Analysis of changing risk factors and explanation of risk predictions with machine learning for improved hamstring strain injury prevention in Australian football

Sim, Aylwin Chun Wei

Analysis of changing risk factors and explanation of risk predictions with machine learning for improved hamstring strain injury prevention in Australian football

Submitted by

Aylwin Chun Wei Sim (CompSci Honours)

A thesis submitted in total fulfilment of the requirements of the degree of

Master of Philosophy

Human-Centred Intelligent Learning and Software Technologies Research Lab

Peter Faber School

Australian Catholic University

Submitted on 3 March 2023

Statement of Authorship and Sources

This thesis contains no material that has been extracted in whole or in part from a thesis that I have submitted towards the award of any other degree or diploma in any other tertiary institution.

No other person's work has been used without due acknowledgment in the main text of the thesis.

All research procedures reported in the thesis received the approval of the relevant Ethics/Safety Committees (where required).

Aylwin Chun Wei Sim

3 March 2023

Supervisory Panel

The work of this thesis was supervised by the following ACU staffs:

Principal supervisor	Associate Prof. Haifeng Shen
Co-Supervisors	Dr Kewen Liao, Dr David Opar

Acknowledgements

I would like to express my deepest appreciation to my supervisors Prof. Haifeng Shen and Dr. Kewen, for their invaluable guidance and support throughout the course of this research. Especially this work was conducted during a difficult time due to COVID-19 pandemic, which brought challenges to the research process. I also would like to thank Dr. Joshua Ruddy and Dr. David Opar for their generosity in offering their time and expertise to help me completing this research.

Lastly, I would like to thank my family for their unwavering support throughout this academic journey. Their love, encouragement, and understanding have been essential in making this research possible.

Table of Contents

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ROC	Receiver operating characteristic
AUC	Area under the curve
ML	Machine Learning
LIME	Local interpretable model-agnostic explanations
SHAP	SHAPley additive explanations
HSI	Hamstring Strain Injury
ACL	Anterior Cruciate Ligament
UCL	Ulnar Collateral Ligament
BFlh	Biceps femoris long head
NHE	Nordic Hamstring Exercise
AFL	Australian Football League
NFL	National Hockey League
NBA	National Basketball Association
ANN	Artificial neural network
NLP	Natural Language Processing
BERT	Bidirectional Encoder Representations from Transformers
SMOTE	Synthetic minority oversampling
CF	Counterfactual
DiCE	Diverse counterfactual explanations

ABSTRACT

Professional athletes and organizations can face significant consequences as a result of injury incidents in sports. Therefore, an abundance of studies has been conducted to identify the risk factors in the hope of preventing injuries from occurring in the first place. Hamstring strain injuries (HSIs) are the most frequent injuries in Australian Football League (AFL). Many studies had shown that there are several prominent risk factors for HSIs. However, this finding cannot be identified with any consistency through assessing the risk factors at a single time point, typically the beginning of a season (e.g., in the pre-season) or more frequently throughout the season (e.g., in the pre-season, early in-season and late in-season). Nonetheless, these studies did not consider the potential variability of risk factors across the season. In light of this, it was hypothesised that risk factors may vary depending on the time of the season.

This thesis aims to answer if the risk of hamstring strain injuries in Australian Football can be reduced through a better understanding of the changing risk factors over the course of the season. Despite the study, identifying HSI risk at individual-level remains a challenge. This study aims to explore whether the risk of HSI for individual players can be better understood by explaining the predictions of machine learning (ML) models.

The study utilised recursive feature selection and cross-validation to provide a holistic understanding of important risk factors at different points. Subsequently, counterfactual explanations were effectively generated for players at risk of sustaining HSI.

The study found that non-modifiable risk factors were primarily linked to pre-season injuries, whereas modifiable risk factors were mostly associated with early in-season injuries.

Counterfactual explanations and ML models offer a novel perspective in interpreting risk and finding potential solutions.

Overall, this study provides new insights into risk factors associated with HSIs at different time points, as well as offers a solution for interpreting risk at individual-level using ML models and counterfactual explanations. The findings have important implications for researchers and practitioners who seek to mitigate the risk of HSI in the future.

THESIS STRUCTURE

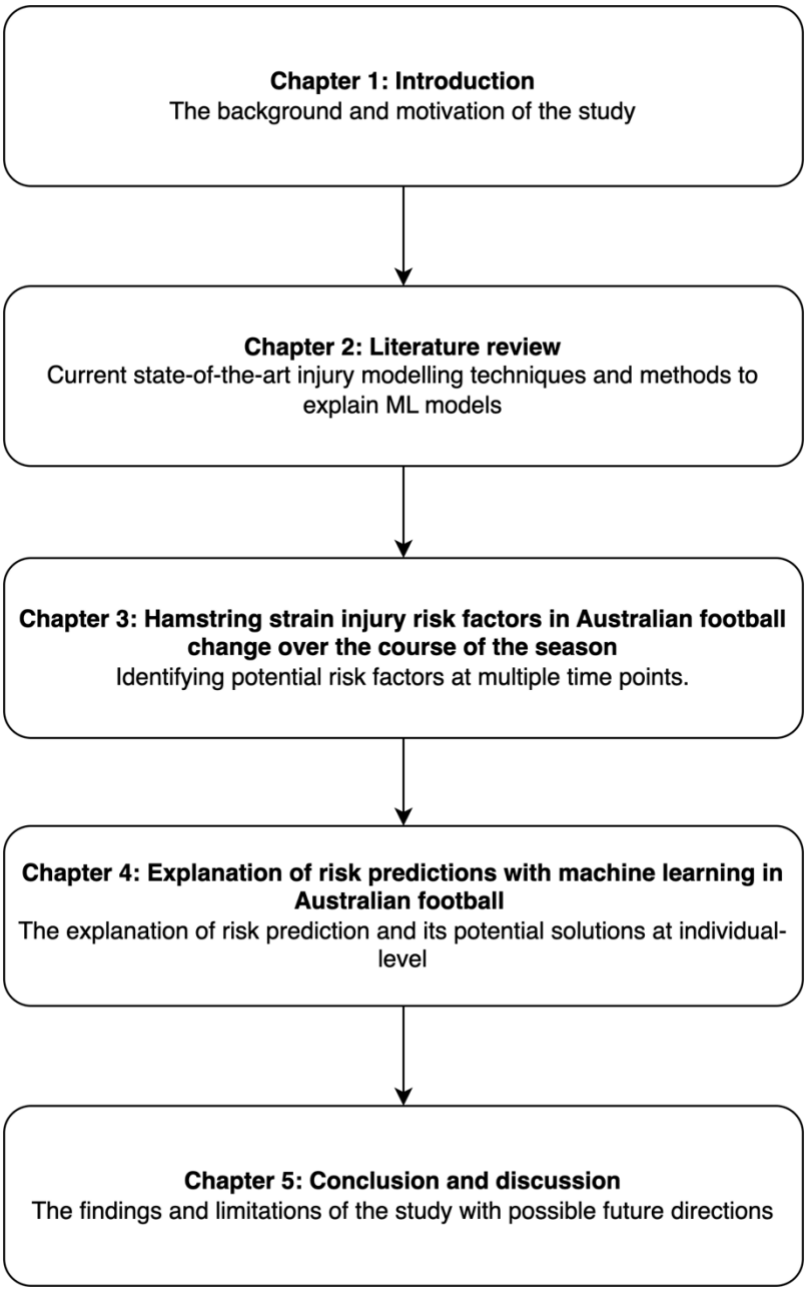


Figure 1: The chapters included in this thesis.

CHAPTER 1: INTRODUCTION

Injuries are common in professional sports. They can cause serious problems to an athlete's well-being, along with financial implication [1, 2]. On average, a game missed by an Australian Football player costs $15,000. As few as 10 absences can cost $150,000 lost in performance value [2]. Normally, injuries require a recovery time of up to 4 weeks before a player can return to play (RTP) [3]. Not to mention any injury that could impact the performance of the athlete permanently [4].

Hamstring strain injuries (HSIs) are the most frequent non-contact injuries in Australian football, American football, professional soccer, and rugby union which often require players to perform high-speed running, jumping, acceleration and deceleration [5-10]. HSI occurs when the muscles at the back of the thigh (hamstrings) were stretched beyond their limit, which results in sharp pain in the posterior thigh [6].

Considering the pervasive occurrence and the adverse outcomes that HSI brings, an abundance of studies has been conducted to investigate the factors that lead to the increased risk of HSI. It was reported older age and a history of HSI are prominent risk factors [6, 11, 12]. Individuals with a prior incidence of HSI are 2.7 times more likely to experience a subsequent HSI compared to those who do not have a history of HSI [11]. Additionally, the risk of HSI increases by 1.3-fold for each additional year of age in Australian rules football players and by 1.9-fold with each increasing year of age in soccer players [6]. Nonetheless, these risk factors cannot be altered, which limits their applicability for future prevention. Studies had shown several modifiable risk factors, which can be addressed through intervention, were associated with HSI. These include biceps femoris long head (BFlh) architectures and eccentric hamstring strength [13-15]. According to a study, hamstring strain injuries are more common in

professional soccer players who have short BFlh fascicles and weak eccentric knee flexor strength [15]. The study proposed that longer BFlh fascicles and stronger eccentric strength can help senior players who sustained HSI lower their risk of injury [15].

In spite of the aforementioned evidence, the findings cannot be identified with any consistency [10, 16, 17]. An earlier research on Australian Football came to the conclusion that the risk of future HSI cannot be predicted by age, history of HSI, and eccentric hamstring strength, despite various machine learning models being utilised to account for the complex interactions between these variables [16]. A subsequent study attempted to address this finding by assessing modifiable risk factors more regularly throughout the season came to the same conclusion that there were no advances in identifying the risk of future HSI [17].

The purpose of this thesis is to determine if the risk of HSI in Australian Football can be mitigated through a better understanding of the changing risk factors over the course of the season. Despite the study, it remains difficult to comprehend risk at individual-level. While previous studies attempted to explain risk factors that contribute to the prediction for individual athletes [18, 19]. Their studies did not identify the potential solution to mitigate risk. As a result, this study was extended to explore whether the risk of HSI for individual players can be better understood by explaining the predictions of machine learning (ML) models, as well as identifying the potential risk mitigation solutions.

It was hypothesised that HSIs occurred in preseason, early in-season and late in-season were associated with different risk factors. This thesis consists of two studies. The first study (Chapter 3) aimed to identify a group of features that can optimise the predictive ability of HSIs in individual time points. Recursive feature elimination and cross validation (RFECV)

were utilised to identify important risk factors. The following study (Chapter 4) aimed to explain the model's prediction for individual players at multiple time points. With modifiable risk factors, the use of counterfactual explanation and ML offers a solution for interpreting risk at individual-level, as well as identifying potential solutions to prevent HSI [20].

The study discovered that pre-season HSIs were strongly associated with age, history of HSI, and height, which are non-modifiable risk factors, whereas early in-season HSIs were significantly linked to modifiable BFlh fascicle length and pennation angle. Conversely, late in-season HSIs did not present any strong associations with either modifiable or non-modifiable risk factors examined in this study. The magnitude of change in modifiable risk factors across pre-season did not improve the identification of in-season HSIs. To the best of my knowledge, this study is the first to demonstrate the effectiveness of counterfactual explanations in interpreting risk and discovering potential solutions to mitigate the risk of future HSI for high-risk players.

Ethical approval

Australian Football League data was used as a case study throughout this research [17]. The data were collected with the approval from ACU Human Research Ethics Committee (approval number: 2017-208H) and with written consent were provided by participating players. This dataset consists of both non-modifiable risk factor and modifiable risk factors collected from individual players over the period of 2018 and 2019 Australian Football League (AFL) seasons.

CHAPTER 2: LITERATURE REVIEW

Injury risk factors

Many studies have been undertaken to identify the potential risk factors of various sports injuries. The initial step for these studies is collecting data. This includes collecting data from athletes through specialised instruments or devices [14, 21], gathering demographic data (weight, height, age etc) from the participants, or conducting surveys on the wellness of athletes through a series of customized questionnaires [17, 22].

There are numerous musculoskeletal injuries in professional sports, with some commonly reported ones including hamstring strain injuries (HSIs), injuries to the lower extremities, anterior cruciate ligament (ACL) injuries, and ulnar collateral ligament (UCL) injuries. The most frequent injury is the hamstring strain injury (HSI) [23]. Football, soccer, and running activities are common to experience this type of injury. Some studies had shown that age is a prominent risk factor for lower limb muscle injuries in sports that involved sprinting [6, 8, 24]. For example, older athletes have been found to be more susceptible to hamstring and calf strain injuries [8, 24]. Studies had suggested muscle atrophy and weaker muscle fibres in senior athletes are [25, 26] making them more prone to injury. This can also be triggered by muscle contractions when competing with younger athletes. Similarly, the presence of prior injuries is also a significant risk factor in many studies [6, 8, 24, 27]. Athletes with a history of injury on the muscle significantly increased risk of a number of muscle strains [24]. This may be due to changes in muscle proprioception, strength and kinematics which lead to the decrease in neuromuscular functions [27-29].

Recent studies had shifted their focus to modifiable risk factors that can be addressed through exercise and intervention [13, 14]. The commonly reported factors are training loads, muscle

imbalance [30], muscle strength [15] and muscle architectures [15, 31]. Physical training helps athletes to improve their physical fitness and overall performance. However, the amount and intensity of external training loads may have an impact on the injury risk [32]. Heavy workloads may result in fatigue, which can lead to injury. Conversely, a light workload may reduce the fitness and preparedness of athletes and subsequently increase the risk of injury [33]. There is an abundance of studies investigating scientific approaches to monitoring training loads in various sport domains [32, 34, 35]. Furthermore, prior studies reported that high between-limb imbalance and eccentric weakness may have contributed to an increased risk of hamstring strain injuries (HSIs) in professional rugby teams [30] and elite Australian footballers [14]. The injury risk may be reduced by enhancing eccentric hamstring strength with the Nordic hamstring exercise (NHE) [36, 37]. Others claimed that having short biceps femoris long head (BFlh) fascicle length can elevate the risk of HSI in soccer players [15]. NHE was also reported as an effective method in lengthening the BFlh fascicles [38].

Injury modelling

Machine learning (ML) models are known for discovering patterns and relationships in complex data. Recent years have seen a substantial surge in interest in leveraging machine learning (ML) models to identify athletes at risk of getting injured [16, 19, 21]. With the availability of data sourced from sensors and internet. Some studies have shown that complex ML models significantly outperformed traditional statistical modelling [21]. In another study, it was claimed that there is no significant improvement in predictive performance when the dataset is relatively small [16]. Nonetheless, logistic regression is the most widely used technique in predicting sport injuries [39]. This is likely due to its simplicity and transparency which can be understood by practitioners [40]. However, the model struggles to capture non-

linear dynamics and complex interplay between risk factors, potentially leading to an inaccurate representation of the risk model [2].

Sports injuries are known to be caused by the complex interactions of various risk factors and inciting events [41]. Many researchers have turned to machine learning models to predict injury risk at the individual-level [1]. One of the most predominantly used methods is tree-based ensemble technique. This is due in part to their ease of interpretation and the ability to improve predictive performance through boosting or bagging methods. [1]. In a particular study, XGBoost was employed to forecast the injury risk in the National Hockey League (NHL) for the upcoming season [19]. It was reported that XGBoost outperformed all other models with an average AUC of 0.948 for position players and a mean AUC of 0.956 for goalies [19]. A similar study was conducted to predict injury risk among middle-distance and long-distance professional runners. Although no other models were compared in this study, XGBoost showed promising results with an average AUC of 0.724 in predicting weekly injury events. It should be highlighted that these studies were carried out with substantially larger datasets that contain at least 6000 observations.

Besides tree-based ensembles, a study proposed a deep-learning based method to predict various types of injuries in NBA basketball [42]. The injury prediction system drew inspiration primarily from a deep learning model known as Bidirectional Encoder Representations from Transformers (BERT), which is commonly used for natural language processing (NLP) tasks. Without any data balancing technique, the system performed significantly better than other models with a resulting AUC of 0.8. It was found that the contusion injury was associated with a number of muscle injuries. Similar research studies have utilized artificial neural networks (ANNs) to quantify the risk of injury in the Australian Rules Football [2, 16].

One of the challenges faced in injury prediction is highly unbalanced data [16, 43], where the number of healthy events significantly outnumbered injury events. This can introduce problems as the model only learns from the majority of healthy events. A common solution to this problem is data random over-sampling and under-sampling. Random Oversampling works by randomly increasing the injury events so that the number of both classes is equal [16, 21]. Under-sampling is used less frequently as this may result in loss of information. Alternately, synthetic minority oversampling (SMOTE), which selectively creates synthetic samples from the minority class, can be utilised to remedy class imbalance. [16]. Another challenge in modelling sports injury is limited data. It was reported that most injury prediction studies were carried out on small datasets with a median sample size of 152 observations [39]. Almost all datasets used in these studies are not publicly available and cannot be cross-validated. Additionally, many of these data are expensive to collect, which requires special devices and consent from professional athletes. In real word, injuries occur because of complex reasons. These include environmental factors, inciting events and psychological factors that are not well captured by the data [33, 41]. Moreover, the risk factors and types of injuries differ across various sports. All these contributors add up to the challenge of predicting injuries accurately in the domain.

