAILS), *Am. J. Sports Med.* **46**(9): 2211–2221.

Moudy, S., Richter, C. and Strike, S. (2018). Landmark registering waveform data improves the ability to predict performance measures, *J. Biomech.* **78**: 109–117.

Moulavi, D., Jaskowiak, P. A., Campello, R. J., Zimek, A. and Sander, J. (2014). Density-based clustering validation, *SIAM Int. Conf. Data Min. 2014, SDM 2014*, Vol. 2, pp. 839–847.

Paquette, M. R., Milner, C. E. and Melcher, D. A. (2017). Foot contact angle variability during a prolonged run with relation to injury history and habitual foot strike pattern, *Scand. J. Med. Sci. Sport.* **27**(2): 217–222.

Pataky, T. C. (2010). Generalized n-dimensional biomechanical field analysis using statistical parametric mapping, *J. Biomech.* **43**(10): 1976–82.

Pataky, T. C. (2012). One-dimensional statistical parametric mapping in Python, *Comput. Methods Biomech. Biomed. Engin.* **15**(3): 295–301.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, É. (2011). Scikit-learn: Machine learning in Python, *J. Mach. Learn. Res.* **12**: 2825–2830.

Phinyomark, A., Osis, S., Hettinga, B. A. and Ferber, R. (2015). Kinematic gait patterns in healthy runners: A hierarchical cluster analysis, *J. Biomech.* **48**(14): 3897–3904.

Pizzuto, F., Rago, V., Bailey, R., Tafuri, D. and Raiola, G. (2016). The importance of foot-strike patterns in running: A literature review, *Sport Sci.* **9**: 87–96.

Pohl, M. B., Mullineaux, D. R., Milner, C. E., Hamill, J. and Davis, I. S. (2008). Biomechanical predictors of retrospective tibial stress fractures in runners, *J. Biomech.* **41**(6): 1160–1165.

Ramsay, J. O. J. O. and Silverman, B. W. (2005). *Functional data analysis*, 2nd edn, Springer, Verlag New York.

Richter, C., King, E., Strike, S. and Franklyn-Miller, A. (2019). Objective classification and scoring of movement deficiencies in patients with anterior cruciate ligament reconstruction, *PLoS One* **14**(7): e0206024.

Richter, C., O'Connor, N. E., Marshall, B. and Moran, K. (2014a). Analysis of characterizing phases on waveforms: An application to vertical jumps, *J. Appl. Biomech.* **30**(2): 316–321.

Richter, C., O'Connor, N. E., Marshall, B. and Moran, K. (2014b). Clustering vertical ground reaction force curves produced during countermovement jumps, *J. Biomech.* **47**(10): 2385–2390.

Richter, C., O'Connor, N. E., Marshall, B. and Moran, K. (2014c). Comparison of discrete-point vs. dimensionality-reduction techniques for describing performance-related aspects of maximal vertical jumping, *J. Biomech.* **47**(12): 3012–3017.

Roche, N., Pradon, D., Cosson, J., Robertson, J., Marchiori, C. and Zory, R. (2014). Categorization of gait patterns in adults with cerebral palsy: A clustering approach, *Gait Posture* **39**(1): 235–240.

Rodríguez, J., Medina-Pérez, M. A., Gutierrez-Rodríguez, A. E., Monroy, R. and Terashima-Marín, H. (2018). Cluster validation using an ensemble of supervised classifiers, *Knowledge-Based Syst.* **145**: 134–144.

Rothschild, C. E. (2012). Primitive running: A survey analysis of runners' interest, participation, and implementation, *J. Strength Cond. Res.* **26**(8): 2021–2026.

Ruxton, G. D. and Beauchamp, G. (2008). Time for some a priori thinking about post hoc testing, *Behav. Ecol.* **19**(3): 690–693.

Sawacha, Z., Guarneri, G., Avogaro, A. and Cobelli, C. (2010). A new classification of diabetic gait pattern based on cluster analysis of biomechanical data, *J. Diabetes Sci. Technol.*, Vol. 4, pp. 1127–1138.

Scheerder, J., Breedveld, K. and Borgers, J. (2015). *Running across Europe: The rise and size of one of the largest sport markets*, Palgrave Macmillan UK.

Stiffler-Joachim, M. R., Wille, C. M., Kliethermes, S. A., Johnston, W. and Heiderscheit, B. C. (2019). Foot Angle and Loading Rate during Running Demonstrate a Nonlinear Relationship, *Med. Sci. Sports Exerc.* **51**(10): 2067–2072.

Tang, J., Alelyani, S. and Liu, H. (2014). Feature selection for classification: A review, *Data Classif. Algorithms Appl.*

Taunton, J. E., Ryan, M. B., Clement, D. B., McKenzie, D. C., Lloyd-Smith, D. R. and Zumbo, B. D. (2002). A retrospective case-control analysis of 2002 running injuries, *Br. J. Sports Med.* **36**(2): 95–101.

Van Gent, R. N., Siem, D., Van Middelkoop, M., Van Os, A. G., Bierma-Zeinstra, S. M. and Koes, B. W. (2007). Incidence and determinants of lower extremity running injuries in long distance runners: A systematic review, *Br. J. Sports Med.* **41**(8): 469–480.

Warr, B. J., Fellin, R. E., Sauer, S. G., Goss, D. L., Frykman, P. N. and Seay, J. F. (2015). Characterization of foot-strike patterns: Lack of an association with injuries or performance in soldiers, *Mil. Med.* **180**(7): 830–834.

Watari, R., Osis, S. T., Phinyomark, A. and Ferber, R. (2018). Runners with patellofemoral pain demonstrate sub-groups of pelvic acceleration profiles using hierarchical cluster analysis: An exploratory cross-sectional study, *BMC Musculoskelet. Disord.* **19**(1): 120.

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

Wolpert, D. H. (1996). The Lack of a Priori Distinctions between Learning Algorithms, *Neural Comput.* **8**(7): 1341–1390.

Xu, D. and Tian, Y. (2015). A Comprehensive Survey of Clustering Algorithms, *Ann. Data Sci.* **2**(2): 165–193.

Yeow, C. H., Lee, P. V. S. and Goh, J. C. H. (2011). An investigation of lower extremity energy dissipation strategies during single-leg and double-leg landing based on sagittal and frontal plane biomechanics, *Hum. Mov. Sci.* **30**(3): 624–635.

Zhao, Z. and Liu, H. (2007). Spectral feature selection for supervised and unsupervised learning, *ACM Int. Conf. Proceeding Ser.*