NOVA IMS Information Management School

# MDSAA

Master Degree Program in

**Data Science and Advanced Analytics**

# LEVERAGING MACHINE LEARNING FOR INJURY PREDICTION AND PREVENTION IN PROFESSIONAL FOOTBALL

A Machine Learning-Based Framework Using Design Science Research

Francisco Alves Lopes Ramos de Campos

Master Thesis

presented as partial requirement for obtaining a Master's Degree in Data Science and Advanced Analytics

**NOVA Information Management School**

**Instituto Superior de Estatística e Gestão de Informação**

Universidade Nova de Lisboa

**NOVA Information Management School**

**Instituto Superior de Estatística e Gestão de Informação**

Universidade Nova de Lisboa

# LEVERAGING MACHINE LEARNING FOR INJURY PREDICTION AND PREVENTION IN PROFESSIONAL FOOTBALL

A Machine Learning-Based Framework Using Design Science Research

by

Francisco Alves Lopes Ramos de Campos

Master Thesis presented as partial requirement for obtaining the Master's degree in Data Science and Advanced Analytics, with a specialization in Data Science

## Supervised by

Vítor Manuel Pereira Duarte dos Santos, PhD, NOVA IMS

July, 2025

# STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledged the Rules of Conduct and Code of Honor from the NOVA Information Management School.

Francisco Campos

Lisbon, July 2025

# ABSTRACT

The occurrence of injuries remains a major problem in professional football which affects both team performance and financial planning and athlete career duration. The research established a functional injury prediction system through machine learning that combined workloads and physiological factors with pitch conditions and weather elements and competition levels. The research used a systematic approach to unite a comprehensive literature review with experimental work on a well-organized football dataset. The research team performed systematic preprocessing through feature engineering and encoding and SMOTE application to handle class imbalance before testing multiple algorithms. XGBoost achieved the optimal balance between recall and F1-score through GridSearchCV tuning which proved essential for identifying actual injury risks because missing a case would result in substantial costs. The approach became applicable in elite sports through SHAP interpretability tools which provided clear explanations about the factors influencing each prediction to help coaches and medical staff make decisions. The final model both identified players at risk and provided clear explanations to enable specific interventions instead of general approaches. The study shows that using internal player metrics together with external match conditions within a strong machine learning framework produces improved injury forecasting results. The research creates a strong base for additional improvements including real-time tracking data integration and team-wide model deployment to advance individualized injury prevention in professional football.



# KEYWORDS

Machine Learning; Football, Injury Prediction; Player Health.


Sustainable Development Goals (SDG):

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# LIST OF ABBREVIATIONS AND ACRONYMS

| | |
|---|---|
| **AI** | Artificial Intelligence |
| **DSR** | Design Science Research |
| **GBM** | Gradient Boosting Machine |
| **GDPR** | General Data Protection Regulation |
| **GPS** | Global Positioning System |
| **KNN** | K-Nearest Neighbors |
| **LR** | Literature Review |
| **ML** | Machine Learning |
| **PRISMA** | Preferred Reporting Items for Systematic Reviews and Meta-Analyses |
| **RQ** | Research Question |
| **SDG** | Sustainable Development Goals |
| **SHAP** | SHapley Additive exPlanations |
| **SLRQ** | Systematic Literature Review Question |
| **SMOTE** | Synthetic Minority Over-sampling Technique |
| **SVM** | Support Vector Machine |

# 1. INTRODUCTION

## 1.1. CONTEXT AND PROBLEM IDENTIFICATION

The modern football environment demands that teams maintain player wellness while preventing injuries to achieve success. The game's evolution has resulted in players facing greater physical challenges. Football clubs must develop efficient methods to track and control player fitness because they need to maintain their competitive edge with more matches and demanding training programs and intense play. As a result, injury prediction and prevention have become critical areas where technology and data analytics can offer substantial advantages to teams looking for an edge (Majumdar et al., 2022; Piłka et al., 2023).

The financial impact of injuries extends beyond match performance because it affects both on-field results and team budgets. Football clubs spend substantial funds on their players yet losing essential players through injury results in major financial losses because of missed matches and medical expenses. Football clubs now use machine learning technology to handle their risks better. Sports science teams use machine learning as their primary tool to develop better strategies for player workload management and recovery periods and training approaches (Chang et al., 2024).

The recent progress in machine learning has led to the creation of predictive models which use player-specific data including training intensity and recovery times and match loads and physical condition to forecast injuries. The models use data from wearable devices and real-time monitoring tools to collect metrics including heart rate and distance covered and sprint frequency and biochemical markers. Machine learning algorithms analyse this data to identify patterns which signal increased injury risk for players. Teams can use this information to modify their training plans and match strategies in advance for injury prevention purposes (Majumdar et al., 2022; Piłka et al., 2023).

These models enable personalized management of players through their main advantage. Machine learning technology enables the creation of individualized training plans and recovery strategies because each player's body reacts uniquely to physical stress. The fitness levels and injury history of players determine their need for extended recovery periods and decreased training intensity. The predictive models enable coaching staff and medical personnel to predict player responses to different training intensities which helps teams prevent injuries and achieve better performance (Van Eetvelde et al., 2021).

Machine learning models also have the potential to make data-driven decisions based on a wide range of factors, including a player's physiology, previous injuries, and even environmental conditions like pitch quality or weather. These models can evolve and improve as more data becomes available, enabling increasingly accurate predictions. This means clubs can fine-tune their training schedules and match preparations to help players stay in optimal condition throughout the season (Chang et al., 2024; Piłka et al., 2023).

Despite significant advancements, several challenges remain in applying machine learning for injury prediction in football. A recent scoping review of 113 sport-science ML studies (Leckey et al., 2025) underscores that these challenges are not football-specific but endemic across disciplines. One key issue is the difficulty in accounting for individual variability, such as differences in player physiology, biomechanics, and responses to physical stress. Sex-specific cohorts show additional variability; for example, models trained on elite Norwegian women's teams required different workload windows and features (Mohaiminul & Emon, 2024). Personalizing models to each player's unique characteristics, such as their injury history, physical condition, and recovery capacity, remains an area in need of further development (Piłka et al., 2023). The process of developing universally applicable models becomes more complex because different teams and leagues use non-standardized data collection methods through their various wearable devices and software platforms. The existing models fail to consider important factors which include pitch quality and travel schedules and sleep patterns that substantially impact injury risk. Factors like pitch quality, travel schedules, and sleep patterns, which significantly affect injury risk, are often overlooked in existing models. Addressing these challenges will require stronger collaboration between sports scientists, data analysts, and football clubs to develop more accurate and practical models (Majumdar et al., 2022; Piłka et al., 2023). To bridge these gaps, future research must focus on personalizing machine learning models and improving the precision of injury prediction. This leads to the formulation of the following research question: **How can machine learning models be personalized to account for individual variability in player physiology, workload, and match conditions to enhance the accuracy and effectiveness of injury prediction and prevention in football?**

## 1.2. OBJECTIVES

The goal of the research would be to develop a model for predicting football player injuries.

In order to achieve the main goal, the following intermediate objectives were defined:

- Conduct a comprehensive study on the role of data analytics and machine learning in sports, injury prediction and prevention, examining current practices and future trends in using player-specific data for injury management.
- Explore how historical injury data and recovery outcomes can train machine learning models to improve decision-making in real-time for player health management, allowing clubs to take important decisions based on similar past incidents.
- Examine the limitations of current machine learning models in injury prediction
- Develop a refined predictive model that enhances injury forecasting by integrating multiple real-time and historical data sources, ensuring adaptability to individual player characteristics. This model will account for workload fluctuations, environmental conditions, and player-specific recovery rates to improve precision in injury prevention.
- Evaluate the model's effectiveness by applying it to real or simulated football data, measuring its accuracy, usability, and practical applicability. The evaluation will assess key performance metrics such as prediction accuracy, false positive/negative rates, and its ability to provide actionable insights for coaching and medical staff. Based on these findings, recommendations for further refinement and real-world implementation will be proposed.

## 1.3. STUDY OUTCOMES AND CONTRIBUTIONS

The results of this research will have a significant impact on various fields, including sports, society, and science. By improving injury prediction methods through personalized machine learning models, this study can contribute to more effective injury prevention strategies in football and other sports.

In sports and society, reducing the number of injuries in football will not only enhance team performance but also contribute to the longevity of players' careers. Injuries have long-term consequences for athletes, affecting both their physical and mental well-being. A better understanding of how personalized data can predict injury risks will allow teams to proactively manage workloads and recovery times, preventing injuries before they happen (Majumdar et al., 2022). This approach will ultimately benefit athletes at all levels, from professionals to youth players, promoting safer participation in sports and reducing the financial burden associated with injuries, such as medical costs and player downtime (Piłka et al., 2023). Additionally, healthier athletes contribute to greater team consistency and success, which generates a positive ripple effect across the broader sports ecosystem, benefiting sponsors, fans, and the overall brand value of clubs (Ekstrand, 2011).

From a scientific and academic perspective, this study contributes to the growing body of research in sports science, particularly in how machine learning can be applied to complex physiological data. One of the key contributions of this research is the development of machine learning models that are more personalized, addressing a significant gap in current models that do not account for individual differences in player physiology and biomechanics (Piłka et al., 2023). This research will also address the lack of standardization in data collection methods across teams and leagues, a major limitation in current injury prediction studies. By proposing ways to harmonize data from different wearable devices and platforms, this study can help advance future research in injury prevention and sports analytics (Majumdar et al., 2022).

In addition to its sports-specific applications, this research has broader implications for the fields of machine learning and artificial intelligence. Personalized machine learning models are increasingly being used across industries, and this research will provide a real-world example of how these models can be fine-tuned to improve predictive accuracy. The ability to tailor predictive models to individual characteristics, such as a player's training load, recovery patterns, and even environmental conditions, showcases how machine learning can be adapted to meet the unique needs of specific use cases (Chang et al., 2024). This could inform applications in healthcare, fitness, and other sectors where personal data is crucial for accurate predictions.

In summary, the contributions of this research extend beyond football, offering new insights and practical solutions for injury prevention in sports, advancing knowledge in personalized machine learning, and encouraging the development of more standardized approaches to sports data collection and analysis.

# 2. LITERATURE REVIEW

This chapter sets out to build a clear picture of the topics at the heart of this research. It starts by looking at how common injuries are in football, what types occur most frequently, and the reasons some players might be more at risk than others. It also reviews how clubs try to prevent these problems, from tailored training routines to nutritional and lifestyle strategies. The chapter then moves into how injuries have traditionally been predicted, and how the use of machine learning is starting to reshape this area, offering more precise and individualized insights. Finally, it walks through the systematic review carried out for this work, explaining how the most relevant studies were selected and what they reveal about the current landscape. This step-by-step exploration not only grounds the study in existing knowledge but also shows why the chosen methods and focus areas make sense for addressing the research question.

