School of Computer Science and Informatics – CMT400, CMT403, CMT404
Dissertation Project Ethics Form

1. PROJECT SUMMARY					
Project Title:	**Injury Prediction in Professional Football: A Machine Learning (ML) Approach.**				
Student Name:	Gerard Glynn-Rivers				
a) Approved methods to be used:	[Secondary data analysis (no primary human data collection) ☒]				
b) Please provide a summary of your project, including its aims and the methods you will be using:					
This dissertation aims to develop a data-efficient and interpretable machine learning model to predict whether professional football players will sustain injuries. Using an existing dataset from Kaggle (Biswas, 2023), titled *Player Injuries and Team Performance*, the project applies supervised learning techniques—specifically, a decision tree classifier—to historical injury and performance data from professional footballers. The goal is to offer an alternative to current prediction models, enhancing interpretability for real-world football settings.					
c)	Do you have any potential conflicts of interest to declare? Information and guidance on conflicts of interest is contained in the Research Integrity Online Training Programme and the Research Integrity and Governance Code of Practice.	Yes ☐	No ☒		
d)	If you have answered 'yes' to b, please confirm the action you propose to take to address such conflicts:				
N/a					
e) What do you consider to be the ethical issues or considerations specific to your project? How do you intend to address these?					
Although the dataset relates to professional athletes it is intended for public academic and analytical use.					
f)	Are you planning to collect data from any users from outside of your group? This includes collecting data from social media. If No, skip to Section 2.	Yes ☐	No ☒		
g)	If working with human subjects, please confirm your participant group and how you will identify and recruit such participants:				
N/a					
h) How many participants are you aiming to recruit?		10 or less ☐	11-20 ☐	21-49 ☐	50+ ☐
i) Please describe the consent procedures/arrangements for the project – will informed consent be obtained from participants, and if so how? (If informed consent will not be obtained, then your project is NOT eligible for supervisor sign off.)					

2. IS YOUR PROJECT ELIGIBLE FOR SUPERVISOR-ONLY REVIEW?		
Does your project involve any of the following:	Yes	No
a) Activity within, or involving, the NHS or health and social care systems (e.g. patients, service users, staff, resources or data).	☐	☒
b) The collection or use of Human Tissue, as defined by the Human Tissue Act 2004	☐	☒
c) The use or involvement of animals.	☐	☒
d) The subject areas of terrorism, extremism or radicalisation. Note: these terms are defined in the University's Security Sensitive Research Policy.	☐	☒

e) Primary data collection without the prior consent of participants* *The following are specifically excluded from this criteria: (1) observation of people in a public space where no identifiable information is being captured; (2) use of publicly available information or secondary data where no identifiable information is being captured.	☐	☒
f) An element of deception, including covert activity.	☐	☒
g) Children under the age of 18 or 'at risk' (vulnerable) adults or groups.	☐	☒
h) Topic areas likely to be considered highly sensitive for participants. Note: examples include sexual behaviour, illegal activities, political, religious or spiritual beliefs, race or ethnicity, experience of violence, abuse or exploitation, and mental health.	☐	☒
i) Access to records of a sensitive or confidential nature, including 'Special Category Data' or criminal offence data.	☐	☒
j) Intrusive or invasive procedures Note: examples include the administration of substances, vigorous exercise, procedures which may involve pain or more than mild discomfort to participants (including the risk of psychological distress or anxiety – more than the risks encountered in everyday life).	☐	☒
k) More than a minimum risk of harm to the safety and wellbeing of participants and/or the student conducting the project	☐	☒
l) Detailed demographics or background information collected beyond age or gender. Age question asks for a specific age (as opposed to using age ranges).	☐	☒
m) Any other activity which clearly falls outside of the parameters of the Module Ethics Framework or raises other material ethical concerns?	☐	☒

If you have selected "Yes" to any of the above, your project is NOT eligible for Supervisor sign-off and must be reviewed by the SREC. Please do not continue with this form. You must instead complete the SREC's 'Application Form for Ethical Review'. See COMSC Ethics website for details.

If you have selected "No" to all of the above, your project is eligible for Supervisor sign-off and you must complete the rest of this form. Your project does not require a SREC application.

3. DECLARATIONS – For use where project IS eligible for Supervisor sign-off.	
STUDENT DECLARATIONS	
a. The information in this form is accurate to the best of my belief. I take full responsibility for it.	☒
b. I have read, and will adhere to, the Module Ethics Framework.	☒
c. I have completed the University's Research Integrity Training and believe we have the necessary skills and knowledge to conduct the project as proposed.	☒
d. I am familiar with the Ethics Policy for Human Research, Research Integrity and Governance Code of Practice and Academic Integrity Policy and will abide by all standards relevant to our project.	☒
e. [If collecting user data] I have submitted all final version participant-facing documents and data collection tools to our supervisor or the module leader for approval prior to commencing data collection.	☒
f. [If collecting user data] I will notify our supervisor or the module leader of any amendments to our project and seek approval of any substantial changes to participant-facing documents or data collection tools.	☒
g. I will notify our supervisor or the module leader of any ethical issues that arise during the project.	☒
Signed: G. Rivers	
Print names: Gerard Glynn-Rivers	

Date: 26/6/2025	
SUPERVISOR DECLARATIONS	
a. I have reviewed this form and I am satisfied that the project, as proposed, meets ethical standards.	☒
b. I have reviewed all participant-facing documents and data collection tools prepared by the group to ensure they meet ethical standards and the Module Ethics Framework.	☒
c. I have completed the University's Research Integrity Training within the last three (3) years and believe I have the necessary skills and knowledge to review the documents referenced above and to offer appropriate supervision in matters of ethics.	☒
d. I am familiar with the Ethics Policy for Human Research, Research Integrity and Governance Code of Practice and Academic Integrity Policy and I will provide support to the group in following all standards relevant to their project.	☒
e. I will encourage the group to discuss with me, and reflect on, any ethical issues that arise during the project and, where relevant, I will ensure such issues are notified to the Module Leader.	☒
Signed: A.Gibson	
Print name: Annelies Gibson	
Date: 02/07/2025	

Injury Prediction in Professional Football: A Machine Learning (ML) Approach

Gerard Glynn-Rivers

21073869

School of Computer Science and Informatics

This dissertation is submitted to
Cardiff University in partial fulfilment of
the requirements for the

MSc Computing 2025

Supervisor: Annelies Gibson

Declaration

I hereby declare that this dissertation is the result of my independent investigation except where I have indicated my indebtedness to other sources.

I hereby certify that this dissertation has not already been accepted in substance for any other MSc, nor is it being submitted concurrently for any other MSc.

I hereby give consent for my dissertation, if accepted, to be available for photocopying and consultation by future students and colleagues, and for the title and summary to be made available to outside organisations.

Signature:

CANDIDATE: Gerard Glynn-Rivers

Date: 01/09/2025

Abstract

Injuries in professional football have a plethora of profound consequences, from diminished on-pitch performance to financial and psychological strain (Bangert et al. 2024; Crossman 1997; Ekstrand 2013). Therefore, effective risk management strategies that emphasise prediction can help to mitigate these impacts (Spikin 2013). This research project aims to develop a Machine Learning (ML) model that predicts whether an English Premier League player will sustain an injury in a Premier League season. This study aims to (1) critically review current literature on football injury risk factors and existing prediction methodologies; (2) design and implement an interpretable Decision Tree Classifier capable of predicting whether a player will sustain an injury; and (3) evaluate the model's predictive performance, interpretability, and potential application for injury prevention.

A secondary data analysis was conducted using a publicly available dataset titled Player Injuries and Team Performance that contained a vast range of information such as a players age, playing position as well as a count of their previous injuries from the 2019/2020 to the 2023/24 English Premier League seasons. A Decision Tree Classifier was chosen and GroupKFold cross-validation and GridSearchCV were used to optimise hyperparameters to mitigate overfitting and underfitting (Dremio 2025; Guerra 2024).

I managed to create a ML model that achieved an accuracy of 96.3%, 0.955 precision, 1.000 recall, and 0.977 F1-score, that was able to correctly identify injured players with minimal false positives. This research directly contributes to existing literature as it helps to address limitations previously identified by Claudino et al. (2019), as this model offers practical interpretable recommendations that coaches and medical staff can act upon. Furthermore, by incorporating temporal separation between training and test data, a limitation noted by Bullock et al. (2022) and Colby et al. (2014) has been overcome, providing a more realistic evaluation of model performance in applied settings.

Acknowledgements

I would like to take a moment to express gratitude to my dissertation supervisor, Annelies Gibson, who helped to make this dissertation project possible. Annelies provided me with a lot of great guidance and support throughout the whole project and without her I know that this project would have not been completed.

Contents

Declaration i
Abstract ii
Acknowledgements iii
List of Tables vi
List of Figures vi
List of Abbreviations vii

1. Introduction 1

2. Literature review 3

2.1. Injuries in Football 3
2.2. Intrinsic Risk factors 4
2.2.1. Biomechanical and Neuromuscular Risk Factors 4
2.2.2. Genetic Susceptibility 5
2.2.3. Psychological and Cognitive Factors 6
2.2.4. Age 8
2.2.5. Previous Injury history 9
2.3. Extrinsic risk factors 10
2.3.1. Training Load and Session Types 10
2.3.2. Matches and In-Match Situational Risk 12
2.3.3. Environmental Conditions and Playing Surfaces 13
2.3.4. The impact of a Player's Position in Football in Sustaining Injuries. 14
2.4. The Impact of Sustained Injuries 15
2.4.1. The Financial and Performance Impact That Injuries Have on Clubs 16
2.4.2. The Developmental and Career Impacts Injuries Have on Players 16
2.4.3. The Psychological Consequences of Injuries 17
2.4.4. The Long-Term Effects of Sustaining an Injury for a Player 18
2.5. Traditional Methods of Injury Prediction 19
2.5.1 Current/ Traditional Injury Prediction Methods 19
2.6. Machine Learning in Football Injury Prediction: Current Applications 20
2.7. Justification 21

3. Methodology 22

3.1. Research Design 22
3.2. Dataset Overview and Preprocessing 23
3.3. Feature Engineering and Variable Transformation 30
3.4. Model Selection and Justification 30
3.5. Model Evaluation Metrics 32

3.6. Ethics 34
3.7. Limitations 34
3.8. Reflexivity 35

4. Findings 36

5. Discussion 39

5.1. Summary of My Findings 39
5.2. Limitations of My Research 40
5.3. Future Research Directions 40

6. Conclusion 41

7. Reflection 41

8. References 43

List of Tables

Table title	Page
Table 1: The key features retained after dataset cleaning.	29

List of Figures

Figure title	Page
Figure 1: A snapshot of the initial Dataset by Biswas (2023) prior to cleaning titled Player Injuries and Team Performance.	24
Figure 2: A cross-section of the dataset after preprocessing and cleaning, where irrelevant features have been removed, ambiguous injury types have been excluded, and binary target variables have been created for classification	25
Figure 3: Number of injuries by player position in the training dataset	26
Figure 4: Number of injuries by age group in the training dataset (2019/20–2022/23)	27
Figure 5: Number of injuries by age group in the test dataset (2023/24)	28
Figure 6: Number of injuries by position in the test dataset (2023/24)	28
Figure 7: Code showing the use of Scikit-learn's LabelEncoder() to convert categorical variables (player position) into numerical format	30
Figure 8: Python code implementing hyperparameter tuning using GridSearchCv	31
Figure 9: Confusion matrix heatmap showing true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) based on predictions made by the final Decision Tree model on the 2023/24 test set	33
Figure 10: Performance metrics (Accuracy, Precision, Recall, F1-Score) and the results of the final model	36
Figure 11: Confusion matrix heatmap based on predictions made by the Decision Tree model	37
Figure 12: Terminal Output Displaying the Three Misidentified Cases	37
Figure 13: Visual Representation of the Final Decision Tree Model for Injury Prediction	37

List of Abbreviations

Abbreviation	Explanation
ML	Machine Learning
AI	Artificial Intelligence
ACL	Anterior Cruciate Ligament
H: Q	Hamstring-to-Quadriceps ratio
ACWR	Acute: Chronic Workload Ratio
SVM	Support Vector Machine
GPS	Global Positioning System
TP	True positives
TN	True negatives
FP	False positives
FN	False negatives
IIR	Injury Incidence Rate
CTE	Chronic Traumatic Encephalopathy
CSV	Comma-Separated Values
PFA	Professional Footballers' Association
FIFA	Fédération Internationale de Football Association
AMPD1	Adenosine Monophosphate Deaminase 1
ACTN3	Alpha-actinin-3
ACE I/D	Angiotensin-Converting Enzyme Insertion/Deletion polymorphism
COL5A1	Collagen Type V Alpha 1 Chain

1. Introduction

“Football has become a more demanding sport in recent years… as a result, injury prevention is more important than ever” Inside FIFA (2025, para.1).

In recent years professional football has become a more demanding sport according to Inside FIFA (2025) because players cover around 10–13 km (about 8.08 miles) per match (Di Mascio and Bradley 2013). Where high-intensity movements, rapid changes of direction, sprints and more will be performed consistently (Gurau et al. 2023). All of which contribute to injuries (Gurau et al. 2023). Though perceived to be a safe sport, Venturelli et al. (2011) evidence how football has a higher injury rate per hour of exposure than both Rugby and American football (sports considered dangerous), as players are 1000x more likely to suffer injuries playing than workers in the most dangerous industrial professions (Drawer and Fuller 2002).

Impacts from injuries are wide-ranging and can result in a reduction in a team’s overall performance (Hägglund et al. 2013), a player’s career progression being hindered (Bangert et al. 2024) and a player’s long-term health being affected (Salzmann et al. 2017). Financially the implications are equally as severe, as Ekstrand (2013) estimates that a single injury can cost a club up to £500,000 per month, a figure driven by continued salary payments during a player’s unavailability (PFA 2023). Because of this Genovard et al. (2025) highlights how injury prevention and prediction are no longer just medical concerns, but strategic imperatives for football clubs.

Because of the wide-ranging impacts (of injuries), the importance of injury prediction methods in football cannot be understated, but despite the urgency, many methods remain limited in their effectiveness. Traditional approaches based on regression assume linearity and thus fail to capture the complex and non-linear interactions between multiple factors (Bittencourt et al. 2016; Bullock et al. 2022). A key component of predicting injuries, given that injuries result from range of physical, environmental, socio-cultural, and psychological variables (Wiese-Bjornstal et al. 1998; Ruddy et al. 2019).

In contrast, ML models offer an adaptive, non-linear data driven alternative that can detect non-linear patterns in large, multidimensional datasets while simultaneously improving its own predictive power as it learns from new data (Chen et al. 2024: Rossi et al. 2021). Yang et al. (2022) showed ML models outperforming logistic regression in a large medical risk prediction context, supporting the broader methodological premise, yet Bullock et al. (2022) reports that 60% of injury prediction studies still rely on traditional techniques.

Despite promising results from ML in sports injury prediction, several reviews and empirical studies highlight several limitations. For instance, Claudino et al. (2019) note that many models lack interpretability, producing outputs that coaches and medical staff cannot readily act upon. Likewise, Colby et al. (2014) emphasise that most studies fail to incorporate temporal validation, limiting their ability to forecast season-long injury risk. Bullock et al. (2022) report that models still rely on traditional statistical techniques, constraining their practical application in professional football. Together, these findings reveal a clear gap in the literature: the absence of injury forecasting tools that are both statistically robust and applicable in professional football contexts.

The aims of this study are therefore to: (1) critically review current literature on football injury risk factors and existing prediction methodologies; (2) design and implement a Decision Tree Classifier capable of predicting whether a player will sustain an injury in a Premier League season; and (3) evaluate the model's predictive performance, interpretability, and potential application for injury prevention strategies, while situating the findings within the broader academic and professional discourse on football injury prevention.

2. Literature Review

This literature review will explore the multifaceted nature of football injuries by first examining the various definitions that constitute an injury. It will then investigate the multifactorial nature of injury risk, considering both intrinsic and extrinsic factors. I will evaluate the current approaches to injury prediction in football, highlighting their limitations while stating how ML may be a more beneficial substitute. Finally, this chapter will end by exploring how ML has already been applied in football, where gaps will be identified that this dissertation research project will look to address.

2.1. Injuries in football

Before injury prediction methods can be explored, a consistent definition of the term injury must be first defined. However, there is no universal definition for the term as many variations exist, which present significant challenges. For example, Junge et al. (2000) define a football injury to be an absence either from matches or training, which is then followed by the need for medical treatment(s). While this is a concise definition for the term, it can be misleading. As absences from matches and training can be caused by several other factors like rest, personal reasons, or even in some cases tactical rotation from a manager (Junge et al. 2000).

In contrast, the council of Europe cited in van Vulpen (1989) provides a broader and more inclusive definition as they define the term to be any physical harm resulting from participation in official football activity, where the consequences range from reduced performance and medical treatment(s) to socio-economic disruptions. Though this is more holistic, this definition introduces subjectivity especially when evaluating injuries that may be less severe.

Despite the lack of consensus on a definition, football injuries are classified by whether they were caused by contact with another player or through non-contact means. A contact injury arises from a physical interaction between players (Gurau et al. 2023). Peterson et al. (2000) found that half of all injuries in their sample were contact related, with Drawer and Fuller (2002) finding that football's injury rate is over 1,000 times higher than the most dangerous Industrial jobs.

Conversely, non-contact injuries occur when a player sustains an injury without any impact from another player or object(s) (Yu and Garrett 2007). Beaumont Health (2024) estimates that 70% of all football injuries sustained are non-contact, with anterior cruciate ligament (ACL) tears taking up a considerable proportion (Boden et al. 2000). Gurau et al. (2023) evidence that even high intensity movements, rapid changes of direction, accelerations and even kicking

the ball can result in injury. These injuries are relevant to predictive modelling, as they may be associated with biomechanical, physiological, or even load-related markers (López-Valenciano et al. 2020; Massidda et al. 2020).

Injuries sustained in football can also vary in their severity. Severity is classified by the number of days between the date of the injury to the date of a player's return to matches and training (Fuller et al. 2006). According to Fuller et al. (2006) the severity of an injury is grouped into the following categories: slight (0 days out), minimal (1–3 days out), mild (4–7 days out), moderate (8–28 days out), severe (>28 days out), and then finally career-ending.

A thorough understanding of injury epidemiology is crucial for interventions to be introduced (van Mechelen et al. 1992). Ekstrand, Hägglund and Waldén (2011) found that 11% to 35% of football injuries sustained are classed as severe and exceed three to even four weeks. With reoccurring injuries not only being common but also associated with a much greater severity (Ekstrand et al. 2011; Ekstrand, Hägglund and Waldén 2009).

Injuries whether sustained from contact or not are driven by a complex interplay of intrinsic and extrinsic factors. Intrinsic factors refer to player-specific factors like age, physical fitness, and injury history (Meeuwisse et al. 2007), while extrinsic factors are external conditions, including training loads and environmental elements (Meeuwisse et al. 2007). The multifactorial nature of injuries underpins the need for an ML model that has multidimensional input features.

2.2. Intrinsic Risk Factors

A player's risk of sustaining an injury is not solely determined by external factors like training loads or pitch conditions, as a substantial body of literature names intrinsic factors as having a key role (Taimela et al. 1990; Meeuwise et al. 2007). Factors that significantly increase injury susceptibility include age, earlier injury history, physical fitness, as well as genetic predisposition (Taimela et al. 1990).

2.2.1. Biomechanical and Neuromuscular Risk Factors

Biomechanical and Neuromuscular risk factors have been identified as contributors to a player's risk of injury, as muscle imbalances, joint laxity or deficits in Neuromuscular control have been highlighted as critical predictors of soft tissue and ligamentous injuries in footballers (Prys et al. 2023).

The most common problem in players is asymmetry between antagonistic muscles as Sugimoto et al. (2015) found that players who experience both Neuromuscular and

Biomechanical deficiencies were likely to obtain ACL injuries. Furthermore, those with weak hip strength had an increased risk of groin injury (Markovic et al. 2020), which is supported by DeLang and Ishøi (2024) who report that groin pain accounts for up to 90% of all groin-related problems in football.

Players found to have joint laxity issues were more susceptible to knee injuries, as joint laxity is a significant predictor of traumatic leg injuries (Östenberg and Roos 2000; Söderman et al. 2001). However, Baumhauer et al. (1995) found no significant correlation between generalised laxity and ankle sprains stating that joint specific factors as well as demands may play a moderating role. Highlighting the need for player specific approaches rather than a one size fits all approach.

Football related movements are predominantly quadriceps-dominant meaning they can lead to a muscular strength imbalance (Brophy et al. 2007). Brophy et al. (2007) found that repeated high load quadricep engagement suppresses hamstrings activation, which increases strain during eccentric contractions. Leading to more non-contact knee injuries, which coincides with Ekstrand et al. (1983) and Read et al. (2016) as they found that non-contact knee injuries were more prevalent in players who had lower hamstring activation in comparison to their quadriceps. In addition, Söderman et al. (2001) noted that both excessively low and high hamstring-to-quadriceps (H: Q) ratios were problematic (the former with traumatic injuries, the latter with overuse issues), suggesting that injury prevention strategies must aim for muscular symmetry and not just strength gains. Myontec (2023) further emphasises this as muscle imbalances greater than 15% may serve as early warning signs for knee injuries.

While these asymmetries go unnoticed in standard fitness assessments, ML models trained on real-time biomechanical items can be used to flag muscle imbalances which offers clubs a proactive tool for injury risk predictions. However, collecting such data requires technology that is not available for all players across multiple seasons in public datasets, which is why this variable was excluded.

2.2.2. Genetic Susceptibility

Though biomechanical and neuromuscular risk factors dominate football injury literature, there has been a shift toward investigating genetic predisposition as a contributor to a player's injury risk (McCabe and Collins 2018). As genetics can influence a plethora of things within the human body, all of which alter a player's vulnerability (McCabe and Collins 2018; Varillas-Delgado 2024).

An abundance of literature has investigated how genetics play a role in a player's injury susceptibility (Massidda et al. 2020; Prys et al. 2023; Lv et al. 2017). For instance, Massidda et al. (2020) found that footballers who had the ACTN3 II (Alpha-actinin-3) genotype (associated with slower muscle fibre contractions) were likely to have non-contact muscle strains. Likewise, the ACE I/D (Angiotensin-Converting Enzyme Insertion/Deletion) polymorphism (which affects levels of the angiotensin-converting enzyme) has been associated with increased post-exercise inflammation and impaired tissue recovery, especially in athletes with the I/D variant (Massidda et al. 2020; Prys et al. 2023). These biological differences impair recovery and increase the likelihood of soft tissue strains or tears.

