Machine Learning–Driven Prediction of Injury Risk in Elite Football: A Tuned Ensemble Model With SHAP-Based Explainability

Submitted By

Minhazul Islam

ID: 221-35-932

Department of Software Engineering

Daffodil International University

Supervised By

Mr. Md. Khaled Sohel

Assistant Professor

Department of Software Engineering

Daffodil International University

Department of Software Engineering

Daffodil International University

December 2025

APPROVAL

This thesis titled on "**Machine Learning–Driven Prediction of Injury Risk in Elite Football: A Tuned Ensemble Model With SHAP-Based Explainability**", submitted by **Minhazul Islam (ID: 221-35-932)** to the Department of Software Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Bachelor of Science in Software Engineering and approval as to its style and contents.

BOARD OF EXAMINERS

Dr. A. H. M. Saifullah Sadi
Professor
Department of Software Engineering
Faculty of Science and Information Technology Daffodil International University

Chairman

Dr. Rubaiyat Islam
Associate Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Internal Examiner 1

Dr. Md. Abdul Kader
Associate Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Internal Examiner 2

Nuruzzaman Faruqui
Assistant Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Internal Examiner 3

Md. Mostafiz Khan
Managing Director
Tecognize Solutions Limited

External Examiner

Machine Learning–Driven Prediction of Injury Risk in Elite Football: A Tuned Ensemble Model With SHAP-Based Explainability

Minhazul Islam

Bachelor of Science

DAFFODIL INTERNATIONAL UNIVERSITY

DAFFODIL INTERNATIONAL UNIVERSITY

DECLARATION OF THESIS AND COPYRIGHT

Author's Full Name: Minhazul Islam

Date of Birth: 28/01/2002

Title: Machine Learning-Driven Prediction of Injury Risk in Elite Football: A Tuned Ensemble Model With SHAP-Based Explainability

Academic Session: 2022-2025

I declare that this thesis is classified as:

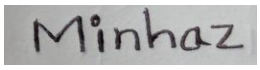
☐ CONFIDENTIA (Contains confidential information under the Official Secret Act 1997)*

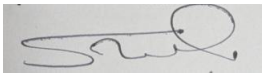
☐ RESTRICTED (Contains restricted information as specified by the organization where research was done)*

☐ OPEN ACCESS I agree that my thesis to be published as online open access I acknowledge that Daffodil International University reserves the following rights:

1. **The Thesis is the Property of Daffodil International University.**
2. **The Library of Daffodil International University has the right to make copies of the thesis for the purpose of research only.**
3. **The Library of Daffodil International University has the right to make copies of the thesis for academic exchange.**

Certified by:





(Student's Signature)	**(Supervisor's Signature)**
Minhazul Islam	**Mr. Md. Khaled Sohel**
Student ID: 221-35-932	**Name of Supervisor**
Date: 24/12/2025	**Date: 24/12/2025**

NOTE : * If the thesis is CONFIDENTIAL or RESTRICTED, please attach a thesis declaration letter.

THESIS DECLARATION LETTER (OPTIONAL)

Librarian,

Daffodil International University,

Daffodil Smart City, Ashulia, Dhaka, Bangladesh

Dear Sir,

CLASSIFICATION OF THESIS AS RESTRICTED

Please be informed that the following thesis is classified as RESTRICTED for a period of three (3) years from the date of this letter. The reasons for this classification are as listed below.

Author's Name: Minhazul Islam

Title: Machine Learning–Driven Prediction of Injury Risk in Elite Football: A Tuned Ensemble Model With SHAP-Based Explainability

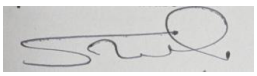
Reasons: (i)

(ii)

(iii)

Thank you.

Yours faithfully,



(Supervisor's Signature)

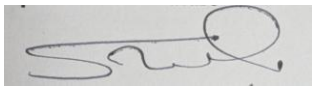
Date: 24/ 12/ 2025

Stamp:

Note: This letter should be written by the supervisor and addressed to the Librarian. *Daffodil International University,* with its copy attached to the thesis.

SUPERVISOR's DECLARATION

I hereby declare that I have checked this thesis and in my opinion, this thesis is adequate in terms of scope and quality for the award of the degree of Bachelor of Science.

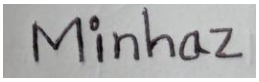


(Supervisor's Signature)

Full Name: Mr. Md. Khaled Sohel
Position: Assistant Professor
Date: 24/12/2025

STUDENT'S DECLARATION

I hereby declare that the work in this thesis is based on my original work except for quotations and citations, which have been duly acknowledged. I also declare that it has not been previously or concurrently submitted for any other degree at Daffodil International University or any other institution.



(Student's Signature)

Full Name: Minhazul Islam

ID Number: 221-35-932

Date: 24/ 12/ 2025

Machine Learning–Driven Prediction of Injury Risk in Elite Football: A Tuned Ensemble Model With SHAP-Based Explainability

Machine Learning–Driven Prediction of Injury Risk in Elite Football: A Tuned Ensemble Model With SHAP-Based Explainability

Minhazul Islam

Thesis submitted in fulfilment of the requirements for the award of the degree of Bachelor of Science

Department of Software Engineering

DAFFODIL INTERNATIONAL UNIVERSITY

DECEMBER 2025

ACKNOWLEDGEMENTS

Alhamdulillah, Allah (SWT), Most Generous is most merciful, and we praise Him only. His inexhaustible blessings, guidance, and strength are only under His infinite blessings that I managed to continue my education and do this thesis. All this academic process has been steered by His wisdom, and without His assistance, patience, and mercy, it would have been impossible to make this step.

I want to say that, after Allah (SWT), I love my dear parents, love and respect them greatly, and I am very thankful to them. They have made me successful because of their unconditional love, numerous sacrifices, and constant prayers, as well as their unwavering faith in me. They suffered, sacrificed their comfort, and helped me emotionally and morally in such a way that I would reach my objectives. They never lost trust in me, even during hard moments, and their support made me have the strength to continue. I am eternally grateful to them, and I hope Allah compensates them highly and rewards their faith with peace, well-being, and Jannah at the top ranks.

I would also like to personally express my many thanks to my noble supervisor, **Mr. Md. Khaled Sohel**, because of his support, invaluable guidance, and patience in this research. His knowledge, positive feedback, and regular encouragement were crucial in the development of this piece of work and quality improvement.

I also recognize the support given by faculty members, instructors, and staff of the Software Engineering Department to make my learning experience productive and contrived towards my growing academic knowledge.

I want to inform my friends and well-wishers that they have greatly helped me in this journey through their encouragement, understanding, and moral support. Lastly, I would like to thank all the people who explicitly or implicitly helped me in accomplishing this thesis. May Allah (SWT) bless and have a successful life for all of them in this world and the hereafter. Ameen.

ABSTRACT

In the top-tier football, particularly in any of the high-demand leagues such as the English Premier League (EPL) and the Spanish LaLiga, injuries among players continue to significantly affect performance and revenue. This paper is based on the use of multidimensional data on 669 professional soccer players and can introduce a novel ensemble-based machine learning system that is capable of identifying the likelihood of an individual suffering an injury in 2024/25. The data set will contain data on match workload, GPS-computed locomotor measures, physiological measures, recovery and wellness measures, and historical injury data. Some of the learning algorithms that we tested include LightGBM, XGBoost, CatBoost, Random Forest, and Logistic Regression. Soft Voting Ensemble. The best-discriminating (AUC = 0.883) and best-calibrating probability models were tuned gradient-boosting models, which outperformed any individual model. Conversely, the LightGBM model on the baseline provided risk estimates that were overconfident and thus not helpful when making medical decisions in practice. SHAP demonstrates significant variables causing injuries, including the number of seasons, the frequency of previous injuries, stride length, sleep quality, and high intensity accelerations. The proposed model creates a powerful, intuitive means by which elite football clubs can forecast injury risk on a real-time basis. This contributes to evidence-based load management and medical interventions of the EPL and LaLiga.