## 2.1. INJURIES IN FOOTBALL

### 2.1.1. Overview

Injuries are a constant issue in football, impacting players of all levels. The sport's high-intensity actions, such as sprinting, jumping, tackling, and sudden changes of direction, place significant strain on players' bodies. Studies suggest that football has one of the highest injury rates among team sports, with between 5 and 9 injuries per 1,000 hours of play (Ekstrand, 2011). These injuries not only disrupt players' performance but also impose significant financial and logistical challenges for teams, including treatment costs and the absence of key players during crucial matches (Della Villa, 2021).

A large proportion of football injuries occur during matches rather than training sessions, with contact injuries being a major cause (Waldén, 2015). Elite players face an even greater risk due to the demanding schedules of domestic leagues, international tournaments, and continental competitions. Fatigue resulting from limited recovery time is a well-known factor contributing to these injuries (Bengtsson, 2013). The combination of physical strain and lack of rest highlights the importance of effective injury prevention strategies in football.

### 2.1.2. Type of injuries

Football injuries are commonly categorized into acute injuries, which occur suddenly due to trauma, and overuse injuries, which develop gradually over time due to repetitive strain.

#### 2.1.2.1. Acute Injuries

Acute injuries in football include sprains, strains, and fractures. For example, ankle sprains often result from abrupt changes in direction or poor landings, while hamstring strains are frequent during high-speed sprints (Hägglund, 2009) and affected almost one-third of players in a recent ML-based cohort study of 284 professionals (Pierre-Eddy Dandrieux, 2024). Additionally, contact injuries, such as ACL tears and concussions, often occur during tackles or collisions and can require long recovery periods (Pfirrmann, 2016). Research suggests that ACL injuries are particularly severe due to their long rehabilitation process, often affecting a player's performance even after recovery (Della Villa, 2021).

#### 2.1.2.2. Overuse Injuries

Overuse injuries, such as tendinitis, stress fractures, and shin splints, are typically caused by excessive training loads, inadequate recovery, or poor biomechanics. These injuries are especially common during periods of increased physical demand, such as preseason training or congested match schedules (Van der Horst, 2017). Overuse injuries often go unnoticed in their early stages but can worsen without proper intervention.

The risk of specific injuries also depends on a player's position. Goalkeepers are prone to upper-body injuries from diving and collisions, while midfielders often sustain lower-body injuries due to their high running demands. External factors, such as poor pitch conditions and inadequate footwear, can further exacerbate the likelihood of both acute and overuse injuries (Waldén, 2015).

### 2.1.3. Biological conditioning

Certain biological factors make some individuals more prone to injuries than others. Variations in anatomy, muscle strength, and biomechanics are key contributors. For example, players with uneven leg lengths or poor core stability are more likely to develop stress-related injuries (Krosshaug, 2016). Additionally, weaker muscles or imbalanced muscle groups can increase vulnerability during high-intensity activities (Zemková, 2020).

Genetics also play a role. Specific genetic markers, such as COL5A1 variations, are linked to an increased risk of tendon injuries (Collins, 2009). Hormonal differences can further influence injury susceptibility. For instance, female athletes have a higher risk of ACL injuries, partly due to hormonal fluctuations affecting ligament strength (Hewett, 2006).

Age is another factor that affects injury risk. Young players are more likely to experience growth-related injuries, while older players are at greater risk for degenerative conditions like arthritis and chronic tendinopathy. Recovery times also tend to lengthen with age, increasing the risk of re-injury (Larruskain, 2018). Moreover, the accumulation of physical stress over the years can lead to chronic conditions, highlighting the importance of early injury prevention strategies.

### 2.1.4. Injuries prevention

Preventing injuries in football requires a comprehensive approach, addressing training methods, nutrition, and lifestyle habits. Effective prevention strategies not only reduce injury rates but also enhance player performance and longevity.

#### 2.1.4.1. Training Methods

Proper training is essential to reducing injury risk. Warm-up routines like the FIFA 11+ program, which focuses on neuromuscular control and strength, have been shown to significantly lower injury rates (Bizzini, 2015). Strength training, particularly for the hamstrings and core muscles, has proven effective in preventing muscle strains and improving overall stability (Van Dyk, 2019). Additionally, wearable technology is increasingly used to monitor training loads and ensure players do not exceed their physical limits (Halson, 2014).

Recovery routines are equally important in preventing injuries. Practices such as active recovery, stretching, and foam rolling help maintain muscle flexibility and reduce the risk of overuse injuries. Clubs are also adopting advanced methods like cryotherapy and compression therapy to accelerate recovery times (Mah, 2011).

#### 2.1.4.2. Nutrition

Diet plays a crucial role in injury prevention by supporting muscle recovery and bone health. Adequate protein intake is necessary for muscle repair, while calcium and vitamin D help maintain bone density (Heaton, 2017). Proper hydration is also critical, as dehydration can impair performance and increase the risk of muscle cramps or strains (Sawka, 2007). Additionally, supplements like omega-3 fatty acids and antioxidants may support recovery by reducing inflammation and oxidative stress.

#### 2.1.4.3. Lifestyle Habits

Lifestyle factors such as sleep and stress management have a significant impact on injury prevention. Lack of sleep has been linked to slower recovery and a higher risk of injury (Mah, 2011). Players are encouraged to adopt consistent sleep routines and include recovery practices, such as stretching, foam rolling, and ice baths, in their daily habits. Mental health also plays a role, as high stress levels can lead to fatigue and poor decision-making, further increasing injury risk (Gouttebarge, 2015).

By integrating these strategies, football clubs can create a safer environment for players, reducing injuries and enhancing overall performance. Furthermore, a focus on individualized prevention strategies ensures that specific needs based on a player's position, age, and physical condition are met effectively.

### 2.1.5. Injuries prediction

Injury prediction in football has traditionally relied on expert judgment, injury history, and physical assessments such as musculoskeletal screenings and standardized fitness tests (Bahr, 2003). While useful, these methods often lack precision and fail to account for real-time workload fluctuations and external factors.

With technological advancements, GPS tracking, wearable sensors, and real-time physiological monitoring have improved injury risk assessment by providing objective data on workload, fatigue, and movement patterns (Johann Windt, 2017). More recently, machine learning models have been introduced, analysing large datasets to identify patterns and predict injuries more accurately (Freitas et al., 2025; Piłka et al., 2023). These AI-driven approaches allow for personalized risk assessments, helping teams take preventive measures before injuries occur.

Despite progress, challenges remain in standardizing data collection and improving model accuracy across different players and teams (Van Eetvelde et al., 2021). Refining injury prediction models with personalized data and optimized machine learning techniques will be essential for enhancing injury prevention strategies in football.

## 2.2. Related Work

### 2.2.1. PRISMA Protocol

The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) methodology is a widely accepted framework used to conduct structured and reproducible literature reviews. It provides a set of tools and guidelines designed to help researchers identify, screen, and select relevant scientific studies based on predefined criteria. Originally developed by a multidisciplinary group of experts in 2005, PRISMA introduced a 27-item checklist and a four-phase flow diagram that together form a robust structure for conducting systematic reviews (D. Moher et al., 2009).

At its core, the methodology helps ensure transparency and rigor throughout the literature selection process, reducing the risk of bias and increasing the reliability of the final results. The four main phases of the PRISMA workflow are:

1. Identification – Articles are identified through database searches using a carefully designed strategy and selected keywords relevant to the study's topic.
2. Screening – Based on inclusion and exclusion criteria, duplicate and irrelevant papers are removed.
3. Eligibility – Articles are assessed in full to determine if they meet the eligibility standards established for the review.
4. Included – The final set of studies that are relevant and valid for analysis are included in the review.

This method ensures that only studies of appropriate quality and relevance are considered, creating a dependable foundation for the theoretical and analytical components of the research.

### 2.2.2. PRISMA Execution

Sections 1 and 2 of this thesis established the theoretical context around football injuries and the emerging use of machine learning in injury prediction and prevention. Drawing from that foundation, a set of keywords was identified and used to construct the search string for retrieving scientific articles relevant to the study’s objectives.

The aim of this review is to conduct a comprehensive study on the role of data analytics and machine learning in sports, injury prediction and prevention, examining current practices and future trends in using player-specific data for injury management and to achieve a well-rounded and current understanding of how artificial intelligence and machine learning are being utilized to forecast and prevent injuries in football. To guide the selection process, the Systematic Literature Review Questions (SLRQs) presented in Table 2.1 were defined.

Table 2.1 - Systematic Review's Research Questions

| | |
|---|---|
| **SLRQ1** | What is the current status of research in using data analytics and machine learning in football, injury prediction and prevention? |
| **SLRQ2** | What kind of AI techniques are currently useful in injury prediction and prevention? |
| **SLRQ3** | What are the major limitations of current machine learning models in injury prediction? |

To answer these questions, and in accordance with the PRISMA methodology, the most relevant and scientifically sound papers were selected from established research databases. A set of specific, English-language keywords was chosen, reflecting the central themes of the study. These keywords were derived directly from the conceptual pillars of the thesis, covering two domains: Football and Machine Learning. The list included:

Table 2.2 – Systematic Review's Keywords

| | **Football** | **Machine Learning** |
|---|---|---|
| **Keywords** | Injuries Prediction | Artificial Intelligence |
| | Injuries Prevention | Deep Learning |
| | | Data Analytics |
| | | Models |

Using these keywords, a Boolean search string was developed to locate relevant terms in titles, abstracts, and keywords of academic articles. The string used was: **("Football") AND ("Machine Learning" OR "Artificial Intelligence" OR "Deep Learning" OR "Data Analytics" OR "Models") AND ("Injuries" OR "Injuries Prevention" OR "Injuries Prediction")**

To ensure the material reflected the current state of the field, **only peer-reviewed scientific articles published between 2020 and 2025** were included. Furthermore, all selected articles had to be written in **English** and published in **journals or conferences** recognized for academic rigor. Articles not matching these requirements, such as non-peer-reviewed reports, magazine articles, books, or publications outside the selected timeframe, were excluded.

The search was conducted in January 2025 on the following scientific information resource databases:

Table 2.3 - Systematic Review's Resource Databases

| Resource Database | Resource URL |
|---|---|
| Scopus | https://www.scopus.com/home.uri |
| Web of Science | https://www.webofknowledge.com/ |
| IEEE | https://www.ieee.org/ |

These databases provided access to a comprehensive collection of articles in areas such as sports science, biomedical engineering, data science, and applied AI. The search results were then filtered and organized using the PRISMA flow diagram (to be included in the following subsection), which visually outlines the number of papers at each review phase.

The ultimate goal of this process is to ensure that the literature review reflects a balanced, objective, and updated view of how AI and ML are transforming injury management in football, with particular attention to models that have been evaluated through real-world or high-quality experimental data.

Following the PRISMA methodology, the next step was to define the inclusion and exclusion criteria for the articles resulting from the mentioned search.

Table 2.4 - Systematic Review's inclusion and exclusion criteria

| Inclusion Criteria | Exclusion Criteria |
|---|---|
| Any scientific article showing evidence of AI & ML utilization in injury prediction and prevention | Papers focusing on injury prediction and prevention but without focusing on AI & ML techniques utilization |
| Paper must be a peered reviewed conference or journal paper written in English | Articles not in English and duplicate papers |
| Paper is published between 2020 and 2025 | Articles published before 2020 |
| | Non-academic or non-scientific papers (e.g., websites, magazines reports, newspapers, consulting articles, books, citations) |
| | Papers with titles outside the scope of this work |

After entering the search into databases such as Scopus, IEEE Xplore, and Web of Science, an initial total of 1,274 records were retrieved. This marks the identification phase of the PRISMA framework.