Further evidence highlights how variants in genes that handle collagen synthesis such as COL5A1 (Collagen Type V Alpha 1 Chain), have been linked to injuries in both the ligaments and the tendons (Lv et al. 2017). Lv et al. (2017) found a correlation between the COL5A1 variant and a higher risk of tendon ruptures, while Brazier et al. (2025) highlighted that the T-C haplotype at rs12722 and rs3196378 increased the likelihood of a soft tissue injury. Likewise, Collins (2017) found that the gene AMPD1 (Adenosine Monophosphate Deaminase 1) was linked with a need for increased recovery after hard activity, with those who had extra rest performing significantly better.

However, it is still important to note that genetic predisposition is not deterministic as it merely interacts with other factors such as training load, recovery practices, and playing surfaces according to Bıçakçı, Cięszczyk and Humińska-Lisowska (2024).

These findings raise opportunities as well as concerns, as genetics could serve as a stable non-volatile input feature for injury risk prediction (McCabe and Collins 2018). But on the other hand, the sensitivity of genetic data presents a limitation for real-world implementation as the reliability, safety, and ethical use of ML in clinical settings is paramount to supporting patient trust and achieving fair healthcare outcomes (Olawade et al. 2025). Because of this, the study will exclude genetic features from the ML model, not because of their lack of relevance, but due to the practical inaccessibility and the ethical complexity surrounding their use.

2.2.3. Psychological and Cognitive Factors

Beyond physical factors, psychological and cognitive aspects are also recognised as key contributors to injury susceptibility. The sport demands not only an elevated level of technical skill but also robust psychological regulation due to it's high-pressured nature (Maddison & Prapavessis 2007). Players who struggle to cope with stressors, experience anxiety, or display emotional volatility may suffer from compromised decision-making, lapses in concentration, and altered motor control all of which increase susceptibility to injury (Ivarsson & Johnson 2010).

Williams and Andersen (1999) show how players that are affected by or have previously experienced elevated levels of stress will be more prone to injury, especially when their coping resources are limited. Supported by Ivarsson and Johnson (2010) who say that players with elevated stress and poor coping strategies sustained significantly more injuries across a season. Likewise, Dvorak et al. (2000) noted that previously injured players scored higher on life-event stress scales compared to those uninjured peers, implying a potential loop between stress and injury recurrence. Junge et al. (2000) further highlights how footballers who have had at least seven previous injuries had more worries about their performance, more competitive anxiety, more peaking under pressure, a higher anger trait, and had increased outward anger expression.

Anxiety was shown as a key mediator in the player's susceptibility to injury, with Junge (2000) linking pre-competition anxiety to a significant increase in injury incidence, while Gracz and Sankowski (2000) suggested that players under pressure may either overexert themselves or lose focus at critical moments leading to injury. Kaplánová (2024) expands on this, highlighting how anxiety can impair cognitive processes such as attention, memory, and decision-making, all of which are crucial for performance and injury avoidance in fast-paced scenarios (MBP 2025).

While the connection between stress and injury is well supported in literature (Junge 2000; Kaplánová 2024; Gracz and Sankowski 2000; Andersen and Williams 1999; Ivarsson and Johnson 2010; Maddison and Prapavessis 2007), the causality is still up for debate as many studies rely on self-reported stress levels which can be inherently subjective and influenced by bias (Lira et al. 2022). Maddison and Prapavessis (2007) call attention to the fact that the directionality of the relationship is one that is complex, as psychological stress may increase injury risk but sustaining an injury can also elevate stress and anxiety (Ivarsson and Johnson 2010; Junge et al. 2000), which creates a reciprocal cycle.

Acknowledging psychological risk factors may be crucial to building a holistic, athlete-centric ML model that can predict injuries, as the analysis of behavioural and physiological signals may prove to be a key indicator of predicating injuries. But mental states can vary dramatically from one day to the next (Cherry 2025), complicating their predictive value in the ML model I am looking to create, so they will be left out.

2.2.4. Age

Age is another intrinsic risk factor in football injuries frequently cited in literature (Lindenfeld et al. 1994; Inklaar et al. 1996; Backous et al. 1988; Östenberg and Roos 2000), but the association seems to vary across developmental stages and even in gender (Backous et al. 1988; Alahmad, Kearney, and Cahalan 2020).

For those younger, there is evidence that the risk of injury increases as you age, with Backous et al. (1988) revealing that players aged 14 and above were twice as likely to sustain injuries compared to those younger. Comparably, Inklaar et al. (1996) evidence that players aged 17-18 experience injury rates comparable to or even higher than those found in adults. This is attributed to rapid musculoskeletal growth during puberty (Wild, Steele and Munro 2013), the ever-increasing training and match intensities (López-Valenciano et al. 2020) as well as poor neuromuscular control (Lehnert et al. 2017).

While those in their youth are at risk, emerging data indicates that players in their twenties are at increased risk too. According to Howden (2022), this age group accounted for the highest volume of football injuries worldwide, with 1875 reported cases, an increase of 324 from the year before. To many this will appear strange, as this age is when a footballer reaches their physical peak (considered to be between the ages of 27 and 29 [Carter 2014]) but the effects of intense training and match loads, increased physical competition and multi season fatigue are known to elevate injury susceptibility (Halvorsen Wik 2022; Faigenbaum and Myer 2010). Argibay-Gonzalez et al. (2022) support this, as they found that players aged 26-30 had the highest rates of injury in the Spanish La Liga League and the English Premier League, representing 47.9% and 51.9% of all total injuries, with the most frequently injured body regions being the thigh (35.6–44.4%), the knee (13.3–13.4%), and the ankle (10.6–17.8%).

Though injury prevalence peaks between the ages of 21 and 30 and then declines among those over 30 it may be due to the smaller proportion of over 30-year-old players present in professional squads rather than indicating a lower susceptibility (Argibay-Gonzalez et al. 2022). Sonesson, Lindblom and Hägglund (2023) highlight how players that start a season with a pre-existing injury are significantly more likely to sustain another injury, with that risk becoming greater as they age. Furthermore, Older players are subjected to higher match demands, including more intense accelerations, decelerations, and total distance covered all of which can increase a player's risk (Buchheit, Mendez-Villanueva and Simpson 2010; Della Villa et al. 2018; Gurau et al. 2023).

Likewise, those over 30 are more likely to have the burden of previous injuries, which when combined with elevated levels of match and training exposure only places further strains on their bodies compared to someone younger with no previous injury (Read et al. 2018).

Currently, the injury incidence rates (IIRs) in adult footballers can reach up to 36 injuries per 1,000 hours of match play, which is particularly impactful for older players with reduced physiological capacity for recovery (López-Valenciano et al. 2020).

Gender adds a further layer of complexity, as Alahmad, Kearney, and Cahalan (2020) report that older players in professional women's football experience a higher incidence of injuries compared to those younger. Conversely, in professional men's football, age-injury correlations are less consistent, as the association between age and injury tends to vary per study (Argibay-Gonzalez et al. 2022; Árnason et al. 2004). Whereas Östenberg and Roos (2000) found that players over the age of 25 were at a significantly higher risk of injury, other studies reported conflicting results (Lindenfeld et al 1994). This inconsistency underscores the fact that age may act more as a proxy variable shaped by many underlying factors rather than being a standalone indicator (Costa E Silva, Teles and Fragoso 2022).

Age was retained as a variable, as it has been identified as a significant factor influencing injury risk. Age was consistently available in publicly accessible datasets, and the player's ages are known making it a practical and reliable variable to incorporate.

2.2.5. Previous Injury History

Although age is a determining factor in whether a player will sustain an injury, several authors have identified previous injury(s) as a being a critical contributing factor for determining whether a player will be at risk of sustaining a future injury (Fulton et al. 2014). Furthered by Hägglund, Waldén and Ekstrand (2006) and Arnason et al. (2004) who established that players who have a history of hamstring injuries were 4 to 6 times more likely to experience a recurrence compared to those with no previous injury. Similarly, Arnason et al. (2004) reported comparable recurrence risk for hamstring injuries, showing how incomplete rehabilitation, residual weakness, or even scar tissue can predispose players to reinjury.

Previous injury has a predictive value that is not limited to certain injuries, as Ekstrand and Gillquist (1983) state how players who have experienced previous ligament injuries (particularly those who have experienced past ankle and knee sprains) were at a significantly higher risk of a similar future injury occurring because of lingering joint instability (that arose from that prior injury). Similarly, Delaney et al. (2000) found that players with a history of concussion were more likely to sustain subsequent concussions later, pointing to a long-term neurological vulnerability because of the initial impact.

Hawkins and Fuller (1999) say that reoccurring injuries account for around 22% and 42% of all football-related injuries. Muscle strains and ligament sprains are the most common reoccurring injuries of all (6–61%) and account for approximately 30% of all injuries sustained

(Inklaar 1994; Hawkins and Fuller 1999). Likewise, ankle sprains are the second most common reoccurring injuries, accounting for around 32–56% of all ankle-related injuries (Ekstrand and Gillquist 1983; Nielsen and Yde 1989).

Previous injury history was included as a feature in the model, as it is one of the most cited injury risk factors. This variable was available in the publicly accessible datasets, enabling its direct integration into the model.

2.3. Extrinsic Risk Factors

While intrinsic factors describe what predisposes an athlete to injury internally, extrinsic risk factors refer to the external forces and contextual demands imposed on the player (Mahmood and Pettinato 2021). These include training loads, match congestion, pitch conditions, environmental factors, and tactical roles (Mahmood and Pettinato 2021). A plethora of literature shows that extrinsic risk factors play a crucial role in injury susceptibility (Meeuwise et al. 2007; Saragiotto, Di Pierro and Lopes 2014).

2.3.1. Training Load and Session Types

Training load is one of the most cited extrinsic risk factors because of the fatigue, overuse, and inadequate recovery time placed on a player (Lathlean et al. 2020). The acute: chronic workload ratio (ACWR) is a commonly used framework to monitor this balance helping to compare short-term training intensity (7-day acute load) against longer-term averages (28-day chronic load) (White 2025). Spikes in this ratio, when acute load dramatically exceeds chronic conditioning, have been shown to correlate with higher injury (Andrade et al. 2020).

Despite this the effectiveness of the ACWR ratio as a predictive method has been debated, as while it can be useful in highlighting risks, the one-size-fits-all approach does not appear to prevent health problems in athletes (Dalen-Lorentsen et al. 2021). Lathlean et al. (2020) says that the volume of training matters less than how load is distributed and adapted to over time suggesting that models based on workload must be contextualised by a player's individual metrics.

The type of training performed can also impact a player's susceptibility to injury (Lathlean et al. 2020). According to Soccer Supplement (2022), football training can be divided into three broad categories: physical, technical, and tactical. Each type of training plays a unique role in the development and performance of a player but if mismanaged, can contribute to overuse, fatigue, and an increased likelihood of injury (Lathlean et al. 2019).

Physical training focuses on developing a player's core physiological attributes like Cardiovascular fitness, Muscular strength and endurance and includes activities like resistance training, sprint drills, and high-intensity intervals (Soccer Supplement 2022). Though essential for performance this kind of training can pose various risks, with weight training yielding 1–7.5 injuries per 1,000 hours and running training reaching 12.1 injuries per 1,000 hours, much of which stems from cumulative stress and poor recovery (Keogh and Winwood 2017; van Mechelen 1992).

Then there's technical training which differs from physical, as it introduces injury risk through high-repetition movements such as rapid accelerations, direction changes, and ball striking (Soccer Supplement 2022). While less intense these movements can still cause various traumatic and overuse injuries (Gurau et al. 2023).

Subsequently, tactical training concentrates on team structures, set plays, and is about in-game decision-making (Soccer Supplement 2022). It involves full-pitch simulations, transitions, and pressing drills, which place many neuromuscular and cognitive demands on players (Skala and Zemková 2023). Although this training is vital for competitive performance, these sessions can raise injury risk due to the physical stress (Taylor et al. 2020).

Critically the timing and structure of these sessions can lead to a build-up of cumulative fatigue which increases a player's injury risk (Faigenbaum and Myer 2010). Inadequate rest periods and poor load management between high load training can lead to symptoms of overtraining syndrome, including emotional exhaustion (Djaoui et al. 2017). Further, Bowen et al. (2017) found that high acceleration counts were predictive of soft-tissue injuries, especially in the youth.

Those who are younger may be particularly at risk, as Jones and Williams (2024) report that 34% to 72% of all football injuries in players that are in age groups ranging from under-9 to under-21 occur in training. This reflects both inadequate load management and underdeveloped recovery systems that Rodrigues et al. (2023) and Carfagno and Hendrix (2014) suggest are prevalent in many football academies around the world. 35% of football academy athletes show signs of overtraining by adulthood underscoring the need for tailored training plans based on age, biological maturation, and positional demands (Rodrigues et al. 2023; Carfagno and Hendrix 2014).

ACWR, acceleration counts, and training type frequency could all serve as useful inputs for forecasting models but the variability in how clubs measure and define these metrics along with the fact that this data is not available in public datasets means that this variable had to be excluded.

2.3.2. Matches and In-Match Situational Risk

While training places a player at risk of sustaining an injury, matches present one of the highest injury burdens on players (Gurau et al. 2023), as injury rates during competition/ matches are 4–6 times higher than in training (Gurau et al. 2023), with approximately 36 injuries per 1,000 match hours compared to just 3.7 per 1,000 in training (López-Valenciano et al. 2020; Ekstrand et al. 2011). Reflecting a complex interplay of factors including fixture congestion, intensity spikes, fatigue accumulation, and situational unpredictability (Dinnery 2024; Bengtsson et al. 2013; López-Valenciano et al. 2020).

One reason matches place a high injury burden on players is because professional footballers will compete and play in 40–80 matches over a 40-week season, participating in two to three games per week because of overlapping domestic, continental, and international commitments (Carling et al. 2016). For instance, in the 2020-21 football season, Bruno Fernandes (current Manchester United footballer) played 73 games, several of which went to extra time (Belev 2022). In extra time a game continues for 30 minutes and has the potential to continue further if the score is still tied, where it will then be decided in a penalty shootout (Unisportstore 2025). Bengtsson et al. (2013) and Ekstrand et al. (2011) found that short recovery windows (fewer than four days between matches) significantly elevated the risk of muscle injury, while players who have 6–10 days of recovery exhibited 20% fewer injuries. Because of this Bengtsson et al. (2013) believe that a schedule where there are fewer matches, but more training sessions could lead to less injuries.

In addition to the high frequency of matches, the intense and demanding nature of a match is another reason players sustain more injuries (Titans Football Academy 2023). Players consistently perform high-speed runs, rapid changes of direction and powerful movements, all of which elevate injury risk (Bangsbo, Nørregaard and Thorsø 1991; Bangsbo, Mohr and Krustrup 2006). Klein et al. (2021) observed that over 70% of non-contact injuries stem from structural overexertion, where unlike in training matches are less controlled, and players may push beyond their physical limits (London Bridge Orthopaedics).

Matches also introduce situations where risk of an injury is heightened, as certain match situations cannot be simulated in controlled environments like training (López-Valenciano et al. 2020). In a match each player's performance is directly influenced by the environment (e.g. teammates, opponents, and the ball), so a player is always adapting and reacting to assure their situational efficiency (Mijatovic et al. 2022). This reactivity increases the likelihood of unplanned turns, landings, and collisions, all of which increase the likelihood of injury (Grippy Sports 2023). This reactive style of performance has been intensified by the increasing pace of the game, as data from the Premier League has shown that the sprinting distance (above 25 km/h) has increased by 35% over the past decade, paralleling a 15% rise in injury rates in the 2022–23 Premier League season (Mijatovic et al. 2022; Dinnery 2024).

Furthermore, contact events like tackles, duels, and aerial challenges are also some of the most hazardous match actions. 46.5% of all contact injuries come from collisions with opponents, including aggressive tackles, aerial challenges, and even off-the-ball collisions (Mijatovic 2022). Tackles alone are responsible for up to 50% of all injuries, particularly to the lower body according to Saal (1991). Lower body injuries account for 70.7% of all injuries that occur with thigh (21.3%) and knee injuries (15.8%) being the most frequent (Klein et al. 2020).

Aiello et al. (2024) found that 3% of all duels (players from opposite teams competing, below elbow height, to gain possession [Wyscout 2024]) resulted in injury, while 0.4% required medical intervention. Aerial duels (defined as contests for the ball above shoulder height [FIFA 2021]) involving head-to-head collisions or impacts with elbows or knees, are a leading source of concussion, cranial trauma and injuries to the elbow, foot, forearm, knee, and shoulder (Andersen et al. 2004). Despite growing awareness of concussion protocols, these remain difficult to predict (López-Valenciano et al. 2020).

Interestingly, even non-contact ball-related actions pose a substantial risk of causing injuries (Aiello et al. 2024). Aiello et al. (2024) noted that 25% of injuries occur while passing or receiving the ball, and this risk is heightened when ball-related actions are performed at a high velocity (e.g. receiving a pass while running, passing the ball while running, and progressing with the ball at speed). Challenging the assumption that contact is the primary cause of match-related injury.

From the perspective of ML, contact events present a modelling challenge. While factors like tackles and duels are less predictable, event-based data (e.g. dual frequency, high-speed running during possession, aerial engagement rate) may provide indirect proxies for risk estimation. However, in this study, match-derived variables will not be used, as their inclusion would require access to detailed performance-tracking data that was not available.

2.3.3. Environmental Conditions and Playing Surfaces

Besides the physical demands of a football match, the environmental conditions significantly influence the risk of injuries (Prys et al. 2023). Gould et al. (2023) and Burton-Hughes (2016) found that slippery surfaces increase the incidence of lower extremity injuries (such as ankle sprains and non-contact knee ligament tears) as they lead to non-contact falls because players make quick turns and unexpected shifts (Grippy Sports 2023). While artificial turf is used to mitigate the effects of the weather, some studies report higher rates of injuries on synthetic surfaces compared to grass, especially when wet (Gould et al. 2023; Scranton et al. 1997). Moyes (2024) highlights how a considerable amount of football injuries stem from a direct impact like a fall, with the most common being caused by slipping in the rain (Wermert 2019).

Likewise, temperature conditions will also exacerbate the physiological implications on the players, as cold temperatures can reduce muscle elasticity and perfusion while increasing stiffness and the likelihood of muscular strain (Blomstrand et al. 1984). Conversely, elevated temperatures will increase a player's risk of injury as playing in the heat will raise the temperature of the body faster which simultaneously increases the risk of heat illness (Brito et al. 2014; Nybo et al. 2017 cited in Bangsbo 2016). In the heat the body must balance oxygen delivery between the working muscles and the skin for thermoregulation which leads to earlier fatigue and a diminished performance (Hosokawa et al. 2018). Many muscles in the body are susceptible to injury in the heat as the muscle temperatures will rise when playing football in hot conditions (William Hill 2022).

Many professional players mitigate these risks through methods like heat acclimatisation, hydration strategies, and modified pacing (Middleton 2025; Nassis et al. 2015; Adams, Hosokawa, and Casa 2016; Racinais et al. 2015), but this is less feasible at non-professional or even youth levels where these methods are unavailable.

While weather conditions and pitch surface type have been linked to variations in injury incidence, these variables are rarely recorded in publicly available datasets. The absence of consistent, environmental, and surface data across all seasons in this study meant that these features could not be incorporated.

2.3.4. The Impact of a Player's Position in Football in Sustaining Injuries

While environmental conditions play a critical role in a player's susceptibility to injury, a substantial body of research has shown that a player's position plays a key role on their injury incidence rate (IIR) (Sarmento et al. 2024; Hall et al. 2022). Midfielders have shown to experience the highest injury frequency during matches (43.6%), followed by defenders (30.0%), then forwards (17.9%), and then finally by goalkeepers (8.6%) (Thema et al. 2025). Central defenders also demonstrated a high training-related IIR (Hawkins and Fuller 1996) despite covering a less distance (Castro 2023) which is correlated to the fact that defenders perform a high number of lateral and backward movements, tackling actions, and sliding motions (Castro 2023).

Despite the differences in injury prevalence rates across the positions, soft tissue injuries are common across all, especially in players who are midfielders (78.0%) and defenders (67.6%) (Thema et al. 2025). While goalkeepers experience a greater proportion of upper limb and bone-related injuries (44.4% and 66.7%) because of actions like aerial duels, dives, and reactive movements (Thema et al. 2025).

Although defenders have the highest IIR in training, Midfilders have a high rate of injury per match because. they cover the most distance on the football pitch and do so at high intensity, where accelerations, decelerations, and directional changes are consistently performed (Thema et al. 2025). Every movement places them at a higher risk to overuse injuries, particularly to the hamstrings, groin, and hip flexors (Della Villa et al. 2018; Rahnama et al. 2002). According to Ekstrand et al. (2006) and Thema et al. (2025) midfielders account for up to 40% of all overuse injuries in European football because of the cumulative load and the complex decision-making processes they must manage under sustained physical exertion (Iskra, Voigt and Raab 2024).

Forwards (which consists of strikers and wingers) were also found to be at an increased risk of match injuries (Andersen et al. 2004). Attributed to the fact that most football incidents (duels and tackles) occur in the mid-defensive zone and in the penalty area, the two zones where attackers usually reside (Della Villa et al. 2018). Faude et al. (2006) agrees with this as they found that strikers were at increased risk in contact situations with 4.5 injuries per 1000 hours exposure. Della Villa et al. (2018) states how forwards are at an increased risk of injury because they consistently perform fast kicking movements as well as acceleration/deceleration movements which predispose them for thigh muscle injuries, with these accounting for 25% of the total injured time in football (Ueblacker, Müller-Wohlfahrt and Ekstrand 2013).

Despite these findings, Chomiak et al. (2000) found that football injuries were distributed evenly across the different pitch zones, with 33% occurring in the defensive half, 21% occurring near the halfway line, and 46% in the attacking third. Similarly, Rahnama et al. (2002) found that 30% of injuries occurred in the defending area, 40% in the midfield area and 30% in the attacking area. Therefore, suggesting the spatial location and tactical behaviours during a match may affect injury outcomes rather than a player's actual position.

The player's position was an included feature because of its influence on injury risk. Many publicly available datasets recorded positional data, allowing for reliable encoding and inclusion in the model to capture positional risk patterns observed in literature and previous match data.

2.4. The Impact of Sustained Injuries

When injuries occur, the consequences can be extensive. These consequences are multifaceted, spanning financial losses, on-field underperformance, developmental delays, and even psychological distress (Ekstrand 2013; Bangert et al. 2024; Crossman 1997).