Table of Contents

List of Figures

List of Tables

CHAPTER 1

INTRODUCTION

The chapter presents the scientific basis, background, and justification of the research in injury-prediction risk of injury in elite football. It is a mixture of the recent sports epidemiological research, sports monitoring, sports physiology, and data-driven analytics. In addition, this chapter also shows how intricate the mechanisms of injuries are, how poorly traditional modeling solutions are, and how machine learning and interpreted AI are growing in popularity in organizations characterized by high performance levels. The chapter also ends with the definition of the motivation and goals of the current research.

1.1 Background and Context

Injuries constitute a chronic and disruptive threat in highly competitive football, and significant outcomes in the performance of the team, investment, and wellness of players. According to major epidemiological researches, professional football players tend to receive two to three injuries annually, among which are the most frequent types of injuries, such as the strains of muscles, ligament ruptures, and overuse traumas [33 35]. Time-loss injuries that prevent players to train and compete in competitive games complicate it much more by making clubs less stable when it comes to tactics preparation and increasing medical and rehabilitation expenses.

These injuries have multisomatic and dynamic underlying mechanisms, which occur as a result of interactions at physical, physiological, psychological, contextual, and historical levels. Examples of external demands to work that cause mechanical stress on the body and increase the risk of developing soft-tissue injuries include high-speed running, sprinting, rapid accelerations, and decelerations [5]. Contemporary football is becoming more reliant on high intensity transition practices and pressing behaviors that result in an incremental load of the neuromuscular system and increased fatigue during

match cycles [4]. The recovery demands are further exacerbated by the congestion of a fitter, as that body finds it more difficult to self-heal.

Recent advances in monitoring athletes have seen the possibility to both externally and internally measure external and internal loads in a very high temporal resolution. The external loads include total distance, sprint data, mechanical load, and movement patterns, which are measured through GPS and inertial sensor systems [23, 22]. In addition to this, physiological and wellness measurements, including heart rate variability (HRV), creatine levels (CK), sleep measurement, perceived stress scale, and a subjective fatigue scale, can provide information on internal load and recovery condition [16, 19]. With the mentioned advances, one of the key scientific issues is the possibility to detect high-risk states prior to suffering injuries, which encourages the elaboration of more complex analytical frameworks.

This paper aims to provide evidence-based research on elite football by defining problems in football and identifying potential remedies to address them.

1.2 Problem definition in Elite Football

The purpose of the paper is to establish evidence based research on the football game with an aim of defining a problem in football and which solutions that may be adopted to address the issue.

Leading football associations collect big multimodal data that incorporates data regarding training weights, physiological and wellness measurements, match actions, and previous morbidity. Nevertheless, it is hard to convert such data into useful findings related to injury prevention. The research problem stated in the thesis is the following:

What is the most effective way of predicting the risk of injury in elite football players through multimodal workload, physiological, wellness, and historical data points?

This problem is problematic in a number of ways. First, injury incidents are relatively low compared to the number of training exposures, resulting in uneven data distributions. Second, effects of workload are time-varying: a certain spike in intensity can be light enough to be an acceptable load to one player but a serious risk to another, based on fitness level, fatigue accumulation, recovery status, and history of injuries. Third, the

topic of physiological variables, including HRV, CK, and sleep quality, is extremely individual and dependent on psychology and environment-related issues [16, 18]. Fourth, the lines of training load and fatigue, recovery, and injury risk are not simple; they interact in such a complicated way that they can hardly be observed in simple models [2].

These difficulties render the need to have predictive systems that are in a position to represent complex, player-specific dynamics and generalization across seasonal cycles and competition scenarios.

1.3 What Is the Rationale of Predicting the Risk of Injury?

Correct injury-risk forecasting has important practical and scientific implications for elite football. In an applied sense, injuries of the time loss variety would yield benefits such as increased player availability, absence of tactics, and reduction in the necessity to have disruptive squad rotation. Prevention of injuries also helps to save money on healthcare, subsequent treatment, and loss of valuable players.

In sports science, the ability to predict potential injury risk comprises the connection between the monitoring of load, recovery management, and performance improvement. Predictive signals that are validated will help practitioners to customize the intensity of daily training, develop rotation plans during busy fixture schedules, examine recovery kinetics, and detect signs of overreaching and other maladaptive stress reactions in the initial stages. These models have the ability to present complex multidimensional data into easy-to-understand risk summaries, and this eliminates the subjective judgment that may be inconsistent in competitive pressures [1].

More so, the elaboration of transparent and decipherable predictive systems is in concert with the current-day priorities in elite sport, wherein organizations are placing a greater priority on injury prevention, welfare of their athletes, and ethical practices regarding player-care.

1.4 Weakness of Traditional Method.

Previously, the methods of predicting injuries was limited to basic heuristic measurements, to single workload ratios, and regression-based statistical models. The most common example is the acute:chronic workload ratio (ACWR), which is intended to contrast between acute load spikes and the training history. The early studies found correlations between higher ACWR and injury rate, but subsequent meta-analyses demonstrate that such studies had serious methodological errors and inconsistent predictive validity [13–15]. These are limitations that may be summed up as oversimplifying the loadadaptation mechanisms, inadequate consideration of nonlinearities, changeable smoothing windows, and not being grounded in physiological processes. Regression models such as linear regression and logistic regression assume independence, linearity and constant variance, which are rarely met in heterogeneous datasets to track athletes [4]. Such models find it difficult to respond to nonlinear interactions, large-dimensionality feature spaces, and variation in the ability of the people to sustain injuries. Moreover, traditional techniques often lack external validity; a set of rules and concepts tailored to one club rarely translate effectively to a different club, due to differences in training approach, player qualities, and effects related to the context.

These limitations indicate that future studies need to utilize more versatile and data-driven approaches that would be able to model complex interactions and individualized patterns of susceptibility to injuries.

1.5 The Part Machine Learning Plays in Sports Injury Analytics

Machine learning has emerged as an attractive option since it can indirectly acquire nonlinear and multivariate relationships through data. In the latest studies, the machine learning models (random forests, gradient boosting machines, support vector machines, and neural networks) have been used to predict muscle injuries, non-contact injuries, fatigue states, and readiness-to-train outcomes in elite football and team sports [6, 12]. The systematic reviews always reported that predictive discrimination is better in ML models than traditional regression models, particularly when using multimodal inputs [7]. Machine learning methods make it beneficial in many ways compared to traditional statistical methods. They combine heterogeneous sources of data (e.g., external load, internal load, sleep, stress, and biomechanical indicators), identify nonlinear threshold

effects and high-order interactions that are hard to explicitly model. The ensemble learning approaches, specifically, enhance cement in that they combine numerous models, which in many cases leads to the enhancement of generalization performance [30, 31]. Nevertheless, the main problem of the practical integration in the elite sport is that it is hard to comprehend. Clear explanations should be given to coaches and other medical staff to facilitate high-risk designations. Regardless of statistical accuracy, black-box predictions do not tend to be actionable. The explanation artificial intelligence systems, in particular SHAP (Shapley Additive exPlanations) systems, have become indispensable in converting risk scores derived by ML into the form of interpretable and physiologically relevant information [27, 28]. SHAP breaks down individual predictions into feature-level contributions, enabling practitioners to gain more insight into the underlying reasons a player is at risk, given the current time point used in workload, physiological, or contextual factors. ML + XAI is a promising trend towards constructive and clear-cut injury prediction models.

1.6 Research Motivation & Objective

Motivation

Despite the rising number of elite clubs that embrace high-end monitoring technologies, the number of injuries has not reduced considerably in the last 20 years. This implies that the current surveillance mechanisms and decision-making procedures do not adopt the complexity of the injury mechanisms comprehensively. It is required to have predictive systems that can combine multimodal data, such as external workloads, internal physiological phenotypes, wellness indicators, and injury history, and simulate nonlinearly interacting dynamics. Meanwhile, the increasing emphasis of the evidence-based practice in elite sports requires easy-to-understand, reliable, and provide useful feedback predictive tools. These types of tools are useful in assisting coaches, sports scientists, and doctors in making intelligent decisions on how to adjust the amount of training, plan recovery, and proceed with the workload of each individual. The main aim of the study is to develop a data-driven and open, and scientifically sound injury prediction system with purposes tailored exclusively to the needs of elite football.