As it is presented in Figure 2.1, during the initial screening phase, duplicate entries (across databases and publication types) were identified and removed, reducing the dataset by 682 records, leaving 592 unique articles. Subsequently, a first screening of titles and abstracts was conducted to assess basic relevance and compliance with inclusion criteria (e.g., post-2018 publications, English language, availability of full text, and relation to injury prediction in football). Based on this step, 400 records were excluded.

This led to a pool of 192 full-text articles assessed in detail. These were evaluated based on the presence of empirical evidence, the application of machine learning techniques, relevance to football-specific injury prediction (as opposed to general sports or biomechanical studies), and methodological quality. As a result, 176 articles were excluded, with reasons ranging from being too general, lacking predictive modelling, or focusing on unrelated domains.

In the final stage, 16 articles met all the inclusion criteria and were retained for the qualitative synthesis presented in this literature review. These studies offered diverse perspectives on ML-based injury prediction methods in football and formed the core knowledge base for the following sections of the thesis. This process is represented in the following workflow picture:

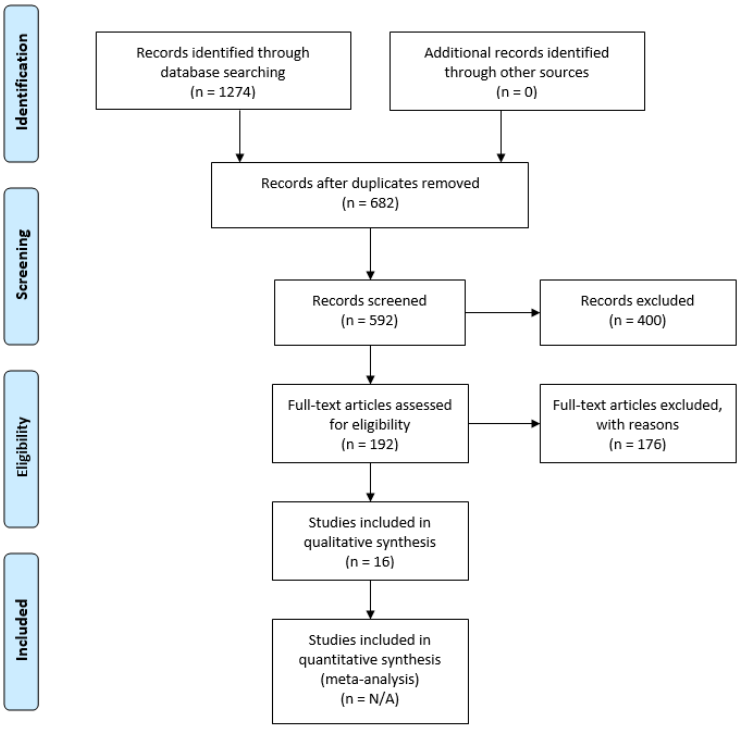


Figure 2.1 - PRISMA Flow Diagram of the Systematic Review Process- source (Moher et al., 2009)

The outcome of this systematic review led to the inclusion of 16 peer-reviewed journal articles, adall of which directly addressed the research questions concerning the use of machine learning for injury prediction in football. These articles are summarized in the table below, with a brief description of their contributions, key conclusions, and identified research gaps or suggestions for future work.

Table 2.5 - PRISMA results table – included articles

| # | Authors | Article | Contribution | Publication Type |
|---|---|---|---|---|
| #1 | (Rommers et al., 2020) | A Machine Learning Approach to Assess Injury Risk in Elite Youth Football Players | Develops an ML model to assess injury risk in youth football players, focusing on training load and movement data. | Journal Article |
| #2 | (Oliver et al., 2020) | Using Machine Learning to Improve Our Understanding of Injury Risk and Prediction in Elite Male Youth Football Players | Evaluates ML applications for identifying injury risk factors in elite football and suggests model improvements. | Journal Article |

| # | Authors | Article | Contribution | Publication Type |
|---|---|---|---|---|
| #3 | (He, 2021) | Prediction Model of Juvenile Football Players' Sports Injury Based on Text Classification Technology of Machine Learning | Proposes a text-classification ML model for injury prediction in youth soccer and evaluates its effectiveness. | Journal Article |
| #4 | (Pu et al., 2023) | Football Player Injury Full-Cycle Management and Monitoring System Based on Blockchain and Machine Learning Algorithm | Utilizes **machine learning** to analyse injury data, predict recovery times, and enhance injury management through AI-driven insights. | Journal Article |
| #5 | (Benjaminse et al., 2024) | Application of Machine Learning Methods to Investigate Joint Load in Agility on the Football Field: Creating the Model, Part I | Investigates machine learning approaches to predict knee joint loading during sport-specific agility tasks in female football players, aiming to enhance ACL injury prevention strategies. | Journal Article |
| #6 | (Majumdar et al., 2022) | Machine Learning for Understanding and Predicting Injuries in Football | Examines the application of machine learning in injury prediction for football players, analysing key risk factors and model accuracy. | Journal Article |
| #7 | (Freitas et al., 2025) | Predicting Noncontact Injuries of Professional Football Players Using Machine Learning | Investigates the effectiveness of machine learning in predicting noncontact football injuries, focusing on biomechanical and workload-related factors. | Journal Article |
| #8 | (Saberisani et al., 2025) | Prediction of Football Injuries Using GPS-Based Data in Iranian Professional Football Players: a machine learning approach | Uses GPS-based data and machine learning techniques to predict injury risk in professional football players. | Journal Article |
| #9 | (Xu, 2025) | Light Sensors and Infrared Radiation Images Based on Artificial Intelligence Data Mining for football performance evaluation and prediction | Explores the use of light sensors and infrared imaging, analysed through AI, to enhance injury detection and prevention in football. | Journal Article |
| #10 | (Valle et al., 2022) | Return to Play Prediction Accuracy of the MLG-R Classification System for Hamstring Injuries in Football Players: A Machine Learning Approach | Developed a reliable hamstring muscle injury classification system based on magnetic resonance imaging, showing excellent results in terms of reliability, prognosis capability, and objectivity. The study utilized machine learning approaches to assess the importance of each factor in determining return to play and offered predictions of expected return to play. | Journal Article |

| # | Authors | Article | Contribution | Publication Type |
|---|---|---|---|---|
| #11 | (Hecksteden et al., 2023) | Forecasting football injuries by combining screening, monitoring and machine learning | This study aimed to forecast non-contact time-loss injuries in male professional football players by combining screening and monitoring data, analysed using machine learning techniques. The findings suggest that integrating these data sources may improve predictive accuracy. | Journal Article |
| #12 | (Chang et al., 2024) | Football Analytics: Assessing the Correlation between Workload, Injury and Performance of Football Players in the English Premier League | Investigates the relationships between player workload, personal traits, match-related factors, performance, and injuries in the English Premier League using statistical and machine learning techniques, providing insights for injury prevention and performance enhancement. | Journal Article |
| #13 | (González et al., 2024) | Predicting Injuries in Elite Female Football Players with Global-Positioning-System and Multiomics Data | Developed an injury prediction model for elite female football players by integrating GPS-derived workload metrics with genomic and metabolomic data, enhancing personalized injury prevention strategies. | Journal Article |
| #14 | (Martins et al., 2022) | Predictive Modelling of Injury Risk Based on Body Composition and Selected Physical Fitness Tests for Elite Football Players | Developed predictive models using body composition and physical fitness variables to assess injury risk in elite football players, aiding in targeted injury prevention strategies. | Journal Article |
| #15 | (Piłka et al., 2023) | Predicting Injuries in Football Based on Data Collected from GPS-Based Wearable Sensors | Developed decision-making models, including rule-based, fuzzy rule-based, and XGBoost algorithms, to predict non-contact lower body injuries in football players using data from GPS-based wearable sensors, achieving high prediction accuracy. | Journal Article |
| #16 | (Robles-Palazón et al., 2023) | Predicting Injury Risk Using Machine Learning in Male Youth Soccer Players | Developed a machine learning screening model that identified six field-based measures influencing injury risk in male youth soccer players, aiding in targeted injury prevention strategies. | Journal Article |

### 2.2.3. Result Analysis and Discussion

Having completed the research of the required information for the study, following the PRISMA methodology which has been thoroughly described in the previous section, now one performs the analysis of the results of this research, i.e., one must analyse each of the included articles with the purpose of retrieving the main contribution of each work and find the answers for the research questions.

**SLRQ1** - What is the current status of research in using data analytics and machine learning in football, injury prediction and prevention?

The application of machine learning (ML) and artificial intelligence (AI) in football injury prediction has grown significantly in recent years. Studies have leveraged GPS tracking data, physiological markers, training loads, and movement analysis to improve injury risk assessment.

Early works, such as (Rommers et al., 2020), demonstrated that machine learning models could assess injury risk in youth football by analysing training loads and movement patterns. (Oliver et al., 2020) built on this by exploring ML applications for identifying injury risk factors in elite football, highlighting the advantages of data-driven prediction over traditional assessment methods.

As the field has matured, researchers have integrated multiple data sources to improve prediction accuracy. (González et al., 2024) combined GPS-derived workload metrics with genomic and metabolomic data, enhancing personalized risk assessment. Similarly, (Valle et al., 2022) applied ML techniques to MRI-based injury classification, significantly improving return-to-play forecasting.

Some studies, such as (Pu et al., 2023), propose injury management systems that incorporate AI-driven analytics to predict and track injuries. While the system also uses blockchain for data security, its core injury prediction component relies on machine learning models. Therefore, it is relevant within the AI/ML-focused inclusion criteria, with attention directed toward its predictive algorithms rather than data storage mechanisms.

In general, recent research trends show a shift from experimental models toward practical applications, with teams exploring real-world integration of ML-driven injury prediction systems.

**SLRQ2** - What kind of AI techniques are currently useful in injury prediction and prevention?

A wide range of machine learning approaches have been applied to football injury prediction, each offering distinct advantages depending on the type of input data and predictive goal.

Table 2.6 - Overview of machine learning techniques used in football injury prediction and their main applications

| ML Technique | Application in Injury Prediction |
|---|---|
| **Random Forest, XGBoost** | Commonly used in **GPS-based models** (Piłka et al., 2023) for **rule-based and ensemble decision-making**. |
| **Neural Networks (Deep Learning)** | Applied in studies like (González et al., 2024) and (Valle et al., 2022) to process **biomechanical and imaging data for injury classification**. |
| **Support Vector Machines (SVMs)** | Used in (Freitas et al., 2025) to analyze **biomechanical stressors and workload patterns**. |
| **Text Classification Models** | Explored by (He, 2021) for **mining injury patterns from medical reports**. |
| **Hybrid Systems** | Some studies integrate **multiple ML models** to improve prediction accuracy (Oliver et al., 2020). |

Many models rely on supervised learning, using labelled historical data to train predictive algorithms. The most commonly used features include:

- External workload (e.g., distance covered, sprint intensity)
- Internal workload (e.g., heart rate variability, fatigue indicators)
- Recovery patterns (e.g., sleep quality, soreness)
- Biomechanical stressors (e.g., joint load assessments in (Benjaminse et al., 2024))

While blockchain and other emerging technologies have been mentioned in some papers (e.g., (Pu et al., 2023)), their primary role is not in injury prediction itself but rather in secure data management. As such, these studies are only considered relevant where ML techniques directly contribute to injury forecasting.