2.4.1. The Financial and Performance Impact That Injuries Have on Clubs

Ekstrand (2013) revealed that injuries can cost a club up to £500,000 a month, due to the salaries for unavailable players, combined with the effects of them being absent (Telfer and Stone 2023). Telfer and Stone (2023) reported that Manchester City sustained 40 injuries in a season at a cost of £11.9 million, while Chelsea's 68 injuries during the 2022–23 season led to £40.1 million loss — a season that coincided with their decline to 12th in the English Premier League.

Despite this, the effects of injuries encompass more than just player wage expenses as they affect the team's performance, their broadcast value as well as their commercial leverage (Eliakim et al. 2020). Availability associates to the number of goals scored, the number of points earned, and a club's overall league position (Hägglund et al. 2013; Eliakim et al. 2020; Eirale et al. 2013) which is key for a club's on pitch success. But availability also strengthens their off-pitch success like the number of sponsorships they receive, the amount of matchday income they generate and whether they will receive media rights (Philippou and Maguire 2022).

Keeping their best players available is key for marketability, so that they can attract sponsors and spectators to the stadiums, and drive merchandise sales (Scelles et al. 2016). When a key player is replaced it often leads to a decline in the quality of player on the field which impacts the fans engagement, leading to lower ticket sales, and diminished commercial revenues (Walia and Boudreaux 2020). Thus, injury management strategies become a cornerstone of a clubs sporting and financial trajectory (Genovard et al. 2025).

2.4.2. The Developmental and Career Impacts Injuries Have on Players

Injuries also lead to detrimental long-term development and career consequences for footballers (Bangert et al. 2024), as injuries interfere with the physical, technical, and tactical development of a player stalling their trajectory as every match and training session contributes to the refinement of these skills (Bangert et al. 2024; Morgans et al. 2014). Larruskain et al. (2021) even found a direct link between injury severity, loss of training, and future selection outcomes at professional levels.

Injuries that affect a player for more than 28 days significantly reduced the likelihood of a players' career advancing, as only 10% of seriously injured youth secured a professional contract compared to 35% of those uninjured (Bangert et al. 2024). In academies long term injuries often lead to release, because of the short-term contracts that are offered, the small squads as well as the intense competition for places (Swainston, Wilson and Jones 2021; Jones and Williams 2024).

2.4.3. The Psychological Consequences of Injuries

Although injuries cause physiological damage and problems to a player's athletic development, injuries also have profound psychological impacts (Crossman 1997). Many players will experience depression, anxiety, frustration, and a sense of a reduced self-worth as an injury feels like a threat to their self-concept (Crossman 1997; Smith 1996: Daley et al. 2021; Christino, Fantry, and Vopat 2015).

Wiese-Bjornstal et al. (1998) created a framework called The Integrated Model of Psychological Response to explain how athletes' cognitive appraisals (e.g. beliefs about the severity or career impact of an injury) interact with emotional responses (such as fear, frustration, or helplessness) and behavioural changes (e.g. withdrawal from training). Brewer (2017) says that players who see their injury as one that could harm their career are likely to experience psychological responses like fear of reinjury, helplessness, and emotional withdrawal. Park, Fury, and Wong (2023) state that negative ideas are linked with an increase in emotional distress, whereas a more positive outline fosters the opposite (Thompson et al. 2018).

Individual and situational factors can impact a player's psychological state. Players under extreme external pressures, whether from coaches, fans or contract clauses have a large emotional burden (McLean 2024). Ahern and Lohr (1997) evidence that low social support and pressure to return from an injury can escalate a player's anxiety, whereas a robust support structure can help them (Yang et al. 2022).

Psychological states have reciprocal effects on injury risk according to Williams and Andersen (1999), as athletes under high stress have reduced focus, more muscle tension and impaired decision-making, all of which elevate injury susceptibility. Life-event stress is a consistent predictor of an individual's injury incidence (Singh and Singh 2021), as a player's likelihood of injury over the course of a season doubled if they had been found to have pre-season anxiety. Likewise, psychological states can affect a player's recovery, as players' with high stress levels and a fear of reinjury were 13 times more likely to suffer an ACL re-injury within two years (Barker 2021).

Craig Bellamy, the Current Wales manager has spoken openly about the mental torment he experienced following a serious knee injury, admitting that even after making a full physical recovery, he struggled with the fear of sprinting, changing direction, and training intensely (George and Brand 2022). Applying Wiese-Bjornstal's Integrated Model (1998) to this situation, Bellamy's appraisal was dominated by fear and uncertainty, reinforcing maladaptive responses like avoidance and mental withdrawal.

The ML model that Iam looking to create may benefit from integrating psychological data as it is a determinant of an athlete's risk to sustaining injury as well as how they will recover. But mental states can vary dramatically (Cherry 2025), complicating their predictive value in the ML model that I am looking to create so I will not include it.

2.4.4. The Long-Term Effects of Sustaining an Injury for a Player

While injuries in football have profound short term affects, they also have long-term consequences, extending far beyond the immediate impact. Some effects may be chronic, career-defining, and persist well into retirement (Koch et al. 2021). Dhinsa (2022) evidence that repeated ankle sprains or fractures lead to chronic joint instability and post-traumatic arthritis, conditions that reduce athletic performance and diminish a player's mobility.

Head injuries are associated with long-term cognitive impairments as Tysvaer and Lochen (1991) found that that 81% of retired professional players displayed some form of cognitive impairment, including reduced memory, attention span, and judgement. Tennessee Orthopaedic Alliance (2024) highlight how repetitive head trauma can contribute to more severe neurological conditions like chronic traumatic encephalopathy or CTE (a degenerative brain disease linked to long-term football participation).

For instance, English football legend Norbert "Nobby" Stiles' long-term exposure to heading the football has been correlated to neurodegenerative damage (Wilts and Gloucestershire Standard 2024). Post-retirement, Nobby developed advanced dementia, and his family and legal teams believe that repeated head impacts with the ball is the cause (Wilts and Gloucestershire Standard 2024).

Furthermore, injuries involving the joints, muscles, and ligaments, have lasting effects if not treated properly and lead to chronic pain and mobility issues (Tennessee Orthopaedic Alliance 2024). For example, a ligament tear in the knee, such as an ACL tear, can result in long-term instability for a player if it is not repaired or rehabilitated correctly (Tennessee Orthopaedic Alliance 2024).

Sustaining an injury will impact a footballer's future financial prospects due to reduced match availability (Calleja-González et al. 2022) limiting their exposure, transfer value, and commercial appeal (Daily Emerald 2025; Moreno-Perez et al. 2020). An injured player may lose out on contract renewals or opportunities to move to higher-tier leagues (Beckles 2023), all of which significantly influence their lifetime earnings as their economic contribution goes beyond their performance encompassing brand partnerships, merchandise sales, and media appearances (Scelles et al. 2016).

For example, Jack Wilshere was once considered to one of England's brightest footballing prospects after becoming Arsenal's youngest league debutant in 2008 aged just 16 (Walker 2023). At 19 he had a Champions League performance against Barcelona that made everyone know his name, but sadly reoccurring ankle injuries severely disrupted his development and almost led to him never being able to walk again (Kenmare 2023). Wilshere retired at prematurely at 30, showing how early-career injuries can derail long-term athletic development and potential.

Post-retirement careers can also be affected by injuries. Former players often have a difficult transition into post-career roles according to Koch et al. (2021), because of poorer health and lower life satisfaction because of injury. When players are forced to retire prematurely, they have less time to build financial security or gain qualifications that would support a life after football (Swainston, Wilson and Jones 2021). In this context, injuries become more than just medical concerns and become career-defining events with prolonged economic and professional consequences.

2.5. Traditional Methods of Injury Prediction

Given the impacts that stem from sports injuries, the importance of approaches to injury prediction cannot be understated. The following are some of the most widely used methods at current.

2.5.1 Current/Traditional Injury Prediction Methods

GPS tracking devices are one of the more popular tools adopted by professional clubs across the world so external load metrics such as sprint distance, top speed, and high-intensity efforts can be monitored (STAT Sports 2024; Walker 2025). By implementing a five-week GPS-guided training programme Bertschy et al. (2021) found that injury rates fell from 8.1 to 4.6 injuries per 1000 exposure hours while the prevalence of injury dropped from 92.6% to 55.2%. Clubb (2023) further highlights how a proactive role in injury prevention can be made when combined with careful periodisation and individualised load management.

Biomechanical analysis is another commonly used method, and this studies the movement of an athlete so asymmetries, instabilities, or inefficient mechanics can be identified (Hulatt 2024). Motion capture tools and 3D kinematics are utilised to identify movement inefficiencies, like poor landing mechanics or asymmetrical joint loading, so those at an increased risk of ACL tears, hamstring injuries, and groin strains are seen (Wan and Shan 2016; Liao 2020). Rokoko (2022) explains that motion capture systems can identify subtle deficiencies in joint range of motion, posture, and muscle coordination, offering insights that go beyond what can be observed naturally.

While these insights are used to predict injury risk through analysing movement patterns most associated with injury-prone mechanics (Lee 2025), they are often interpreted via linear regression or rule-based systems. Both of which overlook the influence of psychological readiness, previous injury, or other external contexts (Ruddy et al. 2019; Wiese-Bjornstal et al. 1998). As such, a shift toward ML approaches, that can handle complex, interacting variables while being able to adapt over time (as more data becomes available) (Chen et al. 2024; Rossi et al. 2021) may be more beneficial.

2.6. Machine Learning in Football Injury Prediction: Current Applications

As traditional injury prediction methods continue to have its inadequacies, there has been a growing shift toward ML approaches (Leckey et al. 2024). ML's capacity to model non-linear interactions across large datasets makes it well suited for this environment, particularly in sport where marginal gains have outsized impacts on player welfare, team performance, and long-term athlete development (Rossi et al. 2021; Kalkhoven et al. 2021).

At its core ML is a subset of artificial intelligence (AI) that enables systems to learn from historical data and improve its performance (IBM). Now ML already underpins everyday items that we take for granted like Netflix recommendations and even social media algorithms (Bin Rashid and Kausik 2021), but its application in sport is still in its preliminary stages according to Ruddy et al. (2019). Unlike regression-based approaches, which assume linearity and independence between variables (Bittencourt et al. 2016), ML models can detect hidden patterns and relationships across multiple, interdependent inputs like workload, injury history, psychological readiness, and match density (Chen et al. 2024).

ML's ever-growing utility in sport is often evidenced, for instance studies reviewed by Rossi et al. (2021) report ML models achieving high accuracy in injury risk classification; however, reported performance varies by cohort, features, and validation design. Similarly, Yang et al. (2022) compared ML models to logistic regression (the current method of predicting injuries) and consistently found superior performance in ML methods for forecasting soft-tissue injuries. These findings show that the shift from observational, experience-led assessments to data-driven, adaptive systems will refine predictions in real time aligning with the performance demands of professional football (Rossi et al. 2021; Leckey et al. 2024).

For example, Liverpool, Brighton and Hove Albion, and Manchester City have utilised data science and ML to elevate their game in all aspects (Celik 2024). Celik (2024) states how Brighton used ML in their recruitment strategy to uncover undervalued talent, helping them to qualify for European football for the first time in their history (Hanson 2022). Whereas

Liverpool utilise ML models to track players training loads, monitor physiological strain, and predict optimal rest periods, helping to reduce injury incidence (Celik 2024) allowing them to win multiple trophies.

Despite this several gaps still remain, as (1) several ML studies are limited by small sample sizes and overfitting, which hinders generalisability to larger more diverse populations (Bullock et al. 2022).(2) ML models often lack external validation because many are tested only on internal datasets without being deployed or tested in live sporting environments (Majumdar et al. 2022). (3) Despite their strengths in predicting, ML models often fail to offer practical, interpretable recommendations that coaches and medical staff can easily act upon (Claudino et al. 2019), creating barriers to adoption.

So, to fulfil its full potential ML must evolve to be (1) easily interpretable, (2) valid in live environments, and (3) co-designed with practitioners so that their needs, knowledge, and requirements are considered in its design (Chen et al. 2022).

2.7. Justification

Current applications of ML in football have their limitations, thus there is a clear need for an ML model that is not only data-efficient but that is interpretable and contextually relevant to football injury prevention. Bittencourt et al. (2016) and Bullock et al. (2022) state how current methods lack the ability to capture the complex, multifactorial nature of injury risk a key component of predicting injuries, given that injuries result from range of physical, environmental, socio-cultural, and psychological variables (Wiese-Bjornstal et al. 1998; Ruddy et al. 2019). So, this study aims to address a critical gap in the literature by creating a ML model that prioritises interpretability over complexity, while still delivering meaningful, personalised insights that produces outputs that coaches and medical staff can act upon.

The justification for this study rests on three core pillars:

1. **The multifaceted impact of football injuries:** Injuries in football affect far more than just a player's availability, as they can incur substantial financial losses for a club, compromise team performance, a player's development, as well as their mental health (Ekstrand 2013; Eliakim et al. 2020; Bangert et al. 2024; Crossman 1997; Hägglund et al. 2013). So, this model can help clubs and players minimise time-loss injuries, protect their revenue, and keep their key players available.

2. **The limitations of current predictive tools:** Traditional/current methods rely on linear assumptions that do not reflect real-world injury dynamics (Claudino et al. 2019) as they can result from a range of physical, environmental, socio-cultural, and psychological

variables (Wiese-Bjornstal et al. 1998; Ruddy et al. 2019). While useful they are unable to adapt, so ML models offer a more sophisticated alternative by learning patterns to adjust over time (Yang et al. 2022).

3. **The need for an interpretable ML solution in sport science:** Although promising results have been achieved using ML models (Yang et al. 2022; Rossi et al. 2021), their adoption into daily workflows remains limited due to issues of interpretability (Claudino et al. 2019). Coaches and sports scientists need transparent, actionable outputs, so this project aims to address this gap by addressing the limitations identified by Claudino et al. (2019) and Sarabia et al. (2021), ensuring that high predictive accuracy is achieved without sacrificing practical applicability that produces outputs that coaches and medical staff can act upon.

3. Methodology

Football is a fast-paced, high-intensity sport where injuries are an inevitable and a recurrent occurrence (Peebles 2023). Injuries have a multitude of wide-ranging consequences on both a player and their club, from diminished on-pitch performance to financial and psychological strain (Bangert et al. 2024; Crossman 1997; Ekstrand 2013). Given these implications, predicting injuries before they occur is a critical area of research.

This chapter will outline the methodological framework that was followed to design, train and evaluate a ML model capable of predicting injuries from secondary data. The choice of a secondary data analysis approach, using the publicly available Player Injuries and Team Performance dataset spanning the 2019/20–2023/24 seasons will be justified in this section. The preparation process of the initial dataset will be discussed (including cleaning, encoding, and feature engineering) along with the final feature set comprising age, playing position, and previous injury count, all of which was supported by previous literature (Fulton et al. 2014; Inklaar et al. 1996; Hall et al.2022). Then the election of a Decision Tree Classifier will be explained, and why evaluation metrics were employed. This section will then conclude by addressing ethical considerations and acknowledging key methodological limitations found while conducting this research.

3.1. Research Design

This research project will adopt a secondary data analysis methodology, as it allows me to generate new insights from already existing datasets without the time or resource demands of primary data collection (Szabo and Strang 1997). This approach was most suitable for a student-led project as I was able to gain access to a broader and more diverse dataset than would be feasible to collect independently (Lexis Nexis 2025).

The primary aim of this project was to develop a ML model that can predict injury occurrence from publicly available data on players. The dataset I utilised from Kaggle, created by Biswas (2023) is titled Player Injuries and Team Performance. This dataset includes a wide range of variables such as injury type, duration, player position, minutes played, match involvement, and contextual team data, spanning multiple professional clubs in the English Premier League. I selected this dataset as it aligned closely with my main aim, which was to find patterns that contribute to injury occurrence based on data. The longitude nature and the season-by-season layout provided a solid foundation for both temporal validation and generalisability testing which made it a solid fit for this project.

So, the models' outputs would be easily interpretable (by medical staff and coaches) I decided to utilise a Decision Tree Classifier. Decision Trees are known for their transparency, low computational cost, and ability to handle both categorical and continuous variables (Scikit Learn 2025). In football, decisions need to be understood and trusted by non-technical users, so the interpretability of my ML model is essential. The visual and rule-based nature of Decision Trees supports this need by aligning closely with the logical frameworks used in practitioner decision-making according to Sarabia et al. (2021).

While Random Forests or Gradient Boosting offer superior predictive accuracy, they sacrifice interpretability which is why I decided against them. In sport where stakeholder trust and transparency are vital, the Decision Tree approach were the most appropriate (Bullock et al. 2022; Data Heroes 2023).

3.2. Dataset Overview and Preprocessing

The dataset utilised for this project was sourced from Kaggle and was compiled by Biswas (2023). The initial dataset included five full English Premier League seasons (2019/20 to 2023/24), making it appropriate for longitudinal analysis and time-based model validation. Detailed player demographics (e.g. age, height, and position), performance metrics (e.g. appearances, minutes played, goals, assists), injury details (e.g. type and duration), and contextual team data (e.g. club name and league finish) were a part of this dataset which is why it was chosen.

Preprocessing was utilised upon the dataset as effective preprocessing is essential to prevent model degradation, as raw datasets often include inconsistencies, missing values, and non-informative features (Yasar 2025). Given the focus on injury prediction, any variable with no clear relationship to injury risk was excluded during preprocessing. Variables such as FIFA (Fédération Internationale de Football Association) card traits and subjective player ratings, which could introduce any further noise without offering medical or biomechanical value were removed (Sharma 2023). Further, ambiguous injury labels such as Pain, Unknown Injury, and other illness-based absences (e.g. Flu, Virus, and Coronavirus) were also removed to ensure relevance.