Objectives

The primary objectives of this thesis are

1. To collect and compile a multimodal dataset comprising information regarding external workloads, physiological, wellness, and historical injury data.
2. To structure a modelling pipeline that is methodologically sound, that grasps both temporal order and class imbalance, as well as concerns nonlinear reactions.
3. This is aimed at deploying and evaluating machine learning models applicable to multivariate nonlinear tasks of using the predicted injury on high-performance settings.
4. In this study, explainability methods, especially SHAP, will be used to reveal the global as well as individual factors which affect the risk of injury.
5. To evaluate the predictive and interpretative effectiveness of the proposed framework in real-life competitive elite football monitoring situations.

2.7 Chapter Summary

The scientific background of predicting injuries in elite football was discussed in this chapter. It emphasized the ways in which injuries may occur due to a wide variety of reasons, how the old approach to modeling has its restrictions, and how machine learning and explainable AI are gaining relevance. This chapter clarifies the purpose and objectives of the research as it introduces the basis on which new methods are discussed in the following chapters. The following chapter consolidates the existing research on injury analytics, athlete monitoring, and machine learning, citing notable gaps in the existing research, which justify the approach to the proposed methodology.

CHAPTER 2

LITERATURE REVIEW AND RESEARCH GAPS

This chapter integrates previous research on injury-related epidemiology, athlete monitoring, workload science, physiological biomarkers, and machine learning in elite football. It provides the theoretical and empirical basis of the data-driven prediction systems of injury development. This chapter also reviews both the older and recent modeling techniques in an evaluative manner, synthesizes findings of both the previous and the current research, and highlights the apparent weaknesses in the current literature. These gaps guide the scientific contribution and methodology of the current study.

2.1 Injury Epidemiology and Load Monitoring in Elite Football

The occurrence of injury in elite football is widely studied using long term surveillance designs that were implemented both in domestic competitions and in international events. Established epidemiological results indicate that time-loss injuries are very commonly perceived among professional football players, with strains in muscles (mostly in the hamstrings, the groin, the quadriceps, and the calf) comprising the greater part [33] -[35]. Such injuries have enormous sports and economic costs that affect the continuity of a squad, availability of match, and also lead to more medical and recovery charges. Physical demands of contemporary football have been very high in the past decades. According to match-analysis studies, there are significant rises in the high-speed running, sprint rate, accelerations, and decelerations that have been induced by the tactical trend that dictates high-pressing, rapid transitions, and tight defensive constructions [4], [5]. Cumulative muscle exhaustion is also enhanced by the additional factor of only having a limited number of matches to rest, a factor that is driven by congestion caused by fixtures, especially at the high-end leagues and international tournaments [40]. Consequently, the risk of injury is not introduced by the single exposures of loads, but rather the repetitive mechanical stress in combination with insufficient recovery. In order to deal with such requirements, high-profile clubs use elaborate athlete-tracking equipment. The external

load is conventionally assessed by means of GPS and inertial sensors, which would present objective measurements, which include the total distance, maximal distance during running, number of sprints, and mechanical load indices [22], [23]. These are measures of the mechanical stresses exerted on musculoskeletal structures and have been again and again linked with injury abundance on exposure exceeding the individual tolerance levels [1], [21]. Nevertheless, epidemiological data also indicate that external load by itself is not enough to justify risk of injuries and it would be necessary to consider the combination of monitoring methods.

2.2 Internal Load, Recovery, and Physiological Stress Indicators

Internal load indicates the psychological and physiological reactions of the athlete to the external training and match demands. Internal load is unlike external load, which measures what the athlete behaves whereas internal load measures the response of the body through adaptation and recovery. Heart-rate variability (HRV) is widely considered an internal load indicator and biochemical indicator of fatigue, e.g., creatine kinase (CK), sleep Delays and quality, subjective fatigue, and perceived stress scores are all commonly used indicators of internal load [16], [17], [19]. HRV is internationally accepted as the measure of the balance and the recovery ability of the autonomic nervous system. Fatigue and neuromuscular preparedness on the one hand, low-intensity workloads on the other are linked to lower HRV [16]. Sleep disorders also contribute to these effects, having adverse effects on cognitive functions, hormonal functions, and tissue repair mechanisms [17], [36]. Sleep and physiological fatigue have the interaction with psychological stress, increasing the vulnerability to injury in the context of prolonged competitive stress [18], [37]. The CK and other biochemical indicators can give some insight to muscle damage and the slow healing process after intense loading [20]. The study showed that higher levels of CK increases the risk of injuries associated with overload with the absence of adequate regulation of the intensity of training. Irrelevant or inconsistently employed in the PD of injuries, even though relevant, internal load indicators are fraught with issues of frequency of measurement, missing data, and inter-subject variability. This said, the literature indicates a solid reason in favor of the combination of internal and external measures of expected loads to ascertain the multifactorial character of injury risk.

2.3 Traditional Injury Modelling Approaches and Their Limitations

Before machine learning was adopted, the use of machine learning in predicting injuries in football was mainly dominated by conventional statistical methods, which encompassed the use of linear or logistic regression, mixed-effects model as well as survival analysis. Such methods gave preliminary information on workload-injury links, but, anyway, they are limited with striking suppositions about linearity, autonomy, and stationarity [4]. These assumptions are seldom applicable in datasets of athlete-monitoring, which has nonlinear interactions, multi collinearity, time dependence, and heterogeneity in individuals. One of the most popular traditional methods applied is the acute-to-chronic workload ratio (ACWR). It is proposed that early research believed that the risk of injury escalation was greater with the sudden increases in acute workload compared with chronic load. Nevertheless, later meta-analyses and methodological criticism found significant conceptual and statistical flaws, such as arbitrary window choice, regression to the mean, and low predictive validity [13]-[15]. This is leading to the perception of ACWR as a descriptive monitoring tool that is not an absolutely predictive model. On the whole, conventional statistical techniques have difficulties modeling dynamic, nonlinear, and individual processes that cause injury development. Such shortcomings have prompted the search for more versatile, data-driven solutions capable of being able to describe complex interactions across a variety of data domains.

2.4 Machine Learning and Explainable AI in Injury Prediction

The use of machine learning (ML) in the study of football injuries has become of growing interest because it can be used to estimate the nonlinearity of relationships, compound the capability of data because it is high-dimensional, and it can be used to capture the interaction effect among multiple variables. Articles that implemented the ML methods continuously claim enhanced predictive accuracy as opposed to conventional statistical formulations, especially when external load, internal load, and historical injury data are

merged [6]–[11]. Random forests and gradient boosting machines are examples of tree-based algorithms that are often used due to their resistance to multicollinearity and feature interactions that are challenging to describe [30], [31]. The ensemble learning techniques also increase predictive stability by aggregating different classifiers, which minimizes variability and improves generalization. According to Table 2.1, the majority of studies conduct their performance in threshold-independent measures, including ROC -AUC or overall discrimination power, to demonstrate that the focus is on risk stratification and not fixed binary decision-making. Indicatively, there is a large number of studies that do not state precision. This is mainly due to the fact that research concerning injury prediction in the past has focused much on discrimination and prioritization of risk as opposed to classification results of a threshold dependent nature. Accuracy is hence only presented where they are directly provided in the original research, and N/A is presented in other sections of Table 2.1 to prevent misrepresentation. One of the problems linked to the deployment of ML models in elite football is interpretability. The black box predictions cannot be easily operationalized in the medical and performance conditions of requiring transparent reasoning. The explainable artificial intelligence (XAI) techniques work around this shortcoming by showing what the model does. SHAP (SHapley Additive exPlanations) has become one of the most popular XAI methods, with the ability to provide both global and participant-characterized explanations by defining the level of contributions of features to specific predictions [27], [28]. Evidence of the use of SHAP in studies indicates that it is useful in identifying risk drivers that can be acted upon, i.e., accrued workloads, lack of recovery, and prior exposure to injury [23].