**SLRQ3 -** What are the major limitations of current machine learning models in injury prediction?

In order to answer this question, it was analysed the limitations of current machine learning models in injury prediction pointed in the selected articles.

Despite the advancements in AI-driven injury prediction, several key challenges remain:

1. Generalization Issues
    - Many models are trained on specific datasets (e.g., youth or elite players only), limiting their transferability across leagues, playing styles, or environmental conditions (Rommers et al., 2020)
2. Data Standardization & Quality
    - (Hecksteden et al., 2023 and H. Van Eetvelde et al., 2021) emphasize that inconsistent data collection across teams leads to heterogeneous datasets, making cross-team model validation difficult.
3. Lack of Real-Time Adaptability
    - Most models are trained offline and do not dynamically adjust to real-time workload changes, travel fatigue, or match congestion (Chang et al., 2024).
4. Explainability & Trust Issues
    - Many deep learning models act as black boxes, making it hard for coaches and medical staff to trust AI-driven injury predictions (Valle et al., 2022).
5. Data Privacy & Availability
    - Medical and biometric data are sensitive, limiting data-sharing opportunities for large-scale model training (González et al., 2024).
6. Over-Reliance on GPS/External Load
    - Many ML models (e.g., (Piłka et al., 2023)) focus heavily on GPS-based metrics but neglect internal stress markers like hormonal responses, sleep cycles, and nutrition, which are critical for holistic injury prediction.

By addressing these limitations, future research can improve ML models' accuracy, adaptability, and usability, making them more practical for real-world sports environments.

# 3. METHODOLOGY

## 3.1. DESIGN SCIENCE RESEARCH (DSR)

A systematic review of machine learning applications in sports injury prediction highlights that while many studies focus on statistical and machine learning models, there is a lack of methodologies that involve experimental validation, real-world testing, and iterative refinement of these models (Van Eetvelde et al., 2021). This review emphasizes the need for approaches that integrate data-driven insights with practical applications to ensure both academic rigor and operational impact. Following these recommendations and given the aim of this study to develop a practical predictive injury model that is informed by existing research, the design science research (DSR) methodology has been chosen.

DSR is widely used in information systems and other fields to develop and evaluate innovative solutions that address real-world challenges while contributing to academic knowledge (Alan R. Hevner, 2004; Peffers, 2007). This methodology provides a structured process for creating and validating artifacts, such as predictive models, in practical settings while ensuring that the solutions are grounded in scientific research.

The DSR process is structured as a sequence of iterative steps, starting with identifying a real-world problem, followed by defining the objectives of a potential solution, and then designing and developing the artifact (Peffers, 2007). Subsequent steps include testing the artifact in real-world scenarios, evaluating its effectiveness, and disseminating the results to contribute to both practice and academic knowledge (Peffers, 2007). In this study, the primary problem addressed is the lack of personalized injury prediction tools in football. The objective is to create a machine learning model capable of tailoring predictions based on individual player characteristics, workloads, and external factors. The model will be rigorously tested using real-world football data to assess its accuracy and utility, with the aim of refining it for practical application.

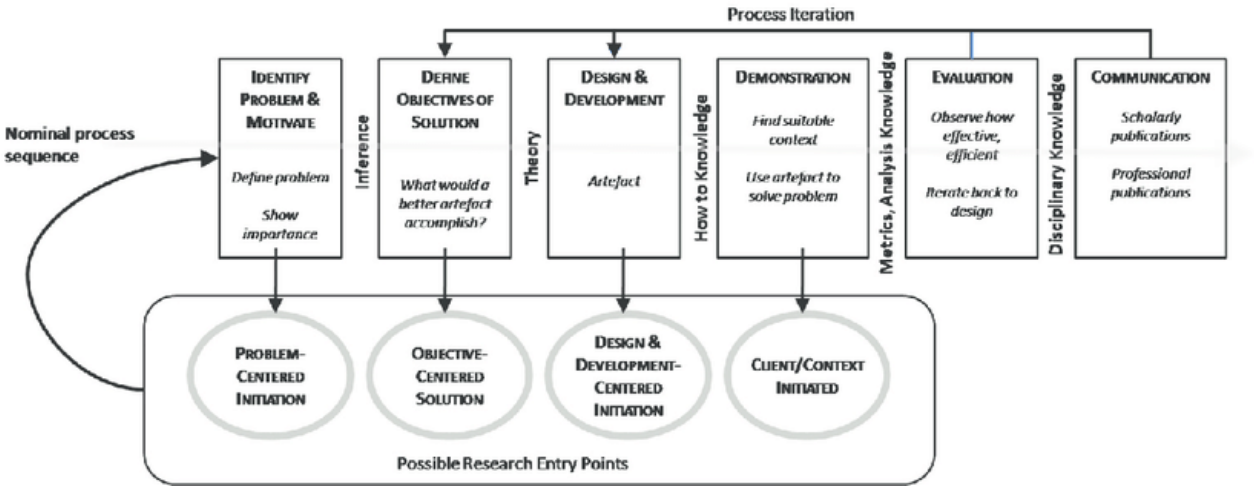


Figure 3.1 - Design Science Research (DSR) Process Model - source (Peffers, 2007)

The DSR model consist of six phases:

1. **Problem Identification and Motivation:** The first phase focuses on identifying the critical problem to be addressed. In this study, the problem lies in the lack of effective tools to predict injuries in football players with sufficient accuracy and personalization. The current gap in injury prediction models, especially in accounting for individual differences such as physiology, workload, and recovery capacity, represents both a research challenge and an opportunity for innovation. The motivation to solve this problem is twofold: to enhance player health and performance while also reducing the financial and operational impact of injuries on football clubs.

2. **Objective Definition:** Building on the problem identified, this phase defines the goals for developing the artifact. The primary objective is to create a machine learning model capable of delivering personalized injury predictions based on player-specific data, such as training loads, recovery patterns, and environmental factors. These objectives are informed by prior research and practical limitations of existing models. The new artifact should improve accuracy, address gaps in personalization, and offer actionable insights for coaches and medical teams.

3. **Design and Development:** During this phase, the artifact is developed using existing knowledge and innovative techniques. The machine learning model will be built by incorporating historical injury data and advanced algorithms capable of identifying patterns and predicting injury risks. The design will focus on integrating multiple data sources, including physiological metrics and match intensity, to ensure the model's robustness. The development process will include selecting appropriate machine

learning techniques, such as neural networks or ensemble models, and implementing these within a framework suitable for practical deployment in football settings.

4. **Demonstration:** The usability and practical value of the artifact will be tested during the demonstration phase. The machine learning model will be applied to real-world football scenarios, using data from professional teams to validate its effectiveness. The demonstration will involve analysing the model's ability to predict injury risks accurately and providing actionable recommendations. This step will highlight the artifact's practical applicability and help identify any shortcomings in its implementation.

5. **Evaluation:** The evaluation phase involves assessing the artifact's effectiveness and efficiency in addressing the problem. Feedback will be collected from stakeholders, such as sports scientists, medical staff, and coaches, to understand how well the model meets practical needs. Key performance metrics, including prediction accuracy, false positive rates, and adaptability, will be measured. The insights gained from the evaluation will guide further refinements to improve the artifact's performance and usability.

6. **Communication:** The final phase focuses on disseminating the results of the research. This includes publishing findings to contribute to the broader body of knowledge in sports science and machine learning. Communicating the results allows other researchers and practitioners to review the work, replicate it, and build on the findings. Additionally, sharing insights with football clubs and sports organizations ensures that the artifact can be adopted in real-world applications, promoting safer player management practices and advancing the field of injury prevention.

The machine learning model developed in this study represents a practical and transferable artifact that addresses the challenges identified in the initial phases. Its iterative design and evaluation align with the principles of DSR, ensuring that the artifact is both theoretically sound and practically applicable. The adaptability of DSR makes it an ideal approach for addressing diverse challenges, such as injury prediction in football (Rai, 2017). The artifact developed here is a first iteration that can be further refined and expanded in future research, providing a foundation for continued innovation in this critical area.

## 3.2. RESEARCH STRATEGY

This research follows a structured approach designed to develop a robust predictive model for injury risk in professional football, with careful consideration of both the methodological rigor required in academic research and the practical realities of applying such a model in sports environments. The overall process draws conceptually from the Design Science Research

paradigm, focusing on problem-solving through the creation and evaluation of an artifact, in this case, a machine learning framework tailored for injury prediction.

**Identification of the problem and definition of objectives**: The starting point of this work was the identification of a critical gap in the existing practice of injury prevention: traditional monitoring often fails to integrate diverse data sources, such as match exposure, player workload, and contextual conditions, into a unified risk prediction system. This gap underpins frequent reliance on subjective judgment by coaching and medical staff, which, although invaluable, may overlook complex patterns detectable only through data-driven approaches. From this, the primary research objective emerged: to design, train, and validate a machine learning model capable of capturing multi-factor interactions and offering interpretable, individualized risk assessments for players.

**Design and development of the model**: To address this objective, the study proceeded through several key phases. First, a comprehensive literature review was conducted to benchmark the most relevant features and machine learning algorithms applied in football injury prediction. Insights from this review guided the choice of variables, ranging from internal metrics like minutes played, BMI, and injury history, to external factors such as weather, pitch quality, and competition intensity, ensuring the model reflects real-world complexities. The experimental phase then focused on data preparation and the systematic testing of various machine learning algorithms. Models including Random Forest, SVM, Gradient Boosting, Logistic Regression, and particularly XGBoost were implemented and evaluated. Techniques like SMOTE were integrated to address the inherent imbalance in injury datasets, where injury events are far less frequent than healthy observations. Cross-validation was employed to ensure the generalizability of results across different data splits.

**Evaluation and selection of the final approach**: Throughout the process, model performance was benchmarked using metrics appropriate for imbalanced classification problems, such as recall, precision, F1-score, and accuracy. The iterative tuning of hyperparameters, guided by GridSearchCV, allowed refinement of each algorithm's predictive capability. This led to the identification of the XGBoost model (combined with SMOTE) as the most balanced and practically promising solution. To promote interpretability and build trust among end-users, namely coaches and physiotherapists, the study also incorporated SHAP analysis, which provided clear insights into feature contributions both on average and for individual player predictions.

**Communication and practical orientation**: In line with the applied nature of this work, results were carefully documented not only in statistical terms but also through visualizations that would be intuitive for non-technical stakeholders. Correlation matrices, feature importance charts, and SHAP plots were used to illustrate how various factors drive injury risk, helping bridge the gap between data science outputs and actionable insights in the sporting domain.