Name	Team Name	Position	Age	Season	FIFA rating	Injury	Date of Injury	Date of return	Match1_b	Match1_before	Match1_	Match1_before_	Match2_befor	Match2_before_injury	Match2_befor	Match2_before_inj	Match3_befor	Match3_before_inj	Match3_befor	Match3_befor	Match1_mis	Match1_missed_m
Jamaal Lascelles	Newcastle	Center Back	26	2019/20	77	Knee injury	Nov 9, 2019	Jan 13, 2020	draw	Wolves	0	7.4	win	West Ham	1	6.1	win	Bournemouth	1	5.8	lose	Aston Villa
Fabian Schär	Newcastle	Center Back	28	2019/20	79	Knee injury	Oct 20, 2019	Nov 24, 2019	lose	Leicester	-5	5.8	win	Man United	1	7.5	lose	Chelsea	-1	6.1	draw	Wolves
Fabian Schär	Newcastle	Center Back	28	2019/20	79	Hamstring strain	Jan 2, 2020	Jan 17, 2020	lose	Man United	-3	4.9	lose	Everton	-1	7.7	lose	Leicester	-3	5.6	draw	Wolves
Fabian Schär	Newcastle	Center Back	28	2019/20	79	Shoulder injury	Jul 16, 2020	Sep 28, 2020	lose	Man City	-5	5.2	lose	Watford	-1	6.3	lose	Tottenham	-2	6.3	draw	Brighton
Paul Dummett	Newcastle	Center Back	28	2019/20	75	Groin injury	Dec 22, 2019	Jan 10, 2020	win	Southampton	2	5.8	lose	Burnley	-1	6	win	Crystal Palace	1	6.2	lose	Man United
Ciaran Clark	Newcastle	Center Back	30	2019/20	75	Calf injury	Dec 9, 2019	Jan 10, 2020	draw	Man City	0	6	win	Sheffield	2	6.4	win	Southampton	1	6.4	lose	Burnley
Ciaran Clark	Newcastle	Center Back	30	2019/20	75	Ankle injury	Feb 16, 2020	Sep 14, 2020	draw	Everton	0	6.3	draw	Norwich City	0	7	lose	Arsenal	-4	5.6	lose	Crystal Palace
Jetro Willems	Newcastle	Left Back	26	2019/20	77	Knee injury	Sep 22, 2019	Oct 5, 2019	draw	Watford	0	6.7	lose	Liverpool	-2	6.5	draw	Brighton	0	6.3	lose	Leicester
Jetro Willems	Newcastle	Left Back	26	2019/20	77	Knee injury	Dec 15, 2019	Dec 25, 2019	win	Sheffield	2	6	win	Southampton	-1	5.8	lose	Burnley	-1	6.2	win	Crystal Palace
Jetro Willems	Newcastle	Left Back	26	2019/20	77	Cruciate ligament tear	Jan 18, 2020	Sep 20, 2020	lose	Leicester	-3	5.8	draw	Wolves	0	6.3	win	Chelsea	1	6.1	draw	Everton
DeAndre Yedlin	Newcastle	Right Back	26	2019/20	74	Groin problems	Apr 28, 2019	Sep 28, 2019	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	draw	Watford
DeAndre Yedlin	Newcastle	Right Back	26	2019/20	74	Hip injury	Nov 27, 2019	Dec 10, 2019	win	West Ham	1	5.6	win	Bournemouth	1	7.3	lose	Aston Villa	-2	5.6	draw	Man City
DeAndre Yedlin	Newcastle	Right Back	26	2019/20	74	Hand injury	Jan 5, 2020	Jan 23, 2020	lose	Man United	-3	5.9	lose	Everton	-1	5.7	lose	Leicester	-3	5.7	draw	Wolves
Emil Krafth	Newcastle	Right Back	25	2019/20	75	Ankle injury	Jan 21, 2020	Jun 20, 2020	win	Chelsea	1	6.6	draw	Everton	0	5.7	N.A.	N.A.	N.A.	N.A.	draw	Norwich City
Emil Krafth	Newcastle	Right Back	25	2019/20	75	Ankle injury	Jul 21, 2020	Sep 11, 2020	lose	Watford	-1	6(S)	lose	Tottenham	-2	6.1	draw	Brighton	0	6.6	lose	Liverpool
Javier Manquillo	Newcastle	Right Back	26	2019/20	75	Hamstring strain	Jan 2, 2020	Feb 28, 2020	lose	Man United	-3	6.1	lose	Leicester	-3	5.2	N.A.	N.A.	N.A.	N.A.	draw	Wolves
Isaac Hayden	Newcastle	Defensive Midfielder	25	2019/20	78	Hamstring strain	Jul 5, 2020	Jul 25, 2020	draw	Aston Villa	0	6.3	win	Bournemouth	3	6.1	draw	West Ham	0	5.9	lose	Man City
Sean Longstaff	Newcastle	Central Midfielder	22	2019/20	74	Ankle injury	Sep 1, 2019	Sep 28, 2019	lose	Norwich City	-2	6.5	win	Tottenham	1	6	draw	Watford	0	7.1	draw	Brighton
Sean Longstaff	Newcastle	Central Midfielder	22	2019/20	74	Hamstring injury	Jul 1, 2020	Jul 25, 2020	win	Southampton	1	6(S)	win	Bournemouth	3	7	N.A.	N.A.	N.A.	N.A.	draw	West Ham
Jonjo Shelvey	Newcastle	Central Midfielder	28	2019/20	79	Hamstring injury	Sep 22, 2019	Oct 5, 2019	lose	Liverpool	-2	5.9	draw	Brighton	0	6.6	N.A.	N.A.	N.A.	N.A.	lose	Leicester
Jonjo Shelvey	Newcastle	Central Midfielder	28	2019/20	79	Calf injury	Dec 9, 2019	Dec 20, 2019	draw	Man City	0	7.5	win	Sheffield	2	7.6	win	Southampton	1	7.4	lose	Burnley
Jonjo Shelvey	Newcastle	Central Midfielder	28	2019/20	79	Hamstring strain	Jan 18, 2020	Feb 28, 2020	lose	Leicester	-3	6	win	Chelsea	1	6.9	N.A.	N.A.	N.A.	N.A.	draw	Norwich City
Matt Ritchie	Newcastle	Right Midfielder	30	2019/20	76	Ankle injury	Aug 28, 2019	Jan 3, 2020	lose	Arsenal	1	6	lose	Norwich City	-2	7	win	Tottenham	1	6.9	draw	Watford
Allan Saint-Maximin	Newcastle	Left winger	23	2019/20	79	Hamstring strain	Sep 22, 2019	Sep 28, 2019	lose	Arsenal	-1	6.5	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	draw	Norwich City
Miguel Almirón	Newcastle	Right winger	26	2019/20	76	Hamstring injury	Dec 9, 2019	Dec 20, 2019	draw	Man City	0	6.8	win	Sheffield	2	6.1	win	Southampton	1	6.2	lose	Burnley
Dwight Gayle	Newcastle	Center Forward	30	2019/20	75	Calf injury	Aug 8, 2019	Oct 5, 2019	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	lose	Arsenal
Dwight Gayle	Newcastle	Center Forward	30	2019/20	75	Hamstring strain	Jan 2, 2020	Jan 10, 2020	lose	Man United	-3	5.8	lose	Everton	-1	5.7	draw	Wolves	0	6.3	win	Chelsea
Andy Carroll	Newcastle	Center Forward	31	2019/20	74	Groin injury	Oct 20, 2019	Nov 1, 2019	lose	Leicester	-5	6	win	Man United	1	5.7	lose	Chelsea	-1	5.7	draw	Wolves
Andy Carroll	Newcastle	Center Forward	31	2019/20	74	Bruised ribs	Nov 26, 2019	Dec 3, 2019	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	lose	Aston Villa	-2	6.3	draw	Man City
Andy Carroll	Newcastle	Center Forward	31	2019/20	74	Knock	Jan 15, 2020	Jun 15, 2020	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	draw	Wolves	0	5(S)	win	Chelsea
Andy Carroll	Newcastle	Center Forward	31	2019/20	74	Muscle injury	Jul 6, 2020	Jul 17, 2020	draw	Aston Villa	0	6.4	win	Bournemouth	3	6.4	N.A.	N.A.	N.A.	N.A.	lose	Man City
Jamaal Lascelles	Newcastle	Center Back	27	2020/21	77	Hip injury	Sep 28, 2020	Oct 16, 2020	win	West Ham	2	6.6	lose	Brighton	-3	5.6	draw	Tottenham	0	6.3	win	Burnley
Jamaal Lascelles	Newcastle	Center Back	27	2020/21	77	Muscle fatigue	Nov 22, 2020	Jan 17, 2021	win	Everton	1	6.4	lose	Southampton	-2	6.1	lose	Chelsea	-2	6.3	win	Crystal Palace
Jamaal Lascelles	Newcastle	Center Back	27	2020/21	77	Hamstring strain	Jan 31, 2021	Feb 14, 2021	lose	Aston Villa	-2	5.6	lose	Leeds	-1	6.1	win	Everton	2	6.2	lose	Crystal Palace
Jamaal Lascelles	Newcastle	Center Back	27	2020/21	77	Foot injury	Apr 4, 2021	Jun 1, 2021	draw	Aston Villa	0	7.4	lose	Brighton	-3	5.8	draw	Tottenham	-3	5.9	win	Burnley
Fabian Schär	Newcastle	Center Back	29	2020/21	79	Virus	Nov 29, 2020	Dec 25, 2020	lose	Southampton	-2	6.1	lose	Chelsea	-2	5.8	N.A.	N.A.	N.A.	N.A.	win	West Brom
Fabian Schär	Newcastle	Center Back	29	2020/21	79	Knee injury	Feb 7, 2021	Apr 25, 2021	win	Everton	2	7	lose	Crystal Palace	-1	6.5	win	Southampton	1	5.4	lose	Chelsea
Ciaran Clark	Newcastle	Center Back	31	2020/21	76	Calf injury	Jan 19, 2021	Jan 29, 2021	lose	Leicester	-1	6.4	lose	Sheffield	-1	6.6	lose	Arsenal	-3	5.4	lose	Aston Villa
Ciaran Clark	Newcastle	Center Back	31	2020/21	76	Knock	Feb 4, 2021	Feb 14, 2021	win	Everton	2	6.2	lose	Crystal Palace	-1	6.5	N.A.	N.A.	N.A.	N.A.	win	Southampton
Javier Manquillo	Newcastle	Right Back	27	2020/21	75	Ankle injury	Feb 7, 2021	Mar 10, 2021	win	Everton	2	6.5	lose	Crystal Palace	-1	6.2	win	Southampton	1	6.1	lose	Chelsea
Isaac Hayden	Newcastle	Defensive Midfielder	26	2020/21	78	Coronavirus	Nov 25, 2020	Dec 11, 2020	N.A.	N.A.	N.A.	N.A.	lose	Chelsea	-2	6.3	N.A.	N.A.	N.A.	N.A.	win	Crystal Palace
Isaac Hayden	Newcastle	Defensive Midfielder	26	2020/21	87	Knee injury	Mar 20, 2021	Jun 1, 2021	draw	West Brom	0	6.7	draw	Aston Villa	0	6.9	lose	Brighton	-3	6.2	draw	Tottenham
Jonjo Shelvey	Newcastle	Central Midfielder	29	2020/21	78	Groin injury	Oct 19, 2020	Nov 20, 2020	draw	Tottenham	0	5.9	win	Burnley	2	7.1	lose	Man United	-3	6.7	draw	Wolves
Jonjo Shelvey	Newcastle	Central Midfielder	29	2020/21	78	Hamstring injury	Jan 4, 2021	Jan 17, 2021	lose	Leicester	-1	5.8	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	lose	Sheffield
Matt Ritchie	Newcastle	Right Midfielder	31	2020/21	76	Shoulder injury	Sep 27, 2020	Nov 25, 2020	draw	Tottenham	0	6	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	win	Burnley
Allan Saint-Maximin	Newcastle	Left winger	24	2020/21	79	Ankle injury	Sep 21, 2020	Oct 1, 2020	win	West Ham	2	6.7	lose	Brighton	-3	5.3	N.A.	N.A.	N.A.	N.A.	draw	Tottenham
Allan Saint-Maximin	Newcastle	Left winger	24	2020/21	79	Calf injury	Nov 22, 2020	Jan 22, 2021	win	Everton	1	6.2	lose	Southampton	-2	6.1	lose	Chelsea	-2	5.9	win	Crystal Palace
Allan Saint-Maximin	Newcastle	Left winger	24	2020/21	79	Groin injury	Feb 28, 2021	Apr 4, 2021	lose	Chelsea	-2	6.1	lose	Man United	-2	7.4	draw	Wolves	0	7	draw	West Brom
Miguel Almirón	Newcastle	Right winger	27	2020/21	76	Knee injury	Feb 28, 2021	Mar 15, 2021	lose	Chelsea	-2	6.1	lose	Man United	-2	5.7	draw	Wolves	0	6.2	draw	West Brom
Dwight Gayle	Newcastle	Center Forward	31	2020/21	75	Inner ligament injury	Sep 5, 2020	Dec 11, 2020	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	win	West Ham
Jamal Lewis	Newcastle	Left Back	23	2020/21	75	Knee problems	Dec 31, 2020	Jan 15, 2021	N.A.	N.A.	N.A.	N.A.	lose	Man City	-2	5.9	N.A.	N.A.	N.A.	N.A.	lose	Leicester
Ryan Fraser	Newcastle	Left winger	27	2020/21	77	Hamstring injury	Nov 2, 2020	Dec 15, 2020	lose	Man United	-3	5.1	draw	Wolves	0	5.7	win	Everton	1	5.7	lose	Southampton
Ryan Fraser	Newcastle	Left winger	27	2020/21	77	Groin injury	Apr 1, 2021	May 31, 2021	draw	West Brom	0	5.8	draw	Aston Villa	0	6.4	lose	Brighton	-3	6.1	draw	Tottenham
Callum Wilson	Newcastle	Center Forward	29	2020/21	79	Hamstring strain	Nov 7, 2020	Nov 26, 2020	draw	Wolves	0	5.7	win	Everton	1	7.8	lose	Southampton	-2	5.2	lose	Chelsea
Callum Wilson	Newcastle	Center Forward	29	2020/21	79	Hamstring injury	Feb 7, 2021	Apr 4, 2021	win	Everton	2	8.5	lose	Crystal Palace	-1	6.9	win	Southampton	1	6.1	lose	Chelsea
Callum Wilson	Newcastle	Center Forward	29	2020/21	79	Hamstring injury	May 8, 2021	May 25, 2021	draw	Liverpool	0	5.7	lose	Arsenal	-2	5.5	win	Leicester	2	8.1	lose	Man City
Jamaal Lascelles	Newcastle	Center Back	28	2021/22	77	Ill	Feb 11, 2022	Feb 16, 2022	draw	Watford	0	7	win	Leeds	1	7	win	Everton	2	6.6	win	Aston Villa
Paul Dummett	Newcastle	Center Back	30	2021/22	75	Calf injury	Jul 28, 2021	Jan 1, 2022	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	lose	Liverpool
Joelinton	Newcastle	Central Midfielder	25	2021/22	79	Knock	Mar 7, 2022	Mar 15, 2022	draw	West Ham	0	6.4	win	Brentford	2	7.7	win	Brighton	1	6.6	win	Southampton
Jonjo Shelvey	Newcastle	Central Midfielder	30	2021/22	77	Calf injury	Aug 15, 2021	Oct 15, 2021	lose	West Ham	-2	7	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	lose	Aston Villa
Jonjo Shelvey	Newcastle	Central Midfielder	30	2021/22	77	Muscle injury	Dec 17, 2021	Dec 26, 2021	win	Burnley	1	7	lose	Leicester	-4	6	lose	Liverpool	-2	7	lose	Man City
Jonjo Shelvey	Newcastle	Central Midfielder	30	2021/22	77	Calf problems	May 1, 2022	May 25, 2022	win	Leicester	1	7	win	Crystal Palace	1	6	lose	Liverpool	1	6	lose	Man City
Allan Saint-Maximin	Newcastle	Left winger	25	2021/22	79	Ill & knock	Feb 14, 2022	Mar 12, 2022	win	Leeds	1	8	win	Everton	2	9	win	Aston Villa	1	5	draw	West Ham
Joe Willock	Newcastle	Central Midfielder	22	2021/22	75	Coronavirus	Dec 21, 2021	Dec 31, 2021	lose	Leicester	-4	5.9	N.A.	N.A.	N.A.	N.A.	lose	Man City	-4	5.4	draw	Man United
Joe Willock	Newcastle	Central Midfielder	22	2021/22	75	Knee injury	Apr 4, 2022	Apr 10, 2022	win	Southampton	1	6.9	lose	Everton	-1	6.5	lose	Tottenham	-4	5.9	win	Wolves
Joe Willock	Newcastle	Central Midfielder	22	2021/22	75	Knee injury	May 1, 2022	May 25, 2022	win	Crystal Palace	1	6(S)	win	Norwich City	3	6	lose	Liverpool	-1	5.1	lose	Man City
Ryan Fraser	Newcastle	Left winger	28	2021/22	77	Hamstring injury	Apr 8, 2022	May 9, 2022	lose	Everton	-1	5.5	lose	Tottenham	-4	5.8	N.A.	N.A.	N.A.	N.A.	win	Leicester
Callum Wilson	Newcastle	Center Forward	30	2021/22	78	Hamstring injury	Aug 30, 2021	Oct 17, 2021	lose	West Ham	2	6.9	lose	Aston Villa	-2	5.3	draw	Southampton	0	6.8	lose	Man United

Figure 1: A snapshot of initial Dataset by Biswas (2023) prior to cleaning titled Player Injuries and Team Performance.

However, the dataset had to be adapted as it failed to include any non-injured players for any of the Premier League seasons that were represented. So, this could be addressed, a sample of non-injured players was sourced for each season from the publicly available squad data page on the Transfermarkt (2025) website. By doing so, class imbalances and the risk of overfitting were mitigated, both of which can compromise a model's generalisability and inflate performance metrics (Leckey et al. 2024).

A binary classification target variable (Injured) was created so that players who sustained at least one injury in each season were labelled as 1 (injured), and those without were labelled as 0 (not injured). To further enhance my model's predictive accuracy, a binary feature representing earlier injury history was also utilised, enabling the model to learn from recurrence patterns which is a key factor in injury susceptibility according to previous literature (Fulton et al. 2014; Hägglund, Waldén and Ekstrand 2006; Arnason et al. 2004). This structure allowed my ML model to clearly distinguish between injured and non-injured players, allowing it to learn patterns associated with each outcome class, a fundamental requirement for effective injury modelling (H2O.ai 2025).

Using label encoding the position variable was converted into a numerical format, where each playing position (e.g. Centre Back, Central Midfielder and Forward) was given their own distinct number. By doing so the Decision Tree Classifier could process the data and incorporate positional information as a predictive feature of injury, aligning with earlier literature that emphasises the role of player position in injury risk.

Name	Season	Season Start	Team Name	Position	Age	Injured Season	Previous Injuries	Injured	Position
Aaron Hickey	2023/2024	2023	Brentford	Right Back	22	1	1	1	16
Alex Iwobi	2023/2024	2023	Everton	Left Winger	27	1	2	1	15
Alexander Isak	2023/2024	2023	Newcastle	Centre Forward	24	1	0	1	3
Amadou Onana	2023/2024	2023	Everton	Defensive Mid	22	1	1	1	9
Ameen Al-Dakhil	2023/2024	2023	Burnley	Centre Back	21	1	0	1	2
Anthony Gordon	2023/2024	2023	Newcastle	Left Winger	23	1	1	1	15
Ashley Young	2023/2024	2023	Everton	Left Back	38	1	0	1	12
Ben Mee	2023/2024	2023	Brentford	Centre Back	33	1	3	1	2
Ben White	2023/2024	2023	Arsenal	Right Back	26	1	1	1	16
Beto	2023/2024	2023	Everton	Centre Forward	25	1	0	1	3
Boubacar Kamara	2023/2024	2023	Aston Villa	Defensive Mid	24	1	1	1	9
Bryan Gil	2023/2024	2023	Tottenham	Left Winger	22	1	1	1	15
Bryan Mbeumo	2023/2024	2023	Brentford	Right Winger	22	1	0	1	19
Bukayo Saka	2023/2024	2023	Various	Right Winger	25	0	0	0	18
Callum Wilson	2023/2024	2023	Newcastle	Centre Forward	32	1	3	1	3
Christian Nørgaard	2023/2024	2023	Brentford	Defensive Mid	30	1	0	1	10
Dan Burn	2023/2024	2023	Newcastle	Centre Back	32	1	0	1	2
Declan Rice	2023/2024	2023	Various	Defensive Mid	25	0	0	0	9
Diego Carlos	2023/2024	2023	Aston Villa	Center Back	32	1	1	1	2
Diogo Dalot	2023/2024	2023	Various	Right Back	25	0	0	0	16
Dominik Szot	2023/2024	2023	Various	Central Midfielder	25	0	0	0	5
Dwight McNeil	2023/2024	2023	Everton	Left Winger	23	1	1	1	15
Elliot Anderson	2023/2024	2023	Newcastle	Central Midfielder	18	1	0	1	5
Erling Haaland	2023/2024	2023	Various	Center Forward	25	0	0	0	8
Ethan Pinnock	2023/2024	2023	Brentford	Center Back	29	1	0	1	2
Ezri Konsa	2023/2024	2023	Aston Villa	Center Back	26	1	1	1	2
Fabian Schär	2023/2024	2023	Newcastle	Center Back	32	1	0	1	2
Gabriel Jesus	2023/2024	2023	Arsenal	Center Forward	27	1	1	1	3
Gabriel Martinelli	2023/2024	2023	Arsenal	Left Winger	23	1	3	1	15
Giovani Lo	2023/2024	2023	Tottenham	Central Midfielder	27	1	3	1	6
Harvey Barnes	2023/2024	2023	Newcastle	Left Winger	26	1	0	1	15
Hjalmar Ekdal	2023/2024	2023	Burnley	Centre Back	23	1	0	1	2

Figure 2: A cross-section of the dataset after preprocessing and cleaning, where irrelevant features have been removed, ambiguous injury types have been excluded, and binary target variables have been created for classification.

To enable a proper evaluation of my ML model, the dataset—once cleaned and pre-processed—was divided into two separate files: one containing a training dataset from the 2019/20–2022/23 English Premier League seasons, and the other containing data exclusively from the 2023/24 season. By testing on unseen future data, my model's generalisability could be evaluated, and real-world future risk forecasting could be simulated (Austin et al. 2016). Two CSV (Comma-Separated Values) files were used and loaded independently into the model, which further reinforced the integrity of temporal validation.

Figure 3: Number of injuries by player position in the training dataset

Figure Three shows the distribution of injuries by playing position in my cleaned training dataset. This includes 272 injury records from 170 players representing 8 different English Premier League clubs between the 2019/20 and 2022/23 seasons. Injuries were distributed across eleven distinct playing positions, with centre backs (92 injuries), centre forwards (53), and central midfielders (50) forming the majority. Other positions included left wingers (29), right backs (28), defensive midfielders (27), left backs (26), right wingers (21), attacking midfielders (17), and goalkeepers (17), with right midfielders accounting for just 2 cases. Notably, the distribution was skewed toward central positions, with centre backs, centre forwards, and central midfielders collectively responsible for over 70% of all recorded injuries.

These findings are consistent with Thema et al. (2025) who found that players positioned centrally experience significantly higher levels of both physical load and injury exposure. Likewise, Hawkins and Fuller (1996) establish that central defenders have a high training-related IIR due to the high number of lateral and backward movements, tackling actions, and sliding motions they perform consistently (Castro 2023). While Thema et al. (2025) report midfielders as having highest match injury frequency, centre backs were most represented in this dataset, which aligns more with Chomiak et al. (2000) who indicates that injuries are distributed evenly across the pitch zones rather than by position, with 33% occurring in the defensive half thus why defenders may have experienced more injuries. Despite this, the distribution found in this dataset aligns with previous literature, suggesting that the training dataset is a good representative of the impact of a player's position in sustaining injuries.

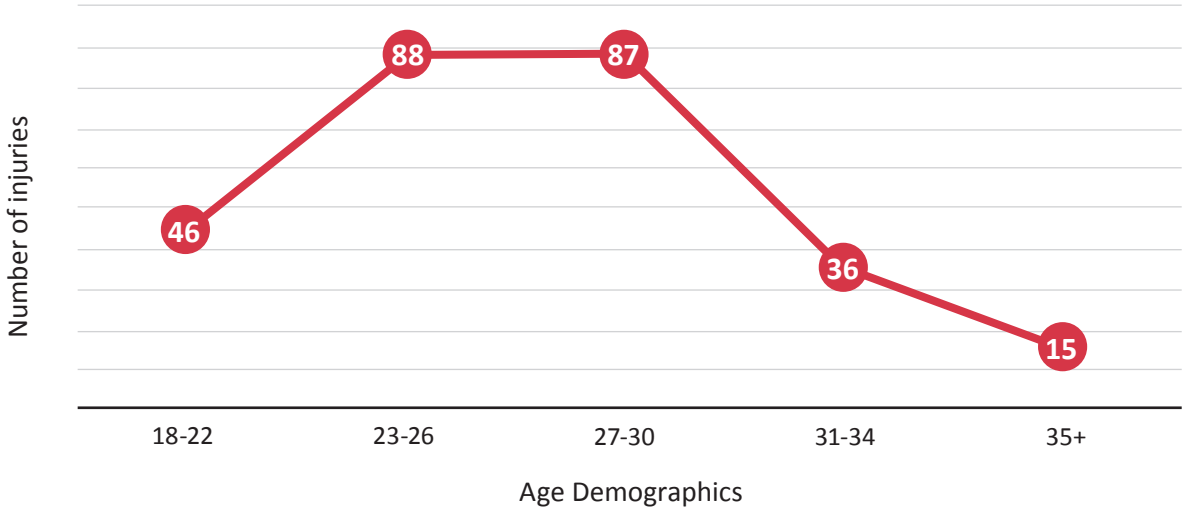

Figure 4: Number of injuries by age group in the training dataset (2019/20–2022/23)

This figure shows the distribution of sustained injuries across five age groups (18–22, 23–26, 27–30, 31–34, and 35+) within my cleaned training dataset. The dataset covers four English Premier League seasons (2019/20–2022/23) and has a total of 272 injuries. Players aged 23–26 sustained the most injuries (88), closely followed by those aged 27–30 (87). The youngest group aged 18–22, recorded 46 injuries, while the 31–34 group recorded 36 injuries, and players aged 35 and over accounted for 15 injuries. This shows that players in their mid-to late-twenties are the most injury-prone demographic, as they were collectively responsible for approximately 64% of all recorded injuries in the dataset.

Figure Four aligns with earlier literature on the topic, as those older had a higher injury rate compared to those younger (Sonesson, Lindblom and Hägglund 2023), with Östenberg and Roos (2000) saying that those over the age of 25 are at a significantly higher risk of injury. Players who are older are stronger, faster as well as heavier as they have grown, matured, and have more experience (Halvorsen Wik 2022). But at those ages, training and matches have more intensity and carry more significance as competitions have intensified (Halvorsen Wik 2022). While those under 24 were not immune to sustaining an injury, they had significantly less compared to their older counter parts, differing from Inklaar et al. (1996) who said that younger players experience injury rates comparable to, or higher than adults.