Modelling metrics

Most sports science datasets are highly imbalance [44]. It is crucial to understand the advantages and limitation of existing metrics prior making decision based on the trained classifier.

ROC AUC – Area under the ROC Curve

AUC is a commonly used metric in the domain of predicting sports-related injuries. It is a threshold-independent metric that can be used to measure the performance of any classifiers without the need to define a cut-off threshold [45]. This feature of AUC makes it particularly useful in the field where the proportion of healthy events and injury events are imbalanced, as it evaluates the classifier across different thresholds, making it a more reliable metric for comparing different models. AUC evaluates the ability of the model to correctly discriminate between positive and negative classes [45]. The value ranges from 0 to 1, with a higher score indicating better performance. A perfect classification is indicated by an AUC of 1.0, while a value less than 0.5 suggests that the classifier is performing worse than random guessing. [46]. Another advantage of AUC is its scale in-invariant nature [47]. AUC score is not affected by changes in the scale of the scores and probabilities output by the model as it is calculated through ranking the prediction outputs of the model.

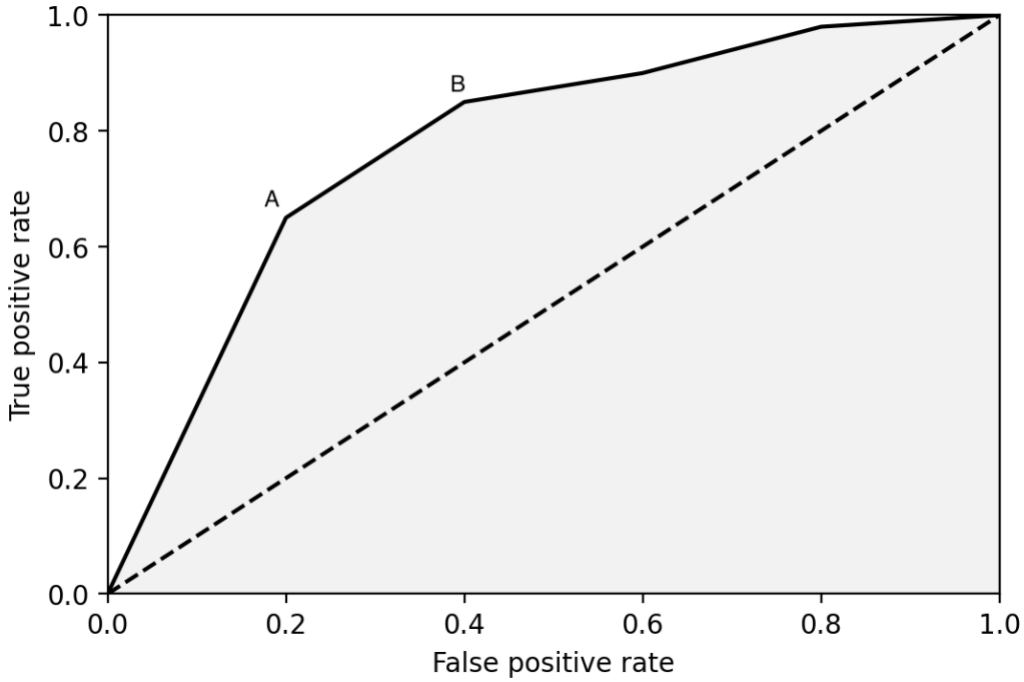


Figure 2: An illustration of Receiver Operating Characteristic (ROC) curve and Area Under the Curve (AUC). Y-axis indicates true positive rate (TPR), also known as sensitivity. X-axis indicates false positive rate (FPR). The diagonal dotted line indicates a random guess where AUC = 0.5. The solid

line is the probability curve of the tested classifier, also known as the ROC curve. The shaded area indicates AUC. Label A and label B indicates different decision thresholds.

True positive rate (TPR), also known as sensitivity and recall, is the proportion of injured events that were identified correctly. It is calculated as the number of injury events that were correctly identified (TP) divided by the actual number of positive samples (actual injury events = TP + FN) as shown in the equations below:

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

False positive rate (FPR), also referred to as false alarm rate, is the proportion of healthy events that were incorrectly identified. It is calculated as the number of healthy events that were incorrectly identified (FP) divided by the actual number of negative samples (actual healthy events = FP + TN) as shown in the equations followed:

$$FPR = \frac{FP}{FP + TN}$$

Figure 2 illustrates how AUC is calculated with the ROC curve. For a given prediction output from any binary classifier, TPR and FPR were calculated across all possible cut-off thresholds to obtain a probability curve, known as the ROC curve. The area under the ROC curve is known as ROC AUC. Ideally, a good classifier has high TPR and low FPR, which yields high AUC score.

Since AUC only measures how well output predictions were ranked on every possible threshold, it neglects the importance of calibrated probability [48]. Furthermore, it was argued that AUC is not suitable in certain cases where the cost of false positive and false negatives should be treated differently [48]. When there is a large increase in false positives (amount of

injury events incorrectly classified as healthy events), the false positive rate is not sensitive enough [49]. This leads us to discuss an alternative metric in the following.

PR Curve – Precision-recall curve

Precision-recall curve was proposed to complement the limitation of the AUC [49]. PR curve measures the trade-off between precision and recall of a classifier across all possible thresholds. A high PR score means the classifier can correctly identify injury events while maintaining a low number of false positives (healthy events misclassified as injury events). The advantage of PR curve is when classes are highly imbalanced, it can reflect the minority class more accurately.

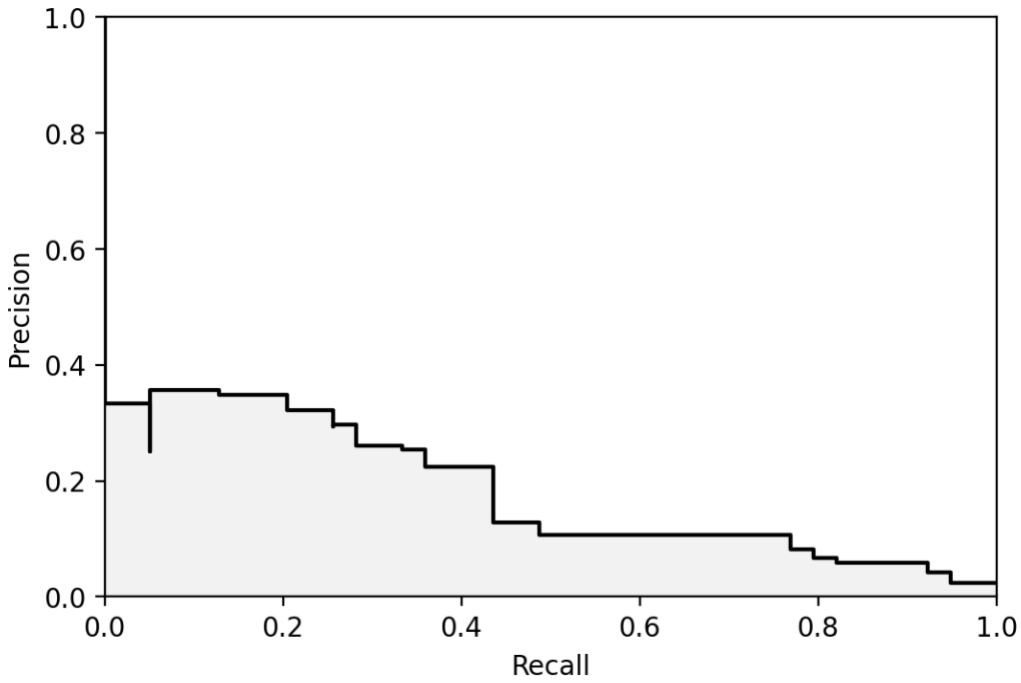


Figure 3: An illustration of precision-recall curve. y-axis represents precision and x-axis represents recall. The solid black line is the precision-recall curve, and the shaded area is the PR score.

Unlike ROC AUC, the aim of PR curve is to achieve the upper-right-hand corner [49]. As indicated in Figure 2, the y-axis indicates precision, it is calculated as the number of injury events that were correctly classified out of all predicted events. As opposed to false positive

rate in ROC AUC, precision can accurately capture the effect when there is an increase in healthy events [49]. Similar to the y-axis of ROC AUC, the x-axis measures how well the classifier can correctly identify injury events. A PR score of 1 indicates the classifier can identify injury events and healthy events perfectly. It is worth noting that high PR value indicates high precision and high recall, whereas a high ROC AUC indicates high recall and low false positive rate.

Precision

Precision measures the proportion of correctly identified injury events among all predicted events. It is calculated as the number of injury events that were correctly identified divided by all events that were predicted as injury (TP + FP).

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

In real world, precision and recall often have a trade-off relationship. That is, increase one often reduce another.

Sensitivity and specificity

Sensitivity, also referred as recall, is the measure of how well a classifier can identify injury events. The value ranges between 0 and 1. High sensitivity value indicates the model can correctly identify large proportion of injury events. It is determined by dividing the number of correctly identified injury events (TP) by the actual number of injury events (TP + FN) as shown below:

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

On the other hand, specificity is the proportion of healthy events that were correctly classified. It is calculated by dividing the number of correctly identified healthy events (TN) by the actual number of healthy events (TN + FP). The equation is as followed:

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

Interpretable Machine Learning

Interpretable machine learning methods help practitioners and clinicians to understand how a prediction is made by the machine. The degree to which a person can comprehend the logic behind a decision in machine learning is known as interpretability [50]. When people understand how a model is making decisions, they are more likely to trust and accept the decisions [51]. This is especially true in real world where decisions made by blind faith in domains such as healthcare and criminal justice may result in catastrophic consequences [51]. In addition, relying solely on the metrics (e.g. classification accuracy) is often inadequate in many real-world tasks [52]. As a result, many techniques were proposed to explain machine predictions.

The scope of interpretability can be divided into two categories: Global and local. Global interpretation focuses on understanding the entire model [53]. For example, it is easier for humans to understand logistic regression than deep neural network. The positive coefficients in logistic regression indicate that an increase in the feature is associated with an increase in the probability of the positive class and the magnitude of coefficients reflects the strength of the relationship. The simplicity of some machine learning models can offer global interpretability, these models are often called white box. However, this is usually hard to achieve in practice. For instance, when the number of input features increases, the interaction of the features can only be understood by holding other feature values constant. On the contrary,

local interpretation aims to comprehend why did the model come up with a decision for certain or a particular instance. This was demonstrated by SHAPley value [54], which was initially proposed to measure the contribution of individual players from a cooperative game theory. The same theory was later used to explain the contribution of individual predictors that lead to the final prediction of an instance [55]. Few studies utilized SHAPley values to explain the individual risk predictions of ML models in professional basketball and soccer players [18, 19].

The following sections discuss the difference between intrinsic and model-agnostic explanations and their respective methods. Table 1 summarised the advantages, drawbacks, and mechanisms of these techniques.

Table 1: The advantages and disadvantages of different interpretable machine learning methods.

Type of interpretation	Models	Mechanism	Scope of interpretability	Advantages	Disadvantages
Intrinsic	Logistic regression	Feature weight	Global	Probabilistic output; highly interpretable for linear features	Multicollinearity; Does not account for interaction between features; oversimplify complex relationship
	Decision trees	Cut-off threshold	Global + Local	Account for interaction between features	Not suitable for linear and complex relationship between features
	Naïve Bayes	Class probabilities	Global	Robust to irrelevant features; probabilistic output	Required all features to be independent of each other
	RuleFit [56]	Decision rules	Global	Account for interaction between features, Suitable for linear relationship	Less interpretable for high dimensional data; May generate additional unhelpful features
	Explainable Boosting Machine [57]	GAM (Generalized additive model)	Global + Local	Account for pairwise interaction; State-of-the-art performance	Require longer time to train
Model-agnostic	LIME (Local interpretable model-agnostic explanations) [51]	Local surrogate	Local	Compatible with diverse types of data (tabular, text, images); Simplify explanation for complex model	Precision of explanation rely on surrogate models; different samplings may yield different explanations for the same data; Assume linearity for local model
	SHAP (SHapley Additive exPlanations) [55]	Shapley values	Global + Local	Provide highly precise (faithful) explanation.	Computationally expensive for high dimensional data; Assume all features have an equal contribution; Not suitable for non-tabular data
	Counterfactual explanations [20, 58]	Contrastive instance	Local	Does not require to access data and model; Provide good explanation for data with fewer features.	Explanation may not be useful in high dimensional data.

	Anchors [59]	Decision rules (anchors)	Local	Account for precision accuracy; Highly interpretable for layperson	Required fine-tuning to obtain good explanation.
	GraphLIME [60]	Local surrogate	Local	Work well on graph data with intricate connections between features (molecular structures etc)	It is computationally expensive to generate explanation for complex graph neural network.
	ASV (Asymmetric Shapley Values) [61]	Shapley values	Global + Local	Improve explanation by considering unequal causal relationship between features; provide incremental explanations for time-series data	Computationally expensive for high-dimensional data; Required expertise to create a causal graph;
	SHAPley Flow [62]	Shapley values	Global + Local	Provide a comprehensive view of a model by considering direct and indirect causal relationship.	Require knowledge to generate explanations.
	MMD-critic [63]	Prototypes and criticisms	Global	Effectively facilitate human on understanding model with complex data distribution.	The optimal number of prototypes and criticism for explanation are not known.
	Saliency maps [64]	Gradients	Local	Easy to implement; Can be generated in almost real-time	Hard to distinguish importance features from noise in saliency map (Saturation problem)
Attribution method	Integrated Gradients (IG) [65]	Gradients	Global + Local	Worked on complex deep neural network architectures; Easy to implement;	Only worked on neural networks and differentiable models; Unequal attributions may be given to features with same contribution; May suffer from gradient shattering problem.

Intrinsic models

Intrinsic models are models that are explainable by itself. They are referred as white-box models, sometimes known as glassbox models. This review covers some common intrinsic models in sports injury prediction tasks.

Logistic regression

Logistic regression is the most used predictive model in sports injury prediction [39]. It is easy to understand and does not require complex hyperparameter tuning. Many studies use logistic regression to explain the significance of input variables on the outcome [66]. It works by extending linear regression with a logistic function, so that the output falls between the value of 0 and 1. Logistic regression is highly interpretable for a number of reasons. Firstly, the coefficient of a trained logistic regression, also known as weight, can provide global interpretation. Specifically, the exponential of its weight can be conveniently interpreted as an odds ratio [67]. For example, suppose the coefficient of age is 0.6, the odds ratio is 1.82 ($=e^{0.6}$). This indicates an increase in the age of an athlete increases the odds of injury by a factor of 1.82, given all other variables are held constant. In addition, the output of logistic regression yields probability. This is useful as it is important to know that an athlete with a chance of 0.95 sustaining an injury is riskier than a person with a chance of 0.6, even though both were classified as injury (≥ 0.5).

Logistic regression falls short when there are multiple correlated variables [66], which makes interpreting the effect of individual variables difficult [66]. Furthermore, logistic regression does not consider the interactions between features, these interactions need to be added manually to take effect [40]. Finally, logistic regression does not perform well when the relationship between variables and outcome is non-linear [40].

Decision trees

Decision tree is a ML model favoured by many researchers in sport science domains [39, 68]. Besides it can provide interpretable results, it also has the ability to capture the interaction between variables [40]. A decision tree consists of nodes and branches. A node represents a condition while a branch represents the outcome of the condition. The tree is built by recursively splitting the data into subsets based on an optimal cut-off value. In classification, Gini index is used as a criterion for splitting nodes and creating branches (optimal cut-off value). A Gini index of 0 indicates the leaf node is complete pure, where only one class is present. Gini index is minimized during training until the tree reaches its maximum tree depth or the minimum number of instances in a leaf node. When using decision tree, cross validation is often required for hyperparameter tuning to prevent overfitting.

In contrast to logistic regression, decision tree does not perform well when the relationship between independent variables and dependent variable is linear. Additionally, it is common for a minor alteration in the data to cause a significant variation in the sequence of splits, results in a completely different tree structure. This makes decision tree inherently unstable when it comes to explanation [40, 66]. A common solution to this problem is to use bagging to average predictions across multiple trees [66].

Naïve bayes

Naïve bayes is a probabilistic supervised ML algorithm that estimates the probabilities of each class using Bayes' theorem. Despite it assumes the variables are independent of each other, it performs reasonably well in many real-world applications [69] and often outperforms many complex modelling methods [66]. The model can be trained in linear time [70]. The interpretability of Naïve Bayes stems from its ability to provide explanations on the impact of

individual variables [40]. To understand how this works, suppose an athlete with a height of 175 cm and age of 23 years old. The maximum conditional probability of a class (injured / uninjured) can be estimated in the equation as followed:

$$y = \text{argmax}_y \left[P(y) * \prod_{i=1}^{n} P(x_i \mid y) \right]$$

Where y is the maximum probability of all classes (injury / healthy), n is the number of variables (height, age), $P(y)$ is the probability for an athlete to be classified as y class (e.g. injury class) without considering its variables, this is usually derived from the training set. $P(x_i \mid y)$ is the probability of x_i given y. For instance, the probability value when height is 175 cm, given an athlete is injured.