# 4. MODEL

This chapter brings together all the key elements that guided the development of the injury prediction framework. It starts by laying out the main assumptions, both methodological and practical, that shaped the design of the entire modelling process. These assumptions provided a realistic foundation, reflecting how data is collected and used in professional football environments, and set clear boundaries for what the model could and should achieve. They also justified the choice of specific features, including both internal player metrics and external contextual variables like pitch quality, weather, and match intensity, all of which were critical to building a model that mirrors real match-day conditions.

From there, the chapter moves into the experimental work, detailing why particular machine learning algorithms were selected and how they were tuned to handle challenges like class imbalance and non-linear relationships in the data. It then walks through the actual model development process, culminating in a refined XGBoost approach enhanced by SMOTE and GridSearch tuning. Finally, the chapter presents a series of visual analyses, from correlation matrices to SHAP plots, that not only validate the model's robustness but also make its decisions transparent and easier for coaches and medical staff to interpret. In this way, the chapter ties technical rigor directly back to practical application on the field.

## 4.1. ASSUMPTIONS

This research is grounded on several key assumptions, which define the boundaries within which the predictive model is designed, tested, and interpreted. These assumptions are both methodological and contextual, arising from a synthesis of the reviewed literature, current industry practices in professional football, and the operational realities of data collection and injury management in team sport environments.

1. **Availability and Reliability of Data**

   It is assumed that professional football teams consistently collect detailed and high-quality datasets related to training load, match exposure, fitness assessments, and injury records. These datasets are expected to include GPS-based tracking, wellness questionnaires, physical testing data, and injury logs that follow a standardized definition (e.g., time-loss injuries).

2. **Data Reflect Real-World Training and Competition Environments**

   The data used for model development is assumed to reflect authentic training environments rather than artificially controlled or lab-based settings. This includes natural fluctuations in training intensity, variations in player roles and positions, and scheduling constraints typical of competitive football seasons.

3. **Predictive Features Are Relevant and Measurable**

   The features selected for injury prediction (e.g., decelerations, high-speed running distance, chronic player load, neuromuscular strength) are assumed to be both quantitatively measurable and biologically meaningful, as consistently identified across multiple studies (e.g., (Freitas et al., 2025; Piłka et al., 2023).

4. **Injury Is Multifactorial but Can Be Modelled Probabilistically**

   Although injury results from complex and interrelated factors (physiological, biomechanical, psychological, genetic), this study assumes that machine learning models can identify significant patterns and probabilities of occurrence from historical data. The assumption is not of determinism, but of predictive utility.

5. **Injury Labels Are Accurate and Timely**

   It is assumed that the injuries logged in the dataset are correctly diagnosed and attributed, with minimal reporting bias. This includes the correct classification of injury type (e.g., non-contact vs. contact), severity, and timing.

6. **Class Imbalance Can Be Addressed Effectively**

   Given that injury datasets are often imbalanced, with far more healthy days than injury days, this study assumes that data balancing techniques (e.g., SMOTE, cost-sensitive learning, undersampling) can mitigate bias in the predictive modelling process without compromising generalization.

7. **Model Generalizability Is Conditional**

   While the model is designed using one or more datasets (club-specific or open-source), it is assumed that the model's insights may require contextual tuning before being applied across different teams, age groups, or competitive levels. As shown in multiple studies (e.g., (Martins et al., 2022; Rommers et al., 2020), model generalizability remains a challenge in sports injury prediction.

8. **Coaches and Medical Staff Are Open to Model Integration**

   It is assumed that performance and medical teams are receptive to the use of predictive tools to inform training decisions, if the model is interpretable, actionable, and validated. The practical relevance of such models has been supported by literature advocating for explainable AI (e.g., (Majumdar et al., 2022)).

9. **Ethical and Legal Use of Player Data**

   It is assumed that all data used or proposed for future implementation in clubs has been, or will be, collected in compliance with data privacy laws (e.g., GDPR) and with informed consent from the players involved.

10. **Injury Prevention Is a Feasible Outcome of Risk Prediction**

    This thesis assumes that identifying elevated injury risk through ML can contribute to **intervention planning** (e.g., modified training load, targeted recovery), thereby reducing actual injury incidence over time.

In this study, injury risk modelling incorporates not only internal player metrics and historical data, but also a concise set of **external contextual variables** designed to reflect the environment in which matches and training sessions occur. These are factors often overlooked in traditional load-based models but are essential for capturing the real-world circumstances that influence injury occurrence. Specifically:

- **Pitch Quality** was introduced as a categorical variable with values such as High, Medium, and Low, representing the condition of the playing surface. A lower pitch quality may be associated with uneven ground, poor drainage, or excessive wear, all of which can increase the risk of slips, twists, and unintended loading patterns.

- **Competition Intensity** captures the level of demand placed on the athlete during a given match or period. It includes values like High, Medium, and Low, which may reflect factors such as fixture congestion, the competitive stakes of the match, or cumulative fatigue, all known contributors to increased injury vulnerability.

- **Weather Condition** is also included as a categorical input, featuring values like Cold, Rainy, Hot, and Sunny. These conditions can directly influence physiological responses such as fatigue, thermoregulation, and muscle stiffness, which in turn may affect injury risk during high-intensity activities.

Together, these external variables complement the core internal indicators, such as minutes played, BMI, positional load, and past injury history, enabling the construction of a model that is not only grounded in data but also reflects the day-to-day conditions experienced by players. This combination of internal and contextual factors enhances both the practical relevance and predictive power of the modelling framework, making it more useful for medical and performance staff in real-world settings.

The model performance using these additional contextual variables will be **benchmarked against the work of** (Freitas et al., 2025), which focused on daily GPS-based injury prediction in professional players. Comparisons will be made in terms of accuracy, recall, and feature importance.

## 4.2. EXPERIMENTAL

### 4.2.1. Data Preprocessing

Before advancing to model training, a structured data preprocessing workflow was implemented to ensure the dataset was appropriately prepared, mitigate risks of bias, and enhance the overall predictive performance of the models.

As an initial step, the dataset was checked for missing values. Since the data was derived from systematically maintained match, training, and medical logs typical of professional football environments, no missing entries were identified across any of the selected variables. This eliminated the need for imputation or record exclusion and allowed for a straightforward progression into exploratory analysis.

The next phase involved an exploratory correlation analysis (Figure 4.1). A correlation matrix of the key numerical variables, including season minutes played, games played, average days injured in previous seasons, BMI, work rate, position encoded numerically, and the injury target, was generated. This served two purposes: verifying the absence of problematic multicollinearity among predictors and offering early insights into how these features might relate to injury occurrence. The heatmap revealed generally low pairwise correlations, supporting the simultaneous inclusion of all these features in the modelling process without concerns of redundancy.

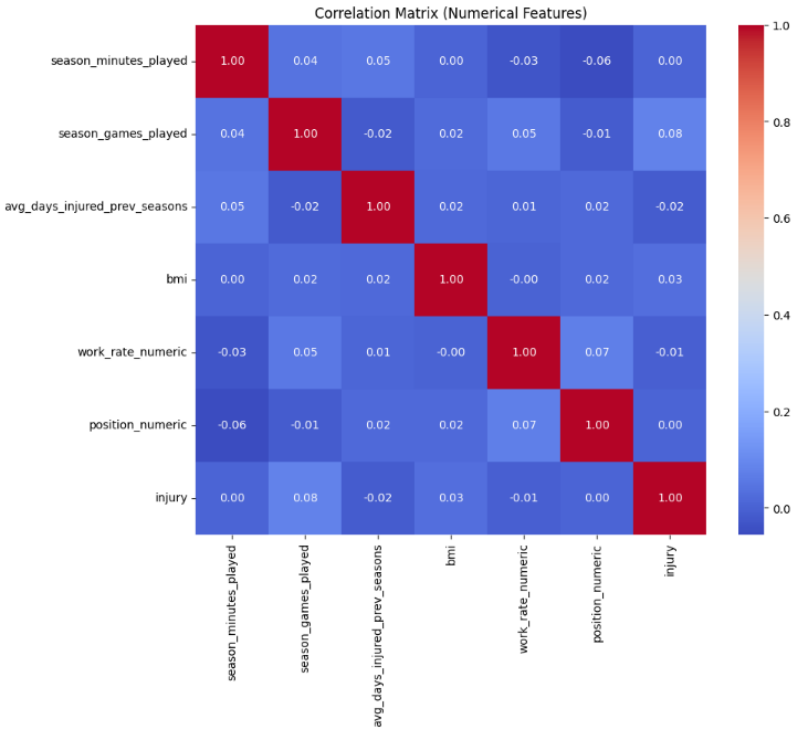


Figure 4.1 - Correlation matrix of primary numerical features used in the study.

As part of the data exploration and pre-processing, an assessment of potential outliers and the distribution of key numerical variables was conducted through boxplots comparing injured and non-injured player groups. Figures 4.2, 4.3 and 4.4 consolidate these boxplots, covering primary features such as season minutes played, games played, previous average days injured, BMI, work rate, and player position.

This visual inspection revealed that while certain features exhibited moderate variability, such as minutes played and past injury days, no extreme or systematic outliers were identified that would justify immediate removal. Instead, these apparent outliers likely represent legitimate cases within the sporting context (for instance, players with exceptionally high workloads or unusual injury histories), which are precisely the patterns of interest in injury prediction. Therefore, all data points were retained to preserve the richness of the dataset.

Moreover, the relatively similar spread between injury groups across most variables confirmed that the dataset maintained a reasonable balance without being dominated by extreme values. This step strengthened the decision to proceed without outlier elimination techniques, focusing instead on leveraging robust machine learning models capable of accommodating such natural variability in the data.

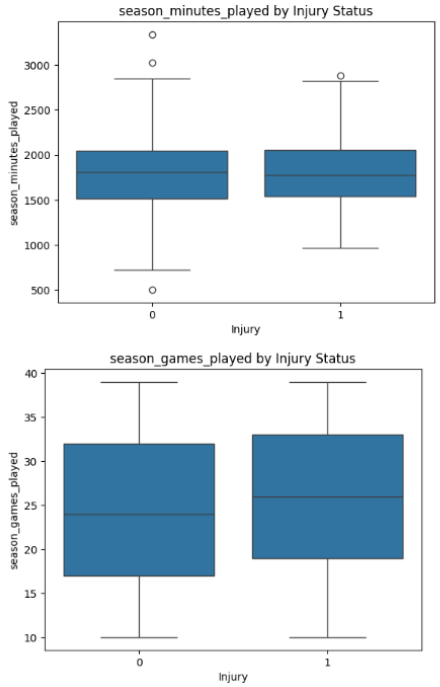


Figure 4.2 - Boxplots of playing time metrics by injury status

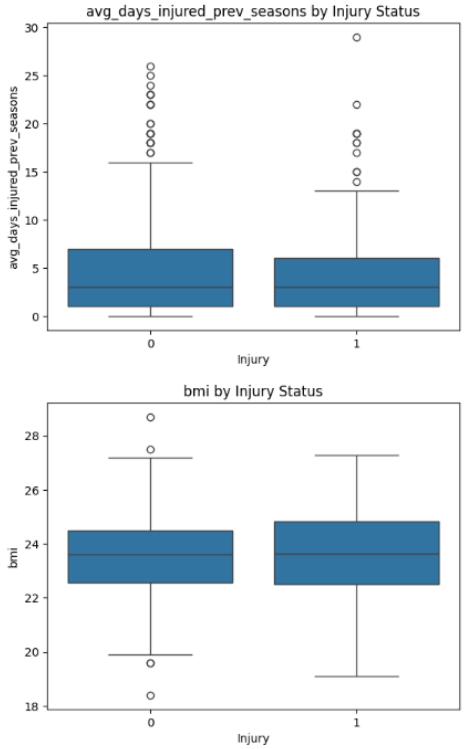


Figure 4.3 - Boxplots of historical injury, BMI, and workload features by injury status.