2.5 Comparative Summary of Related Works

Table 2.1 – Summary of Related Studies

Ref	Title	Dataset / Context	Methodology	Output Performance	Limitations	Precision
[1]	Machine Learning for Understanding and Predicting Injuries in Football SpringerOpen	Elite football	ML classifiers	AUC up to ~0.80	Single-club data	N/A
[2]	ML approaches to injury risk prediction in sport BJSM	38 studies	38 studies	38 studies	38 studies	38 studies
[3]	ML methods in sport injury prediction & prevention Springer	27 studies	Systematic review	ML > traditional stats	Limited prospective data	N/A
[4]	Internal & external load vs injury using ML (meta-analysis)	Pro soccer	Meta-analysis	Combined load improves AUC	Inconsistent injury definition	N/A
[5]	ACWR & injury risk meta-analysis	Multi-sport	Meta-analysis	ACWR AUC < 0.65	Not football-specific	N/A
[6]	ACWR for injury prevention in soccer	Pro soccer	Systematic review	Mixed/inconclusive	No ML evaluation	N/A
[7]	Workload & injury risk in athletes	Elite athletes	Systematic review	Confirms workload relevance	Few ML studies	N/A
[8]	Injury forecasting with GPS + ML	1 club, 1 season	RF & GBM	Good PR metrics	Small, single-club	N/A
[9]	Internal and external loads predicting injuries	40 players	Tree-based ML	Higher AUC vs. external-only	Small dataset	N/A
[10]	Predicting non-contact injuries in football PLOS ONE	4 seasons	Ensemble ML	High discrimination	Missing physiology data	N/A
[11]	Multi-season ML load–injury study	EPL club	RF, GBM	Stable multi-season patterns	Single club	N/A
[12]	Injury risk assessment in soccer via ML	25 players	ML models	Moderate–high accuracy	Very small N	N/A

[13]	ML-based muscle injury prediction (4-year) MDPI	Pro club, 4 seasons	LR, RF, GBM	AUC ≈ 0.75	Limited XAI	N/A
[14]	Internal/external load & injury ML synthesis	Multi studies	Meta-analysis	Combined load > regression	Poor tuning detail	N/A
[15]	Loading/Unloading in elite football	2 seasons	Longitudinal	Load patterns linked to risk	No ML classifier	N/A
[16]	Internal load vs recovery in U19	U19	HR metrics	Complex load–recovery patterns	No injury outcome	N/A
[17]	Time-series ML for injury risk	Multi-sport	Time-series ML	Better than static features	Not football-specific	N/A
[18]	Big-data ML for injury forecasting	Multi-sport	SVM pipeline	Real-time forecasting feasible	Mixed-sport dataset	N/A
[19]	Survey of ML injury prediction methods	2015–2023	Narrative review	Overview of DL & XAI	Not peer-reviewed	N/A
[20]	ML approaches to injury risk in sport	Multi-sport	Review	Highlights key features	No model evaluation	N/A
[21]	Concept for ML football injury system ScienceDirect	Conceptual	System design	Defined ML pipeline	No real data	N/A
[22]	Internal/external load with ML in soccer	Pro soccer	ML synthesis	Non-linear load effects captured	Same data limits as #14	N/A
[23]	SHAP-based injury prediction (baseball)	Baseball	GBM + SHAP	Personalised insights	Non-football	Precision reported (≈0.81)
[24]	Deep learning for injury prediction	Multi-sport	DL vs ML	DL outperforms ML	Preprint	N/A
[25]	GPS/internal load & injury ML meta-analysis	Pro soccer	ML synthesis	ML > threshold models	Needs multi-class dataset	N/A

2.6 Research Gaps

Although the advancement of machine learning-based injury prediction is soaring, some significant gaps were observed in the literature. First, a significant number of studies are based heavily on external load measures, and not many internal load, recovery, and wellness measures are incorporated. Second, temporal structure is frequently poorly dealt with, and methods of random data split are used, which is not representative of true forecasting environments and can exaggerate results on benchmarking.

Third, systematic hyperparameter optimization and ensemble learning are not fully used as they have proved to be beneficial in the structured sports domain. Fourthly, explainability is still under-integrated, which restricts the levels of practitioner confidence and practice adoption. Also, the majority of research is methodologically limited to small data sets of one club with low-scale external validation, which diminishes the results concerning other leagues and seasons. Last but not least, there are often player-specific adaptations and personalized risk profiles that are not paid enough attention to yet, even though there is substantial evidence to support that the susceptibility of athletes to injuries greatly depends on an individual.

To work towards these gaps, an ensemble-based multimodal, temporally aware, interpretable, and multimodal modelling framework is needed that generates calibrated and practical injury-risk estimates that could be useful in elite football settings.

2.7 Chapter Summary

This chapter accessed the literature on the epidemiology of injuries, monitoring of athletes, the science of workload, and predictive modelling in elite football. The evidence shows that risk of injury is due to the interaction between external workload, internal physiological stress, recovery capacity, and exposure to injury in the past in a complex, nonlinear manner. Conventional statistical methods and threshold-related measures cannot adequately reflect these dynamics.

Ensemble-based models show better predictive discrimination because machine learning can provide a promising alternative. Nevertheless, the current literature is limited in terms

of multimodal integration, temporal validation, the lack of explainable AI usage, and lack of consistency in reporting of classification scores, including precision. These research gaps can be linked to define each research gap in the chapter that will in turn lead to the methodology framework that will be provided in the next chapter in the book to fill these research gaps.

CHAPTER 3

METHODOLOGY

This chapter elaborates on the methodological framework that was created to design a machine-learning system that can predict the risk of injury in elite football. The methodology is based on the idea of integrating various data domains into a single predictive architecture, such as workload measures, physiological measures, biomechanical measurements, wellness measures, and the physiological history of longitudinal injuries. Since injuries are nonlinear and emerge as a result of complicated interactions between external loads, internal stress, the phenomenon of tissue adaptation, and individual player predispositions [1, 7, 12], the modeling pipeline was created to take into consideration both the statistical characteristics of the data and the physiological facts of the injury processes. The chapter has been structured with each of the parts of the pipeline. The data collection and aggregation process is presented first, including the combination of English Premier League and LaLiga databases into a multidimensional database of 669 players to play in the competitive season of 2024/25. It is then succeeded by intensive preprocessing, such as that of missing-values, outlier detection, and temporal adjustment, unit normalization, and control of class-imbalance, which are all critical to maintaining data integrity and avoiding leakage during model training.

Later sections give the informed- Feature engineering on a domain basis with special focus to the construction of an injury-history-module with temporal-resolving. This aspect uses season-phase group, mechanism group, recurrence measurement, and cumulative exposure measures, which are all known to be predictive factors of injury susceptibility [2, 4, 28]. The modeling workflow then conveys the development of single machine-learning classifiers, the optimization of cross-validation through the use of hyperparameter search, and the creation of the developed final Soft Voting Ensemble, which combines 7 tuned base learners. The time-conscious train validation test split, three-fold temporal cross-validation, and evaluation procedures are also outlined in the chapter. Finally, the SHAP explainability integration is presented as a needed development of practitioner-oriented interpretation, allowing to easily identify the strongest factors of workload, physiological, and historical nature that lead to the expected risk of injury. Figure 3.1, provided below, presents an overview of the end-to-

end workflow, including data ingestion, preprocessing, feature engineering, model training, ensemble integration, evaluation, and explainability, that will serve as the methodological roadmap of the rest of this chapter.