Now, suppose there is a 50% chance for any given athlete to get injured (derived from the training dataset). The likelihood is 0.6 when the height is 175 cm given an athlete is injured $P(x_{height} = 175 \mid injured)$ and the likelihood is 0.4 when the age is 23 years old given an athlete is injured $P(x_{age} = 23 \mid injured)$ (Calculated through discretization or probability density). The estimated conditional probability for the given athlete to get injured is calculated as equation:

$$\begin{aligned} P(injured|X) &= P(X|injured) \times P(injured) \\ &= 0.5 \times 0.6 \times 0.4 \\ &= 0.12 \end{aligned}$$

The calculation is repeated similarly for healthy class $P(healthy|X)$. The predicted class is determined by selecting the one with the highest conditional probability after comparing the probabilities of all classes.

The biggest limitation of naïve bayes classifier is it assumes all features are independent of each other. This is known as conditional independences, and it is unlikely to find this type of data in real world [71].

Other interpretable models

What if there is a model that has the advantage of linear model, at the same time consider the interactions between variables. RuleFit is the answer to this type of model [56]. RuleFit works by first training a decision tree model with input data. Later, the decision tree is decomposed into multiple decision rules. These decision rules are important as they served as additional features to the original dataset to form a dataset that contains interaction information between features. The final step is to train a sparse linear model (e.g. LASSO) with the dataset. The result is an interpretable linear model that has the effect of original variables, as well as interaction derived from decision rules. Since the final product of RuleFit is a linear model, many of its drawbacks are the same as linear model. Additionally, it may create many unhelpful features which makes explanation more difficult [40].

In interpretable machine learning, there is often a trade-off between model interpretability and performance [72]. Powerful models like neural networks and XGBoost usually perform well by sacrificing interpretability. Recently years, some researchers have attempted to develop high performing models that do not compromise interpretability [57]. One example is EBM (Explainable Boosting Machine) [57]. It is an improved generalized additive model (GAM) which takes account of pairwise interactions. It works by training multiple tree models with individual features using a low learning rate. Each feature is compared against one another in a round-robin fashion to find the best function. This process is repeated for many iterations until individual features can be summarised in a graph. The outcome of the model is the

interpretation of the contribution of each feature in predicting the final outcome. However, the downside of this method is it requires significant computational resources and is time-consuming to train multiple trees across many iterations.

Local model-agnostic methods

Model-agnostic methods are interpretability techniques that can be applied to explain any ML model.

LIME (Local interpretable model-agnostic explanations)

One of the pioneer studies in explaining any opaque ML model is LIME [51]. It is still widely used today due to its human-friendly explanation and compatibility with various forms of data, including text, tabular and images. While it is not always feasible to completely understand a complex model [51], LIME generates explanations that are locally faithful. Locally faithful explanations are explanations that are accurate in local region. The advantage of this type of explanation is high interpretability. To comprehend how a black-box model comes up with a decision for a particular instance, one does not need to understand how the model makes predictions for all kinds of data. To achieve this objective, LIME constructs a dataset centred around a selected instance (input data) and then trains a sparse linear model (white-box model) using the generated dataset. The resulting output is an interpretable model that can explain the instance locally, known as local surrogate model. This can be viewed as an optimization problem as followed:

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}}\, \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

Where x is an input data for explanation, f is any given trained black-box model, g is a subset of sparse linear models and its variants, they are referred as interpretable models. π_x defines the size of local neighbourhood for any given x. The higher the value of π_x, the more dissimilar data would be generated. Lastly, $\Omega(g)$ defines the complexity of selected linear model g. It is

used to balance the trade-off between faithfulness and interpretability. The first term aims to minimize the discrepancy between the output prediction from the complex model and the interpretable model in a defined local neighbourhood, this is referred to as the locality-aware loss [51]. The second term provides a regularization effect for the interpretable model *g*. For example, this can be the regularization parameter of Lasso model. Higher value can increase the number of zero-weighted input features and produce a simplified model.

Since LIME uses simple linear models (e.g. decision trees and linear regression) as local surrogate models, the explanations are highly interpretable explanations for humans. LIME is also model agnostic, which can be utilized to provide explanations for any black-box model. However, there are numerous trade-offs that come with the flexibility and robustness it brings. Firstly, LIME is restricted to provide only local faithful explanations, it cannot provide global understanding for the underlying complex model. Secondly, LIME is a perturbation-based explainer, the explanation relied on the dataset it generates. The problem arises when the generated data used to extrapolate the black-box model falls within an area where the model was trained with limited or insufficient data, leading to deceptive explanations [73]. LIME uses an exponential kernel to calculate the distance between two data instances. The kernel requires a user-defined width to calculate the scope of data instances it generates to influence the surrogate model. However, it is unclear what is the optimal width to produce good explanations [40]. Moreover, whether the distance measure should be treated equally across all input features required further studies [40].

SHAP (SHAPley additive explanations)

Similar to LIME, SHAP is another model-agnostic explainer that is capable of explaining the predictions of any black-box model [55]. Unlike LIME, SHAP can also be used to provide

global interpretation by aggregating the contribution of individual predictions. Additionally, SHAP does not require training additional surrogate models. The idea of SHAP originates from the SHAPley value, which was proposed back in 1951 [54] to calculate the contribution of individual players in a cooperative game theory so that the total payout can be distributed equally across all members. With regards to machine learning, the input features can be viewed as players and the objective is to determine the contribution of each feature to the prediction.

Understanding the calculation of a SHAPley value is necessary to comprehend how SHAP explanation works. Shapley value determines the marginal contribution of a predictor. The calculation of SHAPley value is illustrated in the context of injury prediction as followed:

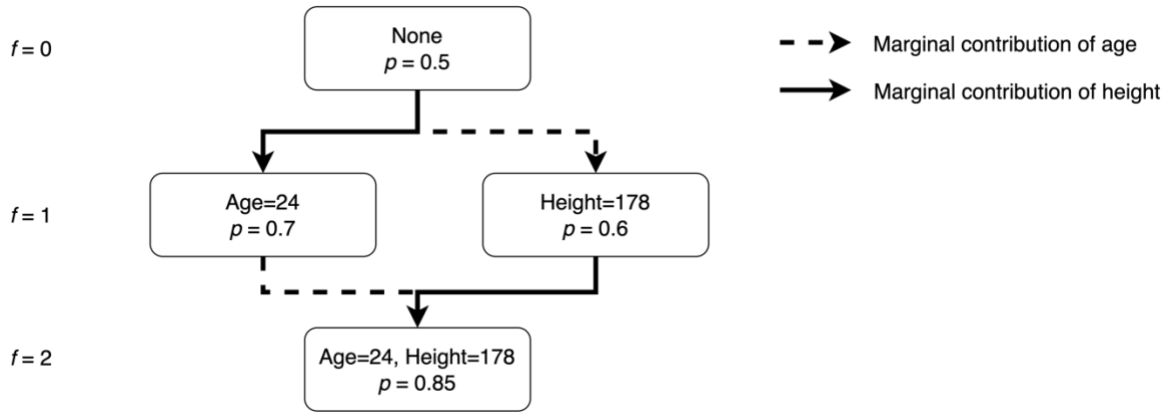


Figure 4: Marginal contribution of age and height, *f* represents the number of features.

Based on Figure 4, assumed age and height are potential risk factors of limb injury. Given a player aged 24 years with a height of 178 cm. A black-box model predicted the player with a 0.85 chance of sustaining an injury. The steps to obtain the marginal contribution of age are described in the following equation:

$$\begin{aligned}\text{SHAP}_{\text{Age}}\ (x_0) &= w_1 \times MC_{\text{Age},\{\text{Age}\}}(x_0) + w_2 \times MC_{\text{Age},\{\text{Age, Height }\}}(x_0)\\ &= \frac{1}{2} \times (0.7 - 0.5) + \frac{1}{2} \times (0.85 - 0.6)\\ &= \frac{1}{2} \times (0.2) + \frac{1}{2} \times (0.25)\\ &= 0.1 + 0.125\\ &= 0.225\end{aligned}$$

Where x_0 is the given player, $MC_{A_{\text{ge}},\{A_{ge}\}}$ is the marginal contribution of age alone, $MC_{\text{Age},\{A_{ge},\text{ Height}\}}$ is the marginal contribution of age when height is held constant. w_1 and w_2 are the weights of the age when number of predictors are *f*=1 and *f*=2 respectively. The calculated value (=0.225) quantifies the marginal contribution of age (24 years) to the final prediction. Likewise, the steps were repeated to calculate the marginal contribution of height in the following equation:

$$\begin{aligned}\text{SHAP}_{\text{Height}}\ (x_0) &= w_1 \times MC_{\text{Height},\{\text{Height}\}}(x_0) + w_2 \times MC_{\text{Height},\{A_{ge},\text{ Height}\}}(x_0)\\ &= \frac{1}{2} \times (0.6 - 0.5) + \frac{1}{2} \times (0.85 - 0.7)\\ &= \frac{1}{2} \times (0.1) + \frac{1}{2} \times (0.15)\\ &= 0.05 + 0.075\\ &= 0.125\end{aligned}$$

Once the shapley values of all predictors were obtained, summing the SHAPley values of age and height yields 0.35. This indicates how both predictors contribute to the final prediction of 0.85. Without these two predictor variables, the initial prediction was 0.5, which is known as base value.

There are a few advantages that make SHAP stands out. SHAP satisfied three properties that are important in interpreting the predictions of ML models [55]. Firstly, local accuracy ensures that the final prediction of the input data is equal to the sum of all feature contributions. Another property is missingness where missing input from original features has no impact on the outcome. Finally, consistency makes sure the explanation of an instance remains unchanged when the training data is increased with similar instances. The main drawback of SHAP is it is computationally expensive to calculate SHAPley values in high dimensional data. This is due to the number of calculations required will increase exponentially when the number of features increases. The author proposed TreeSHAP kernel to solve this problem for tree ensemble

models. However, it losses its model-agnostic property and may suffer from unintuitive feature attributions [40].

Counterfactual explanations

The counterfactual (CF) explanation method involves identifying the smallest possible modification to the input data that would cause an ML model to make a different or opposite prediction. [58]. Given an athlete who was classified as injured by the machine, understanding the underlying risk factors is crucial for clinicians. However, such an explanation cannot provide insights into what actions can be taken to avoid injury. Counterfactual explanations can fulfil this purpose by showing clinicians what can be done to make the athlete re-classified as healthy by the ML model. This is especially important in professional sports domains where coaches and physicians are interested in creating effective training programs and implementing targeted interventions [16].

There are however challenges in seeking good counterfactual explanations. One of these challenges is prolixity, where a series of counterfactual explanations may be produced through random perturbations or search, which are not useful for interpretation [74]. For instance, increase or decrease the feature values in predetermined way to seek counterfactual explanation. One common approach to this problem is to identify the minimal changes (e.g., the nearest unlike neighbour) of the input features required in order to obtain a different prediction from the model [74]. Wachter et al. [58] proposed the following loss function to optimise this problem through gradient descent:

$$L(x, x', y', \lambda) = \lambda \cdot \left(\hat{f}(x') - y'\right)^2 + d(x, x')$$

Where x is the input instance, x' is the counterfactual explanation instance, and y' is the class of counterfactual instance (e.g. healthy class). The parameter λ controls the proximity between

counterfactual instance and input instance. A lower value of λ indicates a counterfactual with feature values that are similar to the input instance. Many counterfactual explanation methods followed this approach to minimise the loss with their proposed objectives [20, 75].

One major challenge of this explanation approach is sparsity, where a large number of features may be altered to create counterfactual explanations [20]. Such explanations can be unintuitive and hard to interpret for humans. Ideally, a sparse counterfactual explanation requires fewer features to obtain a different prediction. A recent study pointed out that a good counterfactual should not have changes in more than 2 features due to the constraint human had on category learning [16]. However, the same study found that this type of counterfactual is rare as it only accounts for less than 1 percent of the total explanations.

CHAPTER 3: Hamstring strain injury risk factors in Australian Football change over the course of the season

NOTE: This chapter has been submitted to MSSE (Medicine & Science in Sports & Exercise) as a journal paper and is currently under review.

Sim A, Timmins RG, Ruddy JD, Shen H, Liao K, Maniar N, Hickey J, Williams MD, Opar DA. Hamstring strain injury risk factors in Australian Football change over the course of the season. Under review.

Introduction

The occurrence of hamstring strain injuries (HSIs) are widespread across several sports [23] including Australian Football [76] and significant work has been conducted to determine the factors that contribute to an elevated risk of future injury. [11]. These risk factors are often categorized as either modifiable or non-modifiable. Older age and a history of HSI are the two most frequently reported non-modifiable risk factors [11]. Modifiable risk factors, which can be addressed through intervention, are extensive, but biceps femoris long head (BFlh) fascicle length [13] and eccentric hamstring strength [14] are current prominent variables.

Risk factors in most prospective cohort studies are assessed at a single time-point, typically at the start of a season (e.g., in the pre-season) [16]). This approach has limitations, as any changes in the measured variables leading up to injury, which may be many months after the pre-season assessment, are not accounted for. Limited studies have been conducted to determine if more frequent risk factor assessments can improve the ability to predict future HSI risk. It was recently reported that more frequent assessments of eccentric knee flexor strength and biceps femoris long head (BFlh) architecture did not improve the ability to predict new HSIs in Australian Football [17]. However, that study did not consider the possibility that risk factors may vary depending on the time of season, nor did it examine if changes in possible risk factors across time (e.g., an increase in eccentric strength from the start to the end of pre-season) at an individual level altered the ability to predict HSI.

Therefore, the primary objective of this study was to identify which factors were most predictive of the risk of HSI during pre-season, early in-season and late in-season in professional Australian Football. The secondary objective was to determine if the magnitude of change in possible risk factors across pre-season was predictive of future in-season HSIs.

Methods

The methods used in this study pertaining to the study design, participants, and data collection have been previously documented [17] but are included in detail here for the ease of the reader.

Study design and participants

Approval for this study was obtained from the ACU Human Research Ethics Committee (approval number: 2017-208H). The study was carried out across two Australian Football League seasons (November 2017 to August 2018 and November 2018 to August 2019, including pre-season but not including finals) across six teams. Written informed consent was acquired from all players before their participation in the study.

Prior to pre-season, the medical staff of the team were responsible for providing details of individual players' history of HSI in the past 12 months and if they had ever sustained an anterior cruciate ligament (ACL) injury. Eccentric knee flexor strength and BFlh fascicle length were assessed at the start of pre-season (November/December), end of pre-season (February/March), and middle of the competitive season (May/June), respectively. Due to scheduling constraints, the actual dates for assessments were not identical across the different teams. The medical staffs were required to complete a standardized injury report form for any player who experienced a HSI during the period of study.

Eccentric knee flexor strength

The evaluation of eccentric knee flexor strength was conducted during the Nordic hamstring exercise similar to previous studies using an instrument device (NordBord, VALD, Queensland, Australia) [13, 14, 76, 77]. The ankle hooks with uniaxial load cells were used to secure the players' ankles immediately once they knelt on the cushioned board. Every player who

underwent this assessment was experienced in performing the Nordic hamstring exercise. The individuals were instructed to gradually move their body forward, using their knee flexor muscles to manage the descent. All players maintained their trunk and hips in a neutral position while holding their hands across the chest throughout the exercise. Players performed a single set of 1-3 maximal repetitions as determined by each team's practices following their warm-up routine. The highest peak force produced by each leg throughout the test was recorded as eccentric knee flexor strength.

Biceps femoris long head architecture

The assessment of BFlh architecture has been reported previously [27, 78-80]. Ultrasound imaging was used to obtain measurements of muscle thickness, pennation angle, and fascicle length of the BFlh. The images were taken along the muscle belly's longitudinal axis using a two-dimensional, B-mode ultrasound (frequency, 12 MHz; depth, 8cm; field of view, 14 x 47mm) (GE Healthcare Vivid-i, Wauwatosa, U.S.A). The scanning site for the BFlh was identified as the midpoint of the line between the sitting bone and knee joint. The architecture assessments were conducted on players lying on a massage plinth after at least 5 minutes of inactivity. The assessor (RGT) adjusted the orientation of the probe accordingly. The reliability of the assessor has been previously established with an intraclass correlation > 0.90 reported for BFlh fascicle length.

Offline analysis was undertaken after the images were collected (MicroDicom, Version 0.7.8, Bulgaria). Muscle thickness was determined by the distance between the superficial and intermediate aponeuroses of the BFlh. Pennation angle was determined by the angle between the intermediate aponeurosis and a fascicle of interest. The angles of superficial and intermediate aponeurosis were defined as the angle between the line marked as the aponeurosis

and an intersecting horizontal reference line across the capture image [81]. Due to part of the fascicle not being visible in the ultrasound probe's field of view, the following equation from Blazevich and colleagues was used for estimation [81]:

$$FL = sin(AA + 90°) \times MT/sin(180° - (AA + 180° - PA))$$

where FL=fascicle length, AA=aponeurosis angle, MT=muscle thickness and PA=pennation angle. Fascicle length was reported in absolute terms (cm) and relative to muscle thickness from a single image. The same assessor (RGT) collected and analysed all scans. The assessor has evidenced reliability in determining measures of BFlh muscle architecture at rest with ICCs >0.95 and %TE <5.0% across the measurement of all architectural variables.

Prospective hamstring strain injury reporting

An HSI was diagnosed if a player experienced posterior thigh pain that prevented them from carrying out subsequent exercise and was verified through a physical examination by the team's doctor [82, 83]. A standard injury report form was filled out by the team's medical staff for each HSI, which gathered information on the affected limb, injured muscle, activity type performed when the injury occurred, and the duration of time required for the player to fully participate in training and competition.