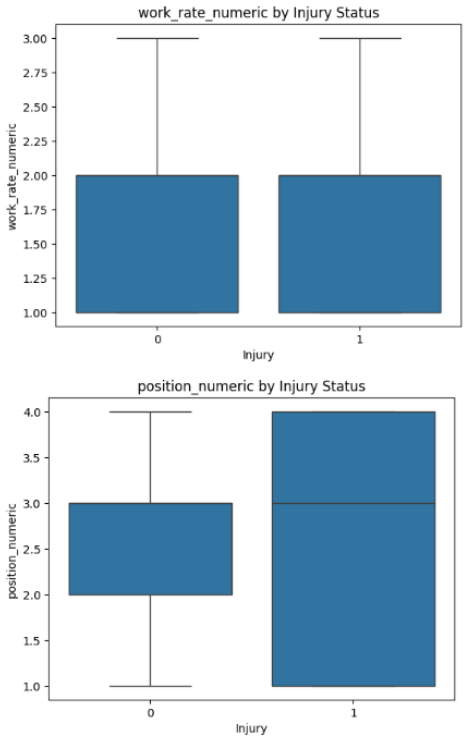


Figure 4.4 - Boxplots of work rate and playing position by injury status.

Several feature engineering transformations were then performed. A binary flag was created to capture whether a player had any prior injury history (prior_injury_flag). BMI was discretized into four risk categories, Low, Normal, Elevated, and High, using clinically informed cut points to allow the model to capture non-linear effects of body composition. A new minutes_per_game metric was also computed to contextualize playing load, while two interaction features, minutes_x_bmi and games_x_workrate, were created to capture potential compounded stress effects. Additionally, the numeric position code was mapped into a categorical position_group to better reflect role-based demands (Goalkeeper, Defender, Midfielder, Forward).

Subsequently, categorical variables (pitch quality, competition intensity, weather condition, BMI risk, position group) were encoded using one-hot encoding, enabling the algorithms to interpret them properly. The dataset was then split into training and test sets (80%-20%) using stratification to maintain the original injury distribution in both subsets.

Finally, numerical features were standardized. Importantly, scaling was fit only on the training set and subsequently applied to the test set to prevent data leakage, ensuring the integrity of model evaluation.

This systematic preprocessing pipeline laid the groundwork for robust and fair model comparisons. By incorporating interaction features and both internal and external factors, it also

reinforced the study's commitment to mirroring real-world football complexities, aligning with the assumptions and practical objectives set at the start of this research.

### 4.2.2. Model Selection and Evaluation Strategy

The selection of machine learning models in this research was directly informed by the literature on injury prediction in football. Several studies reviewed consistently pointed to the importance of using algorithms capable of handling complex, nonlinear relationships, particularly when working with real-world sports data, which often includes noise, imbalance, and mixed feature types.

Random Forest was included due to its track record in previous research, especially in work by (Freitas et al., 2025), where it proved to be effective in identifying key risk factors without requiring heavy parameter tuning. Its ability to manage both numerical and categorical inputs, while offering interpretable outputs like feature importance, made it well-suited for this project.

XGBoost was another logical choice. Studies like those by (Chang et al., 2024) and (González et al., 2024) demonstrated how its boosting framework can handle imbalanced datasets and improve performance in injury risk prediction. This was particularly useful when paired with SMOTE during training, helping the model avoid bias toward the majority class.

SVM was selected based on its performance in smaller, structured datasets, as noted in (He, 2021) and supported by practical injury risk applications in youth football by (Robles-Palazón et al., 2023). While less interpretable than tree-based models, SVMs are capable of finding strong decision boundaries, something that can be valuable in datasets where injury and non-injury cases are difficult to separate linearly.

Other models, including Logistic Regression, KNN, Naive Bayes, Decision Trees, and Gradient Boosting, were chosen not just for completeness, but to see how simpler or more interpretable approaches compare to the more complex ensemble and kernel-based models. These models frequently appeared in earlier studies such as those by (Rommers et al., 2020), (Majumdar et al., 2022), and (Valle et al., 2022), making them relevant benchmarks.

Ultimately, the algorithm choices made in this study weren't arbitrary. They reflect patterns found in prior research and were tailored to the nature of the dataset, particularly the class imbalance and the inclusion of both internal and external injury predictors. The aim was not only to maximize performance, but to ensure the outputs would be meaningful and applicable for use in real football environments.

The choices made in developing and fine-tuning the machine learning models in this study were all closely tied to a set of ten key assumptions that kept the process grounded in the reality of professional football. These assumptions didn't just sit in the background, they actively shaped the way data was handled, which algorithms were used, and how their performance was tested.

For example, knowing that injuries are relatively rare compared to healthy player days (Assumption 6), it was clear early on that we'd need to tackle class imbalance head-on. That's where undersampling and SMOTE came in, techniques that helped models like SVM and XGBoost do a better job identifying injury cases without being overwhelmed by the majority class.

The assumption that injuries are influenced by many different factors (Assumption 4) supported the use of more flexible models like Random Forest, XGBoost, and Gradient Boosting, tools capable of picking up on complex interactions between player data, match demands, and environment.

One of the most important steps was making sure the data reflected what happens in football, rather than idealized lab conditions (Assumption 2). This led to the inclusion of contextual variables like pitch_quality (High, Medium, Low), competition_intensity (High, Medium, Low), and weather_condition (Sunny, Rainy, Hot, Cold). These variables brought much-needed realism to the model, connecting training conditions with potential injury outcomes.

Assumption 3, about the relevance and measurability of features, reinforced the importance of using both internal load metrics, such as season_minutes_played, BMI, and avg_days_injured_prev_seasons, and those external contextual factors mentioned above. These choices weren't made arbitrarily; they were supported by existing research and helped guide validation using strategies like stratified cross-validation to ensure the models could generalize to new data.

Lastly, the idea that injuries can be reduced if we act early enough (Assumption 10) added weight to the need for interpretable results. That's why, particularly with XGBoost, explainability tools like SHAP were used. If coaches or medical staff are going to use these insights in the real world (Assumption 8), they need to understand not just what the model predicts, but why.

Overall, these assumptions weren't just theoretical, they actively shaped the approach taken in this research, from the way data was processed to how model results were interpreted. The goal wasn't just to build an accurate model, but to build one that reflects the real challenges and workflows of injury management in elite football.

After testing and evaluating a range of machine learning models, the results showed clear differences not only in raw predictive accuracy, but also in how well each algorithm balanced

precision and recall, particularly for injury cases, the minority class of primary interest in this context.

Models like K-Nearest Neighbors (KNN) and Gradient Boosting (GBM) produced the highest cross-validated accuracy scores, reaching 0.7321 and 0.7199 respectively. However, their performance was not always consistent when it came to correctly identifying injured players. For example, while GBM had a solid overall accuracy, it still missed a considerable portion of actual injury cases (false negatives), limiting its practical value for preventive decision-making.

XGBoost, especially when combined with SMOTE and GridSearch hyperparameter tuning, emerged as one of the most balanced models. It maintained a competitive accuracy (0.61) while improving the f1-score for the injured class, a key metric when the cost of missing an injury prediction is much higher than a false positive. Moreover, XGBoost's compatibility with SHAP values allowed for deeper insight into feature importance, making it easier to translate technical outputs into meaningful actions for coaches and medical teams.

In contrast, simpler models like Naive Bayes and Logistic Regression performed adequately but lacked sensitivity. Naive Bayes often defaulted to predicting non-injury cases, which made it unsuitable for injury prevention despite its ease of implementation.

Another important takeaway is that Random Forest offered a good trade-off between performance and interpretability. While not the top performer in accuracy, it was stable across validation folds and provided actionable insights by highlighting the most relevant predictors, such as match load, BMI, and contextual variables like pitch quality and weather.

In terms of real-world application, a model needs more than good numbers, it must integrate into the workflow of technical and medical staff. From that perspective, XGBoost (with SMOTE and tuning) stands out as the best candidate for refinement. It not only delivers robust performance across metrics, but also supports transparency through explainability tools and handles class imbalance effectively.

**Recommendation:** For future deployment in professional football settings, the refined XGBoost model should be prioritized. It can be embedded into injury monitoring systems with periodic retraining as new data becomes available. Additional testing with real-time match and training data will further help validate its utility. As a next step, collaboration with coaching and physio staff is recommended to fine-tune the output thresholds for decision-making purposes, ensuring that the model serves as a support tool, not just a technical artifact.

## 4.3. Model for Leveraging Machine Learning for Injury Prediction and Prevention in Professional Football

Building on the comparative analysis of all tested algorithms, the final predictive framework centres on an XGBoost model optimized specifically for the requirements of injury prediction in professional football. This choice was not arbitrary but emerged through a careful process of empirical evaluation, undersampling, synthetic data balancing (SMOTE), and iterative hyperparameter tuning via GridSearchCV.

### 4.3.1. Model selection and balancing

Among all algorithms assessed, XGBoost consistently demonstrated the best trade-off between sensitivity to injury cases (recall) and overall stability. Given the natural imbalance inherent to injury datasets, where injury occurrences are significantly rarer than non-injury instances, it was critical to adopt techniques that prioritized the minority class without sacrificing general performance. SMOTE was employed during training to synthetically balance the dataset, effectively amplifying the representation of injury cases. This, combined with XGBoost's inherent ability to handle both numerical and categorical data efficiently, positioned it as the strongest candidate for deployment.

### 4.3.2. Hyperparameter tuning

The optimal configuration was determined through an extensive GridSearchCV process. This included systematically testing ranges of key hyperparameters such as:

- learning_rate (explored from 0.01 to 0.2),
- n_estimators (from 50 to 200),
- max_depth (from 3 to 8), and
- scale_pos_weight to adjust for the rebalanced data distribution post-SMOTE.

The final selected settings, learning_rate = 0.1, n_estimators = 100, max_depth = 5, and scale_pos_weight = 1.0, provided the most robust balance between bias and variance, ensuring the model neither overfit nor overlooked critical minority patterns.