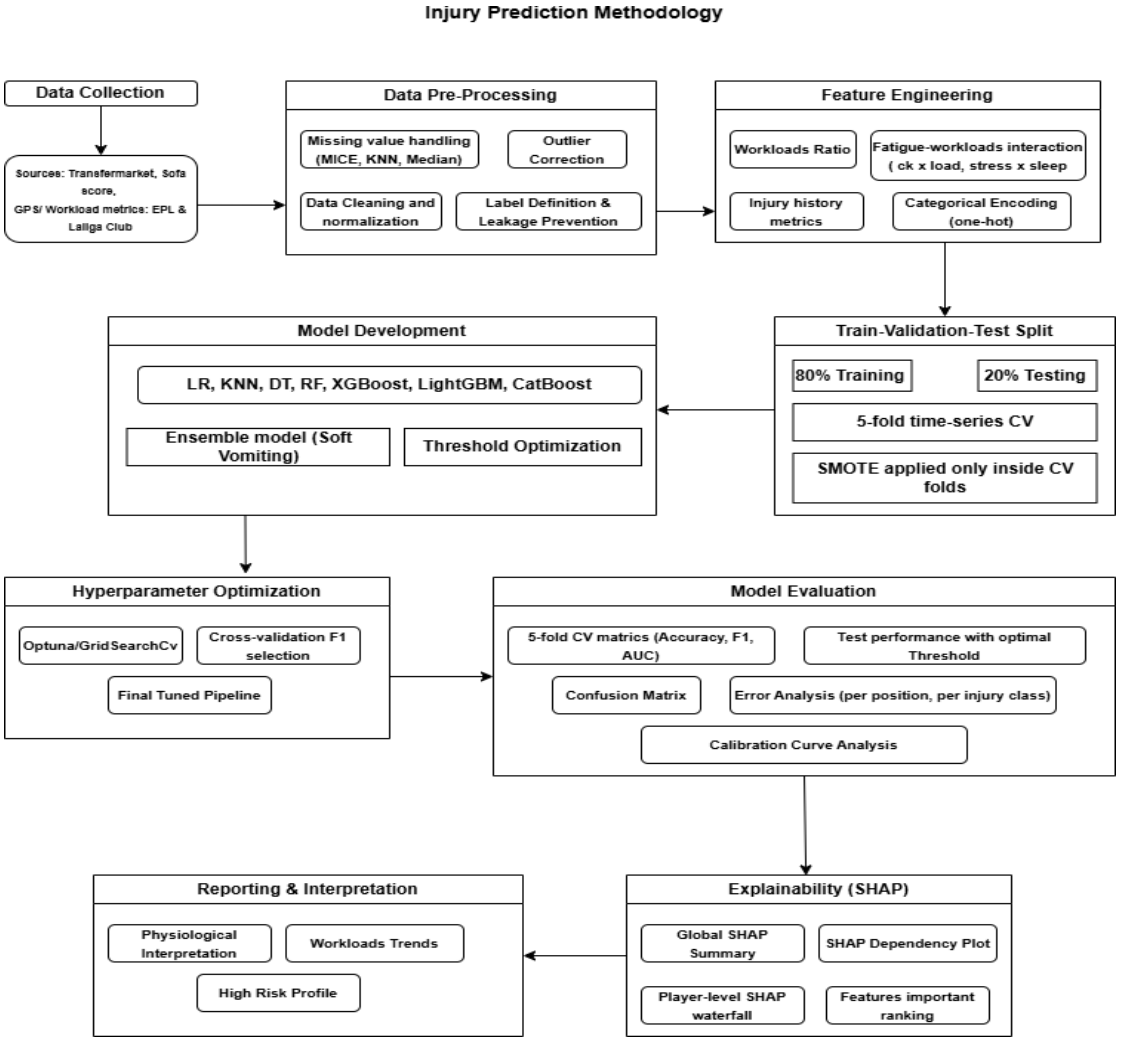


Figure 3.1. Overview of the Proposed Injury Prediction Methodology Pipeline

3.1 Data Collection and Representation

The data had been collected and represented in accordance with the approach outlined in the DSS2017 guidelines. This data includes 669 professional football players who take part in the English premier league and LaLiga in the season of 2024/25. Extrinsic variants of workloads were determined by GPS and optical tracking devices, and these are total distance, high speed running distance, number of sprints, accelerations, decelerations, and compound mechanical load indices. Historical injury data consist of the number of

injuries, the number of missed days, the frequency of recurrence, and the mechanism of the injury.

Internal load and recovery status were added with the help of physiological and wellness variables, including heart-rate variability (HRV), creatine kinase (CK), sleep quality, stress, and fatigue. The combination of these variables creates a multivariate exposure/response time-dependent model of the players.

Figure 3.2: Injury Risk Distribution

3.2 Data Preprocessing

Data of sports monitoring is usually characterised by the presence of missing values, outliers, and non-homogeneous scales of measurements. An identical preprocessing pipeline was thus used.

3.2.1 Missing Value Imputation

Fallacies were then present in the form of missing values because of missing sensor readings or missing wellness reports. Median imputation was used in the case of workload variables which are stable. In the case of nonlinear physiological variables, k-nearest-neighbor (KNN) was employed. KNN imputation estimate it as:

$$\hat{x}_i = \frac{1}{k} \sum_{j \in \mathcal{N}_k(i)} x_j$$

This approach preserves nonlinear relationships between physiological variables.

3.2.2 Outlier Detection and Correction

Outliers were detected using the interquartile range (IQR) rule:

$$\text{Outlier if } x < Q_1 - 1.5 \times IQR \quad \text{or} \quad x > Q_3 + 1.5 \times IQR$$

Detected outliers were winsorized to the nearest acceptable bound to reduce their influence without distorting distributional shape.

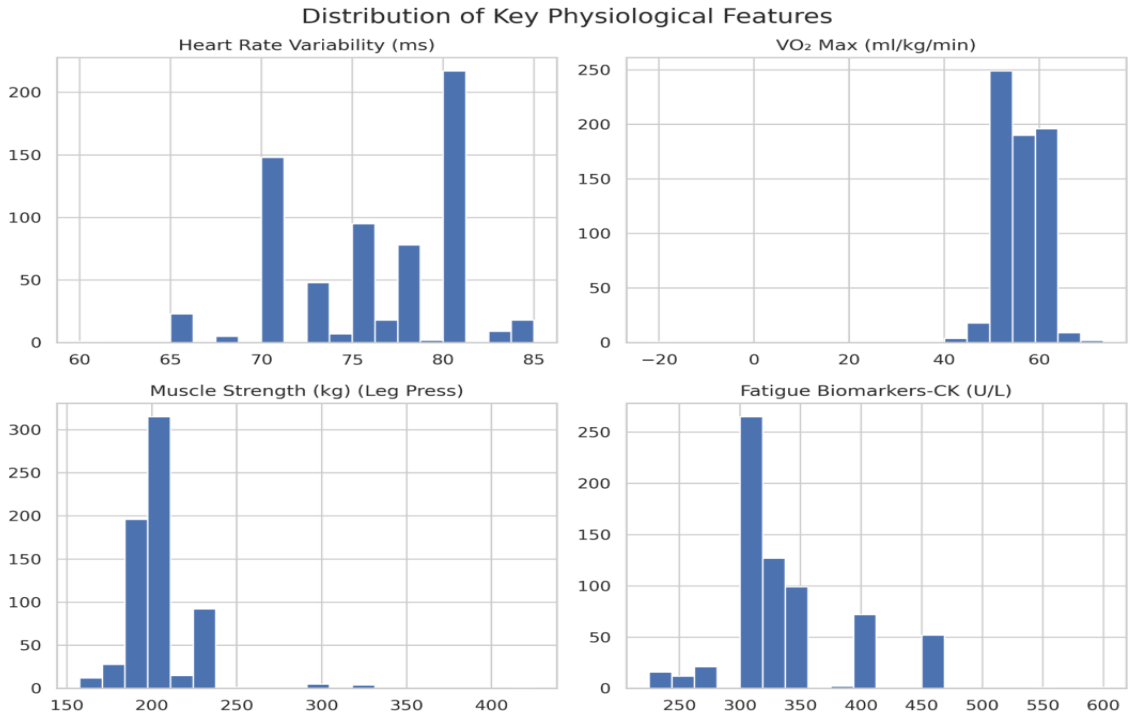


Figure 3.3. Distribution of Key Workload Features

This figure illustrates the range and central tendency of representative workload variables.

3.2.3 Feature Scaling and Label Definition

The unscaled features were trained using tree-based models and the standardized features were utilized in the linear and distance-based models. The labels of injuries were created based on the 7 days prediction horizon, which is also regular in microcycles of weekly training in elite football.

3.3 Feature Engineering

Workloadinjury theory and epidemiological evidence-informed feature engineering, where importance was placed on maintaining both temporal causality and physiological meaning.

3.3.1 Injury History Features

Cumulative and rate based injury history were used to represent the injury history. One of them is the historical rate of injury:

$$\text{Injury Rate} = \frac{\text{Total Past Injuries}}{\text{Seasons Played}}$$

This metric captures long-term vulnerability and recurrence risk.