Statistical Analysis

The Python 3.9.2 programming language (Python Software Foundation, https://www.python.org/) and the following packages were used to conduct statistical analyses: scikit-learn, statsmodel, panda, numpy, matplotlib and seaborn.

General Modelling Approach

The general modeling approach applied to this study can be found in Figure 5a.

Data pre-processing

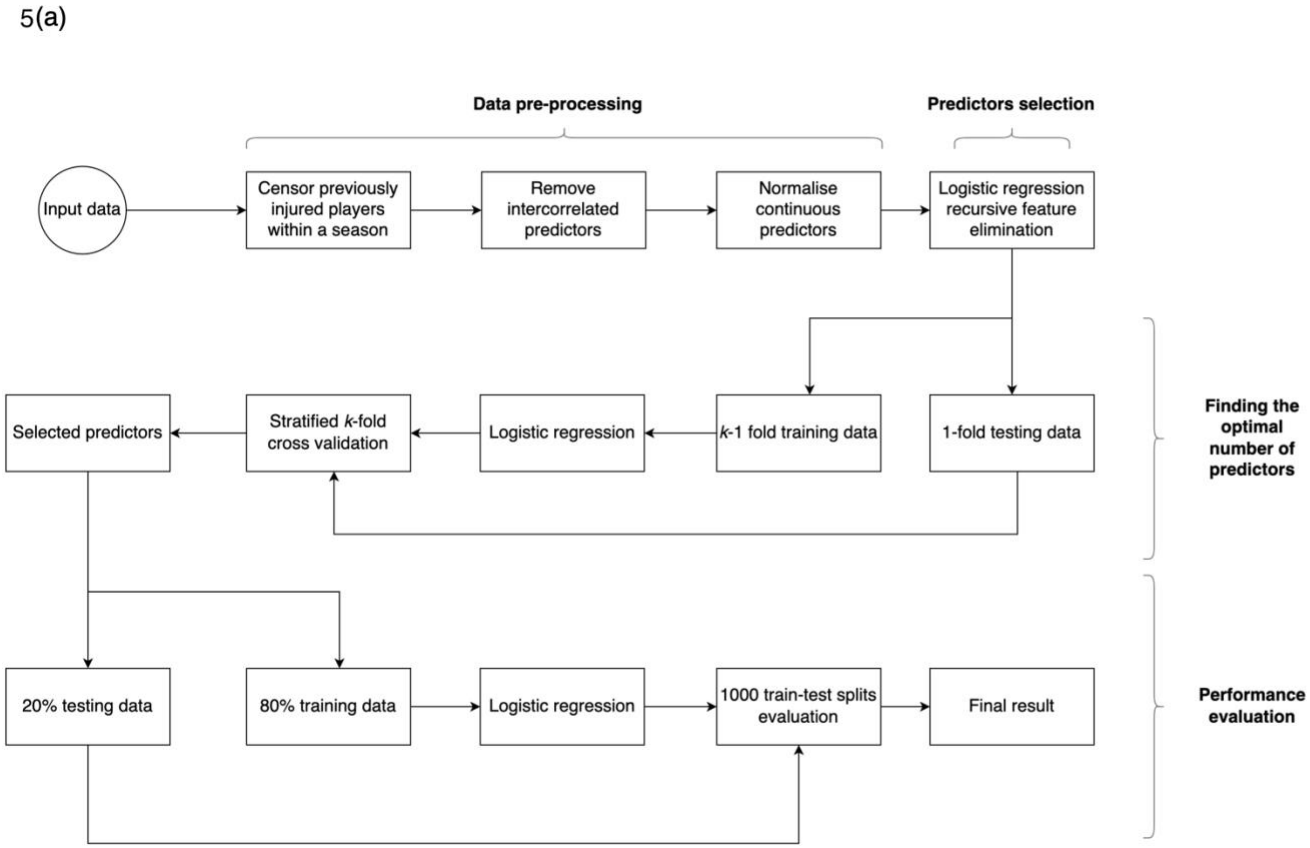
For all analyses, an observation was removed if it consisted of at least one missing value. Additionally, players who sustained an HSI in previous time points within a season were censored from building models to predict HSIs that occurred in later timepoints. For example, a player who sustained an HSI in pre-season was excluded from training models to predict HSIs that occurred in early in-season and later in-season. Likewise, players who sustained an HSI in early in-season were excluded from training models to predict HSIs occurring in late in-season.

Correlation analysis was conducted on input predictor variables to identify redundant predictors. When the Pearson's correlation coefficient between two predictors exceeded the threshold of 0.8, the predictor with the higher mean correlation among other predictor variables was eliminated.

Following this, the remaining input predictor variables were normalized [43] into the range of 0 and 1, using the following equation:

$$x_{\text{norm}} = \frac{x - min(x)}{max(x) - min(x)}$$

where x is the value to scale, $min(x)$ is the smallest value of the predictor, and $max(x)$ is the largest value of the predictor.



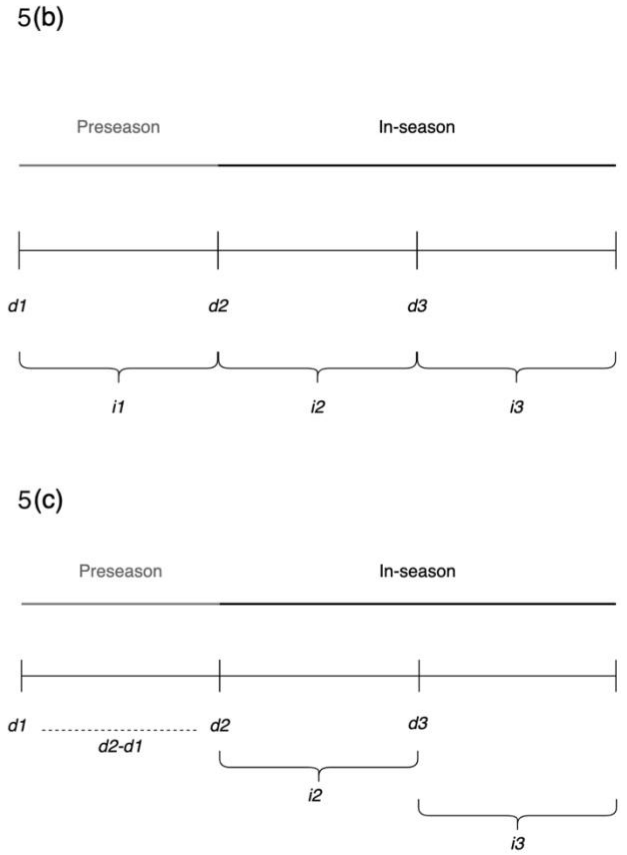


Figure 5a: Adopted workflow process to identify important risk factors and build optimal models for performance evaluation.

Figure 5b: The modelling approach for Analysis 1. *d* represents data assessment at different time points where *d1* are data assessed at the start of pre-season; *d2* are data assessed at the end of pre-season; *d3* are data assessed in the middle of in-season. *i* represents hamstring strain injuries (HSIs) that occurred within individual assessment time frames where *i1* are prospective HSIs that occurred in pre-season; *i2* are prospective HSIs that occurred in early in-season; *i3* are prospective HSIs that occurred in late in-season.

Figure 5c. The modelling approach for the Analysis 2. *d* represents data assessment in different time points where *d1* are data assessed at the start of pre-season; *d2* are data assessed at the end of pre-season, *d2-d1* are magnitude of change of data in preseason. *i* represents hamstring strain injuries (HSIs) that occurred within individual assessment time frames where *i2* are prospective HSIs that occurred in early in-season; *i3* are prospective HSIs that occurred in late in-season; *i2+i3* are prospective HSIs that occurred throughout in-season.

Predictor selection

The aim of the predictor selection process in the current study was to eliminate redundant predictors and identify which subset of risk factors achieved the highest predictive performance across the different time points. A wrapper feature selection method, specifically recursive feature elimination, was used to search through different subsets of risk factors associated with HSI [84]. Recursive feature elimination, which is robust to overfitting [85], was conducted in this study by fitting a logistic regression with all input predictor variables and recursively eliminating predictors that were less important based on the coefficients. Once the predictor with the lowest coefficient was removed, the model was fitted with the remaining predictors to repeat the process. This process was repeated until there was only one remaining predictor, after which the importance of individual predictors was ranked. Preliminary analyses using this dataset showed that models built using recursive feature elimination outperformed models built using all predictors. Recursive feature elimination, however, does not identify the optimal number of predictors.

Finding the optimal number of predictors

Stratified k-fold cross validation was utilised to determine the optimal number of predictors. In this study, $k = 5$ was applied to divide data into 5 stratified folds. For each split, 1 fold of the data was assigned for testing, while and the rest of the data (k-1 folds) were used for training. The number of selected predictors resulting in the highest AUC averaged across 5 folds was chosen as the optimal number of predictors.

Performance evaluation

Once the optimal number of risk factors was determined, the final step was to evaluate the performance of logistic regression with selected risk factors. In practice, data is usually split

into training and testing sets with a ratio of 70%-80% and 20-30%, respectively [16]. Any split within this threshold has been shown to have an accurate estimation of the model's performance [86]. A 20%/80% train-test split was used in this study. Stratified cross validation was utilised to preserve the percentage of injured and uninjured athletes for all iterations. Since the given dataset was relatively small (<455 observations), 1000 iterations of evaluation were performed. The metric used to evaluate predictive performance was area under the curve (AUC) [87]. AUC measures the ability of the models to correctly predict prospectively injured and uninjured players. A value of 0.5 for the AUC indicates the predictive performance is no better than guessing, whereas a value of 1.0 for the AUC indicates perfect prediction.

Analysis 1

The aim of Analysis 1 was to determine which risk factors best predicted HSIs at different time points throughout the season.

The general modelling approach was applied to Analysis 1. The subset of data utilised for Analysis 1 has been illustrated in Figure 5b, where *d1* are data assessed at the start of pre-season, *d2* are data assessed at the end of pre-season, and *d3* are data assessed in the middle of in-season. *i1* is the window following *d1* during which prospective HSIs could have occurred throughout pre-season, *i2* is the window following *d2* during which prospective HSIs could have occurred early in-season, and *i3* is the window following *d3* during which prospective HSIs could have occurred during late in-season.

Analysis 1 utilised all non-modifiable risk factors assessed at the start of pre-season and modifiable risk factors assessed at multiple time points (*d1* or start of pre-season, *d2* or end of pre-season, and *d3* or middle of in-season) as predictor variables. Prospective HSIs that

occurred between individual assessment time frames (*i1* or between the start and end of pre-season, *i2* or between the end of pre-season and the middle of in-season, and *i3* or between the middle of in-season and the end of the in-season before the start of finals) were the target of the prediction models. (Refer to Table 2 for types of input predictor variables and target variables included in each of the individual models).

Analysis 2

Analysis 2 aimed to determine whether the magnitude of change in data between the start and end of pre-season, as well as more frequent assessment during pre-season, improved the ability to predict in-season HSIs, beyond the data collected at the start and end of pre-season alone.

The general modelling approach was applied to Analysis 2. The subset of data utilised for Analysis 2 has been illustrated in Figure 5c, where *d1* are data assessed at the start of pre-season, *d2* are data assessed at the end of pre-season, *d2-d1* is the magnitude of change in the risk factors across pre-season. *i2* is the window during which prospective HSIs could have occurred early in-season and *i3* is the window during which prospective HSIs could have occurred during late in-season.

Analysis 2 utilised all non-modifiable risk factors assessed at the start of pre-season and modifiable risk factors assessed at the start and end of pre-season as predictor variables. Prospective HSIs that occurred during the in-season periods (*i2* and *i3*) were the target of the prediction models. Additionally, the magnitude of change in modifiable risk factors was determined as the absolute difference between values captured at the end of pre-season and values captured at the start of pre-season. (Refer to Table 2 for types of input predictor variables and target variables included in individual modelling approaches).

Table 2: Types of predictor variables and target variables included in individual models for Analysis 1 and Analysis 2.

Model	Input predictor variables					Target variables		
	Non-modifiable risk factors	Modifiable risk factors				HSIs		
	d1	d1	d2	d3	d2-d1	i1	i2	i3
Analysis 1								
d1->i1	✓	✓				✓		
d2->i2	✓		✓				✓	
d3->i3	✓			✓				✓
Analysis 2 **HSI occurred in early in-season (i2)**								
d1	✓	✓					✓	
d2	✓		✓				✓	
d1&d2	✓	✓	✓				✓	
d2-d1	✓				✓		✓	
d1&d2&(d2-d1)	✓	✓	✓		✓		✓	
HSI occurred in late in-season (i3)								

d1	✓	✓						✓
d2	✓		✓					✓
d1&d2	✓	✓	✓					✓
d2-d1	✓				✓			✓
d1&d2&(d2-d1)	✓	✓	✓		✓			✓

d1: data assessed at the start of pre-season, d2: data assessed at the end of pre-season, d3: data assessed in the middle of in-season, d1&d2: data assessed at start and end of pre-season, d2-d1: magnitude of change in data between start and end of pre-season; i1: HSIs occurred in pre-season, i2: HSIs occurred in early in-season; i3: HSIs occurred in late in-season.

RESULTS

During the 2018 and 2019 AFL seasons, a total of 311 male Australian Football players (aged 23.7 ± 3.8 years, height 188.1 ± 7.6 cm, 86.5 ± 8.8 kg) were evaluated at least once, resulting in 455 player seasons. Among these player seasons, 74 (16.3%) resulted in an HSI while the remaining 381 (83.7%) did not.

After the removal of missing values for Analysis 1, the total number of injured and uninjured player seasons during *i1* was 14 and 339 respectively (*d1->i1*; Table 2). For *i2*, the total number of injured and uninjured player seasons with complete datasets assessed at *d2* was 24 and 259 respectively (*d2->i2*; Table 2). For *i3*, the total number of injured and uninjured player seasons (with complete datasets assessed at *d3*) was 11 and 225 respectively (*d3->i3*; Table 2).

For Analysis 2, the total number of injured and uninjured player seasons with complete datasets during early in-season (i.e. *i2*) was 23 and 219 respectively (*i2*; Table 3). For late in-season (i.e. *i3*), the total number of injured and uninjured player seasons with complete datasets was 9 and 210 respectively.

Analysis 1

The performance of the individual models in Analysis 1 can be found in Figure 6. Data that were assessed at the end of pre-season and used to predict HSIs that occurred early in-season displayed the best predictive performance (median AUC = 0.86, interquartile range (IQR) = 16; Table 2) (*d2->i2*, Figure 6). The prediction of pre-season HSIs utilising data assessed at the start of pre-season (*d1->i1*; Figure 6) resulted in a median AUC of 0.83 and an interquartile range (IQR) of 0.16. In contrast, data assessed at the middle of the in-season period and used

to predict HSIs that occurred late in-season (*d3->i3*, Figure 6) resulted in the poorest predictive performance (median AUC = 0.46, interquartile range (IQR) = 0.25; Table 2).

Pre-season HSI

Players with history of HSI are more likely to sustain an HSI in pre-season (Figure 7a-c, $p < 0.01$). Shorter players displayed a higher risk of sustaining HSI in pre-season (Figure 7a). A significantly increased risk of pre-season HSI was observed in older athletes (Figure 7b, $p < 0.05$; Supplemental Material 1) and players who had thicker BFlh muscles were more susceptible to HSI in pre-season (Figure 7c).

Early in-season HSI

Players with a greater BFlh pennation angle and shorter fascicle length were at significantly increased risk of sustaining HSI during the early in-season period (Figure 7d, 3e; $p < 0.05$; Supplemental Material 1).

Late in-season HSI

Although height, age, history of ACL injury, BFlh pennation angle, fascicle length, relative eccentric knee flexor strength, as well as relative eccentric knee flexor strength imbalance were selected as predictive predictors (Figure 7f-k), the overall predictive performance of AUC was below 0.5 (median AUC = 0.46, interquartile range (IQR) = 0.25; Table 2).

Table 3: The results of Analysis 1. The performance of models built with selected predictors assessed and evaluated at start of pre-season and hamstring strain injuries (HSIs) that occurred in pre-season (*d1->i1*), end of pre-season and HSIs that occurred in early in-season (*d2->i2*), and middle of in-season and HSIs that occurred in late in-season (*d3->i3*). The descriptive summary is the outcome of 1000 iterations of train-test splits.

Model	Risk factors*	Frequency			AUC						
		HSI	Non-HSI	Total	Interquartile range	Standard Deviation	Minimum	Lower quartile	Median	Upper quartile	Maximum
d1->i1	prior HSI, height, age, muscle thickness	14	339	353	0.16	0.12	0.40	0.73	0.83	0.89	0.99
d2->i2	pennation angle, fascicle length	24	259	283	0.16	0.11	0.37	0.77	0.86	0.93	1.00
d3->i3	prior ACL, height, age, pennation angle, fascicle length, relative eccentric knee flexor force, eccentric knee flexor force imbalance	11	225	236	0.25	0.17	0.02	0.33	0.46	0.58	0.91

Performance is measured as area under the curve (AUC).

*Risk factors were selected by recursive feature elimination and 5-fold cross validation.