Whereas the test dataset (once cleaned) only comprised of 94 injury records from 64 different players who represent 7 clubs across the English Premier League in the 2023/24 season. The positional distribution showed that midfielders sustained the highest number of injuries (31), followed closely by defenders (24), then forwards (6), with goalkeepers recording the fewest injuries (1). While this pattern differs slightly from the training set, (which was more heavily skewed toward central defensive roles), the positional distribution closely mirrored that of the training set, which indicated a consistency in the injury risk profiles across the five-year period.

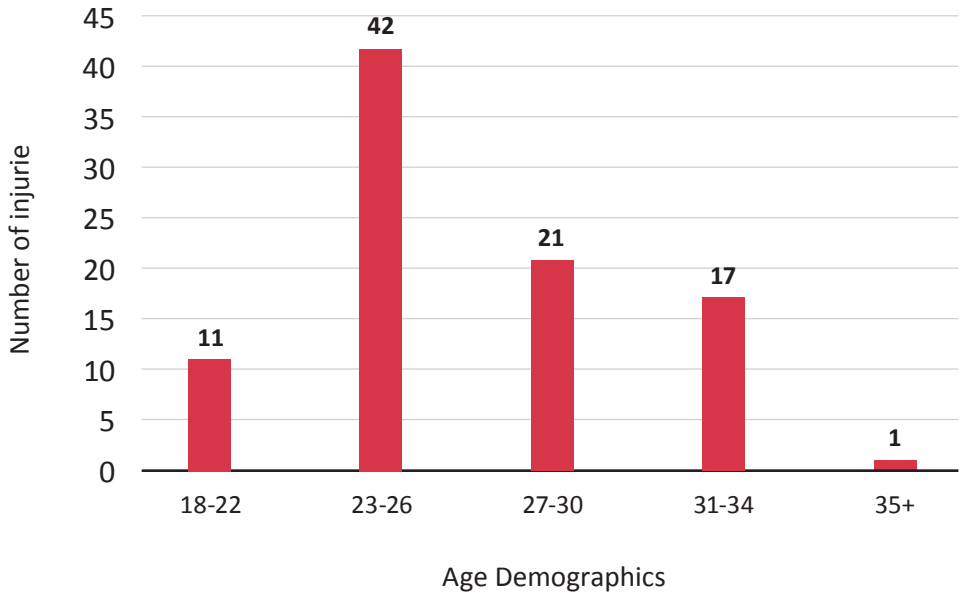

Figure 5: Number of injuries by age group in the test dataset (2023/24)

Figure Five shows the age distribution of all injuries in the cleaned test dataset from the 2023/24 English Premier League season. The highest number of injuries occurred among players aged 23–26 (42), followed by those aged 27–30 (21), 31–34 (17), and 18–22 (11). Players aged 35 and over accounted for a single injury. These results broadly reflect previous literature as Östenberg, and Roos (2000) explains that players over the age of 25 were at a significantly higher risk of injury. However, the high number of injuries among players aged 23–26 and 27–30 differs from Östenberg and Roos (2000) and instead shows how age may be a proxy variable shaped by many factors rather than being a standalone indicator of injury (Costa E Silva, Teles and Fragoso 2022).

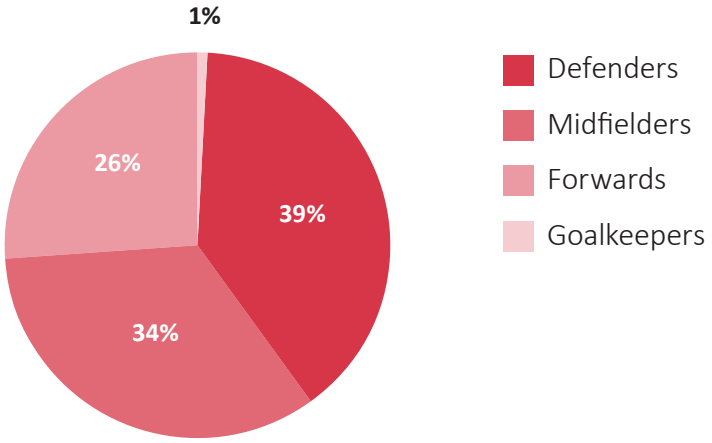

Figure 6: Number of injuries by position in the test dataset (2023/24)

Figure Six illustrates the number of injuries sustained by playing positions during the 2023/24 English Premier League season. In this test dataset, Defenders experienced the highest proportion of injuries, accounting for 36 out of 92 cases, followed by Midfielders with 31 injuries and Forwards with 24. Goalkeepers accounted for a single injury. This distribution aligns with prior literature by Hawkins and Fuller (1996) who found that defensive roles were more susceptible to injury as they perform a high number of lateral and backward movements, tackling actions, and sliding motions (Castro 2023). Likewise, midfielders (who accounted for the second most injuries in this dataset) experience a high injury frequency because they cover the most distance on the football pitch and do so at a high intensity (Thema et al. 2025). Although Andersen et al. (2004) states that forwards are at an increased risk of injuries, the pattern shown in the dataset shows a greater injury burden among defensive players.

To support training and interpretability, only the following features were retained after cleaning and preprocessing: player name, team name, position, age, season, season start and whether they were injured or not. These attributes were selected based on prior literature (Lindenfeld et al. 1994; Inklaar et al. 1996; Sarmento et al. 2024; Hall et al.2022; Fulton et al. 2014; Hägglund, Waldén and Ekstrand 2006) and are shown below in Table One.

Table 1: The key features retained after dataset cleaning

Feature retained:	**Description of the feature retained:**
Name	The full name of the player involved in the injury.
Team Name	The club the player was registered with at the time of injury.
Position	The player's position (e.g., Defender, Midfielder, Forward).
Age	The age of the player at the time of the injury.
Season	The season in which the injury occurred (e.g., 2019/20).
Injury	The type of injury sustained (e.g., Hamstring, ACL tear, Groin strain).
Date of Injury	The date on which the injury was recorded.
Date of Return	The date on which the injury was recorded.

Lastly, justified duplication was deliberately preserved. Recurrent injuries can occur in a single season and are well supported by literature as being a major risk factor for future injury (Hägglund, Waldén and Ekstrand 2006; Arnason et al. 2004; Hawkins and Fuller 1999; Inklaar 1994).

3.3. Feature Engineering and Variable Transformation

Several feature engineering processes were applied to the training dataset before it was applied to the model. For instance, the non-numeric categories were altered into integers like the player's position all of which were label encoded using Scikit-learn's LabelEncoder function. Allowing the Decision Tree Classifier to process them effectively while preserving their categorical nature (Geeks for Geeks 2025). Figure Seven provides an example of this encoding process for the Position variable.

```
# This will change the players football positions into numbers through label encoding them.
le = LabelEncoder()
train_df["Position_enc"] = le.fit_transform(train_df["Position"])
test_df["Position_enc"] = le.transform(test_df["Position"])
```

Figure 7: Code showing the use of Scikit-learn's LabelEncoder() to convert categorical variables (player position) into numerical format.

In addition to encoding categorical variables, the Age feature was retained as a continuous numeric variable to maintain the full range of variation across the player pool. While age could have been grouped into categories or normalised, it was intentionally preserved in its raw form to align with recommendations in earlier literature, which consistently identifies age as an influential predictor of injury risk (Lindenfeld et al. 1994; Inklaar et al. 1996; Backous et al. 1988; Östenberg and Roos 2000). Ensuring that subtle differences in player age are still able to be captured by the model.

3.4. Model Selection and Justification

A Decision Tree Classifier was utilised as the primary ML model due to its simplicity, transparency, and robust performance in classification tasks (Scikit Learn 2025). In the environment of injury prediction (where explainability is crucial for real-world adoption) Decision Trees provide visual and rule-based outputs that can be easily understood by non-technical stakeholders such as coaches, medical staff, and analysts (Sarabia et al. 2021).

It is said that Decision Trees can imitate human decision-making by splitting data based on feature thresholds, producing visual and rule-based outputs that are highly interpretable (Master's in data science 2025). Their ability to handle both categorical and continuous variables, ignore irrelevant features, and tolerate missing values made them well suited for the structure of this dataset (Scikit Learn 2025) and for the goals of my project.

Many authors have criticised traditional injury prediction models for being overly opaque, especially regression-based models that assume linearity and ignore feature interactions (Ruddy et al. 2019; Wiese-Bjornstal et al. 1998). In contrast, Decision Trees are better equipped to capture non-linear relationships between features such as age, playing position, and injury recurrence because they do not assume a linear relationship (Chen et al. 2024; Rossi et al. 2021).

So, the performance of my model would be more efficient I decided to implement hyperparameter tuning using GridSearchCV as recommended by Logunova (2023). This performed a search over combinations of max_depth and min_samples_split using 5-fold cross-validation, where it would find the most effective configuration of parameters based on cross-validated accuracy. By choosing n_splits = 5, a good balance between computational efficiency and model performance estimation was achieved (Chugani 2024), as model stability and model performance can then be verified. These tuning steps were utilised after the first version of my model only achieved 60% accuracy on the unseen 2023/24 test dataset. By applying cross-validation, the model was exposed to multiple training–validation splits, which helped mitigate overfitting and underfitting (Dremio 2025). Likewise, GridSearchCV enabled an automated and exhaustive search for optimal hyperparameters rather than relying on defaults, producing a stable and higher-performing model (Guerra 2024). After tuning, the model's accuracy on the 2023/24 test dataset increased significantly (to over 90%), showing that hyperparameter optimisation was a critical turning point in performance improvement.

```
#This code will create and train the decision tree model using GridSearchCV.
gkf = GroupKFold(n_splits=5)
parameters = {
    "max_depth": [3,5,10,None],
    "min_samples_split": [2,5,10]
}
dt_model = DecisionTreeClassifier(random_state=42)
grid_search = GridSearchCV(
    estimator=dt_model,
    param_grid=parameters,
    cv=gkf.split(X_train,y_train,groups=train_df["Name"]),
    scoring="f1",
    n_jobs=1
)
```

Figure 8: Python code implementing hyperparameter tuning using GridSearchCv(player position) into numerical format.

The model was first trained on the full training dataset covering 2019/20–2022/23, then validated using 5-fold cross-validation so that its robustness could be investigated (Chugani 2024). Once it was valid, the final model was retrained and evaluated on the unseen 2023/24 test dataset, enabling an out-of-sample performance assessment which is critical according to Austin et al. (2016).

Every methodological choice was made with the practical applicability of real-world football settings in mind. For instance, a Decision Tree Classifier was chosen over more complex models due to its transparency and interpretability so that the needs, knowledge, and requirements of practitioners are considered (Chen et al. 2022). Claudino et al. (2019) states many ML models do not offer practical, interpretable recommendations that coaches and medical staff can easily act upon, creating barriers to adoption.

3.5. Model Evaluation Metrics

So, the performance of the ML model could be evaluated, several metrics were utilised.

The first metric used to assess the performance of the model was accuracy, which measures the proportion of correct predictions made on the previously unseen data from the 2023/2024 English Premier League season. This allows me to see how reliable the model is as it investigates how often the model correctly finds both injured and non-injured players. As outlined by EvidentlyAI (2025), an accuracy values closer to 1.0 shows a stronger predictive performance and any lower means it could be improved.

Next, the model's precision was evaluated to determine the proportion of correctly identified injured players among all predicted to sustain an injury, thereby assessing its ability to minimise false positive predictions. EvidentlyAI (2025) states that the precision metric is important in scenarios where the cost of a false positive is high, and in the case of football, where inaccuracies can lead to unnecessary precautions, it is important. Therefore, a strong precision ensures a greater confidence in each prediction made (EvidentlyAI 2025).

Recall was utilised to evaluate the proportion of actual injuries that the model correctly identified. According to EvidentlyAI (2025) this metric specifically measures the model's ability to capture true positives (those who sustained injuries) from the total pool of injured players within the dataset. In professional football, this is key as not flagging an injury (a false negative) is undesirable due to the scarcity of players negatively affecting the performance of a team (Rossi et al. 2021).

The confusion matrix allowed for a comprehensive view of the model's performance as each prediction was broken down into four distinct categories: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). True positives refer to injured players who were correctly identified as injured, while true negatives represent those correctly predicted as not injured, then false positives occur when the model incorrectly predicts an injury for a player who did not sustain one, and false negatives reflect instances where the model failed to identify a player who was in fact injured. This metric is a vital diagnostic tool when evaluating classification models in contexts involving imbalanced classes and real-world implications of prediction errors according to EvidentlyAI (2025).

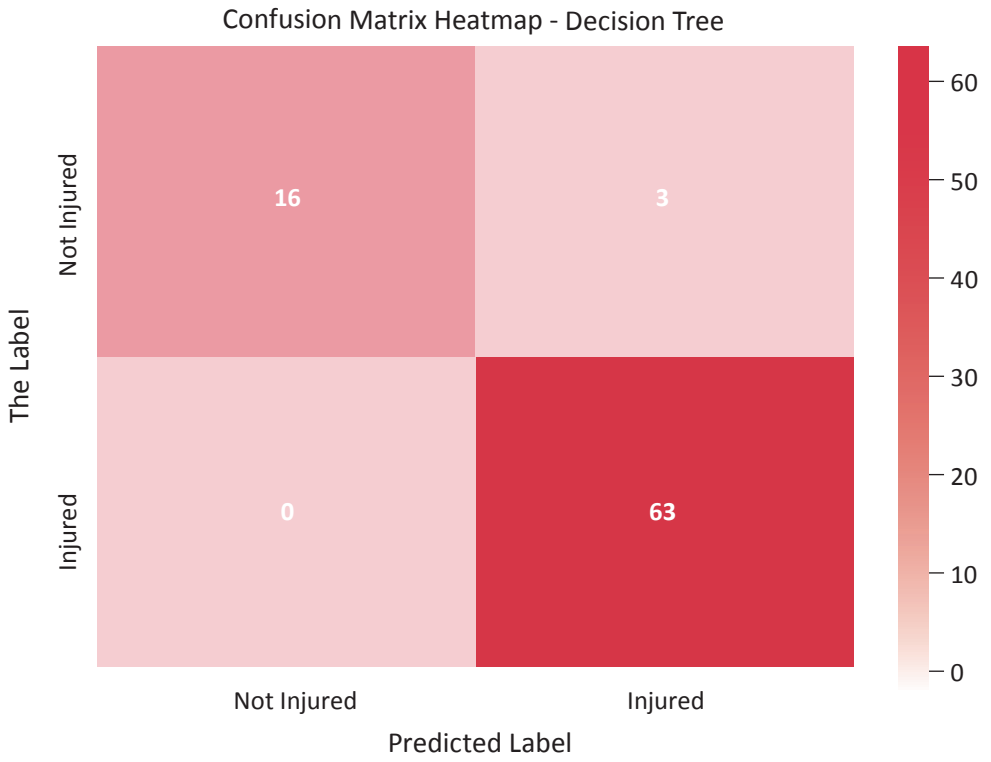

Figure 9: Confusion matrix heatmap showing true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) based on predictions made by the final Decision Tree model on the 2023/24 test set

Finally, the last metric used was F1-score, which provided a balanced evaluation of the model's performance as precision and recall were combined into a singular metric. Sharma (2023) explains that the F1-score is crucial as it captures how well the model balances the trade-off between finding true positives and avoiding false positives. In an environment were missing a genuine injury risk (false negative) and wrongly flagging a healthy player (false positive) carry implications this is invaluable. Sharma (2023) highlights how the F1-score is preferred over accuracy when the cost of false predictions is uneven and when both precision and recall are critical to a model's success.

Aside from these Metrics, 5-fold cross-validation was used during hyperparameter tuning via GridSearchCV to find the optimal combination of model parameters (max_depth and min_samples_split). These according to Singh (2023) are two important concepts in ML, as they not only improve the performance of ML models, but they also provide a robust estimate of a model's performance while simultaneously helping to find the best set of hyperparameters.

3.6. Ethics

This dissertation research project was reviewed to ensure it followed Cardiff University's ethical standards, as it is based entirely on secondary data analysis but still needed careful consideration due to the presence of identifiable information.

The dataset utilised titled Player Injuries and Team Performance by Biswas (2023), was sourced from Kaggle (an open-access platform for sharing datasets used in academic, educational, and non-commercial research). The dataset included information on football players' injury events between the 2019/20 and 2023/24 seasons. Aswell as player names alongside general demographic and performance data (e.g. position, age, injury type, and team). While the players' names are present, the data is presented in a way that is already public knowledge given the professional nature of the those involved.

The dataset is publicly available and should be used explicitly for research and educational purposes only. Nevertheless, care was taken to ensure ethical integrity and respect for data context throughout the research process and all findings will be reported in a responsible and ethical manner.

3.7. Limitations

There are several limitations to this research that should be acknowledged.

The dataset that I used only included players from a small subset of English Premier League clubs, with eight teams in the training set and seven in the test set. The limited scope of representation may have reduced the generalisability of my model across the full Premier League as injury incidence, recovery protocols, and medical support can vary across club. Exposure to various playing styles, training loads, and rehabilitation practices may have been more beneficial as the results could have been generalised as factors now may be underrepresented.

In addition, the positional distribution in both datasets was skewed toward central roles. With positions such as centre backs, central midfielders, and centre forwards being the most common. While this does align with earlier literature on the topic (Thema et al. 2025; Hawkins

and Fuller 1996) my model may be biased toward recognising injury patterns in central positions and may underperform when applied to others.

Finally, while age, position, and previous injury were selected based on fanatical support in the literature, they may not fully capture the complexity of injury risk. There is a plethora of factors that can lead an individual to have increased susceptibility to injury like match load, recovery time, and environmental conditions which should have been included to improve my model's predictive accuracy. While these are equally as important, they were excluded as the much of the data was hard to obtain. Though this was intention, it may result in a lower overall performance.

3.8. Reflexivity

During this research, I recognised the importance of adopting a reflexive approach to critically examine my own assumptions, beliefs, and judgements (Jamieson et al. 2023). This practice not only enhances the rigour of my research but also makes my positionality more transparent (Ryan and Golden 2006).

By taking a reflexive approach, I acknowledge the potential influence of my unconscious biases on the development and outcome of my research project. For instance, my methodological choices may have been influenced by my own subconscious desire to create a simpler ML model, so it was easier to create. As a student who is new to building an ML model, I may have unconsciously opted for simpler approaches, not because they were the most suitable, but because they were the most achievable given my skill level. This decision to use a Decision Tree over more complex algorithms could highlight my novice position and the tension between my own personal academic ambition as well as practical feasibility.

By acknowledging my positionality as a beginner, I aimed to increase the transparency and rigour of this project, demonstrating not only what was achieved but how and why those choices were made.

4. Findings

This findings chapter will outline the findings I obtained from this dissertation research project.

When evaluated the model achieved an overall accuracy of 96.3%, with a precision of 0.955, recall of 1.000, and an F1-score of 0.977, as shown in Figure Ten. Indicating that all players injured in the 2023/24 English Premier League season were correctly identified. However, three non-injured players were incorrectly predicted to be injured. The confusion matrix (Figure Ten) furthers illustrates this, showing 63 true positives, 16 true negatives, three false positives, and no false negatives.

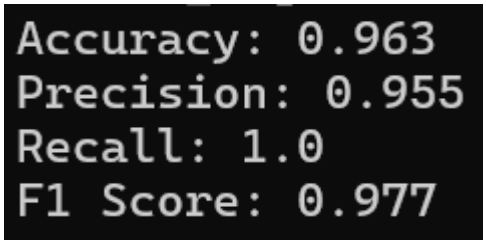

Figure 10: Performance metrics (Accuracy, Precision, Recall, F1-Score) and the results of the final model

The high recall score achieved demonstrates how effective the model is in capturing every injury case in the dataset, while the strong precision value reflects its ability to minimise incorrect positive predictions. Taken together, these results suggest that the model achieved an optimal balance between sensitivity and precision, offering both a strong predictive performance and reliability in identifying those players at risk of injury.

Likewise, the performance highlights the predictive strength of the selected features: Age, Position, and Previous Injury Count as these were chosen due to previous literature on the topic. Nonetheless, the three false positives show that while the model can find high-risk players effectively, refining the feature set and adding more that were identified in the previous literature in future research may help to reduce overprediction.

Figure Eleven (on the next page) presents the confusion matrix heatmap for the Decision Tree model, illustrating its classification performance across the binary injury outcome. The model correctly identified 63 injured players and 16 non-injured players, with no false negatives observed. However, three false positive predictions were made, as shown in Figure Twelve. These misclassifications involved Declan Rice, João Palhinha, and William Saliba each aged 25, who had positional encodings of 9, 9, and 7, respectively. These three players did not have injuries, but the predicted label indicated an injury occurrence. No other further anomalies were seen in the classification outputs, and all other predictions aligned with the ground truth labels.