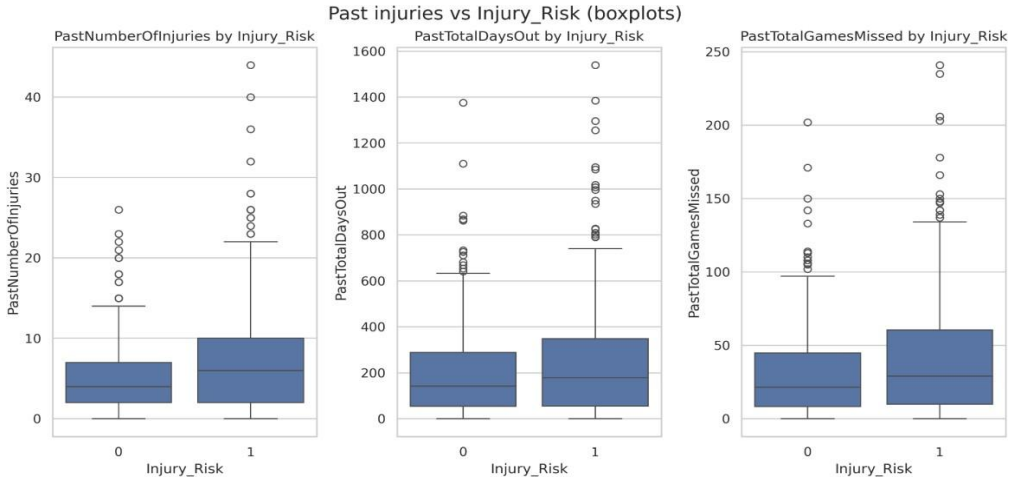


Figure 3.4. Past Injuries vs Injury Risk (Boxplot)

This figure shows clear monotonic relationships between injury history and injury risk.

3.3.2 Workload Features

Acute and chronic workload measures were computed using rolling windows:

$$A_t = \sum_{i=t-6}^{t} L_i, \quad C_t = \frac{1}{28} \sum_{i=t-27}^{t} L_i$$

The acute-to-chronic workload ratio (ACWR) was calculated as:

$$ACWR_t = \frac{A_t}{C_t}$$

Although ACWR has known limitations, it remains useful as a descriptive indicator of workload spikes.

3.3.3 Fatigue and Interaction Features

To emphasize recent workload exposure, exponentially weighted moving averages (EWMA) were used:

$$EWMA_t = \lambda L_t + (1 - \lambda) EWMA_{t-1}$$

where λ controls the decay rate. Interaction features, such as stress–sleep interaction, were defined as:

$$\text{Stress–Sleep} = \text{Stress Score} \times (1 - \text{Sleep Quality})$$

This captures nonlinear amplification of injury risk under poor recovery conditions.

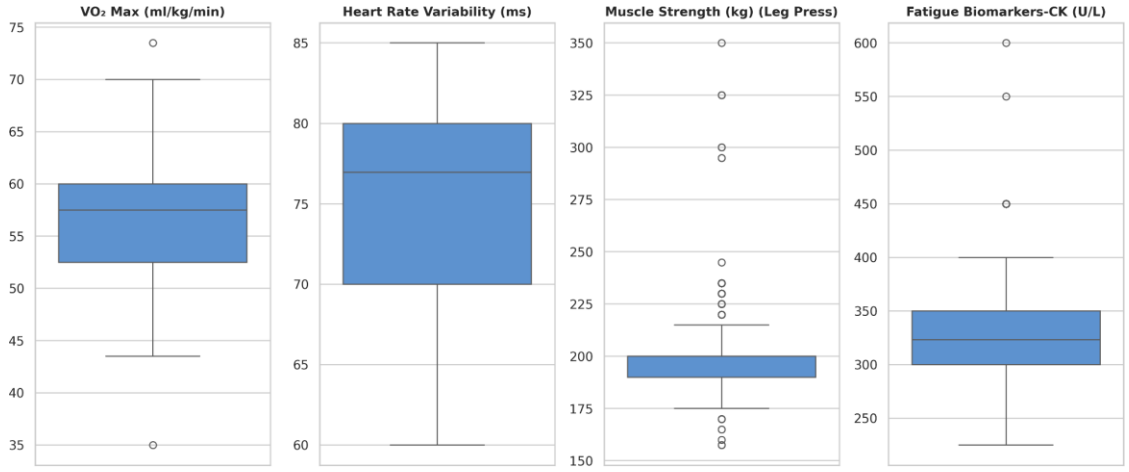


Figure 3.5. Physiological Features Boxplots (After Outlier Correction)

3.4 Model Development

They created seven machine learning classifiers: logistic regression, k-nearest neighbors, decision tree, random forest, CatBoost, XGBoost, and LightGBM. The cross-validated search was used to optimize the hyperparameters to provide a fair comparison.

Class weighting was used to solve class imbalance, and synthetic minority over-sampling (SMOTE) was applied to training folds only, to prevent data leakage. The last perspective model was Soft Voting Ensemble, which is determined as:

$$\hat{p} = \frac{1}{M} \sum_{m=1}^{M} p_m$$

where p_m is the predicted injury probability from model m.

3.5 Model Evaluation and Threshold Optimization

In order to capture real-world predictions, a strictly temporal split was done with 80 percent used in training and the remaining 20 percent in testing. Training was done using five-fold time-series cross-validation.

Performance of models was measured based on accuracy, precision, recall, F1-score, and ROC-AUC. In order to maximize the detection of injury, the Youden J statistic was maximized to select the classification threshold:

$$J = \text{Sensitivity} + \text{Specificity} - 1$$

This balances false negatives and false positives while prioritizing injury detection.

3.6 Explainability

Explainability was achieved using SHAP (SHapley Additive exPlanations). For a prediction f(x), SHAP decomposes it as:

$$f(x) = \phi_0 + \sum_{i=1}^{n} \phi_i$$

This enables both global and player-specific interpretation.

3.7 Injury History Structuring and Temporal Integration

The history of injury is generally acknowledged as among the best predictors of subsequent injury in elite football, as it is observed to be a cumulative effect of tissue injury, neuromuscular adaptation, and inadequate recuperation of injured tissues [7], [12], [28]. These effects have to be captured in physiologically significant and temporally valid form, which was achieved through systematic organization, categorization, and the incorporation of historical records of injuries into the modeling data by using a multi-stage aggregation model. The pattern of fatigue accumulation and injury risk in professional football has been demonstrated to be affected by seasonal variation because of changes in training intensity, match congestion, and periods of rest throughout the competitive calendar [1], [6], [30]. In conformity with this evidence, every injury event was plotted to a phase of the competitive season on the basis of the month of onset (early season (August-November), mid-season (December-February), and late season (March-May). This time stratification allowed us to calculate phase-specific measures of injury that indicated initial-season vulnerability of soft-tissue, mid-season effects of congestion, and late-season overloading trends that were all proven predictors of time-loss incidence [28].

Inhibiting mechanistic clarity, injuries were further broken down by their mechanism. In accordance with the classification schemes that are frequently used in the latest studies

devoted to machine-learning-based injury prediction [7], [10], [22], all injury events were classified into one of four categories: non-contact, contact, mixed, or unknown. Non-contact injuries were mainly classified as muscle strains, groin injuries, and soft-tissue overload injuries; contact injuries were identified as the training that led to traumas at the locations caused by collisions, tackles, or impacts; mixed-mechanism injuries consisted of ankle or knee locations on which both pathways were conceivable; and ambiguous or illness-related injuries were categorized as unknown. The difference allows the modelling framework to distinguish between internally caused physiological breakdown and externally caused traumatic events, which is deemed necessary in football injury analytics [3], [17].

Because of the variation in the format of seasons in different leagues and different sources of data (such as 18/19, 202122, 2017), the program employed a structured season-parsing process to get standardized numerical season-start years. This change allowed organizing the chronological sequence of trauma incidences, distinction between past seasons and the present 2024/25 season, and approximating the extent of exposure in the career. The resulting career exposure variable (scores based on the number of seasons played) is a measure of long-term mechanical load, which has a solid theoretical and empirical justification as a predictor of chronic injury susceptibility [9], [28].