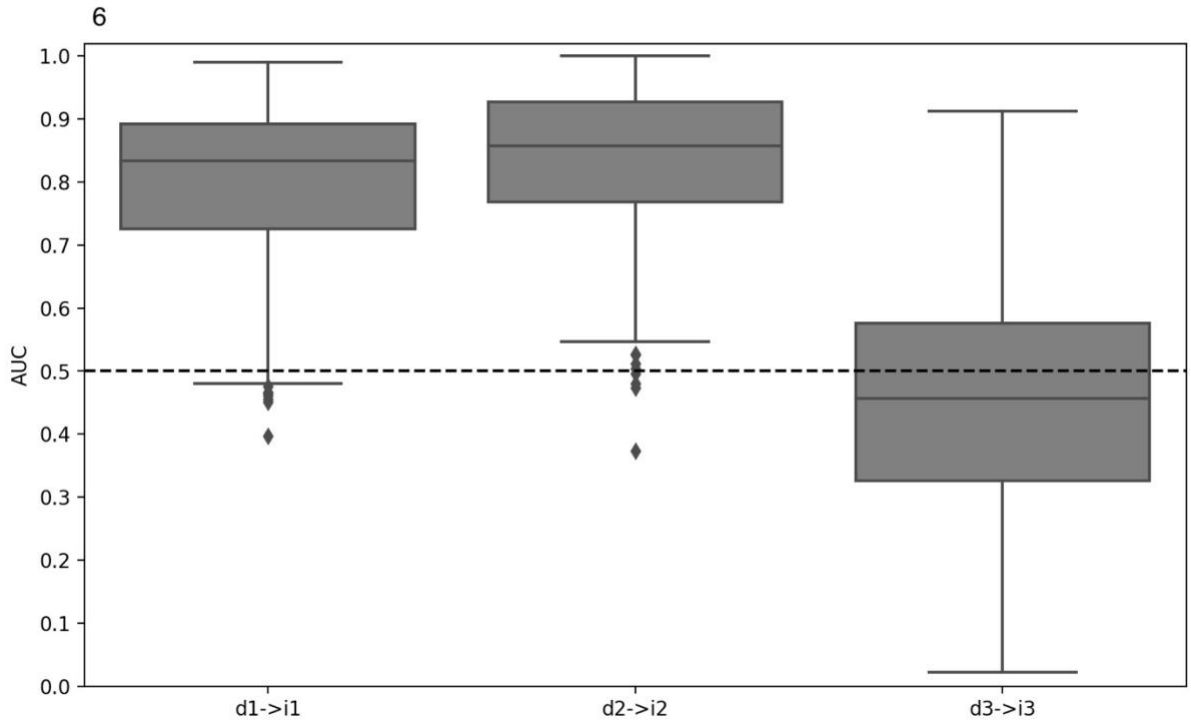


Figure 6: The results of Analysis 1. The performance of models built with selected predictors assessed and evaluated at the start of pre-season and hamstring strain injuries (HSIs) that occurred in pre-season (*d1->i1*), end of preseason and HSIs that occurred in early in-season (*d2->i2*), in the middle of preseason and HSIs that occurred in late in-season (*d3->i3*).

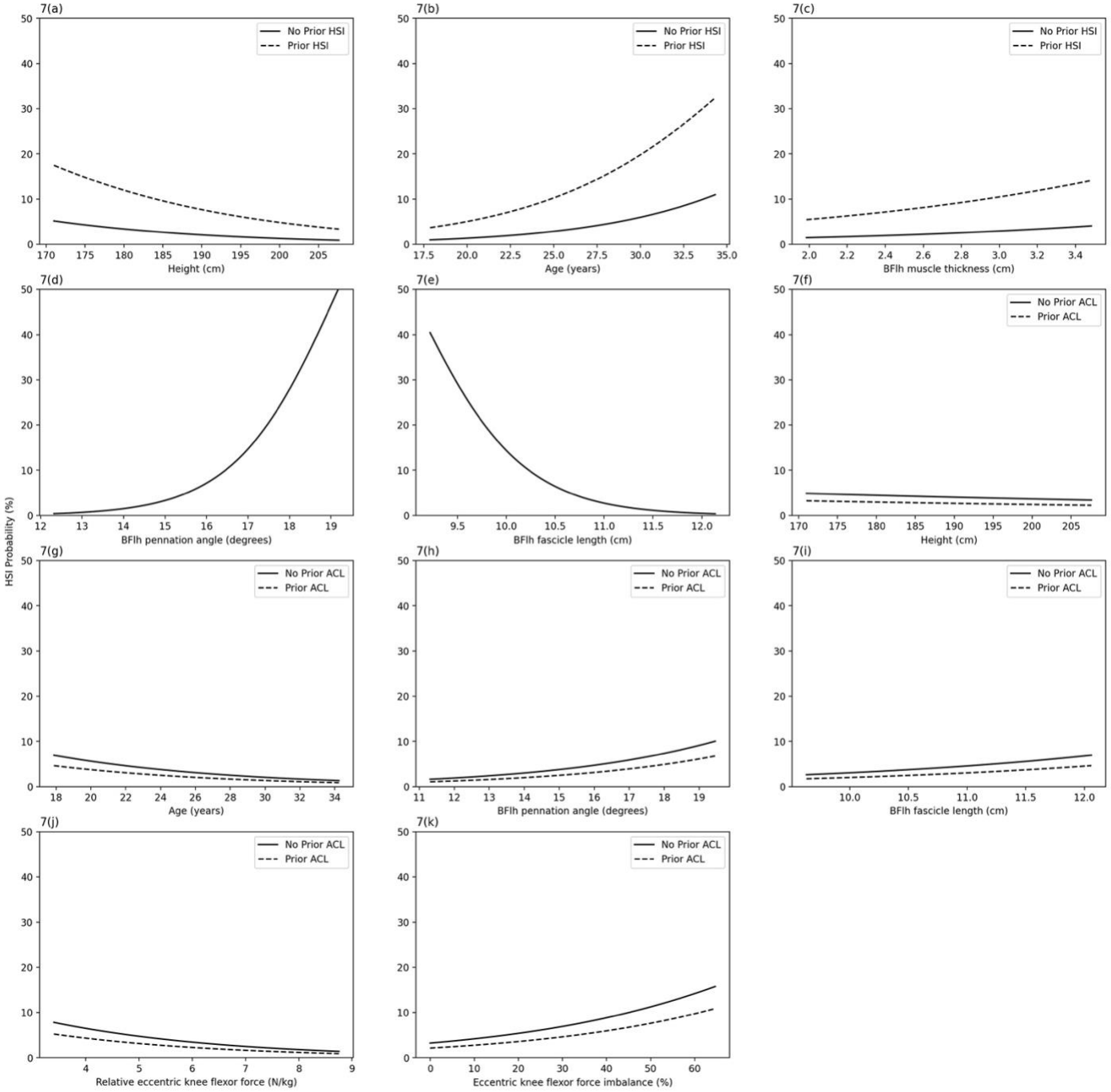


Figure 7a: The impact of change in height on hamstring strain injury (HSI) probability in pre-season with other factors set as mean constants (Age = 23.54 years, Muscle thickness = 2.64 cm).

Figure 7b. The impact of change in age on hamstring strain injury (HSI) probability in pre-season with other factors set as mean constants (Height = 188.07 cm, Muscle thickness = 2.64 cm).

Figure 7c. The impact of change in muscle thickness on hamstring strain injury (HSI) probability in pre-season with other factors set as mean constants (Height = 188.07 cm, Age = 23.54 years).

Figure 7d. The impact of change in pennation angle on hamstring strain injury (HSI) probability in early in-season with fascicle length set as mean constant (Fascicle length = 10.72 cm).

Figure 7e. The impact of change in fascicle length on hamstring strain injury (HSI) probability in early in-season with pennation angle set as mean constant (Pennation angle = 15.37 degrees).

Figure 7f. The impact of change in height on hamstring strain injury (HSI) probability in late in-season with other factors set as mean constants (Age = 23.13 years, Pennation angle = 15.39 degrees, Fascicle length = 10.74 cm, Relative eccentric knee flexor force = 5.45 N/kg, Eccentric knee flexor force imbalance = 9.33%).

Figure 7g. The impact of change in age on hamstring strain injury (HSI) probability in late in-season with other factors set as mean constants (Height = 188.05 cm, Pennation angle = 15.39 degrees, Fascicle length = 10.74 cm, Relative eccentric knee flexor force = 5.45 N/kg, Eccentric knee flexor force imbalance = 9.33%).

Figure 7h. The impact of change in pennation angle on hamstring strain injury (HSI) probability in late in-season with other factors set as mean constants (Height = 188.05 cm, Age = 23.13 years, Fascicle length = 10.74 cm, Relative eccentric knee flexor force = 5.45 N/kg, Eccentric knee flexor force imbalance = 9.33%).

Figure 7i. The impact of change in fascicle length on hamstring strain injury (HSI) probability in late in-season with other factors set as mean constants (Height = 188.05 cm, Age = 23.13 years, Pennation angle = 15.39 degrees, Relative eccentric knee flexor force = 5.45 N/kg, Eccentric knee flexor force imbalance = 9.33%).

Figure 7j. The impact of change in relative eccentric knee flexor force on hamstring strain injury (HSI) probability in late in-season with other factors set as mean constants (Height = 188.05 cm, Age = 23.13 years, Pennation angle = 15.39 degrees, Fascicle length = 10.74 cm, Eccentric knee flexor force imbalance = 9.33%).

Figure 7k. The impact of change in eccentric knee flexor force imbalance on hamstring strain injury (HSI) probability in late in-season with other factors set as mean constants (Height = 188.05 cm, Age = 23.13 years, Pennation angle = 15.39 degrees, Fascicle length = 10.74 cm, Relative eccentric knee flexor force = 5.45 N/kg).

Analysis 2

The performance of the individual models in Analysis 2 can be found in Figure 4a and 4b. Neither the predictions of early in-season HSIs (median AUC = 0.67, interquartile range (IQR) = 0.15; Table 4) nor late in-season HSIs (median AUC = 0.67, interquartile range (IQR) = 0.26; Table 5) were improved by assessing the magnitude of change in data across preseason. For HSIs occurring early in-season, the model with the best predictive performance utilised BFlh fascicle length and pennation angle, which were assessed at the end of pre-season. The resulting median AUC was 0.84 and the interquartile range was 0.16 (Table 4). Predicting late in-season injuries utilising the absolute change in BFlh pennation angle and fascicle length across pre-season, as well as history of ACL displayed the best predictive performance (median AUC = 0.67, interquartile range (IQR) = 0.26; Table 5). However, the predictive performance was not significantly improved when compared to relative BFlh fascicle length and fascicle length, which were assessed at the start of pre-season only (median AUC = 0.65, interquartile range (IQR) = 0.25; Table 5).

Table 4: The results of Analysis 2. The performance of models built with selected predictors assessed at start of pre-season (d1), end of pre-season (d2), start and end of pre-season (d1, d2), the magnitude of change of data in pre-season (d2-d1), data assessed at the start and end of pre-season and the magnitude of change of data in pre-season (*d1, d2, (d2-d1)*) as predictor variables, and hamstring strain injuries (HSIs) occurred in early in-season (*i2*) as target variable.

Models	Risk factors*	Frequency			AUC						
		HSI	Non-HSI	Total	Interquartile range	Standard Deviation	Minimum	Lower quartile	Median	Upper quartile	Maximum
d1	fascicle length (d1), relative fascicle length (d1)	23	219	242	0.15	0.11	0.29	0.60	0.68	0.75	0.96
d2	pennation angle (d2), fascicle length (d2)	23	219	242	0.16	0.11	0.44	0.75	0.84	0.91	1.00
d1&d2	pennation angle (d2), fascicle length (d2)	23	219	242	0.16	0.11	0.44	0.75	0.84	0.91	1.00
d2-d1	prior HSI, pennation angle (c1), fascicle length (c1), eccentric knee flexor force imbalance (c1)	23	219	242	0.15	0.11	0.25	0.59	0.67	0.74	0.98
d1&d2&(d2-d1)	pennation angle (d2), fascicle length (d2)	23	219	242	0.16	0.11	0.44	0.75	0.84	0.91	1.00

Performance is measured as area under the curve (AUC).

*Risk factors were selected by recursive feature elimination and 5-fold cross validation.

d1&d2; models built with non-modifiable risk factors assessed at the start of pre-season and modifiable risk factors assessed at the start and end of pre-season.

d2-d1; models built with non-modifiable risk factors assessed at the start of pre-season and magnitude of change of modifiable risk factors between start and end of pre-season.

c1; magnitude of change of specific risk factor between start and end of pre-season.

Table 5: The results of Analysis 2. The performance of models built with selected predictors assessed at start of pre-season (d1), end of pre-season (d2), start and end of pre-season (d1, d2), the magnitude of change of data in pre-season (d2-d1), data assessed at the start and end of pre-season and the magnitude of change of data in pre-season (*d1, d2, (d2-d1)*) as predictor variables, and hamstring strain injuries (HSIs) occurred in late in-season (*i3*) as target variable.

Models	Risk factors*	Frequency			AUC						
		HSI	Non-HSI	Total	Interquartile range	Standard Deviation	Minimum	Lower quartile	Median	Upper quartile	Maximum
d1	fascicle length (d1), relative fascicle length (d1)	9	210	219	0.25	0.16	0.20	0.54	0.65	0.79	0.98
d2	eccentric knee flexor force imbalance (d2)	9	210	219	0.23	0.16	0.17	0.44	0.55	0.67	0.93
d1&d2	prior ACL, pennation angle (d1), fascicle length (d1)	9	210	219	0.29	0.18	0.12	0.45	0.58	0.74	0.98
d2-d1	prior ACL, pennation angle (c1), fascicle length (c1)	9	210	219	0.26	0.19	0.17	0.50	0.67	0.76	1.00
d1&d2&(d2-d1)	fascicle length (c1)	9	210	219	0.27	0.20	0.14	0.46	0.64	0.73	1.00

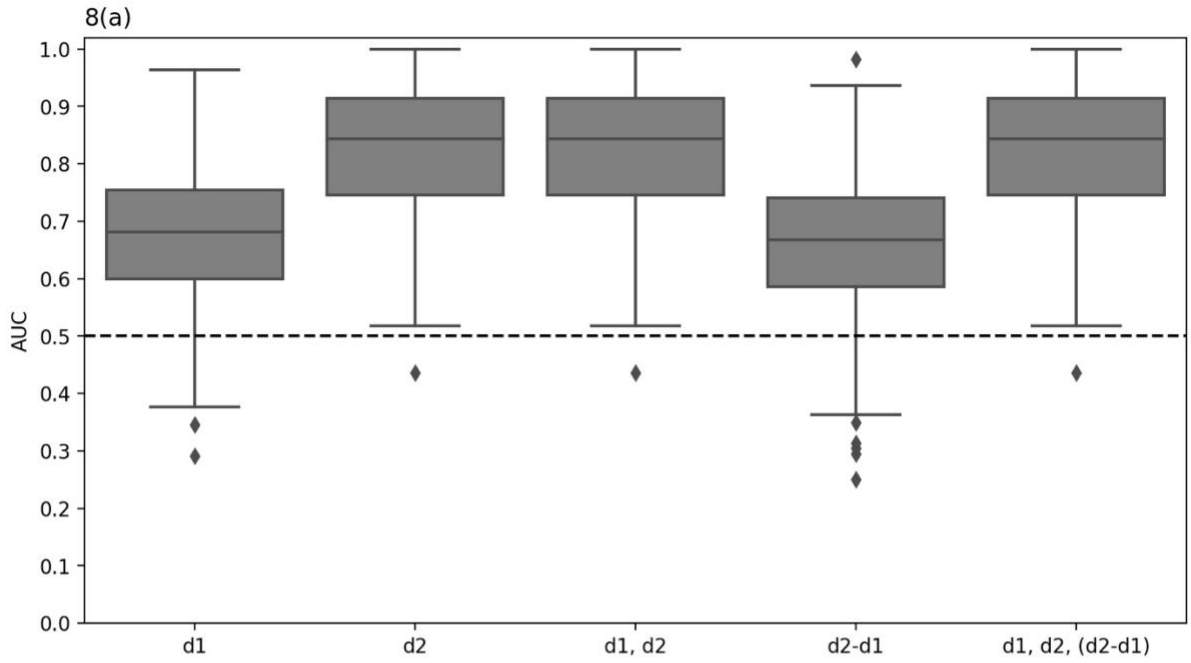
Performance is measured as area under the curve (AUC).

*Risk factors were selected by recursive feature elimination and 5-fold cross validation.

d1&d2; models built with non-modifiable risk factors assessed at the start of pre-season and modifiable risk factors assessed at the start and end of pre-season.

d2-d1; models built with non-modifiable risk factors assessed at the start of pre-season and magnitude of change of modifiable risk factors between start and end of pre-season.

c1; magnitude of change of specific risk factor between start and end of pre-season.