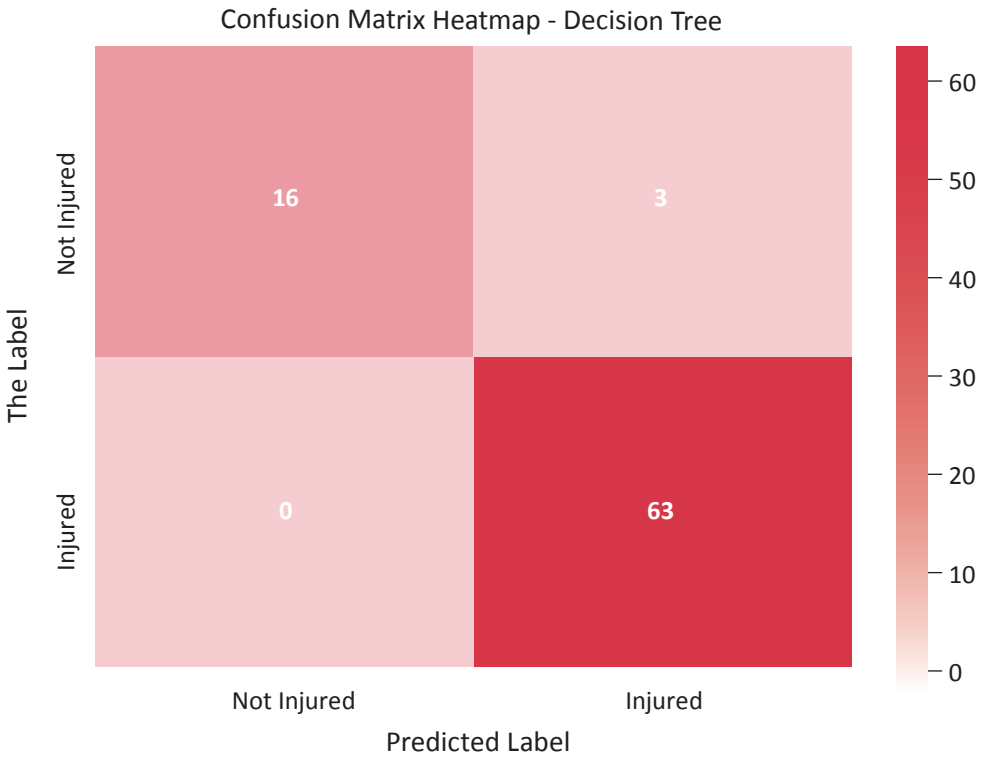

Figure 11: Confusion matrix heatmap based on predictions made by the Decision Tree model

	Name	Age	Position_enc	Actual	Predicted
88	Declan Rice	25	9	0	1
192	Joao Palhinha	25	9	0	1
366	William Saliba	25	7	0	1

Figure 12: Terminal Output Displaying the Three Misidentified Cases

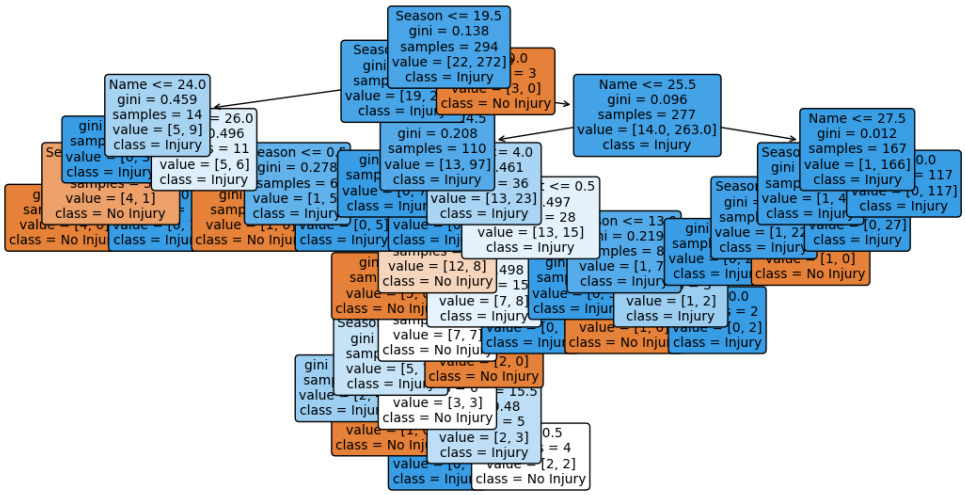

Figure 13: Visual Representation of the Final Decision Tree Model for Injury Prediction

Figure Thirteen highlights how the Decision Tree model made its decisions. As shown in Figure Thirteen the decision tree starts its decision-making process by splitting the data based on the seasons workload values, which helps to separate each player who have distinct levels of exposure. Then the model will utilising the given features (e.g. playing position, age, and previous injury count) to make predictions that are more details as well as accurate. Every node in the decision tree is coloured to show which class it will predict. Dark blue is the non-injury cases while orange means more injury cases. The Gini score at each node shows how mixed the data is at that point, with lower values meaning the split is more accurate. At the end of each branch, the leaf nodes show the final prediction that was made, how many players fall into that group, and how those predictions are distributed.

5. Discussion

The discussion chapter will critically reflect on the findings of the research, evaluating their significance in the context of existing literature, the original aims of the study, and real-world application. So, the research can be positioned in the broader landscape of injury prediction in football.

5.1. Summary of My Findings

The final Decision Tree Classifier model, trained on the 2019/20–2022/23 English Premier League seasons managed to achieve an accuracy score of 96.3%, a precision of 0.955, a recall of 1 and a F1-score of 0.977. By achieving a perfect recall score the model successfully identified every injured player, eliminating false negatives from occurring but three false positive predictions were made.

The literature review helped to prove that injuries are a multifactorial interaction of both intrinsic and extrinsic risk factors that traditional methods fail to capture (Bullock et al. 2022; Claudino et al. 2019). This project aimed to address gaps by training an ML model on temporally separated datasets rather than cross-sectional data, to respond to calls for externally valid, practitioner-oriented systems (Colby et al. 2014). Addressing Colby et al.'s (2014) limitation that ML models neglect temporal variation while aligning with Genovard et al. (2025), as such a model can be embedded into club injury risk management strategies to help players.

Most injuries in both datasets occurred among those aged 23-30, aligning with Argibay-Gonzales (2022) and Sonesson et al. (2023) as Halvorsen Wik (2022) and Faigenbaum and Myer (2010) say that match exposure, fatigue as well as the physical demands placed on a player is some of the reasons why that age experience heightened injury risk. Similarly, the weighting of

previous injury history supports Hägglund, Walden and Ekstrand (2006) as certain reoccurring injuries can increase a player's risk 4-6 times. However, because so many key variables were excluded (because of dataset limitations) the scope may be narrowing than those frameworks proposed by Meeuwisse et al. (2007).

The lack of interpretability in injury ML models identified by Claudino et al. (2019) was addressed through utilising a Decision Tree Classifier, as Scikit learn (2025) highlight how they allow for transparent, rule-based outputs that align with processes used by medical staff and coaches (Sarabia et al. 2021). Thus, high predictive performance does not have to be sacrificed for interpretability, a statement shared by Chen et al. (2024) who also advocate for explaining ML models in sport clinical decisions.

The perfect recall scored has important implications for injury prevention, given the wide-ranging impacts that injuries can have. Although three false positives were identified by the model, the fact that there were no false negatives means that no players at risk were overlooked. When compared to both Rossi et al. (2021) and Bullock et al. (2022) who achieved 83% accuracy with SVMs but had generalisation issues with black box ensemble models, this research shows that transparency and robustness can coexist.

However, every finding from this research must be treated with caution. Injury risk is inherently multifactorial, as a plethora of both intrinsic and extrinsic factors interact alongside each other that lead to injuries (Taimela et al. 1990; Meeuwise et al. 2007) many of which were unavailable in the dataset used. While features were used that were found in a plethora of literature, the model may still fail to capture the full complexity of these interactions. By including a broader range of external context variables in the future (like weather and intensity) it could help to refine the model's predictive accuracy and adaptability in predicting more injuries.

This ML model could be integrated into a club's injury prediction and prevention framework so that players at an elevated risk to sustaining an injury can be identified to help the club mitigate the wide-ranging impacts that stem from injuries. Although the deployment of the model would require consistent data collection, real-time feature updates, and stakeholder training to ensure actionable insights.

5.2. Limitations of My Research

One key limitation of this study is the restricted scope of the dataset, as it contains a limited number of players from a small subset of English Premier League clubs. Though this design reduced the risk of missing or inconsistent records, it inevitably reduced the diversity of playing styles, training regimes, and medical protocols represented. As Bullock et al. (2022) highlights, the model could have benefitted from extra exposure to varied competitive contexts so that it could be generalisable beyond the current data environment; therefore, future research could enhance generalisability by incorporating broader multi-club or multi-league datasets.

Although the model's performance metrics, in particular the recall score, initially seemed like a positive, the results may be misleading. Such high scores could indicate underlying issues such as data leakage, duplicated entries, or a lack of transparency in the data collection process (Calvino, Baker, and Korres 2024). Thus the models true predictive capacity can be concealed, and the results may be overestimated which may prove to be different when used in real-world applications.

5.3. Future Research Directions

If I were to conduct this research project again, I would:

- Utilise datasets from various football leagues around the world so the model's external validity and generalisability would be improved.
- Look to integrate additional risk factor features that affect a player's susceptibility to injury like fixture density, recovery time, and surface type, as suggested by Bengtsson et al. (2013) and Gould et al. (2023).
- Test the model under operational conditions, to align with Bullock et al. (2022)'s emphasis on external validation, to ensure the model's performance is robust prior to implementation.

6. Conclusion

The final conclusions demonstrate that it is possible to build a ML model that is both transparent and interpretable. By utilising a Decision Tree Classifier, the model achieved an overall accuracy of 96.3%, a precision of 0.955, a recall of 1.000, and an F1-score of 0.977—highlighting its strong predictive performance. These results indicate that all players injured during the 2023/24 English Premier League season were correctly identified. However, three non-injured players (Declan Rice, João Palhinha, and William Saliba, each aged 25) were incorrectly predicted to be injured. Their respective positional encodings were 9, 9, and 7. Although these players did not experience injuries, the model assigned them labels indicating injury occurrence.

This research project helped address a critical gap in the literature identified by Claudino et al. (2019), demonstrating that injury prediction models can achieve high performance while remaining transparent and interpretable for coaches and medical staff. Furthermore, by incorporating temporal separation between training and test data, the study overcame a limitation noted by Bullock et al. (2022) and Colby et al. (2014), offering a more realistic evaluation of model performance in applied settings.

7. Reflection

Partaking in this project has been a challenging yet rewarding experience as despite the limitations, I have been able to develop a wide range of technical, analytical and research skills that will benefit me in future instances. As someone with a strong and keen interest in both ML and football, I decided this topic knowing that I would stay engaged. At the beginning, my understanding of ML was extremely limited given that I am a student that is new to building an ML model. Knowing this was the case I decided to enrol in a course to build the technical foundation needed to start this project, which gave me the confidence to apply techniques like data cleaning, feature engineering and hyperparameter tuning to a real dataset that later proved critical. Taking theory and putting it into practice was difficult but it was a situation that only further strengthened my ability to build my final model.

One key lesson learnt was the constraint of working with secondary, publicly available datasets. Although these datasets allowed me to gain access to a broader range of information than I could have collected independently (Lexis Nexis 2025) they lacked key features that were highlighted in the literature that would have improved my model's predictive accuracy. Key variables absent from the dataset by Biswas (2023) include physical, environmental, socio-cultural, and psychological factors despite their relevance. This reinforced to me the importance of early and thorough data preparations so that key gaps could be anticipated, so a pivot could be made if needed.

In addition, this project revealed to me a limitation in my own methodological decision making. By taking a reflexive approach to this research, I now recognise that some of my choices may have been shaped by my own unconscious biases. As a student new to ML, I may have gravitated toward developing a simpler model, not because they were the most appropriate but because they felt more achievable given the time and my experience. For instance, the choice to utilise a decision tree over more complex algorithms show a desire to maintain interpretability and manage complexity. While this was one of my aims for this project, it may have limited the model's predictive power overall. By acknowledging this limitation, it has helped me to better understand the role of reflexivity in research and the importance of considering not only my own biases, but the broader contextual factors in research.

Beyond developing my technical skills, the project also sharpened my academic writing and critical thinking as throughout this project I learnt how to link empirical findings to existing literature that allowed me to be both coherent and reflective. Managing my references meticulously and refining the structure of each argument made helped me to become more confident in articulating and justifying my methodological choices, which I now see as essential components of rigorous research.

The skills I have gained from this project will help to inform any future academic or professional endeavours as they are highly transferable. The ability to navigate complex datasets, engineer features and critically evaluate model outputs are essential across a wide range of data driven domains, from healthcare to finance, and will be invaluable in any future role I may pursue.

Finally, this project has contributed to the ever-growing body of literature in ML and sports injury prediction as it offers a reproducible framework that can be adapted across various contexts while simultaneously demonstrating how data driven methods can augment traditional risk assessment.

8. References

Adams, W.M., Hosokawa, Y., and Casa, D.J. 2016. Body-cooling paradigm in sport: Maximizing safety and performance during competition. *Journal of Sport Rehabilitation*, 25, pp.382–394. doi: 10.1123/jsr.2015-0008.

Ahern, D.K. and Lohr, B.A. 1997. Psychosocial factors in sports injury rehabilitation. *Clinical Sports Medicine*, 16(4), pp.755–768. doi:10.1016/s0278-5919(05)70052-1.

Aiello, F., Avery, L., Gardner, T., Rutherford, H., McCall, A., and Impellizzeri, F.M. 2024. Broadening our understanding of injury mechanisms to include at-risk situations: An overview of potential injuries at the FIFA Men's World Cup Qatar 2022™. *Science and Medicine in Football*. doi:10.1080/24733938.2024.2372304.

Alahmad, T.A., Kearney, P. and Cahalan, R. 2020. Injury in elite women's soccer: A systematic review. *The Physician and Sportsmedicine*, 48(3), pp.259–265. doi:10.1080/00913847.2020.1720548.

Andersen, M.B. and Williams, J.M. 1999. Athletic injury, psychosocial factors, and perceptual changes during stress. *Journal of Sports Sciences*, 17, pp.735–741. doi: 10.1080/026404199365597

Andersen, T.E., Arnason, Á., Engebretsen, L. and Bahr, R. 2004. Mechanisms of head injuries in elite football. *British Journal of Sports Medicine*, 38(6), pp. 690–696. doi: 10.1136/bjsm.2003.009357.

Andrade, R., Wik, E.H., Rebelo-Marques, A., Blanch, P., Whiteley, R., Espregueira-Mendes, J. and Gabbett, T.J. 2020. Is the Acute: Chronic Workload Ratio (ACWR) Associated with Risk of Time-Loss Injury in Professional Team Sports? A Systematic Review of Methodology, Variables and Injury Risk in Practical Situations. *Sports Medicine*, 50(9), pp.1613–1635. doi:10.1007/s40279-020-01308-6

Argibay-Gonzalez, J.C., Vazquez-Estevez, C., Gutierrez-Santiago, A., Parames-Gonzalez, A., Reguera-Lopez-de-la-Osa, X. and Prieto-Lage, I. 2022. Analysis of injury patterns in men's football between the English league and the Spanish league. *International Journal of Environmental Research and Public Health*, 19(18), p. 11296. doi: 10.3390/ijerph191811296.

Arnason, Á., Sigurdsson, S.B., Gudmundsson, Á. and Holme, I. 2004. Risk factors for injuries in football. *The American Journal of Sports Medicine*, 32(1 Suppl), pp. 5–16. doi: 10.1177/0363546503258912

Austin, P.C., van Klaveren, D., Vergouwe, Y., Nieboer, D., Lee, D.S. and Steyerberg, E.W. 2016. Geographic and temporal validity of prediction models: different approaches were useful to examine model performance. *Journal of Clinical Epidemiology*, 79, pp. 76–85. doi: 10.1016/j.jclinepi.2016.05.007.

Backous, D.D., Friedl, K.E., Smith, N.J., Parr, T.J. and Carpine, W.D. Jr. 1988. Soccer injuries and their relation to physical maturity. *American Journal of Diseases of Children*, 142, pp. 839–842.doi: 10.1001/archpedi.1988.02150080045019

Bangert, Y., Jaber, A., Trefzer, R., Zietzschmann, S., Koch, K.A., Kern, R., Spielmann, J., Renkawitz, T. and Weishorn, J. 2024. The impact of injury on career progression in elite youth football: Findings at 10 years. *Journal of Clinical Medicine*, 13(7), p. 1915. doi: 10.3390/jcm13071915.

Bangsbo, J., Krustrup, P., Hansen, P.R., Ottesen, L., Pfister, G. and Elbe, A-M. 2016. Science and Football VIII: The Proceedings of the Eighth World Congress on Science and Football. London: Routledge. doi: 10.4324/9781315670300.

Bangsbo, J., Mohr, M. and Krustrup, P. 2006. Physical and metabolic demands of training and match-play in the elite football player. In: R. J. Maughan, ed. *Nutrition and Football: The FIFA/FMARC Consensus on Sports Nutrition*. 1st ed. London: Routledge, pp. 1–26.doi: 10.1080/02640410500482529

Bangsbo, J., Nørregaard, L. and Thorsø, F. 1991. Activity profile of competition soccer. *Canadian Journal of Sport Sciences = Journal Canadien des Sciences du Sport*, 16(2), pp. 110–116.

Barker, J. 2021. *Not just a physical thing: The psychology of sports injuries and recovery*. Available at: https://answers.childrenshospital.org/psychology-sports-injury-recovery/ [Accessed: 25 May 2025].

Baumhauer, J.F., Alosa, D.M., Renstrom, A.F., Trevino, S. and Beynnon, B. 1995. A prospective study of ankle injury risk factors. *American Journal of Sports Medicine*, 23, pp. 564–570. doi: 10.1177/036354659502300508

Beaumont Health. 2024. *ACL tears*. Available at: https://www.beaumont.org/conditions/acl-tears [Accessed: 23 March 2025].

Beckles, O. 2023. FA's new non-league contract will cut injured players' wages. It must be stopped. *The Guardian*, 30 March. Available at: https://www.theguardian.com/football/2023/mar/30/fa-new-non-league-contract-cut-injured-players-wages-must-be-stopped-pfa [Accessed: 6 June 2025].

Belev, J. 2022. *Bruno burning out? Fernandes leads footballers for most games in 2021*. Available at: https://www.globaltimes.cn/page/202201/1245267.shtml [Accessed: 20 May 2025].

Bengtsson, H., Ekstrand, J. and Hägglund, M. 2013. Muscle injury rates in professional football increase with fixture congestion: An 11-year follow-up of the UEFA Champions League injury study. *British Journal of Sports Medicine*, 47(12), pp. 743–747. doi: 10.1136/bjsports-2013-092383.

Bertschy, M., Howard, J.T., Oyama, S. and Cheever, K. 2021. Reduced injury prevalence in soccer athletes following GPS guided acclimatization. *International Journal of Exercise Science*, 14(7), pp. 1070–1077. doi: 10.70252/ZJJG8017.

Bin Rashid, A., and Kausik, M.A.K. 2024. AI revolutionizing industries worldwide: A comprehensive overview of its diverse applications. *Journal of Emerging Technologies*, 22(4), pp. 101–120.doi: https://doi.org/10.1016/j.hybadv.2024.100277

Biswas, A. 2024. *Player injuries and team performance dataset*. Available at: https://www.kaggle.com/datasets/amritbiswas007/player-injuries-and-team-performance-dataset [Accessed: 3 June 2025].

Bittencourt, N.F.N., Meeuwisse, W.H., Mendonça, L.D., Nettel-Aguirre, A., Ocarino, J.M. and Fonseca, S.T. 2016. Complex systems approach for sports injuries: Moving from risk factor identification to injury pattern recognition. *British Journal of Sports Medicine*, 50(21), pp. 1309–1314.doi: 10.1136/bjsports-2015-095850

Blomstrand, E., Bergh, U., Essén-Gustavsson, B. and Ekblom, B. 1984. Influence of low muscle temperature on muscle metabolism during intense dynamic exercise. *Acta Physiologica Scandinavica*, 120(2), pp. 229–236. doi: 10.1111/j.1748-1716.1984.tb00128.x.

Boden, B.P., Dean, G.S., Feagin, J.A. Jr. and Garrett, W.E. Jr. 2000. Mechanisms of anterior cruciate ligament injury. *Orthopedics*, 23(6), pp. 573–578. doi: 10.3928/0147-7447-20000601-15.

Bowen, L., Gross, A.S., Gimpel, M. and Li, F.X. 2017. Accumulated workloads and the acute:chronic workload ratio relate to injury risk in elite youth football players. *British Journal of Sports Medicine*, 51(5), pp. 452–459. doi: 10.1136/bjsports-2015-095820.

Brazier, J., Antrobus, M.R., Callus, P.C., Herbert, A.J., Stebbings, G.K., Martin, D., Day, S.H., Kilduff, L.P., Bennett, M.A., Erskine, R.M., Raleigh, S.M., Cullen, T., Collins, M., Pitsiladis, Y.P., Heffernan, S.M. and Williams, A.G. 2025. Variants within the MMP3 and COL5A1

genes associate with soft tissue injury history in elite male rugby athletes. *Journal of Science and Medicine in Sport*. doi: https://doi.org/10.1016/j.jsams.2025.05.007

Brewer, B.W. 2017. Psychological responses to sport injury. *Oxford Research Encyclopedia of Psychology*. doi https://doi.org/10.1093/acrefore/9780190236557.013.172

Brito, J., Racinais, S. and Nassis, G.P. 2014. The second Summer Youth Olympic Games in Nanjing, People's Republic of China: Preparing youth athletes to compete in the heat. Open Access *Journal of Sports Medicine*, 5, pp. 205–207. doi: 10.2147/OAJSM.S70278.

Brophy, R., Backus, S., Pansy, B., Lyman, S. and Williams, R. 2007. Lower extremity muscle activation and alignment during the soccer instep and side-foot kicks. *Journal of Orthopaedic & Sports Physical Therapy*, 37(5), pp. 260–268.doi: 10.2519/jospt.2007.2255

Buchheit, M., Mendez Villanueva, A., Simpson, B.M. and Bourdon, P.C. 2010. Match running performance and fitness in youth soccer. *International Journal of Sports Medicine*, 31(11), pp. 818–825.doi: 10.1055/s-0030-1262838

Bullock, G.S., Collins, G.S., Malhotra, A., Eisen, D.P. and Pollock, N. 2022. 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 Medicine*, 52(10), pp. 2469–2482. doi: 10.1007/s40279-022-01698-9.

Burton-Hughes, L. 2016. *Preventing slips, trips, and falls: Control measures & tips*. Available at: https://www.highspeedtraining.co.uk/hub/preventing-slips-trips-falls/ [Accessed: 22 May 2025].

Calvino, C., Baker, N. & Korres, D., 2024. Machine Learning Model Metrics – Can I Trust Them? Available at: https://www.fticonsulting.com/insights/articles/machine-learning-model-metrics-trust-them [Accessed 7 September 2025].

Carfagno, D.G. and Hendrix, J.C. III. 2014. Overtraining syndrome in the athlete: Current clinical practice. *Current Sports Medicine Reports*, 13(1), pp. 45–51. doi: 10.1249/JSR.0000000000000027.

Carling, C., McCall, A., Le Gall, F. and Dupont, G. 2016. The impact of short periods of match congestion on injury risk and patterns in an elite football club. *British Journal of Sports Medicine*, 50(12), pp. 764–768. doi: 10.1136/bjsports-2015-095501.

Carter, B. 2014. *Football injuries: 'I lost my career to a hamstring tear'*. Available at: https://www.bbc.co.uk/news/magazine-28254123 [Accessed: 15 July 2025].

Castro, B. 2023. *Positional demands and the link to soccer injuries*. Available at: https://www.isspf.com/articles/positional-demands-and-the-link-to-soccer-injuries/ [Accessed: 26 June 2025].

Celik, D. 2024. *Data science & machine learning: how the future of football will shape*. Available at: https://www.linkedin.com/pulse/data-science-machine-learning-how-future-football-shape-celik-phd-8l9fe/ [Accessed: 11 June 2025].