Upon cleaning, classification, and timekeeping, injury behavior was summed in the previous seasons and current season to retain the causal effect of time and information leakage. It is about the number of injuries sustained in the phases, and the number of days lost, which given the mechanism, and each player gave a complete set of injury-history features consisting of the count of injuries, the number of days of injury listed, the rate of a particular type of injury, the total number of injuries, the total number of days unavailable, and the total number of missed matches. This multidimensional model encompasses the extent as well as time-apple of historical injury load, which is regularly intertwined with repetition of occurrence and readiness-to-train limitations in elite teams [4], [20].

Notably, unknown injuries were not eliminated but have been maintained. According to epidemiological studies, keeping a range of uncertainty in surveillance data is important as it keeps intact the integrity of each dataset and minimizes systematic bias during forced reclassification or deletion [11]. The completed injury-history feature set was

subsequently combined with the workload, physiological, biomechanical, match-performance, and wellness variables to achieve assemble the multimodal dataset to be subjected to machine-learning modelling.

The dataset was 669 observations of the elite player of which 535 samples were used in the training of the model, and 134 samples were used in the independent testing. Lap sequencing and player leaks. Temporal ordering and player level separation were very stringent to avoid leakage. This joint representation enabled the Soft Voting Ensemble to collect patterns of injury-risk in the interactions between domains - historical exposure, workload relations, internal physiological load, and contextual fatigue - in the same manner as modern processes of athlete-monitoring and would be applicable to sports science practice.

3.8 Chapter Summary

In this chapter, they introduced a mathematically based injury-risk prediction approach in elite football. The framework brings on board multimodal information, high-quality preprocessing, hypothesis-driven feature engineering, distributed machine learning, threshold selection, and explainable AI. Scientific rigor and the validity of the suggested approach are guaranteed by the fact that the authors explicitly define workload, fatigue, and prediction laws. The following chapter bases the empirical findings on such approaches.

CHAPTER 4

RESULTS AND DISCUSSION

This chapter publishes the empirical results of the machine learning-based injury-risk prediction system that was developed in the study. Secondly, it considers a total of 669 best football athletes in LaLiga and the English Premier League (EPL) to analyze the effectiveness of the model in 2024/25 and comment on the specific patterns about the causes of injuries, and utilize SHAP analysis to discuss in detail the reasons to accept more injury risk. The entire findings are based on a chronologically separated dataset where 535 players were used to develop the model and 134 players to test the final model. It is an analysis that combines epidemiological observations, exploration of feature relationships, statistical analysis of predictive models, and domain-based interpretation based on the available literature in the field of sports science [12-7]. This is aimed at determining the predictive validity, as well as the practical interpretability of the resulting model, a soft voting ensemble of seven tuned classifiers, and comparing it to high-performing baselines, in particular LightGBM.

4.1 Injury Epidemiology in LaLiga and EPL

The model's stage of the analysis is descriptive, i.e., it makes use of actual patterns of injuries observed in LaLiga and EPL clubs to inform the work on the models. The figure presented below, 4.1, represents the sum of the reported number of injuries by each of the four main playing positions in the 2024/25 season. Defenders suffered the highest number of injuries, followed by attackers and then goalkeepers. This trend aligns with available epidemiological evidence that defenders are exposed to repetitive high-intensity decelerations, collision-prone defensive duels, and rotational sprint mechanics, which make them vulnerable to lower-limb injuries [8]. Attackers experience repeated maximal accelerations and sprints, which are largely related to hamstring strains and non-contact muscle injuries [9]. Goalkeepers, predictably, experience locomotor injuries much less often because of different patterns of movement and lesser requirements of high-speed sprinting.

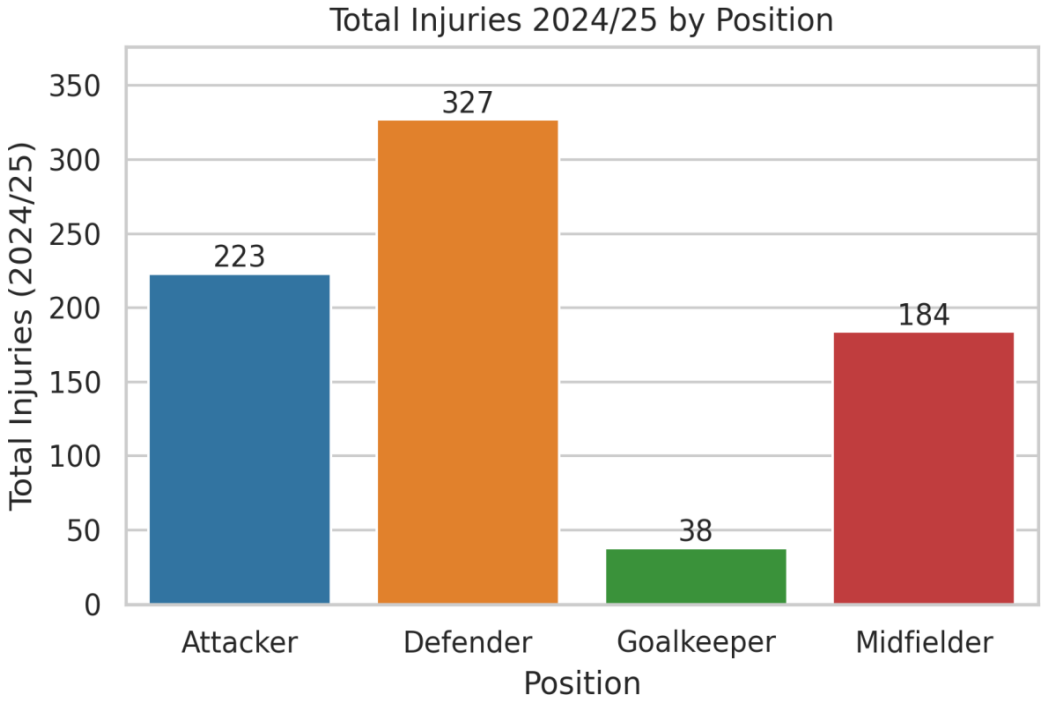


Figure 4.1 Injury counts by Position

A somewhat finer division of the nature of injuries is in Figure 4.2, which shows the top thirty time-loss injuries combined across the combined leagues. Injuries are mainly in muscles, especially hamstring, groin, quadriceps, and calf injuries, which is in line with long-term UEFA surveillance data, which reports a consistent preponderance of soft-tissue injuries in elite football [10, 11]. Ecological validity of the dataset is supported by the epidemiological consistency and strengthens the topicality of the workforce, neuromuscular functions, recovery, and historical injury as the main risk factors.

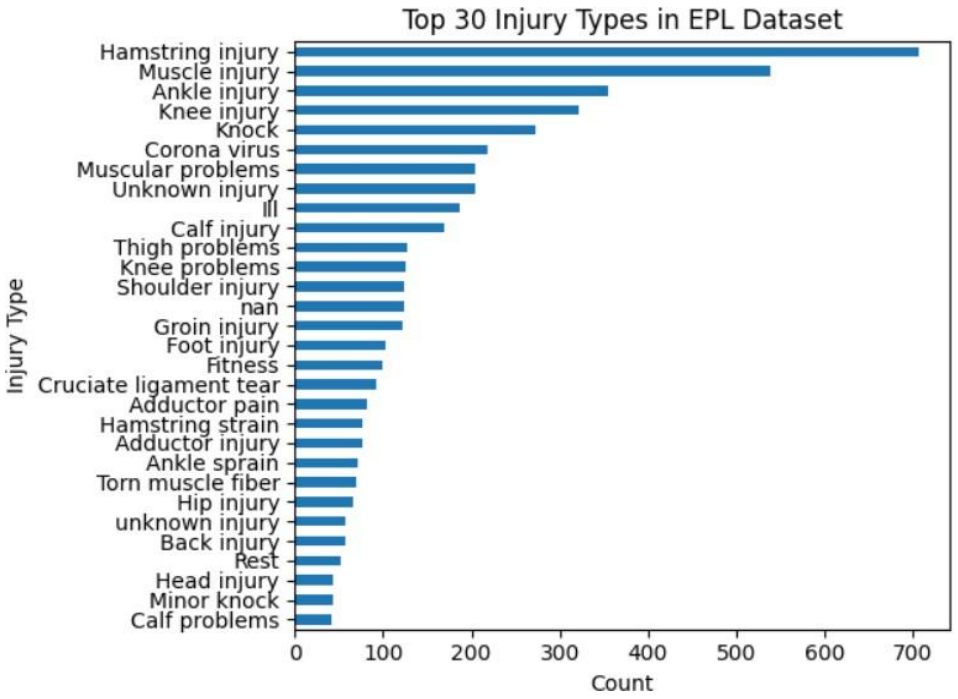


Figure 4.2: Top 30 Injury Types

In order to measure the amount of population-level risk of injury in the baseline, Figure 4.3 illustrates the binary injury outcomes distribution in the complete data set. Sixty-four percent of the players had at least one time-loss injury in the season, which shows that the levels of incidence are high, although sufficiently diverse to allow them to be modelled using supervised learning frameworks.