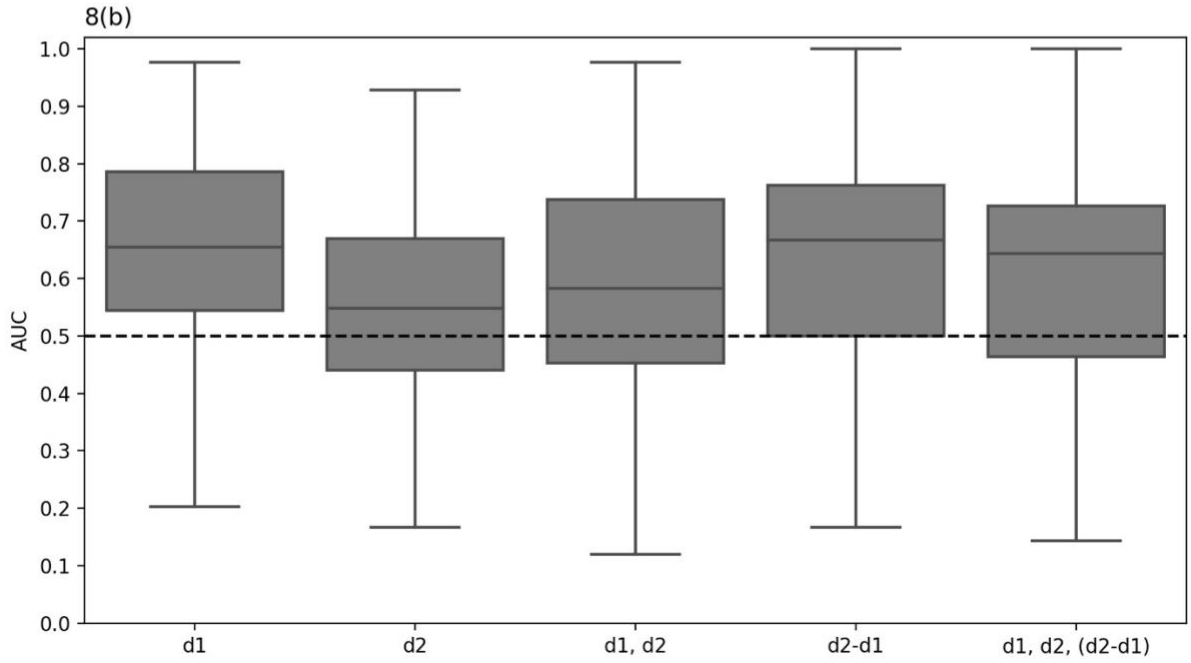


Figure 8a: The performance of models built with selected predictors assessed at start of pre-season (*d1*), end of pre-season (*d2*), start and end of pre-season (*d1, d2*), the magnitude of change of data in pre-season (*d2-d1*), data assessed at the start and end of pre-season and the magnitude of change of data in pre-season (*d1, d2, (d2-d1)*) as predictor variables, and hamstring strain injuries (HSIs) that occurred in early in-season (*i2*) as target variable. AUC = area under the curve.

Figure 8b. The performance of models built with selected predictors assessed at start of pre-season (*d1*), end of pre-season (*d2*), start and end of pre-season (*d1, d2*), the magnitude of change of data in pre-season (*d2-d1*), data assessed at the start and end of pre-season and the magnitude of change of data in pre-season (*d1, d2, (d2-d1)*) as predictor variables, and hamstring strain injuries (HSIs) that occurred in late in-season (*i3*) as target variable. AUC = area under the curve

Discussion

This study aimed to assess whether the factors associated with HSI in professional Australian Football changed across the season. The current study found that the subset of risk factors that best predicted the occurrence of HSI was different between the pre-season and in-season periods. This study also aimed to assess whether the magnitude of change in HSI risk factors across the pre-season period improved the prediction of HSIs sustained in-season beyond using measures taken at the start or end of pre-season alone. The magnitude of change in eccentric knee flexor strength and BFlh muscle architecture variables across the pre-season period generally displayed poorer predictive performance than the absolute measures themselves (particularly those taken at the end of pre-season).

Did more frequent assessment of risk factors improve the prediction of future HSI?

The model that showed the best performance in this study used BFlh fascicle length and pennation angle measured at the end of pre-season to predict HSIs occurring only in the first half of the in-season period. This model predicted prospective HSIs with a median AUC of 0.86. An earlier study aimed to predict HSIs among elite Australian Football players, using data on age, history of HSI, and eccentric hamstring strength obtained during two AFL seasons [16]. When attempting to predict HSIs occurring during the same season, the median AUC values for the 2013 and 2015 AFL seasons were 0.58 and 0.57, respectively [16]. In this previous study, the median AUC for predicting HSIs that occurred during the 2015 AFL season using data from the 2013 AFL season was 0.52 [16]. It was suggested that more frequent measures of the risk factors examined may have improved predictive performance. However, another study showed that taking assessments of modifiable risk factors more frequently did not improve the ability to identify athletes with a high risk of HSI beyond data collected at a single timepoint [17]. In support of these previous findings [17], it was observed that more

frequent measurements did not improve the ability to predict the occurrence of HSI. However, the assessment of different risk factors at different timepoints did improve predictive performance. In addition, recursive feature elimination was utilised to optimise predictive performance and improve the interpretability of built models. Results from preliminary analyses suggest that the selected predictors are likely to deliver better predictive performance than utilising all predictors. The findings of this study suggest that a subset of risk factors, as opposed to all risk factors, used in previous studies, may have been more effective in predicting prospective HSIs.

Does the magnitude of change in risk factor data across pre-season improve the ability to predict HSI throughout the season beyond the absolute values?

In addition to suggesting that more regular measures of the risk factors examined may improve predictive performance [16], prior work has also noted that assessing risk factors at the start of pre-season alone assumes that these factors will remain unchanged throughout the season (or prior to HSI). It has been suggested that changes in HSI risk factors could have a more significant effect on injury risk compared to the absolute values of such factors at a specific time [16, 88]. AUC values of 0.7 and above are regarded as having significant impacts in sport science domains [21]. In the current study, models built with the magnitude of change in risk factors across pre-season were less optimal when attempting to predict HSIs during early in-season, or *i2* (median AUC of 0.66), as well as HSIs during late in-season, or *i3* (median AUC of 0.63). Conversely, models built using the absolute values measured at the end of pre-season, or *d2*, performed better when predicting HSIs during early in-season, or *i2* (median AUC of 0.83). However, the performance of all models attempting to predict late in-season HSIs, or *i3*, were the poorest.

The current results suggest that risk factor data assessed at the end of pre-season provides the strongest performance when predicting in-season HSIs. Despite the magnitude of change in modifiable risk factor data performing poorly from a prediction standpoint, it is important to acknowledge that significant adaptations in eccentric knee flexor strength and BFlh muscle architecture can be elicited in as little as two weeks [31]. Given this, it is likely that athletes saw significant adaptations across the pre-season period and that modifiable risk factor data assessed at the end of pre-season provided a better indication of athletes' physical status during the in-season period compared to data collected at the start of pre-season. In contrast to this, data collected at the midpoint of the in-season period displayed the worst predictive performance when used to predict injuries that occurred during the second half of the in-season period. This suggests that despite this data being more aetiologically relevant, there may exist other factors that influence the risk of HSIs occurring during the latter half of the season to a greater extent than those examined in this study.

In which phase of the season was the predictive performance for HSI best?

The best performing model aimed to predict HSIs during the first half of the in-season period and was built using data collected at the end of pre-season (median AUC of 0.86; Table 3). In contrast, the poorest performing model was built using data collected at the midpoint of the in-season period and aimed to predict HSIs in the second half of the in-season period (median AUC of 0.46; Table 3). A study conducted earlier reported that there is a noticeable increase in BFlh fascicle length among all players in the early in-season period [27]. However, it was observed that players with a history of HSI saw greater decreases in BFlh fascicle length during the latter part of the in-season period when compared to players without a history [27]. This may, to an extent, explain why BFlh fascicle length assessed at the end of pre-season did not

present strong association with late in-season HSIs, when compared to early in-season HSIs [27].

Absolute risk factor data assessed at the end of pre-season may provide practitioners with the most insight regarding HSI risk, and that additional assessments of the studied variables throughout the in-season period may not add further value. The relatively poor performance of the models built to predict late in-season HSIs suggests that there may be additional factors that influence the risk of injury to a greater extent in the latter stages of the season.

Limitations

Due to the length of BFlh fascicles exceeding the ultrasound field of view (14x4.7mm), extrapolation methods were used to calculate BFlh fascicles [89]. Although the extrapolation method was proven to be highly reliable in an earlier study (ICC>0.97) when validated against cadaveric data [81], the drawback is that it may overestimate BFlh fascicle length [90]. Due to the lack of a standardised classification system [91], information about the muscle that was injured was not provided for all HSIs reported in this study. Further subgroup analysis may be conducted if more injury data of the injured muscle were recorded. Due to the absence of player exposure data in this study, the reported incidence of HSIs did not take into account the duration of training and competition. In addition, the use of athlete tracking technologies to account for high-speed running and strength training exposure may offer more insights regarding HSI risk. Warm up procedures were not standardised for strength assessments. Future studies should consider standardising warm-up practices to limit the impact it may have on the strength outcomes. The use of logistic regression in this study assumes linearity between target variable and risk factors. Complex non-linear models may be utilised with proper hyperparameter tuning practice. Although previous studies showed the use of non-linear models outperformed

logistic regression in injury prediction [21, 92], these studies were conducted on a larger dataset. Earlier work showed no improvements in predictive performance when complex modelling approach was used [16]. The absence of a standardised fine-tuning process on small and imbalanced dataset may be the cause, which result in overfitting. Despite this study recording a high number of prospective HSIs in comparison to previous research [93], the relatively low injury rates and the class imbalance problem that this presents remains a limitation of this study and as well as most prospective sports injury studies in general. It is unclear whether predictive performance would be improved if class imbalance was addressed. Furthermore, the presence of missing data results in reduced numbers of player seasons used for the analysis in this study. Although AUC is used in many studies [16, 17, 19], other metrics should be considered thoroughly when evaluating the generalisation of binary classifiers. In addition, future studies should utilise interpretability methods in machine learning to help experts better understand the decisions of trained models beyond predictive performance. Finally, previous work suggests that HSI risk factors are not transferable to different sporting populations [94] so applications of the current findings to other sports (e.g., soccer, rugby) should be done with caution.

Conclusion

This study has demonstrated that the risk factors most associated with prospective HSIs change throughout an Australian Football season. Non-modifiable risk factors (History of HSI, age and height) demonstrated a strong association with pre-season HSIs, whereas early in-season HSIs were better explained by modifiable risk factors. Conversely, late in-season injuries did not present any strong associations with either modifiable or non-modifiable risk factors examined in this study. The magnitude of change in modifiable risk factors across pre-season did not improve the prediction of in-season HSIs. The results of this study suggest that

assessing the same risk factors at multiple time points throughout the season may not be the best approach when identifying athletes at an increased risk of HSI. Instead, assessing different risk factors at specific time points with aetiological relevance may provide practitioners with more insight.

CHAPTER 4: Explanation of risk predictions with machine learning in Australian football

Introduction

Previous chapter had shown that the risk factors of HSI may vary depending on the time of Australian Football season, where non-modifiable risk factors were mostly associated with pre-season HSIs and modifiable risk factors were mostly associated with early in-season HSIs. While it is useful to understand what risk factors are important at a particular time point, however, it cannot explain the risk of injury for individual athletes based on their specific conditions. Individual-level explanation allows practitioners to inform counselling and take preventive measures for injury-prone athletes [19].

Followed by the surge of machine learning models used in predicting sport injuries [21, 42]. Some studies had shown how to explain the risk of injury at individual-level. A recent study conducted on National Basketball Association (NBA) athletes utilised Shapley Additive Explanations (SHAP) to identify the factors that contribute positively and negatively to the prediction of lower extremity muscle strain (LEMS) [19]. Similarly, a study conducted on elite soccer players explained the risk contribution from individual's blood sample features with SHAP [18]. However, the required change of risk factors for a particular athlete to significantly reduce the risk of future injury remains unclear. In real world, practitioners would like to know how to reduce the risk of injury for an athlete who attends the consultation. For example, knowing the optimal range of training load for an athlete to decrease the risk of being classified as injured is important for practitioners to provide advice.

This study aimed to improve injury prevention in Australian Football through explaining machine's prediction. Counterfactual explanations can identify the minimal change in modifiable risk factors required for an athlete to reduce the risk of HSI.

Methods

The modelling pipeline is shown in figure 9. The steps are discussed as followed:

Dataset

The dataset used in this study is Australian Football League (AFL) [17]. 311 Australian Football players were involved in this study. The value range of individual risk factors assessed in preseason, early in-season and late in-season are shown in Table 6.

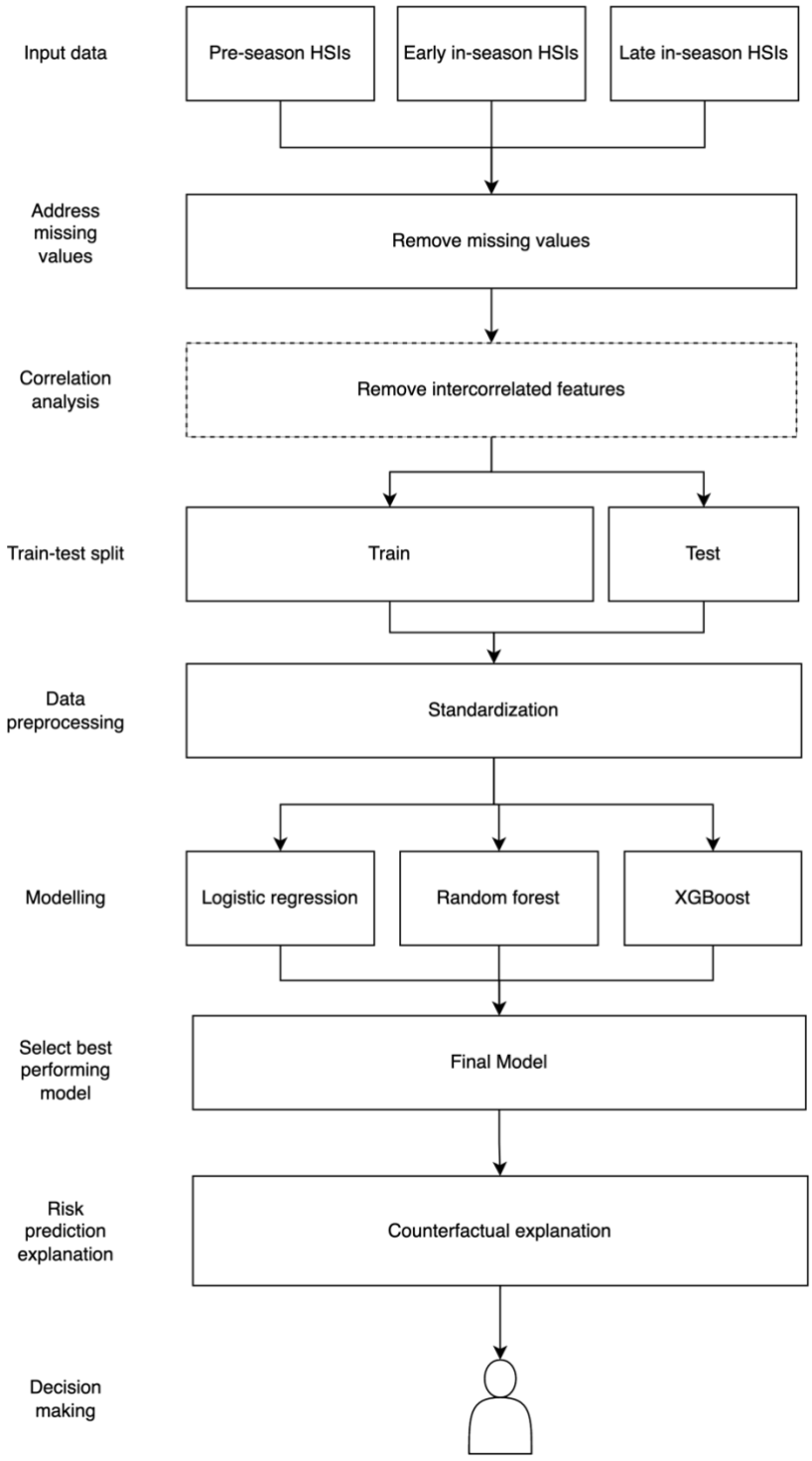


Figure 9: The modelling pipeline to generate counterfactual explanations.

Table 6: The range of individual risk factors across multiple time points in this study.

Factors	Unit	Range [min, max]		
		Pre-season	Early in-season	Late in-season
Prior HSI	Yes/no	[0, 1]	[0, 1]	[0, 1]
Prior ACL	Yes/no	[0, 1]	[0, 1]	[0, 1]
Height	Centimetre	[171, 208]	[171, 208]	[171, 208]
Weight	Kilogram	[65, 114]	[65, 111]	[65, 114]
Age	Years	[17.877, 34.490]	[17.877, 32.411]	[17.877, 34.403]
BFlh muscle thickness	Centimetre	[1.985, 3.500]	[2.041, 3.442]	[1.883, 3.575]
BFlh pennation angle	Degrees	[12.300, 19.500]	[12.320, 19.260]	[11.310, 19.540]
BFlh fascicle length	Centimetre	[8.625, 11.830]	[9.220, 12.166]	[9.630, 12.086]
BFlh relative fascicle length	fascicle length/muscle thickness	[3.072, 4.890	[3.044, 4.990	[3.065, 5.346]
Eccentric knee flexor force	Newton	[237.500, 638.500]	[277.000, 785.000]	[275.375, 801.500]
Relative eccentric knee flexor force	Newton/Kilogram	[2.306, 7.323]	[3.044, 8.820]	[3.400, 8.808]
Eccentric knee flexor force imbalance	Percentage	[0, 119.583]	[0, 67.730]	[0, 65.299]

Data pre-processing

Prior to splitting input data into training and testing sets, any row that consists of at least one missing value was removed. In sport science domains, certain features may be engineered from other features, which can produce highly intercorrelated features [17]. Additionally, some regression models (e.g. Logistic regression) require intercorrelated features to be removed for better predictive performance. For these reasons, the option of removing intercorrelated features with given threshold is provided. Most sports injury datasets had few injury events [16]. 80/20 stratified shuffle split was performed to randomly select 80% of data for training and 20% testing sets while preserving the percentage of non-injury and injury classes. Similar to other studies [16, 21], continuous variables (except prior HSI and prior ACL) were standardized to eliminate the effect of differing scales in variables to make sure they are treated equally by the model during training. This involves scaling individual variables so that it has a mean value of 0 and a standard deviation value of 1 using the equation below:

$$x_{scaled} = \frac{x - \mu}{\sigma}$$

Where x is the input value to scale, μ is the mean of the variable and σ is the standard deviation of the variable.