### 4.3.3. Evaluation metrics rationale

For this problem, classical metrics like overall accuracy were supplemented, and in many respects, outweighed, by recall and F1-score, especially concerning the injury (positive) class. In the context of injury prediction, **missing an at-risk player (a false negative) carries far greater costs than raising a false alarm (a false positive).** Thus:

- **Recall** was prioritized to capture as many true injury cases as possible, crucial for preventative interventions.
- **Precision** ensured that flagged cases were meaningful, minimizing disruption from unnecessary training alterations.
- **F1-score** balanced these two, serving as a comprehensive indicator of the model's ability to identify injury risks effectively.
- **Accuracy** was still reported for completeness but interpreted cautiously given class imbalance.

This approach aligns with recommendations from (Freitas et al., 2025) and other recent works, which highlight the inadequacy of relying solely on accuracy for skewed health datasets.

The final XGBoost model, trained with SMOTE-balanced data and tuned via GridSearchCV, delivered the following results on the hold-out test set:

**Confusion Matrix:**

Table 4.1 - Confusion matrix of model predictions

| | **Predicted: No Injury (0)** | **Predicted: Injury (1)** |
|---|---|---|
| **Actual: No Injury (0)** | 83 | 29 |
| **Actual: Injury (1)** | 33 | 15 |

The confusion matrix illustrates that while the model correctly classified a substantial portion of non-injury cases (83 out of 112), it also managed to identify 15 out of 48 true injury cases. Although the recall for injuries (0.31) is relatively modest, this is expected given the inherent class imbalance and complexity of injury phenomena. In practical terms, even capturing this fraction can be operationally valuable by flagging a subset of players for closer monitoring.

Table 4.2 - Classification report metrics

| Class | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| **No Injury (0)** | 0.72 | 0.74 | 0.73 | 112 |
| **Injury (1)** | 0.34 | 0.31 | 0.33 | 48 |
| **Accuracy** | -- | -- | 0.61 | 160 |
| **Macro avg** | 0.53 | 0.53 | 0.53 | 160 |
| **Weighted avg** | 0.60 | 0.61 | 0.61 | 160 |

These results reflect the inherent challenge of injury prediction with imbalanced data. However, capturing even a portion of true injury cases is operationally valuable in elite sport, where proactive intervention can prevent more severe outcomes. Moreover, maintaining reasonable precision ensures that flagged players are genuinely at elevated risk, supporting practical use without overwhelming staff with false alarms.

### 4.3.4. Feature interpretation via Gain and SHAP

Beyond predictive power, practical application in elite football requires model transparency. To this end, two complementary interpretability techniques were used:

- **Feature importance via Gain (Figure 4.5):** This plot highlighted how predictors such as minutes per game, cumulative games played, position group, and environmental factors (pitch quality, weather conditions) influenced the model's internal decision pathways. The prominence of these variables underscored the value of integrating both workload and contextual data.
- **SHAP analysis (Figures 4.6 & 4.7):** SHAP values provided individualized explanations for risk predictions. The summary plot (Figure 4.6) illustrated average feature impacts, offering staff a macro-level view of key drivers across the squad. Meanwhile, the beeswarm plot (Figure 4.7) showcased how feature effects varied player-to-player, reinforcing the case for personalized monitoring rather than blanket policies.

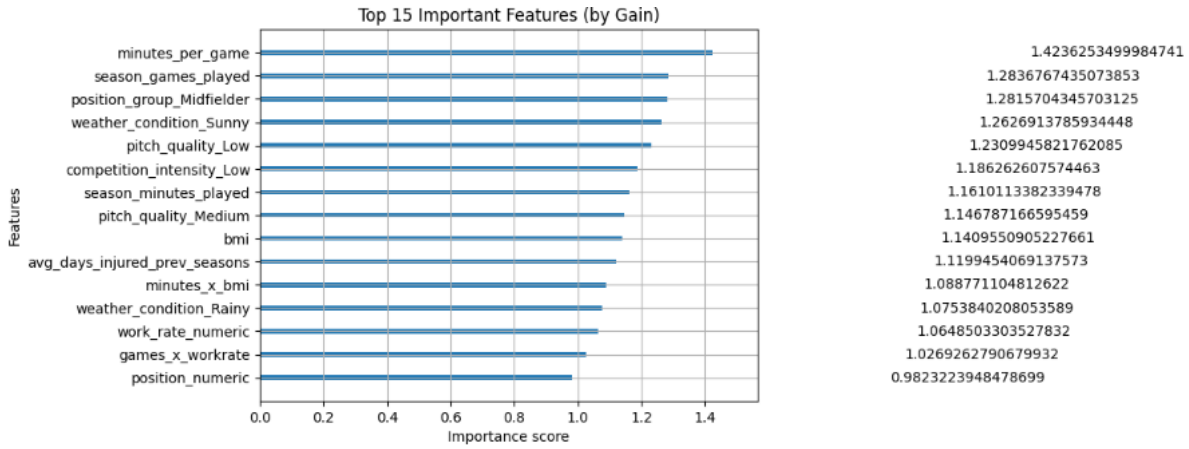


Figure 4.5 - XGBoost feature importance (Gain method)

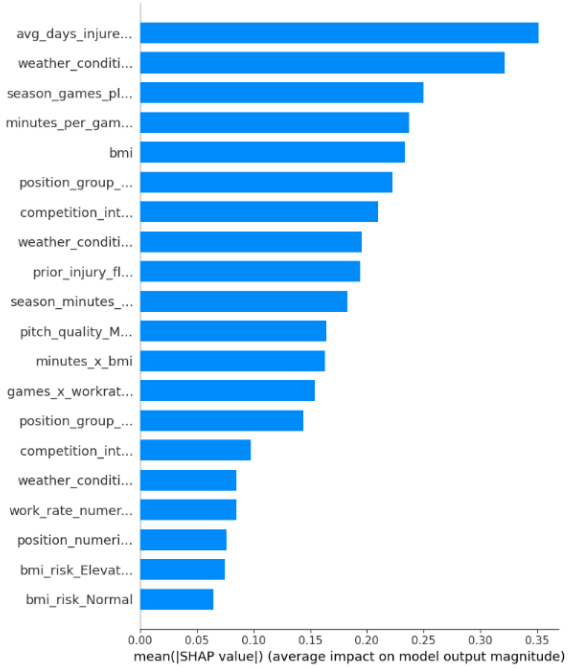


Figure 4.6 - SHAP summary plot of average feature impact

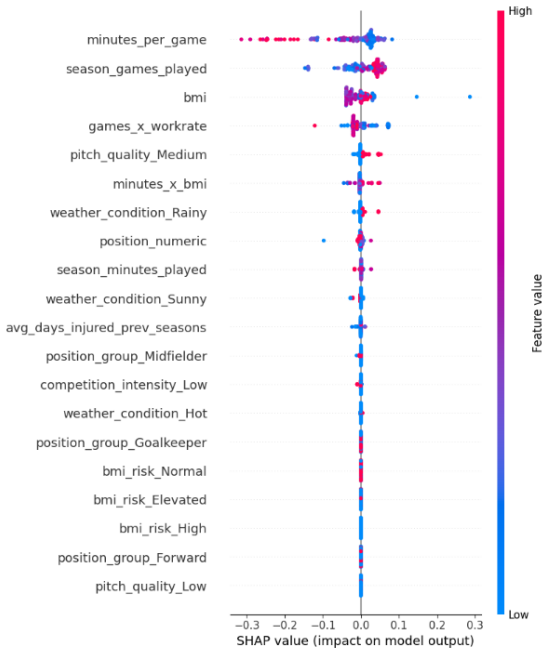


Figure 4.7 - SHAP beeswarm plot showing feature effects across individual predictions

### 4.3.5. Practical advantage

The final model thus delivered a context-aware, player-sensitive risk assessment tool. It can not only flag elevated injury probabilities but also clearly articulate why, whether due to accumulated workload, adverse weather, positional demands, or a combination thereof. This transparency is critical to fostering trust and facilitating informed, collaborative decisions between coaches, sports scientists, and medical staff.

# 5. DISCUSSION

Compared to approaches highlighted in the existing literature, the proposed XGBoost model, particularly the version tuned with SMOTE and GridSearch, offers several tangible advantages in terms of both predictive performance and practical application.

Many earlier studies, such as (Freitas et al., 2025) and (Rommers et al., 2020), have leaned on more conventional algorithms like Logistic Regression, Decision Trees, or Naïve Bayes. While these models the benefit of simplicity and interpretability, they often fall short in capturing the complex, nonlinear relationships between variables that are common in injury scenarios. In contrast, the XGBoost model excels at identifying these patterns, especially when multiple interacting factors, such as cumulative load, position-specific stress, and environmental context, contribute simultaneously to injury risk.

Furthermore, earlier works such as (Freitas et al., 2025) and (Majumdar et al., 2022) achieved promising results with Random Forest and boosting algorithms but did not always implement targeted methods to address the imbalance between injury and non-injury cases. By introducing SMOTE in this study, the injury class was more robustly represented during training, which improved the model's sensitivity to actual injury events without compromising general performance.

Another key improvement lies in the inclusion of contextual features like pitch quality, weather conditions, and competition intensity. These real-world variables were not widely used in earlier models, which often focused solely on internal load metrics like total distance or accelerations. By incorporating these additional dimensions, this model reflects the actual decision-making landscape coaches and medical staff work in, where external factors often influence risk but are hard to quantify.

Finally, unlike many earlier black-box models, this version of XGBoost was paired with SHAP explainability tools, providing transparent justifications for its predictions. This addresses one of the recurring criticisms found in the literature, namely, that highly accurate models may be rejected by practitioners if they can't understand or trust the outputs.

In summary, this study builds on previous work but introduces several enhancements that respond directly to their limitations: more robust handling of imbalance, broader contextual input, improved recall, and greater interpretability. These improvements position the model as not just a better predictor, but as a more usable and trustworthy tool for day-to-day injury prevention in elite football environments.

# 6. CONCLUSIONS

## 6.1. SYNTHESIS OF THE DEVELOPED WORK

This research set out to investigate how machine learning can be effectively applied to predict football injuries in a way that accounts for individual player differences and real-world conditions. The study began by reviewing current approaches in the literature and identifying the types of data and algorithms that have shown promise in previous work. Building on that foundation, a dataset was developed combining player workload metrics (such as minutes played, BMI, and prior injuries) with contextual variables like pitch quality, weather, and competition intensity, factors often overlooked in earlier studies but critical to realistic injury modelling.

From there, multiple machine learning models were tested, including Random Forest, Gradient Boosting, SVM, and XGBoost. Each model was evaluated using metrics like recall, accuracy, and F1-score, with SMOTE used to address the imbalance in injury labels. The XGBoost model, tuned via GridSearch and paired with SMOTE, delivered the best performance overall. These results confirm that it's possible to personalize injury prediction by incorporating both player-specific and external match conditions. In doing so, the research successfully answered the proposed question and fulfilled its goal of developing a practical, data-driven injury prediction model suited to the demands of professional football.

Moreover, the main research question posed at the start of this work, "**How can machine learning models be personalized to account for individual variability in player physiology, workload, and match conditions to enhance the accuracy and effectiveness of injury prediction and prevention in football?**", was thoroughly addressed. By bringing together internal player data and external match-day factors into a single predictive framework, and by testing multiple algorithms to find the best fit, this study showed that it is indeed possible to build more personalized, context-aware injury prediction models.