Chen, H., Gomez, C., Huang, C.M. and Unberath, M. 2022. Explainable medical imaging AI needs human-centred design: Guidelines and evidence from a systematic review. *npj Digital Medicine*, 5, article 156. doi: 10.1038/s41746-022-00699-2.

C Chen, S., Gong, J., Wang, Y., Zhang, X., Xu, H. and Zhou, M. 2024. Integrating machine learning and artificial intelligence in life-course epidemiology: Pathways to innovative public health solutions. *BMC Medicine*, 22, p. 354. doi: 10.1186/s12916-024-03566-x.

Cherry, K. 2025. *Consciousness in psychology: Levels and states of awareness*. Available at: https://www.verywellmind.com/what-is-consciousness-2795922 [Accessed: 6 June 2025].

Chomiak, J., Junge, A., Peterson, L. and Dvorak, J. 2000. Severe injuries in football players: Influencing factors. *American Journal of Sports Medicine*, 28(Suppl 5), pp. S58–S68.doi: 10.1177/28.suppl_5.s-58

Christino, M.A., Fantry, A.J. and Vopat, B.G. 2015. Psychological aspects of recovery following anterior cruciate ligament reconstruction. *JAAOS – Journal of the American Academy of Orthopaedic Surgeons*, 23(8), pp. 501–509.doi: 10.5435/JAAOS-D-14-00173

Chugani, V. 2024. *A comprehensive guide to K Fold Cross Validation*. Available at: https://www.datacamp.com/tutorial/k-fold-cross-validation [Accessed: 15 July 2025].

Claudino, J.G., Capanema, D.d., de Souza, T.V., Serra-Olivares, J., Pereira, A.C. and Nakamura, F.Y. 2019. Current approaches to the use of artificial intelligence for injury risk assessment and performance prediction in team sports: A systematic review. *Sports Medicine*, 5:28. doi: 10.1186/s40798-019-0202-3.

Clubb, J. 2023. *Moving from reactive to proactive training load management*. Available at: https://www.globalperformanceinsights.com/post/moving-from-reactive-to-proactive-training-load-management [Accessed: 11 June 2025].

Colby, M.J., Dawson, B., Heasman, J., Rogalski, B. and Gabbett, T.J. 2014. Accelerometer and GPS derived running loads and injury risk in elite Australian footballers. *Journal of Strength and Conditioning Research*, 28(8), pp. 2244–2252. doi: 10.1519/JSC.0000000000000362.

Collins, C. 2017. Resistance training, recovery, and genetics: AMPD1 the gene for recovery. *Journal of Athletic Enhancement*, 6(2). Available at: https://www.scitechnol.com/peer-review/resistance-training-recovery-and-genetics-ampd1-the-gene-for-recovery-acCT.php?article_id=5980

Costa E Silva, L., Teles, J. and Fragoso, I. 2022. Sports injuries patterns in children and adolescents according to their sports participation level, age, and maturation. *BMC Sports Science, Medicine and Rehabilitation*, 14(1), p. 35. doi: 10.1186/s13102-022-00431-3.

Crossman, J. 1997. Psychological rehabilitation from sports injuries. *Sports Medicine*, 23(5), pp. 333–339. doi: 10.2165/00007256-199723050-00005.

Daily Emerald. 2025. *How sports injuries can impact an athlete's career and subsequently their value*. Available at: https://dailyemerald.com/165357/promotedposts/how-sports-injuries-can-impact-an-athletes-career-and-subsequently-their-value/ [Accessed: 6 June 2025].

Dalen-Lorentsen, T., Bjørneboe, J., Clarsen, B., Myklebust, G. and Andersen, T.E. 2021. Does load management using the acute:chronic workload ratio prevent health problems? A cluster randomised trial of 482 elite youth footballers of both sexes. *British Journal of Sports Medicine*, 55, pp. 108–114. doi: https://doi.org/10.1136/bjsports-2020-103003

Daley, M.M., Griffith, K., Milewski, M.D. and Christino, M.A. 2021. The mental side of the injured athlete. *JAAOS – Journal of the American Academy of Orthopaedic Surgeons*, 29(12), pp. 499–506.doi: 10.5435/JAAOS-D-20-00974

DataHeroes. 2023. *Decision Trees: An overview and practical guide*. Available at: https://dataheroes.ai/blog/decision-trees-an-overview-and-practical-guide/ [Accessed: 7 June 2025].

Delaney, J.S., Lacroix, V.J., Leclerc, S., and Johnston, K.M. 2000. *Concussions during the 1997 Canadian Football League season. Clinical Journal of Sport Medicine*, 10, pp. 9–14.doi: 10.1097/00042752-200001000-00003

DeLang, M. and Ishøi, L. 2024. *Groin injuries in football: The hidden epidemic affecting player performance*. Available at: https://www.sportsmith.co/articles/groin-injuries-in-football-the-hidden-epidemic-affecting-player-performance/ [Accessed: 23 May 2025].

Della Villa, F., Mandelbaum, B.R. and Lemak, L.J. 2018. The effect of playing position on injury risk in male soccer players: Systematic review of the literature and risk considerations for each playing position. *American Journal of Orthopedics*, 47(10).doi: 10.12788/ajo.2018.0092

Devlin, P. 2023. *False starts: Navigating false positives in injury prediction with big data and AI*. Available at: https://www.linkedin.com/pulse/false-starts-navigating-positives-injury-prediction-big-paul-devlin/ [Accessed: 15 July 2025].

Dhinsa, B. 2022. *Ankle sprains and the risk of chronic instability*. Available at: https://www.topdoctors.co.uk/medical-articles/ankle-sprains-and-the-risk-of-chronic-instability/ [Accessed: 25 May 2025].

Di Mascio, M. and Bradley, P.S. 2013. Evaluation of the most intense high intensity running period in English FA Premier League soccer matches. *Journal of Strength and Conditioning Research*, 27(4), pp. 909–915. doi: 10.1519/JSC.0b013e31825ff099.

Dinnery, B. 2024. *EPL 2023–24 season injury review & financial impact*. Available at: https://www.premierinjuries.com/article/epl-2023-24-season-injury-review-financial-impact [Accessed: 14 February 2025].

Djaoui, L., Haddad, M., Chamari, K. and Dellal, A. 2017. Monitoring training load and fatigue in soccer players with physiological markers. *Physiology & Behavior*, 181, pp. 86–94. doi: 10.1016/j.physbeh.2017.09.004.

Drawer, S. and Fuller, C.W. 2002. Evaluating the level of injury in English professional football using a risk-based assessment process. *British Journal of Sports Medicine*, 36(6), pp. 446–451.doi:10.1136/bjsm.36.6.446

Dremio. 2025. *Hyperparameter tuning*. Available at: https://www.dremio.com/wiki/hyperparameter-tuning/ [Accessed: 27 July 2025].

Dvorak, J. and Junge, A. 2000. Football injuries and physical symptoms: A review of the literature. *The American Journal of Sports Medicine*, 28(5 Suppl), pp. S3–S9. doi: 10.1177/28.suppl_5.s-3.

Dvorak, J., Junge, A., Chomiak, J., Graf-Baumann, T., Peterson, L., Rösch, D. and Hodgson, R. 2000. Risk factor analysis for injuries in football players: Possibilities for a prevention program. *American Journal of Sports Medicine*, 28(5 Suppl), pp. S69–S74.doi: 10.1177/28.suppl_5.s-69

Eirale, C., Tol, J.L., Farooq, A., Smiley, F. and Chalabi, H. 2013. Low injury rate strongly correlates with team success in Qatari professional football. *British Journal of Sports Medicine*, 47(12), pp. 807–808. doi: 10.1136/bjsports-2012-091040.

Ekstrand, J. 2013. Keeping your top players on the pitch: The key to football medicine at a professional level. *British Journal of Sports Medicine*, 47(12), pp. 723–724.doi: 10.1136/bjsports-2013-092771

Ekstrand, J. and Gillquist, J. 1983. The avoidability of soccer injuries. *International Journal of Sports Medicine*, 4, pp. 124–128.doi: 10.1055/s-2008-1026025

Ekstrand, J., Gillquist, J., Möller, M., Öberg, B. and Liljedahl, S.O. 1983. Incidence of soccer injuries and their relation to training and team success. *American Journal of Sports Medicine*, 11, pp. 63–67.doi: 10.1177/036354658301100203

Ekstrand, J., Hägglund, M. and Waldén, M. 2009. Injury incidence and injury patterns in professional football: The UEFA injury study. *British Journal of Sports Medicine*, 43(6), pp. 393–399. doi: 10.1136/bjsm.2008.052884.

Ekstrand, J., Hägglund, M. and Waldén, M., 2011. Epidemiology of muscle injuries in professional football (soccer). *The American Journal of Sports Medicine*, 39(6), pp.1226–1232. doi: 10.1177/036354651039587.

Ekstrand, J., Hägglund, M., Kristenson, K., Magnusson, H. and Waldén, M., 2013. Fewer ligament injuries but no preventive effect on muscle injuries and severe injuries: An 11-year follow-up of the UEFA Champions League injury study. *British Journal of Sports Medicine*, 47(12), pp.732–737. doi: 10.1136/bjsports-2013-092394.

Evidently AI, 2025. *Accuracy vs. precision vs. recall in machine learning*. Available at: https://www.evidentlyai.com/classification-metrics/accuracy-precision-recall [Accessed 15 July 2025]

Evidently AI, 2025. *How to interpret a confusion matrix for a machine learning model*. Available at: https://www.evidentlyai.com/classification-metrics/confusion-matrix [Accessed 15 July 2025].

Faigenbaum, A.D. and Myer, G.D., 2010. Resistance training among young athletes: Safety, efficacy, and injury prevention effects. *British Journal of Sports Medicine*, 44(1), pp.56–63. doi: 10.1136/bjsm.2009.068098.

FIFA, 2021. *Aerial duel*. Available at: https://www.fifatrainingcentre.com/en/resources-tools/

football-language/in-contest/aerial-duel/index.php [Accessed 22 May 2025].

Frothingham, S. 2020. *Muscle imbalance: causes, treatment, and prevention*. Available at: https://www.healthline.com/health/muscle-imbalance [Accessed 3 June 2025].

Fuller, C.W., Ekstrand, J., Junge, A., Andersen, T.E., Bahr, R., Dvorak, J., Hägglund, M., McCrory, P., and Meeuwisse, W.H., 2006. Consensus statement on injury definitions and data collection procedures in studies of football injuries. *British Journal of Sports Medicine*, 40(3), pp.193–201. doi: 10.1136/bjsm.2005.025270.

Fulton, J., Wright, K., Kelly, M., Zebrosky, B., Zanis, M., Drvol, C. and Butler, R. 2014. 'Injury risk is altered by previous injury: a systematic review of the literature and presentation of causative neuromuscular factors,' *International Journal of Sports Physical Therapy*, 9(5), pp. 583–595.

Garrick, J.G. and Requa, R.K., 2003. Sports and fitness activities: the negative consequences. *Journal of the American Academy of Orthopaedic Surgeons*, 11(6), pp.439–443. doi:10.5435/00124635-200311000-00008.

GeeksforGeeks. 2025. *Passing categorical data to Sklearn Decision Tree*. Available at: https://www.geeksforgeeks.org/passing-categorical-data-to-sklearn-decision-tree/ [Accessed 7 June 2025].

Genovard, F., Muñoz, J., Petchamé, J. and Solanellas, F., 2025. *Risk management approaches in sports organisations: A scoping review*. doi: https://doi.org/10.1016/j.heliyon.2025.e42270

George, K. and Brand, G., 2022. *How injury carries mental trauma: psychological rehab key for recovery*. Available at: https://www.skysports.com/football/news/11661/11990386/how-injury-carries-mental-trauma-psychological-rehab-key-for-recovery [Accessed 12 June 2025].

Google Developers. 2025. *Accuracy, Precision, and Recall*. Available at: https://developers.google.com/machine-learning/crash-course/classification/accuracy-precision-recall [Accessed 27 July 2025].

Gracz, J. and Sankowski, T., 2000. *Psychologia sportu*. Poznań: Akademia Wychowania Fizycznego.

Guerra, M., 2024. *Automating Model Selection and Hyperparameter Tuning with GridSearchCV*. Available at: https://www.accel.ai/anthology/2024/10/30/automating-model-selection-

and-hyperparameter-tuning-with-gridsearchcv [Accessed: 27 July 2025].

Gurau, T.V., Gurau, G., Voinescu, D.C., Anghel, L., Onose, G., Iordan, D.A., Munteanu, C., Onu, I., and Musat, C.L. 2023. Epidemiology of injuries in men's professional and amateur football (Part I). Journal of Clinical Medicine 12(17), 5569. doi: 10.3390/jcm12175569.

H2O.ai. 2025. What is Binary Classification. Available at: https://h2o.ai/wiki/binary-classification/ [Accessed 3 June 2025].

Hägglund, M., Waldén, M. and Ekstrand, J. 2006. Previous injury as a risk factor for injury in elite football: A prospective study over two consecutive seasons. *British Journal of Sports Medicine*, 40(9), pp.767–772. doi: https://doi.org/10.1136/bjsm.2006.026609

Hägglund, M., Waldén, M. and Ekstrand, J., 2006. Previous injury as a risk factor for injury in elite football: A prospective study over two consecutive seasons. *British Journal of Sports Medicine*, 40(9), pp.767–772. doi: 10.1136/bjsm.2006.026609.

Hägglund, M., Waldén, M., Magnusson, H., Kristenson, K., Bengtsson, H. and Ekstrand, J., 2013. Injuries affect team performance negatively in professional football: An 11-year follow-up of the UEFA Champions League injury study. *British Journal of Sports Medicine*, 47(12), pp.738–742. doi: 10.1136/bjsports-2013-092215.

Hall, E.C.R., Larruskain, J., Gil, S.M., Lekue, J.A., Baumert, P., Rienzi, E., Moreno, S., Tannure, M., Murtagh, C.F., Ade, J.D., Squires, P., Orme, P., Anderson, L., Whitworth-Turner, C.M., Morton, J.P., Drust, B., Williams, A.G. and Erskine, R.M. (2022) 'Playing position and the injury incidence rate in male academy soccer players', *Journal of Athletic Training*, 57(7), pp. 696–703. doi: https://doi.org/10.4085/1062-6050-0346.21

Hanson, C., 2022. *Albion secure ninth spot with final-day win over West Ham*. Available at: https://web.archive.org/web/20220523122504/https://www.brightonandhovealbion.com/news/2629952/albion-secure-ninth-spot-with-final-day-win-over-west-ham/ [Accessed 11 June 2025].

Hawkins, R.D. and Fuller, C.W. 1999. A prospective epidemiological study of injuries in four *English professional football clubs*. British Journal of Sports Medicine, 33, pp.196–203.

Howden, 2022. *European Football Injury Index: September 2022*. Available at: https://www.howdengroup.com/sites/corporate.howdenprod.com/files/2022-09/Football-Injury-Index-September-22.pdf [Accessed 15 July 2025].

Hulatt, L., 2024. *Movement analysis: techniques & meaning*. Available at: https://www.

studysmarter.co.uk/explanations/sports-science/physiotherapy/movement-analysis/ [Accessed 3 June 2025].

IBM, n.d. *Machine learning*. Available at: https://ibm.com/think/topics/machine-learning [Accessed 13 February 2025].

Inklaar, H. 1994. *Soccer injuries*. II: Aetiology and prevention. Sports Medicine, 18, pp.81–93.

Inklaar, H., Bol, E., Schmikli, S.L. and Mosterd, W.L., 1996. Injuries in male soccer players: Team risk analysis. *International Journal of Sports Medicine*, 17, pp.229–234.

Inside FIFA, 2025. *Injury prevention and health promotion*. Available at: https://inside.fifa.com/health-and-medical/injury-prevention [Accessed 23 March 2025].

Iskra, M., Voigt, L. and Raab, M., 2024. Accounting for dynamic cognition–action interaction in decision-making tasks in sports: A scoping review. *Sport, Exercise, and Performance Psychology*. doi:https://doi.org/10.1037/spy0000361

Ivarsson, A. and Johnson, U., 2010. Psychological factors as predictors of injuries among senior soccer players: A prospective study. *Journal of Sports Science and Medicine*, 9(2), pp.347–352.

Jamieson, M.K., Govaart, G.H. and Pownall, M. 2023. Reflexivity in quantitative research: A rationale and beginner's guide. *Social and Personality Psychology Compass* 17(3), pp. 1-15. doi: https://doi.org/10.1111/spc3.12735

Jones, S. and Williams, M., 2024. *Injury prevalence, strength training and performance in youth football*. Available at: https://exercise.research.southwales.ac.uk/blog/injury-prevalence-strength-training-and-performance-youth-football/ [Accessed 21 May 2025].

Junge, A., 2000. The influence of psychological factors on sports injuries: Review of the literature. *American Journal of Sports Medicine*, 28(5 Suppl), pp. S10–S15.doi:10.1177/28.suppl_5.s-10

Junge, A., Dvorak, J., Rösch, D., Graf-Baumann, T., Chomiak, J. and Peterson, L., 2000b. Psychological and sport-specific characteristics of football players. *American Journal of Sports Medicine*, 28(5 Suppl), pp. S22–S28.doi:10.1177/28.suppl_5.s-22

Kalkhoven, J.T., Watsford, M.L., Coutts, A.J., Edwards, W.B. and Impellizzeri, F.M., 2021. Training load and injury: Causal pathways and future directions. *Sports Medicine*, 51(6), pp.1137–1150. doi:10.1007/s40279-020-01413-6.

Kaplan, S., 2025. *Can machine learning models give an accuracy of 100%?* Available at: https://enjoymachinelearning.com/blog/can-machine-learning-models-give-an-accuracy-of-100/ [Accessed 27 July 2025].

Kaplánová, A., 2024. Psychological readiness of football players for the match and its connection with self-esteem and competitive anxiety. *Heliyon*, 10(6), e27608. doi: 10.1016/j.heliyon.2024.e27608.

Kenmare, J., 2023. J*ack Wilshere details the worst injury of his career that a doctor had only seen once before*. Available at: https://www.sportbible.com/football/jack-wilshere-arsenal-injury-ben-foster-premier-league-883704-20230614 [Accessed 11 June 2025].

Klein, C., Luig, P., Henke, T., Bloch, H. and Platen, P. 2021. How do football (soccer) injuries occur? A systematic video analysis of 345 moderate and severe match injuries. *British Journal of Sports Medicine* 55(Suppl 1), pp. A2.3–A3. doi: 10.1136/bjsports-2021-IOC.5.

Klein, C., Luig, P., Henke, T., Bloch, H. and Platen, P., 2020. Nine typical injury patterns in German professional male football (soccer): A systematic visual video analysis of 345 match injuries. *British Journal of Sports Medicine*, Published ahead of print. doi: 10.1136/bjsports-2019-101344.

Koch, M., Klügl, M., Frankewycz, B., Lang, S., Worlicek, M., Popp, D., Alt, V. and Krutsch, W., 2021. Football-related injuries are the major reason for the career end of professional male football players. Knee Surgery, *Sports Traumatology, Arthroscopy*, 29(11), pp.3560–3568. Available at: https://doi.org/10.1007/s00167-021-06684-8

Larruskain, J., Lekue, J.A., Martin-Garetxana, I. and Barrio, I., 2021. Injuries are negatively associated with player progression in an elite football academy. *Science and Medicine in Football*, 6(4), pp.338–344. doi: 10.1080/24733938.2021.1943756.

Lathlean, T., Gastin, P., and Newstead, S.V. 2020. Player wellness (soreness and stress) and injury in elite junior Australian football players over 1 season. *International Journal of Sports Physiology and Performance*, 15(10), pp.1–8. doi: https://doi.org/10.1123/ijspp.2019-0828

Lathlean, T.J., Gastin, P.B., Newstead, S.V. and Finch, C.F. 2019. A prospective cohort study of load and wellness (sleep, fatigue, soreness, stress, and mood) in elite junior Australian football players. *International Journal of Sports Physiology and Performance*, 14(6), pp.829–840.doi: 10.1123/ijspp.2018-0372

Leckey, C., van Dyk, N., Doherty, C., Lawlor, A. and Delahunt, E. 2024. Machine learning approaches to injury risk prediction in sport: a scoping review with evidence synthesis.

British Journal of Sports Medicine 59(7), pp.491–500. doi: 10.1136/bjsports-2024-108576.

Lee, S., 2025. *Biomechanics of injury risk: exploring principles, assessments & prevention*. Available at: https://www.numberanalytics.com/blog/biomechanics-injury-risk-prevention [Accessed 11 June 2025].

Lehnert, M., De Ste Croix, M., Zaatar, A., Hughes, J., Varekova, R. and Lastovicka, O., 2017. Muscular and neuromuscular control following soccer-specific exercise in male youth: Changes in injury risk mechanisms. *Scandinavian Journal of Medicine & Science in Sports*, 27(9), pp.975–982. https://doi.org/10.1111/sms.12705

LexisNexis 2024. *Secondary data analysis*. Available at: https://www.lexisnexis.com/en-int/glossary/data-as-a-service/secondary-data-analysis [Accessed 3 June 2025].

Lindenfeld, T.N., Schmitt, D.J., Hendy, M.P., Mangine, R.E. and Noyes, F.R., 1994. Incidence of injury in indoor soccer. *American Journal of Sports Medicine*, 22, pp.364–371. doi:10.1177/036354659402200312

Lira, B., O'Brien, J.M., Peña, P.A., Galla, B.M., D'Mello, S., Yeager, D.S., Defnet, A., Kautz, T., Munkacsy, K. and Duckworth, A.L., 2022. Large studies reveal how reference bias limits policy applications of self-report measures. *Scientific Reports*, 12, p.19189.doi: https://doi.org/10.1038/s41598-022-23373-9

Logunova, I., 2023. *Hyperparameter tuning and evaluation of ML models*. Available at: https://serokell.io/blog/hyperparameter-tuning-in-ml [Accessed 15 Jul. 2025].

London Bridge Orthopaedics n.d. *Stress and its impact on sports injuries – is there a link?* Available at: londonbridgeorthopaedics.co.uk [Accessed 22 May 2025]

López-Valenciano, A., Ruiz-Pérez, I., Garcia-Gómez, A., Vera-Garcia, F.J., De Ste Croix, M., Myer, G.D. and Ayala, F., 2020. Epidemiology of injuries in professional football: A systematic review and meta-analysis. *British Journal of Sports Medicine*, 54(12), pp.711–718. doi: 10.1136/bjsports-2018-099577.