Modelling

Once the training and testing data were standardized, the training data were input to the following models:

- Logistic regression
- Random forest
- XGBoost (eXtreme Gradient Boosting) [95]

logistic regression was chosen for its popularity in sports injury prediction domains [39]. It is also robust against overfitting. Overfitting occurs when a model performs well in the training data but is unable to generalize on unseen data. In order to capture the potentially complex interactions between input features, two additional models were chosen: random forest and XGBoost [95]. These models had demonstrated to be effective at modelling non-linear relationships and interactions in the data [19, 21], which can be useful in predicting sports injuries where many factors may be involved. To address the class imbalance issue in which the number of healthy events significantly outweighs the number of injury events [16], all three models were configured to have equal class weights. This enabled the models to give equal importance to both classes during training and prediction, thereby mitigating the potential bias towards the majority class. This is similar to over-sampling and under-sampling methods [16, 21] in other studies except no modifications were made to the training data. Due to limited injury events, the models employed in this study were not fine-tuned, all hyperparameters are by default.

After evaluating the predictive performance of all models with the allocated testing data, the model with the best discrimination ability measured in AUC was chosen as the final model for individual-level prediction explanation. This is to ensure the explanation of the prediction are reliable and accurate. The classification threshold is 0.5, instance predicted with a value below 0.5 was classified as healthy and instance predicted with a value above 0.5 was classified as injury.

Counterfactual explanation

Once the model is determined, the final step is to generate actionable counterfactual (CF) explanations [20]. This is an optimization problem where the input is a trained model f and a

randomly selected injured player x. The objective is to produce counterfactual examples $\{c_1, c_2, c_k\}$, which result in a decision different from the input x. Good counterfactual examples must satisfy a few important properties as followed:

- Proximity
- Sparsity
- Diversity

Proximity defines the similarity between input instance x and the counterfactual examples $\{c_1, c_2, c_k\}$. This is important as it defines the quality of the generated counterfactual explanations. For instance, it is not helpful to produce a counterfactual explanation where the height of the player is 300 cm. Proximity aims to find the minimal change in input features that leads to an opposite decision. It can be calculated with the equation as followed:

$$\text{Proximity} := -\frac{1}{k}\sum_{i=1}^{k} \text{dist}\,(\boldsymbol{c}_i, \boldsymbol{x})$$

For continuous features, $dist$ is the $l1$-distance, also known as Manhattan distance between the counterfactual example and the input instance. For categorical variables, it is calculated as the number of categorical features that is not equal to the input instance's categorical features.

The second property is sparsity, it defines the number of input features required to change in order to obtain an opposite prediction. This requires identifying a small set of features that are most relevant to the explanation. Fewer features are better understood by human [58]. It was suggested the number of feature changes should not be more than three due to the constraints human have in category learning [74]. Sparsity can be calculated with the equation as followed:

$$\text{Sparsity} : 1 - \frac{1}{kd}\sum_{i=1}^{k}\sum_{l=1}^{d} 1_{\left[c_i^l \neq x_i^l\right]}$$

where d is the number of features. When the value of the same feature is different between counterfactual example and input instance. The sparsity was decreased.

The third property is diversity. Diversity encourages the generated counterfactual examples to be significantly different from each other. It is not informative if multiple counterfactual explanations are similar. Diversity can be measured as followed:

$$\text{Diversity} : \Delta = \frac{1}{C_k^2} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \text{dist}\left(\boldsymbol{c}_i, \boldsymbol{c}_j\right)$$

Similar to the proximity property, the distance metric is measured as the $l1$-distance between continuous features in the counterfactual examples.

Based on these important properties which define a 'good' counterfactual, counterfactual explanations can be formulated into an optimization problem by combining these properties into a loss function. This study utilised DiCE (Diverse Counterfactual Explanations) for its efficiency and advantage in considering diversity. For more details on generating counterfactual explanations in this study, please refer to the original paper [20]. Injury players were randomly selected in the testing dataset for counterfactual explanation. The top three counterfactual explanations were selected for interpretation.

Results

The predictive performance of individual models are displayed in Table 7. AUC measures the level of discrimination.

Table 7: The predictive performance of individual models. The model with the highest AUC is selected to generate counterfactual explanations with DiCE.

Models	AUC		
	Preseason	Early in-season	Late in-season
Logistic regression (feature selection)	0.838	0.670	0.608
Logistic regression (all features)	0.500	0.648	0.588
Random forest	0.400	0.726	0.392
XGBoost	0.608	0.804	0.510

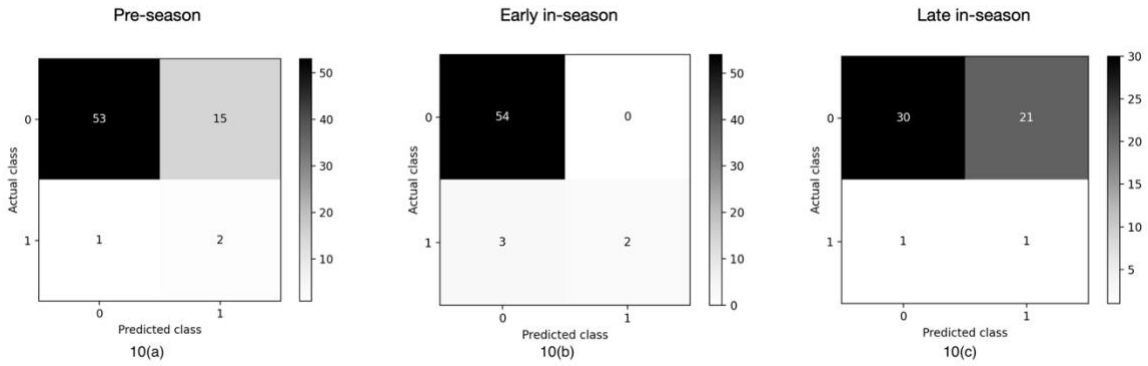


Figure 10: Confusion matrices of the best predictive models in preseason, early in-season and late in-season respectively.

Table 8: The counterfactual explanations for player A and player B who sustained HSIs in pre-season.

Player	Prior HSI	Height	Age	Muscle thickness	HSI
A	1	198	30.071	2.782	1
A1	-	-	-	2.69	0
A2	-	-	23.307	-	0
A3	-	-	19.752	3.040	0

Player	Prior HSI	Height	Age	Muscle thickness	HSI
B	0	176	25.767	2.835	1
B1	-	-	23.327	-	0
B2	-	-	-	2.700	0
B3	-	-	29.405	2.490	0

Table 9: The counterfactual explanations for player C and player D who sustained HSIs in early in-season.

Player	Prior HSI	Prior ACL	Height	Weight	Age	BFlh muscle thickness	BFlh pennation angle	BFlh fascicle length	BFlh relative fascicle length	Eccentric knee flexor strength	Relative eccentric knee flexor strength	Eccentric knee flexor force imbalance	HSI
C	1	0	188	87	21.822	2.539	15.235	9.981	3.932	481	5.529	11.894	1
C1	-	-	-	-	-	-	-	10.416	-	-	-	-	0
C2	-	-	-	-	29.315	-	-	-	-	-	-	-	0
C3	-	-	-	-	-	-	-	-	-	-	-	10.026	0

Player	Prior HSI	Prior ACL	Height	Weight	Age	BFlh muscle thickness	BFlh pennation angle	BFlh fascicle length	BFlh relative fascicle length	Eccentric knee flexor strength	Relative eccentric knee flexor strength	Eccentric knee flexor force imbalance	HSI
D	0	0	185	83	31.455	2.754	16.055	10.261	3.726	349.500	4.211	17.757	1
D1	-	-	-	-	-	-	-	10.623	-	-	-	-	0
D2	-	-	-	-	-	-	-	-	-	-	-	11.711	0
D3	-	-	-	-	-	-	-	-	3.832	-	-	-	0

Table 10: The counterfactual explanations for player E who sustained HSI in late in-season.

Player	Prior ACL	Height	Age	BFlh pennation angle	BFlh fascicle length	Relative eccentric knee flexor strength	Eccentric knee flexor force imbalance	HSI
E	0	186	18.712	17.540	11.511	5.750	13.692	1
E1	-	-	-	-	10.084	-	-	0
E2	-	-	-	11.96	-	-	-	0
E3	-	-	-	-	-	8.240	-	0

Preseason HSI

The model with the best predictive performance for pre-season HSIs is logistic regression with risk factors identified in the previous study. The resulting AUC is 0.838 (Table 7). These risk factors were history of HSI, height, age and muscle thickness. In the testing dataset, there are 68 healthy players and 3 injured players (Figure 10a). 67% of injury players (2 out of 3) were correctly classified.

Based on Table 8, Player is 198 cm tall and has a history of HSI. It was suggested that player A could reduce his muscle thickness by about 0.1 centimetres to be classified as non-injury (Table 8, A1). Although younger age (Table 8, A2) can reduce the risk of HSI, it is not modifiable. Similarly, despite younger age and having thicker muscles (Table 8, A3) can compensate the risk of HSI, they are not modifiable for player A.

In contrast to player A, player B is younger, shorter in height and without a history of HSI. It was suggested player B could reduce his muscle thickness by 0.135 cm to mitigate the risk of HSI (Table 8, B2). Likewise, being younger can reduce the risk of HSI in pre-season but it is not alterable (Table 8, B1). Counterfactual example B3 (Table 8) suggests when player B is 29.4 years old, having muscle thickness reduced to 2.49 cm can mitigate the risk of HSI in preseason (Table 8, B3).

Early in-season HSI

XGBoost with all risk factors displayed the best predictive performance in early in-season HSIs. The resulting AUC is 0.804 (Table 7). Based on Figure 10b, 40% of players (2 out of 5) were correctly classified by the XGBoost model.

According to Table 9, given the condition of player C who sustained an HSI in early in-season. It was suggested that player C could increase his BFlh fascicle length to 10.416 cm to prevent HSI (Table 9, C1). It was also suggested that player C could reduce his eccentric knee flexor force imbalance by 1.868% to avoid HSI (Table 9, C3). Although older age could reduce the risk of sustaining HSI in early in-season, age is not modifiable. (Table 9, C2).

Unlike player C, player D is older and does not have a history of HSI. Counterfactual explanation (Table 9, D1) suggested that player D could increase his BFlh fascicle length to 10.623 cm for reduced risk of HSI. Alternatively, player D can increase his BFlh relative fascicle length to 3.832 or improve his eccentric knee flexor force imbalance to 11.711% (Table 9 D2, D3) to avoid HSI.

Late in-season HSI

Although it was suggested that player E could have shorter BFlh fascicle length or a reduced BFlh pennation angle, or having a greater relative eccentric knee flexor strength to prevent HSI in late in-season (Table 10, E1, E2, E3). It is inconclusive as the best predictive performance of HSIs occurred in late in-season remains poor (Table 7, AUC = 0.608).

Discussion

How trustworthy are counterfactual explanations for clinical studies?

Based on the study conducted, there are a few constraints when generating counterfactual explanations for players who sustained HSIs at different time points. The primary constraint is the predictive ability of ML models, even though counterfactual explanations can be generated for HSIs occurred in late in-season. The level of discrimination remains low (AUC = 0.608, Table 7). It is recommended that counterfactual explanations must be interpreted together with

predictive performance. Practitioners should validate the model thoroughly before relying on counterfactual explanations to make decisions. This study tried to mitigate this issue by training multiple ML models and selecting the model with the best discrimination ability for explanations.

When does counterfactual explanation work best?

Counterfactual explanations work best when the dimension of the dataset is relatively low. It is challenging when multiple risk factors were changed to obtain counterfactual explanations. A study stated that "good" counterfactual explanations should have no more than 2 changing features [74]. However, this is very unlikely in many datasets, as most counterfactuals involved at least more than 5 changing features. As a result, counterfactual explanations should be used in a controlled environment where the risk factors were thoroughly studied and identified.

Conclusion

This study has demonstrated that the application of counterfactual explanations on machine learning models can offer valuable and practical insights for practitioners. By generating counterfactual explanations, practitioners and clinicians not only gain deeper understanding of the factors that contribute to the risk of HSI for injury-prone players, but also identifying potential solutions for risk mitigation. In the future, this study serves as a stepping stone for clinicians to develop a custom risk intervention program for vulnerable athletes based on their conditions.

CHAPTER 5: CONCLUSION

The aim of this research was to identify if risk factors of HSI vary across multiple time points of Australian football season. Chapter 3 has found that preseason HSIs were strongly associated with age, history of HSI, height and muscle thickness (median AUC = 0.83). This finding partially supports previous studies which concluded age and history of HSI are prominent non-modifiable risk factors [11]. Subsequently, multivariate logistic regression had shown that early in-season HSIs were strongly associated with BFlh fascicle length and pennation angle (median AUC = 0.86). This finding also partially supports a previous study which concluded shorter fascicle length were associated with increased risk of HSI [15]. It was found out HSIs in late in-season were not associated with any risk factors in this study and the underlying reasons remained unclear. The investigation was taken further by examining whether the magnitude of change in modifiable risk factors (BFlh muscle architecture and eccentric knee flexor strength) were useful in predicting in-season HSIs (early in-season and late in-season). The result revealed that in-season HSIs did not demonstrate any association with the change of data in preseason.

Chapter 4 aimed to interpret the prediction of machine learning models, so that the risk of HSI can be better understood at individual-level. It has been demonstrated that counterfactual explanation is an effective and novel approach in understanding risks, as well as facilitating practitioners in identifying potential solutions.

Limitations and future directions

Although this research has discovered significant findings. It also comes with several major limitations. Primarily, the dataset in this study is relatively small and consists of missing values. This is also the general limitation in the domain as a recent study stated the average sample

size used to develop musculoskeletal injury prediction models is less than 200 [39]. This is a bottleneck when it comes to developing highly accurate ML models, as these models require a lot of data to discover complex patterns and relationships. This may explain why logistic regression is widely used in the field [39] and why hyperparameter tuning practices are not made known in most studies. Secondly, the dataset in the study is highly imbalanced. The number of healthy events significantly outweighs the number of injury events. Common techniques are bagging, over-sampling and under-sampling [16, 21]. However, fewer injury events mean the models may learn insufficient information from the injury class and outliers can impact the results significantly. When it comes to modelling, the injury outcome is binary and does not account for the severity of injury. It is unclear whether including the severity of injury would improve the overall predictive performance and provide more findings. Due to limited risk factors conducted in this study, the reason for the poor predictive performance of late in-season HSIs remains unclear. Future research may investigate further. Finally, the reliability of counterfactual explanations heavily relies on the predictive performance of machine learning models. Practitioners should evaluate the models thoroughly before making any clinical decision.

REFERENCES

1. Van Eetvelde, H., et al., *Machine learning methods in sport injury prediction and prevention: a systematic review.* J Exp Orthop, 2021. **8**(1): p. 27.
2. Dower, C., et al. *An enhanced metric of injury risk utilizing artificial intelligence*. in *Proceedings of the 13th Annual MIT SLOAN Sports Analytics Conference*. 2018.
3. Ekstrand, J., et al., *Time before return to play for the most common injuries in professional football: a 16-year follow-up of the UEFA Elite Club Injury Study.* British Journal of Sports Medicine, 2020. **54**(7): p. 421-426.
4. Verrall, G.M., et al., *Assessment of player performance following return to sport after hamstring muscle strain injury.* J Sci Med Sport, 2006. **9**(1-2): p. 87-90.
5. Orchard, J., H. Seward, and M.J. Orchard, *2012 AFL injury report*. 2013, Australian Football League, Docklands Australia.
6. Opar, D.A., M.D. Williams, and A.J. Shield, *Hamstring strain injuries: factors that lead to injury and re-injury.* Sports Med, 2012. **42**(3): p. 209-26.
7. Ayala, F., et al., *A Preventive Model for Hamstring Injuries in Professional Soccer: Learning Algorithms.* Int J Sports Med, 2019. **40**(5): p. 344-353.
8. Orchard, J.W., *Intrinsic and extrinsic risk factors for muscle strains in Australian football.* Am J Sports Med, 2001. **29**(3): p. 300-3.
9. Kenneally-Dabrowski, C.J.B., et al., *Late swing or early stance? A narrative review of hamstring injury mechanisms during high-speed running.* Scand J Med Sci Sports, 2019. **29**(8): p. 1083-1091.
10. Tokutake, G., et al., *The Risk Factors of Hamstring Strain Injury Induced by High-Speed Running.* J Sports Sci Med, 2018. **17**(4): p. 650-655.