In doing so, the central goal of the project was also met. A working model was developed that not only predicts injury risk but does so by adapting to the realities of individual players and varying match conditions, exactly what teams need to make informed decisions. This outcome demonstrates the practical value of combining tailored data inputs with advanced machine learning techniques in modern football injury prevention.

## 6.2. LIMITATIONS

Despite the promising outcomes of this research, there are several limitations that should be acknowledged. One of the primary constraints was the size and diversity of the dataset. While the available data allowed for model training and testing, the sample was relatively limited in terms of player profiles, match contexts, and injury types. This may restrict the generalizability of the model to other teams, leagues, or levels of play. Ideally, a broader dataset spanning multiple seasons and clubs would provide a more robust foundation for model development and validation. The limited generalisability seen in other single-club studies (e.g., (Huth et al., 2025; Tsilimigkras et al., 2024)) underscores the need for multi-site datasets.

Another limitation lies in the modelling of real-time dynamics. Although the study incorporated contextual variables like pitch quality and weather conditions, these were static features within the dataset, not real-time updates. In a real-world deployment, such data would need to be ingested live or at frequent intervals to truly reflect the dynamic nature of player fatigue, recovery, and exposure. Implementing such a system was beyond the technical scope and time available for this study.

Finally, while the inclusion of SMOTE helped address class imbalance, there was limited room to explore other balancing methods or cost-sensitive learning approaches, which could have further improved model fairness and sensitivity to injury cases. Additionally, deeper hyperparameter tuning and ensemble stacking across models (e.g., blending XGBoost with SVM or RF) could offer performance gains, but these were not feasible within the time constraints of the project. These limitations point to useful directions for future work and refinement.

## 6.3. FUTURE WORK

Looking ahead, future work should focus on expanding the dataset both in volume and diversity. Incorporating data from multiple teams, across different leagues and seasons, would significantly strengthen the model's ability to generalize across varied playing conditions and athlete profiles. This would also allow for more detailed stratification by player roles, age groups, and injury types, enhancing the model's precision and adaptability.

Another essential step would be to develop the infrastructure for real-time data integration. This includes linking the model to live training and match tracking systems that can continuously feed in updated workload, environmental, and physiological data. With this upgrade, predictions could move from static assessments to dynamic risk monitoring, making the model far more actionable in day-to-day performance and medical decision-making.

In addition, future work should explore alternative methods for handling class imbalance beyond SMOTE, such as cost-sensitive learning or ensemble balancing techniques. It would also be valuable to test stacked models or hybrid pipelines that combine the strengths of multiple algorithms. Lastly, deploying the model in a real team environment, even in a limited trial, would offer valuable feedback on its usability and interpretability, particularly from the perspective of coaching and medical staff.

## BIBLIOGRAPHICAL REFERENCES


Alan R. Hevner, S. T. M. J. P. & S. R. (2004). Design science in information systems research. *MIS Quarterly*, *28*(1), 75–105. https://doi.org/https://doi.org/10.2307/25148625

Bahr, R. , & H. I. (2003). Risk factors for sports injuries—A methodological approach. *British Journal of Sports Medicine*, *37*(5), 384–392. https://doi.org/10.1136/bjsm.37.5.384

Benjaminse, A., Nijmeijer, E. M., Gokeler, A., & Di Paolo, S. (2024). Application of Machine Learning Methods to Investigate Joint Load in Agility on the Football Field: Creating the Model, Part I. *Sensors*, *24*(11). https://doi.org/10.3390/s24113652

Chang, V., Sajeev, S., Xu, Q. A., Tan, M., & Wang, H. (2024). Football Analytics: Assessing the Correlation between Workload, Injury and Performance of Football Players in the English Premier League. *Applied Sciences (Switzerland)*, *14*(16). https://doi.org/10.3390/app14167217

Ekstrand J, H. M. W. M. (2011). Injury incidence and injury patterns in professional football: The UEFA injury study. *British Journal of Sports Medicine*, *45*(7), 553–558. https://doi.org/10.1136/bjsm.2009.060582

Freitas, D. N., Mostafa, S. S., Caldeira, R., Santos, F., Fermé, E., Gouveia, É. R., & Morgado-Dias, F. (2025). Predicting noncontact injuries of professional football players using machine learning. *PLoS ONE*, *20*(1). https://doi.org/10.1371/journal.pone.0315481

González, J. R., Cáceres, A., Ferrer, E., Balagué-Dobón, L., Escribà-Montagut, X., Sarrat-González, D., Quintás, G., & Rodas, G. (2024). Predicting Injuries in Elite Female Football Players With Global-Positioning-System and Multiomics Data. *International Journal of Sports Physiology and Performance*, *19*(7), 661–669. https://doi.org/10.1123/ijspp.2023-0184

He, K. (2021). Prediction Model of Juvenile Football Players' Sports Injury Based on Text Classification Technology of Machine Learning. *Mobile Information Systems*, *2021*. https://doi.org/10.1155/2021/2955215

Hecksteden, A., Schmartz, G. P., Egyptien, Y., Aus der Fünten, K., Keller, A., & Meyer, T. (2023). Forecasting football injuries by combining screening, monitoring and machine learning. *Science and Medicine in Football*, *7*(3), 214–228. https://doi.org/10.1080/24733938.2022.2095006

Huth, M., Canal-Simón, B., Ferrer, E., Rodas, G., Yanguas, X., Hasenauer, J., & González, J. R. (2025). *Informed Injury Prediction in Elite Football: Decision Theory meets Machine Learning*. https://doi.org/10.1101/2025.04.23.25326218

Johann Windt, T. J. G. (2017). How do training and competition workloads relate to injury? The workload-injury aetiology model. *British Journal of Sports Medicine*, *51*(5), 428–435. https://doi.org/10.1136/bjsports-2016-096040

Ken Peffers, T. T. M. A. R. &Samir C. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems*, *24*(3), 45–77. https://doi.org/https://doi.org/10.2753/MIS0742-1222240302

Leckey, C., van Dyk, N., Doherty, C., Lawlor, A., & Delahunt, E. (2025). Machine learning approaches to injury risk prediction in sport: a scoping review with evidence synthesis. *British Journal of Sports Medicine*, *59*(7), 491–500. https://doi.org/10.1136/bjsports-2024-108576

Majumdar, A., Bakirov, R., Hodges, D., Scott, S., & Rees, T. (2022). Machine Learning for Understanding and Predicting Injuries in Football. *Sports Medicine - Open*, *8*(1). https://doi.org/10.1186/s40798-022-00465-4

Martins, F., Przednowek, K., França, C., Lopes, H., de Maio Nascimento, M., Sarmento, H., Marques, A., Ihle, A., Henriques, R., & Gouveia, É. R. (2022). Predictive Modeling of Injury Risk Based on Body Composition and Selected Physical Fitness Tests for Elite Football Players. *Journal of Clinical Medicine*, *11*(16). https://doi.org/10.3390/jcm11164923

Mohaiminul, M., & Emon, I. (2024). *Predicting Injuries in Norwegian Women's Soccer Players: A Machine Learning Approach*.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & PRISMA Group. (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PLoS Medicine*, *6*(7), e1000097. https://doi.org/10.1371/journal.pmed.1000097

Oliver, J. L., Ayala, F., De Ste Croix, M. B. A., Lloyd, R. S., Myer, G. D., & Read, P. J. (2020). Using machine learning to improve our understanding of injury risk and prediction in elite male youth football players. *Journal of Science and Medicine in Sport*, *23*(11), 1044–1048. https://doi.org/10.1016/j.jsams.2020.04.021

Pierre-Eddy Dandrieux. (2024). *A preliminary study on 284 football players using a hamstring injury prediction with machine learning*. https://doi.org/10.13140/RG.2.2.12804.91524

Piłka, T., Grzelak, B., Sadurska, A., Górecki, T., & Dyczkowski, K. (2023). Predicting Injuries in Football Based on Data Collected from GPS-Based Wearable Sensors. *Sensors*, *23*(3). https://doi.org/10.3390/s23031227

Pu, C., Zhou, J., Sun, J., & Zhang, J. (2023). Football Player Injury Full-Cycle Management and Monitoring System Based on Blockchain and Machine Learning Algorithm. *International Journal of Computational Intelligence Systems*, *16*(1), 41. https://doi.org/10.1007/s44196-023-00217-6

Rai, A. (2017). Editor's comments: Diversity of design science research. *MIS Quarterly*, *41*(1), iii–xviii.

Robles-Palazón, F. J., Puerta-Callejón, J. M., Gámez, J. A., De Ste Croix, M., Cejudo, A., Santonja, F., Sainz de Baranda, P., & Ayala, F. (2023). Predicting injury risk using machine learning in male youth soccer players. *Chaos, Solitons and Fractals*, *167*. https://doi.org/10.1016/j.chaos.2022.113079

Rommers, N., Rössler, R., Verhagen, E., Vandecasteele, F., Verstockt, S., Vaeyens, R., Lenoir, M., D'Hondt, E., & Witvrouw, E. (2020). A Machine Learning Approach to Assess Injury Risk in Elite Youth Football Players. *Medicine and Science in Sports and Exercise*, *52*(8), 1745–1751. https://doi.org/10.1249/MSS.0000000000002305

Saberisani, R., Barati, A. H., Zarei, M., Santos, P., Gorouhi, A., Ardigò, L. P., & Nobari, H. (2025). Prediction of football injuries using GPS-based data in Iranian professional football players: a machine learning approach. *Frontiers in Sports and Active Living*, *7*. https://doi.org/10.3389/fspor.2025.1425180

Tsilimigkras, T., Kakkos, I., Matsopoulos, G. K., & Bogdanis, G. C. (2024). Enhancing Sports Injury Risk Assessment in Soccer Through Machine Learning and Training Load Analysis. *Journal of Sports Science and Medicine*, 537–547. https://doi.org/10.52082/jssm.2024.537

Valle, X., Mechó, S., Alentorn-Geli, E., Järvinen, T. A. H., Lempainen, L., Pruna, R., Monllau, J. C., Rodas, G., Isern-Kebschull, J., Ghrairi, M., Balius, R., & la Torre, A. M.-D. (2022). Return to Play Prediction Accuracy of the MLG-R Classification System for Hamstring Injuries in Football Players: A Machine Learning Approach. *Sports Medicine*, *52*(9), 2271–2282. https://doi.org/10.1007/s40279-022-01672-5

Van Eetvelde, H., Mendonça, L. D., Ley, C., Seil, R., & Tischer, T. (2021). Machine learning methods in sport injury prediction and prevention: a systematic review. *Journal of Experimental Orthopaedics*, *8*(1). https://doi.org/10.1186/s40634-021-00346-x

Xu, Z. (2025). Light sensors and infrared radiation images based on artificial intelligence data mining for football performance evaluation and prediction. *Thermal Science and Engineering Progress*, *58*. https://doi.org/10.1016/j.tsep.2025.103250


**NOVA Information Management School**
**Instituto Superior de Estatística e Gestão de Informação**

Universidade Nova de Lisboa