Lv, Z.T., Gao, S.T., Cheng, P., Liang, S., Yu, S.Y., Yang, Q. and Chen, A.M. 2017. Association between polymorphism rs12722 in COL5A1 and musculoskeletal soft tissue injuries: a systematic review and meta-analysis. *Oncotarget*, 9(20), pp.15365–15374. doi: https://doi.org/10.18632/oncotarget.23805

Maddison, R. and Prapavessis, H., 2005. A psychological approach to the prediction and prevention of athletic injury. *Journal of Sport and Exercise Psychology*, 27(3), pp.289–310.

doi: 10.1123/jsep.27.3.289.

Mahmood, I. and Pettinato, M., 2021. Impact of intrinsic and extrinsic factors on the pharmacokinetics of peptides: When is the assessment of certain factors warranted? *Antibodies* (Basel), 11(1), p.1. doi: 10.3390/antib11010001. PMID: 35076485; PMCID: PMC8788552.

Majumdar, A., Bakirov, R., Hodges, D., Scos, S. and Rees, T., 2022. Machine learning for understanding and predicting injuries in football. *Sports Medicine - Open*, 8(1), p.73. doi: 10.1186/s40798-022-00465-4. PMID: 35670925; PMCID: PMC9174408.

Markovic, G., Šarabon, N., Pausic, J. and Hadžić, V., 2020. Adductor muscles strength and strength asymmetry as risk factors for groin injuries among professional soccer players: A prospective study. *International Journal of Environmental Research and Public Health*, 17(14), p.4946. doi: 10.3390/ijerph17144946.

Massidda, M., Miyamoto-Mikami, E., Kumagai, H., Ikeda, H., Shimasaki, Y., Yoshimura, M., Cugia, P., Piras, F., Scorcu, M., Kikuchi, N., Calò, C.M. and Fuku, N., 2020. Association between the ACE I/D polymorphism and muscle injuries in Italian and Japanese elite football players. *Journal of Sports Sciences*, 38(21), pp.2423–2429. doi: 10.1080/02640414.2020.1787683.

Master's in data science, 2025. *What Is a Decision Tree?* Available at: https://www.mastersindatascience.org/learning/machine-learning-algorithms/decision-tree/ [Accessed 12 June 2025].

MBP, 2025. *Decisions in milliseconds: The influence of neuroscience in football*. Available at: https://mbpschool.com/en/decisions-neuroscience-football/ [Accessed 6 June 2025].

McCabe, K. and Collins, C., 2018. Can genetics predict sports injury? The association of the genes GDF5, AMPD1, COL5A1 and IGF2 on soccer player injury occurrence. *Sports* (Basel), 6(1), p.21. doi: https://doi.org/10.3390/sports6010021

McLean Hospital, 2024. *Athlete mental health: What you need to know*. Available at: https://www.mcleanhospital.org/essential/athlete-mh [Accessed 6 June 2025].

Middleton, E., 2025. *How do footballers play in very cold conditions?* Available at: https://www.bbc.co.uk/sport/articles/c791yweyvzdo [Accessed 22 May 2025].

Mijatovic, D., Krivokapic, D., Versic, S., Dimitric, G. and Zenic, N. 2022. Change of direction speed and reactive agility in prediction of injury in football; prospective analysis over one

half-season. *Healthcare* (Basel) 10(3), p.440. doi: 10.3390/healthcare10030440.

Moreno-Perez, V., Paredes, V., Pastor, D., Garrosa, F.N., Vielcazat, S.J., Del Coso, J. and Mendez-Villanueva, A., 2021. Under-exposure to official matches is associated with muscle injury incidence in professional footballers. *Biology of Sport*, 38(4), pp.563–571. doi:10.5114/biolsport.2021.100360.

Morgans, R., Orme, P., Anderson, L. and Drust, B., 2014. Principles and practices of training for soccer. *Journal of Sport and Health Science*, 3(4), pp.251–257.doi:https://doi.org/10.1016/j.jshs.2014.07.002

Moyes, S. 2025. *Treatment for football-related injuries*. Available at: https://simonmoyes.com/sports-medicine-specialists/football-related-injuries/ [Accessed: 27 July 2025].

Myontec, 2024. *Decoding muscle imbalances in elite football players: insights from EMG analysis*. Available at: https://www.myontec.com/post/decoding-muscle-imbalances-in-elite-football-players-insights-from-emg-analysis [Accessed 24 May 2025].

Nassis, G.P., Brito, J., Dvorak, J., Chalabi, H. and Racinais, S., 2015. The association of environmental heat stress with performance: Analysis of the 2014 FIFA World Cup Brazil. *British Journal of Sports Medicine*, 49, pp.609–613. doi: 10.1136/bjsports-2014-094449.

National Institute of Arthritis and Musculoskeletal and Skin Diseases (NIAMS). 2024. *Sports injuries*. Available at: https://www.niams.nih.gov/health-topics/sports-injuries [Accessed 3 June 2025].

Olawade, D.B., Marinze, S., Qureshi, N., Weerasinghe, K. and Teke, J., 2025. The impact of artificial intelligence and machine learning in organ retrieval and transplantation: A comprehensive review. *Current Research in Translational Medicine*, 73(2), p.103493. doi: https://doi.org/10.1016/j.retram.2025.103493

Östenberg, A. and Roos, H., 2000. Injury risk factors in female European football: A prospective study of 123 players during one season. *Scandinavian Journal of Medicine & Science in Sports*, 10, pp.279–285.doi:10.1034/j.1600-0838.2000.010005279.x

Park, A.L., Furie, K. and Wong, S.E., 2023. Stronger athlete identity is a risk factor for more severe depressive symptoms after musculoskeletal injury in pediatric athletes: a systematic review. *Current Reviews in Musculoskeletal Medicine*, 16(5), pp.220–228. doi:10.1007/s12178-023-09828-0.

Peebles, J . 2023. *What sport has the most injuries?* Available at: https://www.

visitcompletecare.com/blog/what-sport-has-the-most-injuries/ [Accessed: 3 June 2025].

Peterson, L., Junge, A., Chomiak, J., Graf-Baumann, T. and Dvorak, J., 2000. Incidence of football injuries and complaints in different age groups and skill-level groups. *The American Journal of Sports Medicine*, 28(5 Suppl), pp. S51–S57. doi: 10.1177/28.suppl_5.s-51. PMID: 11032108.

PFA. 2023. *Significant changes to non-league contracts from next season*. Available at: https://www.thepfa.com/news/2023/3/22/significant-changes-to-non-league-contracts-from-next-season [Accessed 23 March 2025].

Philippou, C. and Maguire, K., 2022. *Still ill? Assessing the financial sustainability of football*. Available at: https://assets.publishing.service.gov.uk/media/62690039e90e07168e3fdc4a/Assessing_the_financial_sustainability_of_football__web_accessible_.pdf [Accessed 25 May 2025].

Prys, M., Rosiński, Ł., Buryta, R., Radzimiński, Ł., Różewski, P. and Rejera, I., 2023. Integrating machine learning for football injury prediction: A concept for an intelligent system. *Procedia Computer Science*, 225, pp.4139–4147. doi: 10.1016/j.procs.2023.10.410.

Racinais, S., Alonso, J.M., Coutts, A.J., Flouris, A.D., Girard, O., Gonzalez-Alonso, J., Hausswirth, C., Jay, O., Lee, J.K.W., Mitchell, N., Nassis, G.P., Nybo, L., Pluim, B.M., Roelands, B., Sawka, M.N., Wingo, J.E. and Périard, J.D., 2015. Consensus recommendations on training and competing in the heat. *Scandinavian Journal of Medicine & Science in Sports*, 25(Suppl 1), pp.6–19. doi: 10.1111/sms.12467.

Rahnama, N., Reilly, T. and Lees, A., 2002. Injury risk associated with playing actions during competitive soccer. *British Journal of Sports Medicine*, 36(5), pp.354–359.doi:10.1136/bjsm.36.5.354

Read, P.J., Oliver, J.L., De Ste Croix, M.B., Myer, G.D. and Lloyd, R.S., 2018. An audit of injuries in six English professional soccer academies. *Journal of Sports Sciences*, 36(13), pp.1542–1548.doi:10.1080/02640414.2017.1402535

Read, P.J., Oliver, J.L., De Ste Croix, M.B., Myer, G.D. and Lloyd, R.S., 2016. Neuromuscular risk factors for knee and ankle ligament injuries in male youth soccer players. *Sports Medicine*, 46(8), pp.1059–1066. doi: 10.1007/s40279-016-0479-z.

Rodrigues, F., Monteiro, D., Ferraz, R., Branquinho, L. and Forte, P., 2023. The association between training frequency, symptoms of overtraining and injuries in young men soccer players. *International Journal of Environmental Research and Public Health*, 20(8), p.5466. doi: 10.3390/ijerph20085466.

Rokoko. 2022. *How motion capture is used in sport*. Available at: https://www.rokoko.com/insights/motion-capture-in-sport [Accessed 3 June 2025]

Rossi, A., Pappalardo, L. and Cintia, P. 2021 'A narrative review for a machine learning application in sports: An example based on injury forecasting in soccer', *Sports*, 10(1), p. 5.doi:https://doi.org/10.3390/sports10010005

Ruddy, J.D., Cormack,S.J., Whiteley, R., Williams, M.D., Timmins, R.G., Opar, D.A., 2019. Modeling the risk of team sport injuries: a narrative review of different statistical approaches. *Frontiers in Physiology*, 10, p.829. Available at: https://www.frontiersin.org/articles/10.3389/fphys.2019.00829/full

Ryan, L. and Golden, A. 2006. 'Tick the box please': A reflexive approach to doing quantitative social research. *Sociology*, 40(6), pp.1191-1200. doi: https://doi.org/10.1177/0038038506072287

Saal, J.A., 1991. Common American football injuries. *Sports Medicine*, 12(2), pp.132–147. doi: 10.2165/00007256-199112020-00005.

Salzmann, G.M., Preiss, S., Zenobi-Wong, M., Harder, L.P., Maier, D. and Dvorák, J., 2017. *Osteoarthritis in football*. Cartilage, 8(2), pp.162–172. doi: 10.1177/1947603516648186.

Sarabia, J.M., Roldan, A., Henríquez, M. and Reina, R., 2021. Using Decision Trees to Support Classifiers' Decision-Making about Activity Limitation of Cerebral Palsy Footballers. *International Journal of Environmental Research and Public Health*, 18(8), p.4320. doi:10.3390/ijerph18084320.

Sarmento, H., Martinho, D.V., Gouveia, É.R., Afonso, J., Chmura, P., Field, A., Savedra, N.O., Oliveira, R., Praça, G., Silva, R. and Barrera-Díaz, J. 2024. 'The influence of playing position on physical, physiological, and technical demands in adult male soccer matches: a systematic scoping review with evidence gap map,' *Sports Medicine*, 54(11), pp. 2841–2864.doi:10.1007/s40279-024-02088-z

Scelles, N., Durand, C., Bonnal, L., Goyeau, D. and Andreff, W., 2016. Do all sporting prizes have a significant positive impact on attendance in a European national football league? Competitive intensity in the French Ligue 1. *Ekonomicheskaya Politika / Economic Policy*, 11, pp.82–107.

Scikit-learn. 2025. *Tree-based models*. Available at: https://scikit-learn.org/stable/modules/tree.html [Accessed: 3 June 2025]

Scranton, P.E., Jr., Whitesel, J.P., Powell, J.W., Dormer, S.G., Heidt, R.S., Jr., Losse, G., and

Cawley, P.W. 1997. A review of selected noncontact anterior cruciate ligament injuries in the National Football League. *Foot & Ankle International*, 18, pp.772–776. doi:10.1177/107110079701801204

Sharma, N., 2023. *Understanding and applying F1 score: AI evaluation essentials with hands-on coding example*. Available at: https://arize.com/blog-course/f1-score/ [Accessed 15 Jul. 2025]

Sharma, R. 2023. *Improving data quality: the foundation for accurate and reliable models*. Available at: https://statusneo.com/improving-data-quality-the-foundation-for-accurate-and-reliable-models/ [Accessed: 3 June 2025]

Singh, S., 2023. *Cross Validation and Hyperparameter Tuning: A Beginner's Guide*. Available at: https://medium.com/@sandeepmaths04/cross-validation-and-hyperparameter-tuning-a-beginners-guide-96d258eedee7 [Accessed 15 Jul. 2025].

Skala, F. and Zemková, E., 2023. Neuromuscular and perceptual-cognitive response to 4v4 small-sided game in youth soccer players. *Frontiers in Physiology*, 14, p.1260096. doi: 10.3389/fphys.2023.1260096.

Smith, A.M., 1996. Psychological impact of injuries in athletes. *Sports Medicine*, 22(6), pp.391–405. doi:10.2165/00007256-199622060-00006.

Soccer Supplement, 2022. *The three branches of football training*. Available at: https://www.soccersupplement.com/blogs/news/the-three-branches-of-football-training [Accessed 20 May 2025].

Söderman, K., Adolphson, J., Lorentzon, R. and Alfredson, H., 2001a. Injuries in adolescent female players in European football: A prospective study over one outdoor soccer season. *Scandinavian Journal of Medicine & Science in Sports*, 11, pp.299–304. doi:10.1034/j.1600-0838.2001.110508.x

Söderman, K., Alfredson, H., Pietilä, T. and Werner, S., 2001b. Risk factors for leg injuries in female soccer players: A prospective investigation during one outdoor season. *Knee Surgery, Sports Traumatology, Arthroscopy*, 9, pp.313–321.doi:10.1007/s001670100228

Sonesson, S., Lindblom, H. and Hägglund, M., 2023. Higher age and present injury at the start of the season are risk factors for in-season injury in amateur male and female football players: a prospective cohort study. *Knee Surgery, Sports Traumatology, Arthroscopy*, 31(10), pp.4618–4630. doi:10.1007/s00167-023-07517-6.

Spikin, I.C. 2013. Risk Management theory: the integrated perspective and its application in the public sector. *Estado, gobierno, gestión pública: Revista Chilena de Administración Pública* (21), pp.89–126.

STATSports. 2024. *Why do athletes wear GPS trackers?* Available at: https://statsports.com/the-locker/why-do-athletes-wear-gps-trackers [Accessed 3 June 2025].

Sugimoto, D., Alentorn-Geli, E., Mendiguchía, J. and Samuelsson, K., 2015. Biomechanical and neuromuscular characteristics of male athletes: Implications for the development of anterior cruciate ligament injury prevention programs. *Sports Medicine*, 45(6), pp.765–778. doi: 10.1007/s40279-015-0311-1

Swainston, S.C., Wilson, M.R. and Jones, M.I. 2020. Player Experience During the Junior to Senior Transition in Professional Football: A Longitudinal Case Study. *Frontiers in Psychology*, 11, p.1672. doi: https://doi.org/10.3389/fpsyg.2020.01672

Szabo, V. and Strang, V.R., 1997. Secondary analysis of qualitative data. *Advances in Nursing Science*, 20(2), pp.66–74. doi: https://doi.org/10.1097/00012272-199712000-00008

Taimela, S., Kujala, U.M. and Osterman, K. 1990. 'Intrinsic risk factors and athletic injuries', *Sports Medicine*, 9(4), pp. 205-215. doi: 10.2165/00007256-199009040-00002.

Taylor, M.K., Hernández, L.M., Sessoms, P.H., Kawamura, C., and Fraser, J.J., 2020. Trauma exposure and functional movement characteristics of male tactical athletes. *Journal of Athletic Training*, 55(4), pp.384–389. doi: 10.4085/1062-6050-111-19.

Telfer, A. and Stone, S. 2023. *Newcastle & Manchester United: Why are Premier League injuries at a new high?* Available at: https://www.bbc.co.uk/sport/football/67417862 [Accessed 3 June 2025].

Tennessee Orthopaedic Alliance, 2024. The long-term effects of football injuries. Available at: https://www.toaeasttn.com/the-long-term-effects-of-football-injuries/ [Accessed 6 June 2025].

Thema, M.T., Jacobs, S., van den Berg, L., Strauss, A., and Mahlangu, M.P. 2025. 'The role of playing position in soccer injury characteristics: evidence from sub-elite athletes,' *Frontiers in Sports and Active Living*, 7, 1542300. doi: https://doi.org/10.3389/fspor.2025.1542300

Thompson, K.A., Bulls, H.W., Sibille, K.T., Bartley, E.J., Glover, T.L., Terry, E.L., Vaughn, I.A., Cardoso, J.S., Sotolongo, A., Staud, R., Hughes, L.B., Edberg, J.C., Redden, D.T., Bradley,

L.A., Goodin, B.R. and Fillingim, R.B., 2018. Optimism and psychological resilience are beneficially associated with measures of clinical and experimental pain in adults with or at risk for knee osteoarthritis. *Clinical Journal of Pain*, 34(12), pp.1164–1172.

Titans Football Academy, 2023. *Football recovery: mastering techniques for success*. Available at: https://thetitansfa.com/football-recovery-mastering-techniques-for-success/ [Accessed: 21 May 2025].

Turner, C.M., Morton, J.P., Drust, B., Williams, A.G., and Erskine, R.M. (2022) 'Playing position and the injury incidence rate in male academy soccer players', *Journal of Athletic Training*, 57(7), pp. 696–703. doi: https://doi.org/10.4085/1062-6050-0346.21

Tysvaer, A.T. and Løchen, E.A., 1991. Soccer injuries to the brain: A neuropsychologic study of former soccer players. *American Journal of Sports Medicine*, 19(1), pp.56–60. doi: 10.1177/036354659101900109.

Ueblacker, P., Müller-Wohlfahrt, H.W. and Ekstrand, J. 2015. Epidemiological and clinical outcome comparison of indirect ('strain') versus direct ('contusion') anterior and posterior thigh muscle injuries in male elite football players: UEFA Elite League study of 2287 thigh injuries (2001–2013). *British Journal of Sports Medicine*, 49(22), pp.1461–1465. doi: https://doi.org/10.1136/bjsports-2014-094285

Unisportstore, 2025. *Extra time in football*. Available at: https://www.unisportstore.com/faq/wiki/extra-time/ [Accessed 21 May 2025].

van Mechelen, W., Hlobil, H. and Kemper, H.C., 1992. Incidence, severity, aetiology, and prevention of sports injuries: A review of concepts. *Sports Medicine*, 14(2), pp.82–99. doi: 10.2165/00007256-199214020-00002.

van Vulpen, A.V., 1989. *Sports for all, sports injuries, and their prevention: Scientific report*. Council of Europe. Oosterbeek: National Institute for Sports Health Care.

Varillas-Delgado, D., 2024. Genes associated with muscle, tendon, and ligament injury epidemiology in women's amateur football players. *Applied Sciences*, 14(5), p.1980. doi: https://doi.org/10.3390/app14051980

Venturelli, M., Schena, F., Zanolla, L., Bishop D. 2011. 'Injury risk factors in young soccer players detected by a multivariate survival model', *Journal of Science and Medicine in Sport*, 14(4), pp. 293-298. doi: 10.1016/j.jsams.2011.02.013.

Walia, B. and Boudreaux, C.J., 2020. The cost of players' injuries to professional sports leagues and other sports organizations. *Managerial Finance*. Doi: https://doi.org/10.1108/MF-06-2020-0301

Walker, O., 2025. *GPS (Wearables): Part 2 – Metrics and Application. Science for Sport*. Available at: https://www.scienceforsport.com/gps-wearables-metrics-and-application/ [Accessed 3 June 2025].

Walker, R., 2023. *Jack Wilshere admits surgeons were worried he wouldn't walk after injury at age 19*. Available at: https://www.dailymail.co.uk/sport/football/article-12195011/Jack-Wilshere-admits-surgeons-worried-wouldnt-walk-injury-age-19.html [Accessed 11 June 2025].

Wermert, K., 2019. *Does rain make you more prone to injury when playing sports?* Available at: https://livewellcc.com/does-rain-make-you-more-prone-to-injury-when-playing-sports/ [Accessed 22 May 2025].

White, R., 2025. *Acute: Chronic Workload Ratio. Science for Sport*. Available at: https://www.scienceforsport.com/acutechronic-workload-ratio/ [Accessed 20 May 2025].

Wiese-bjornstal, D. M., Smith, A. M., Shaffer, S. M., & Morrey, M. A.1998. An integrated model of response to sport injury: Psychological and sociological dynamics. *Journal of Applied Sport Psychology*, 10(1), pp.46-69. doi: 10.1080/10413209808406377.

Wik, E.H., 2022. Growth, maturation, and injuries in high-level youth football (soccer): A mini review. *Frontiers in Sports and Active Living*, 4. doi: https://doi.org/10.3389/fspor.2022.975900

Wild, C.Y., Steele, J.R. and Munro, B.J., 2013. Musculoskeletal and estrogen changes during the adolescent growth spurt in girls. *Medicine & Science in Sports & Exercise*, 45(1), pp.138–145.doi:https://doi.org/10.1249/MSS.0b013e31826a507e

William Hill, 2022. *How Weather Affects Football. William Hill*. Available at: https://news.williamhill.com/football/weather-effect-football/ [Accessed 22 May 2025].

Williams, J.M. and Andersen, M.B., 1998. Psychosocial antecedents of sport injury: Review and critique of the stress and injury model. *Journal of Applied Sport Psychology*, 10(1), pp.5–25.

Wilt and Glos Standard, 2023. 'No safe number times footballers can head ball,' high court told. *Wilt & Gloucestershire Standard*, 26 January. Available at: https://www.wiltsglosstandard.co.uk/news/national/25269440.no-safe-number-times-footballers-can-head-ball-high-court-told/ [Accessed 11 June 2025].

Wyscout, 2024. *Duel*. Available at: https://dataglossary.wyscout.com/duel/ [Accessed 22 May 2025].

Yang, X., Wang, H. and Sun, N., 2022. Machine learning outperforms traditional logistic regression and offers new possibilities for cardiovascular risk prediction: A study involving 143,043 Chinese patients with hypertension. *Frontiers in Cardiovascular Medicine*, 9, p.1025705. doi: 10.3389/fcvm.2022.1025705.

Yasar, K., 2025. *What is data preprocessing? Key steps and techniques*. Available at: https://www.techtarget.com/searchdatamanagement/definition/data-preprocessing [Accessed 4 June 2025].

Yde, J. and Nielsen, A.B. 1990. Sports injuries in adolescents' ball games: soccer, handball, and basketball. *British Journal of Sports Medicine*, 24, pp.51–54. doi: 10.1136/bjsm.24.1.51

Yu, B. and Garrett, W.E. 2007. 'Mechanisms of non-contact ACL injuries', *British Journal of Sports Medicine*, 41(Suppl 1),