11. Green, B., et al., *Recalibrating the risk of hamstring strain injury (HSI): A 2020 systematic review and meta-analysis of risk factors for index and recurrent hamstring strain injury in sport.* British Journal of Sports Medicine, 2020. **54**(18): p. 1081-1088.

12. Pizzari, T., B. Green, and N. van Dyk, *Extrinsic and Intrinsic Risk Factors Associated with Hamstring Injury*, in *Prevention and Rehabilitation of Hamstring Injuries*, K. Thorborg, D. Opar, and A. Shield, Editors. 2020, Springer International Publishing: Cham. p. 83-115.

13. Maniar, N., et al., *Hamstring strength and flexibility after hamstring strain injury: a systematic review and meta-analysis.* Br J Sports Med, 2016. **50**(15): p. 909-20.

14. Opar, D.A., et al., *Eccentric hamstring strength and hamstring injury risk in Australian footballers.* Med Sci Sports Exerc, 2015. **47**(4): p. 857-65.

15. Timmins, R.G., et al., *Short biceps femoris fascicles and eccentric knee flexor weakness increase the risk of hamstring injury in elite football (soccer): a prospective cohort study.* Br J Sports Med, 2016. **50**(24): p. 1524-1535.

16. Ruddy, J.D., et al., *Predictive Modeling of Hamstring Strain Injuries in Elite Australian Footballers.* Med Sci Sports Exerc, 2018. **50**(5): p. 906-914.

17. Opar, D.A., et al., *Screening Hamstring Injury Risk Factors Multiple Times in a Season Does Not Improve the Identification of Future Injury Risk.* Med Sci Sports Exerc, 2022. **54**(2): p. 321-329.

18. Rossi, A., et al., *Blood sample profile helps to injury forecasting in elite soccer players.* Sport Sciences for Health, 2022: p. 1-12.

19. Luu, B.C., et al., *Machine Learning Outperforms Logistic Regression Analysis to Predict Next-Season NHL Player Injury: An Analysis of 2322 Players From 2007 to 2017.* Orthop J Sports Med, 2020. **8**(9): p. 2325967120953404.

20. Mothilal, R.K., A. Sharma, and C. Tan, *Explaining machine learning classifiers through diverse counterfactual explanations.* Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2019.

21. Lövdal, S.S., R.J.R. Den Hartigh, and G. Azzopardi, *Injury Prediction in Competitive Runners With Machine Learning.* Int J Sports Physiol Perform, 2021. **16**(10): p. 1522-1531.

22. Taylor, K., et al., *Fatigue monitoring in high performance sport: a survey of current trends.* J Aust Strength Cond, 2012. **20**(1): p. 12-23.

23. Maniar, N., et al., *Incidence and prevalence of hamstring injuries in field-based team sports: a systematic review and meta-analysis of 5952 injuries from over 7 million exposure hours.* British Journal of Sports Medicine, 2022: p. bjsports-2021-104936.

24. Hägglund, M., M. Waldén, and J. Ekstrand, *Risk factors for lower extremity muscle injury in professional soccer: the UEFA Injury Study.* Am J Sports Med, 2013. **41**(2): p. 327-35.

25. Faulkner, J.A., et al., *The aging of elite male athletes: age-related changes in performance and skeletal muscle structure and function.* Clin J Sport Med, 2008. **18**(6): p. 501-7.

26. Volpi, E., R. Nazemi, and S. Fujita, *Muscle tissue changes with aging.* Curr Opin Clin Nutr Metab Care, 2004. **7**(4): p. 405-10.

27. Timmins, R.G., et al., *Effect of Prior Injury on Changes to Biceps Femoris Architecture across an Australian Football League Season.* Med Sci Sports Exerc, 2017. **49**(10): p. 2102-2109.

28. Bahr, R. and T. Krosshaug, *Understanding injury mechanisms: a key component of preventing injuries in sport.* British Journal of Sports Medicine, 2005. **39**(6): p. 324-329.

29. Fulton, J., et al., *Injury risk is altered by previous injury: a systematic review of the literature and presentation of causative neuromuscular factors.* Int J Sports Phys Ther, 2014. **9**(5): p. 583-95.

30. Bourne, M.N., et al., *Eccentric Knee Flexor Strength and Risk of Hamstring Injuries in Rugby Union: A Prospective Study.* Am J Sports Med, 2015. **43**(11): p. 2663-70.

31. Timmins, R.G., et al., *Architectural Changes of the Biceps Femoris Long Head after Concentric or Eccentric Training.* Med Sci Sports Exerc, 2016. **48**(3): p. 499-508.

32. Halson, S.L., *Monitoring training load to understand fatigue in athletes.* Sports Med, 2014. **44 Suppl 2**(Suppl 2): p. S139-47.

33. Windt, J. and T.J. Gabbett, *How do training and competition workloads relate to injury? The workload-injury aetiology model.* Br J Sports Med, 2017. **51**(5): p. 428-435.

34. Windt, J., et al., *Why do workload spikes cause injuries, and which athletes are at higher risk? Mediators and moderators in workload–injury investigations.* British Journal of Sports Medicine, 2017. **51**(13): p. 993-994.

35. Wallace, L.K., K.M. Slattery, and A.J. Coutts, *The ecological validity and application of the session-RPE method for quantifying training loads in swimming.* J Strength Cond Res, 2009. **23**(1): p. 33-8.

36. Arnason, A., et al., *Prevention of hamstring strains in elite soccer: an intervention study.* Scand J Med Sci Sports, 2008. **18**(1): p. 40-8.

37. Petersen, J., et al., *Preventive effect of eccentric training on acute hamstring injuries in men's soccer: a cluster-randomized controlled trial.* Am J Sports Med, 2011. **39**(11): p. 2296-303.

38. Pincheira, P.A., et al., *Biceps femoris long head sarcomere and fascicle length adaptations after 3 weeks of eccentric exercise training.* J Sport Health Sci, 2022. **11**(1): p. 43-49.

39. Bullock, G.S., et al., *Just How Confident Can We Be in Predicting Sports Injuries? A Systematic Review of the Methodological Conduct and Performance of Existing Musculoskeletal Injury Prediction Models in Sport.* Sports Med, 2022. **52**(10): p. 2469-2482.

40. Molnar, C., *Interpretable machine learning. A Guide for Making Black Box Models Explainable*. 2019.

41. Bittencourt, N.F.N., et al., *Complex systems approach for sports injuries: moving from risk factor identification to injury pattern recognition—narrative review and new concept.* British Journal of Sports Medicine, 2016. **50**(21): p. 1309-1314.

42. Cohan, A., J.G. Schuster, and J. Fernández, *A deep learning approach to injury forecasting in NBA basketball.* Journal of Sports Analytics, 2021.

43. Majumdar, A., et al., *Machine Learning for Understanding and Predicting Injuries in Football.* Sports Med Open, 2022. **8**(1): p. 73.

44. Richter, C., M. O'Reilly, and E. Delahunt, *Machine learning in sports science: challenges and opportunities.* Sports Biomech, 2021: p. 1-7.

45. Bradley, A.P., *The use of the area under the ROC curve in the evaluation of machine learning algorithms.* Pattern Recognit., 1997. **30**: p. 1145-1159.

46. Ruddy, J.D., et al., *Modeling the Risk of Team Sport Injuries: A Narrative Review of Different Statistical Approaches.* Front Physiol, 2019. **10**: p. 829.

47. Fawcett, T. *ROC Graphs: Notes and Practical Considerations for Data Mining Researchers*. 2003.

48. Lobo, J.M., A. Jiménez-Valverde, and R. Real, *AUC: a misleading measure of the performance of predictive distribution models.* Global Ecology and Biogeography, 2008. **17**: p. 145-151.

49. Davis, J. and M.H. Goadrich, *The relationship between Precision-Recall and ROC curves.* Proceedings of the 23rd international conference on Machine learning, 2006.

50. Miller, T., *Explanation in Artificial Intelligence: Insights from the Social Sciences.* Artif. Intell., 2017. **267**: p. 1-38.

51. Ribeiro, M.T., S. Singh, and C. Guestrin, *"Why Should I Trust You?": Explaining the Predictions of Any Classifier.* Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.

52. Doshi-Velez, F. and B. Kim, *Towards A Rigorous Science of Interpretable Machine Learning.* arXiv: Machine Learning, 2017.

53. Lipton, Z.C., *The Mythos of Model Interpretability.* Queue, 2016. **16**: p. 31 - 57.

54. Shapley, L.S. *17. A Value for n-Person Games*. 1953.

55. Lundberg, S.M. and S.-I. Lee, *A unified approach to interpreting model predictions*, in *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 2017, Curran Associates Inc.: Long Beach, California, USA. p. 4768–4777.

56. Friedman, J.H. and B.E. Popescu, *PREDICTIVE LEARNING VIA RULE ENSEMBLES.* The Annals of Applied Statistics, 2008. **2**: p. 916-954.

57. Lou, Y., et al., *Accurate intelligible models with pairwise interactions.* Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, 2013.

58. Wachter, S., B.D. Mittelstadt, and C. Russell, *Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR.* Cybersecurity, 2017.

59. Ribeiro, M.T., S. Singh, and C. Guestrin. *Anchors: High-Precision Model-Agnostic Explanations*. in *AAAI Conference on Artificial Intelligence*. 2018.

60. Huang, Q., et al., *GraphLIME: Local Interpretable Model Explanations for Graph Neural Networks.* ArXiv, 2020. **abs/2001.06216**.

61. Frye, C., I. Feige, and C. Rowat, *Asymmetric Shapley values: incorporating causal knowledge into model-agnostic explainability.* ArXiv, 2019. **abs/1910.06358**.

62. Wang, J., J. Wiens, and S.M. Lundberg. *Shapley Flow: A Graph-based Approach to Interpreting Model Predictions*. in *International Conference on Artificial Intelligence and Statistics*. 2020.

63. Kim, B., O. Koyejo, and R. Khanna. *Examples are not enough, learn to criticize! Criticism for Interpretability*. in *NIPS*. 2016.

64. Simonyan, K., A. Vedaldi, and A. Zisserman, *Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps.* CoRR, 2013. **abs/1312.6034**.

65. Sundararajan, M., A. Taly, and Q. Yan, *Axiomatic Attribution for Deep Networks.* ArXiv, 2017. **abs/1703.01365**.

66. Hastie, T.J., R. Tibshirani, and J.H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd Edition*. in *Springer Series in Statistics*. 2005.

67. Schober, P. and T.R. Vetter, *Logistic Regression in Medical Research.* Anesth Analg, 2021. **132**(2): p. 365-366.

68. Rossi, A., et al., *Effective injury forecasting in soccer with GPS training data and machine learning.* PLoS One, 2018. **13**(7): p. e0201264.

69. Kharya, S. and S. Soni, *Weighted Naive Bayes Classifier: A Predictive Model for Breast Cancer Detection.* International Journal of Computer Applications, 2016. **133**: p. 32-37.

70. Bahel, V., S.K. Pillai, and M. Malhotra, *A Comparative Study on Various Binary Classification Algorithms and their Improved Variant for Optimal Performance.* 2020 IEEE Region 10 Symposium (TENSYMP), 2020: p. 495-498.

71. Zhang, H. *The Optimality of Naive Bayes*. in *The Florida AI Research Society*. 2004.

72. Arrieta, A.B., et al., *Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI.* Inf. Fusion, 2019. **58**: p. 82-115.

73. Hooker, G. and L.K. Mentch, *Please Stop Permuting Features: An Explanation and Alternatives.* ArXiv, 2019. **abs/1905.03151**.

74. Keane, M.T. and B. Smyth, *Good Counterfactuals and Where to Find Them: A Case-Based Technique for Generating Counterfactuals for Explainable AI (XAI).* ArXiv, 2020. **abs/2005.13997**.

75. Dandl, S., et al. *Multi-Objective Counterfactual Explanations*. in *Parallel Problem Solving from Nature*. 2020.

76. Orchard, J.W., H. Seward, and J.J. Orchard, *Results of 2 decades of injury surveillance and public release of data in the Australian Football League.* Am J Sports Med, 2013. **41**(4): p. 734-41.

77. Opar, D.A., et al., *A novel device using the Nordic hamstring exercise to assess eccentric knee flexor strength: a reliability and retrospective injury study.* J Orthop Sports Phys Ther, 2013. **43**(9): p. 636-40.

78. McGrath, T.M., et al., *Determinants of hamstring fascicle length in professional rugby league athletes.* Journal of Science and Medicine in Sport, 2020. **23**: p. 524-528.

79. Pollard, C.W., et al., *Razor hamstring curl and Nordic hamstring exercise architectural adaptations : Impact of exercise selection and intensity.* Scandinavian Journal of Medicine & Science in Sports, 2019. **29**: p. 706-715.

80. Hickey, J.T., et al., *Pain-Free Versus Pain-Threshold Rehabilitation Following Acute Hamstring Strain Injury: A Randomized Controlled Trial.* J Orthop Sports Phys Ther, 2020. **50**(2): p. 91-103.

81. Kellis, E., et al., *Validity of architectural properties of the hamstring muscles: correlation of ultrasound findings with cadaveric dissection.* J Biomech, 2009. **42**(15): p. 2549-54.

82. van Dyk, N., et al., *A comprehensive strength testing protocol offers no clinical value in predicting risk of hamstring injury: a prospective cohort study of 413 professional football players.* British Journal of Sports Medicine, 2017. **51**(23): p. 1695-1702.

83. van Dyk, N., et al., *Hamstring and Quadriceps Isokinetic Strength Deficits Are Weak Risk Factors for Hamstring Strain Injuries: A 4-Year Cohort Study.* Am J Sports Med, 2016. **44**(7): p. 1789-95.

84. Kohavi, R. and G.H. John, *Wrappers for feature subset selection.* Artif. Intell., 1997. **97**(1–2): p. 273–324.

85. Guyon, I., et al., *Gene Selection for Cancer Classification using Support Vector Machines.* Machine Learning, 2002. **46**(1): p. 389-422.

86. Gholamy, A., V. Kreinovich, and O. Kosheleva. *Why 70/30 or 80/20 Relation Between Training and Testing Sets: A Pedagogical Explanation*. 2018.

87. Hajian-Tilaki, K., *Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation.* Caspian J Intern Med, 2013. **4**(2): p. 627-35.

88. Meeuwisse, W.H., et al., *A dynamic model of etiology in sport injury: the recursive nature of risk and causation.* Clin J Sport Med, 2007. **17**(3): p. 215-9.

89. Franchi, M.V., et al., *Muscle Architecture Assessment: Strengths, Shortcomings and New Frontiers of in Vivo Imaging Techniques.* Ultrasound Med Biol, 2018. **44**(12): p. 2492-2504.

90. Franchi, M.V., et al., *Ultrasound-derived Biceps Femoris Long Head Fascicle Length: Extrapolation Pitfalls.* Med Sci Sports Exerc, 2020. **52**(1): p. 233-243.

91. Pollock, N., et al., *British athletics muscle injury classification: a new grading system.* Br J Sports Med, 2014. **48**(18): p. 1347-51.

92. Lu, Y., et al., *Machine Learning for Predicting Lower Extremity Muscle Strain in National Basketball Association Athletes.* Orthop J Sports Med, 2022. **10**(7): p. 23259671221111742.

93. Opar, D.A., et al., *Is Pre-season Eccentric Strength Testing During the Nordic Hamstring Exercise Associated with Future Hamstring Strain Injury? A Systematic Review and Meta-analysis.* Sports Med, 2021. **51**(9): p. 1935-1945.

94. Lee Dow, C., et al., *Prediction of Hamstring Injuries in Australian Football Using Biceps Femoris Architectural Risk Factors Derived From Soccer.* Am J Sports Med, 2021. **49**(13): p. 3687-3695.

95. Chen, T. and C. Guestrin, *XGBoost: A Scalable Tree Boosting System*, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016, Association for Computing Machinery: San Francisco, California, USA. p. 785–794.

SUPPLEMENTAL MATERIALS

Table 11: Supplemental Material 1 - The p-value of individual risk factors determined by multivariate logistic regression models in Analysis 1.

Model	Risk Factors*	p-value
d1->i1	Prior HSI	< 0.01
	Height	0.112
	Age	0.047
	Muscle thickness	0.267
	Intercept	0.948
d2->i2	Fascicle length	< 0.001
	Pennation angle	< 0.001
	Intercept	0.226
d3->i3	Prior ACL	0.999
	Height	0.809
	Age	0.322
	Pennation angle	0.316
	Fascicle length	0.322
	Relative eccentric knee flexor force	0.348
	Eccentric knee flexor force imbalance	0.293
	Intercept	0.448

*Risk factors were selected by recursive feature elimination and 5-fold cross validation.

Prior HSI: prior hamstring strain injury (HSI),

Prior ACL: prior anterior cruciate ligament (ACL) injury,

d1->i1: data assessed at start of pre-season and hamstring strain injuries (HSIs) that occurred in preseason, d2->i2: data assessed at end of pre-season and hamstring strain injuries (HSIs) that occurred in early in-season, d3->i3: data assessed in the middle of in-season and hamstring strain injuries (HSIs) that occurred in late in-season.