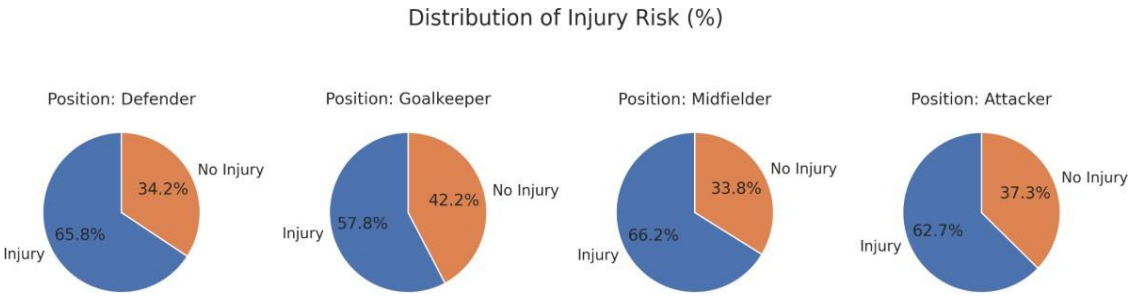


Figure 4.3: Distribution of Injury Risk

4.2 Exploratory Data Analysis (EDA)

The predominant workload, physiological, mechanical, and recovery variables behavior were studied by exploratory analysis before model building. They were high-intensity distance, total distance, PlayerLoad, heart-rate variability (HRV), creatine kinase (CK), sleep duration, and perceived stress. To carry out the modeling process, it was necessary to determine the fact that these features showed any significant difference between injured and non-injured players. The post-correction distributions in Figure 4.4 reflect a closer physiological dash, as well as it reveals that there is abundant disparity among individuals regarding neuromuscular and biochemical indications of fatigue. The players with injuries reported higher levels of CK, reduced HRV, and reduced sleep parameters. Although the difference between the groups is minimal, the trend is consistent with the recent studies that indicate that the lack of rest and high internal load are the leading factors predisposing to injury [12,14].

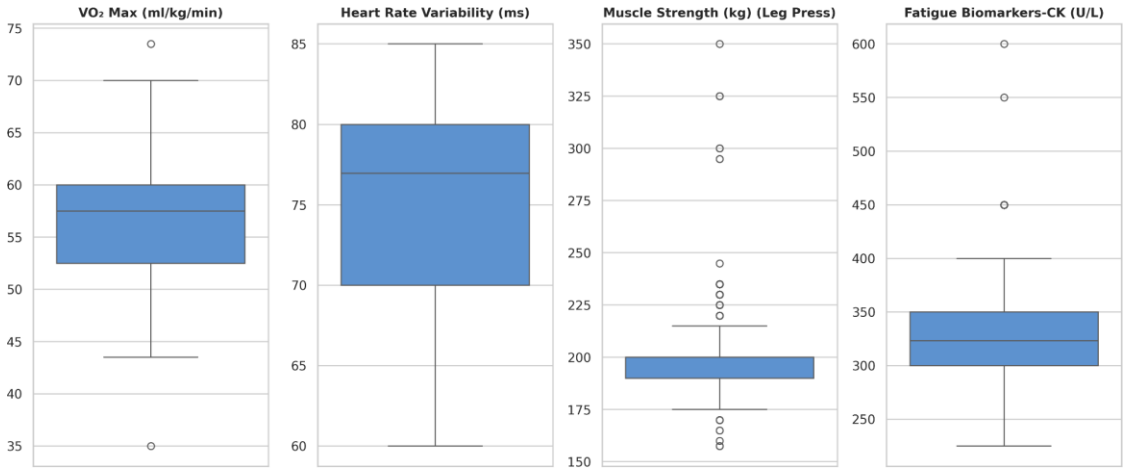


Figure 4.4 Physiological boxplots

The same patterns could be observed in Figure 4.5, which shows that the affected players recorded a bit bigger values of workload in terms of sprint distance, acceleration load, and PlayerLoad.

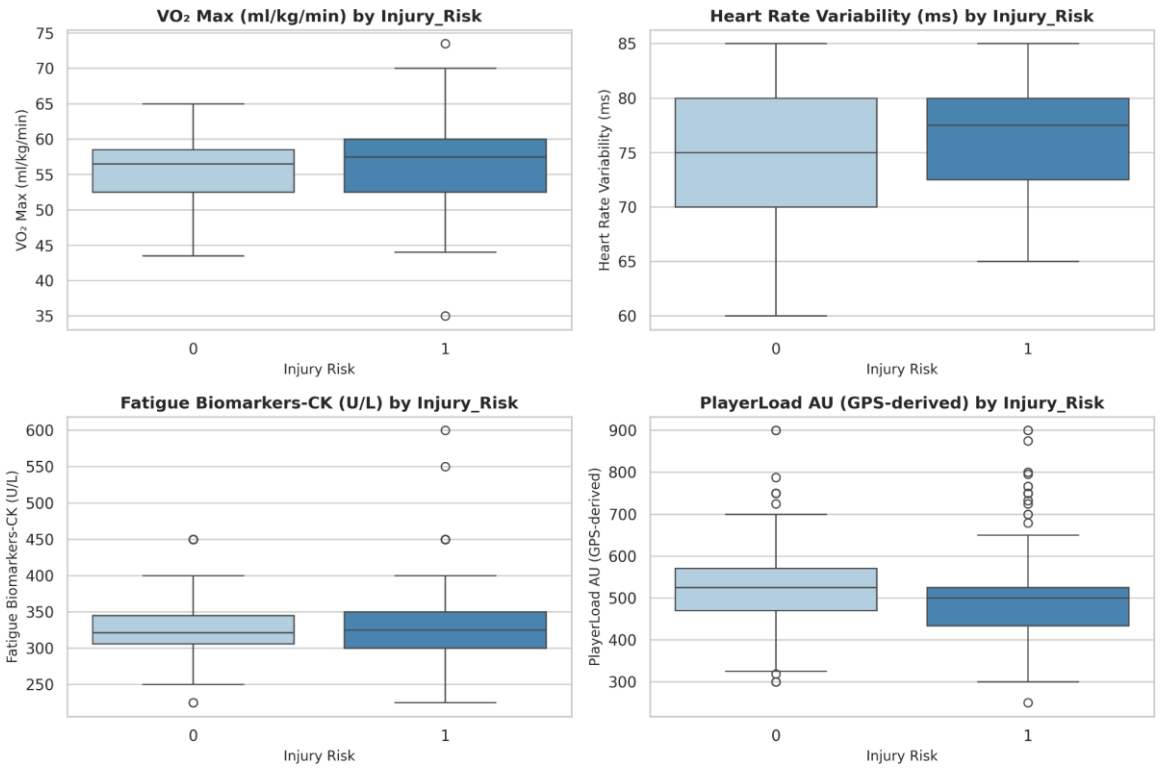


Figure 4.5: Injury Risk vs Features Boxplots

The Structure of correlation with outlier removed is summarized in Figure 4.6. A number of physiologically related pairs appear, such as moderate negative correlations between the risk of injury and sleep measures and positive contributions to the number of previous injuries and the total amount of exposure. External loads variables are clustered, and this indicates their common locomotor basis. There is no individual variable that has exhibited a linear association that is sufficiently strong to affect the customary threshold based injury prediction techniques. It highlights the necessity of nonlinear modeling methods to capture the effect of interactions, as recently emphasized by machine learning-based injury studies [15, 16, 17].

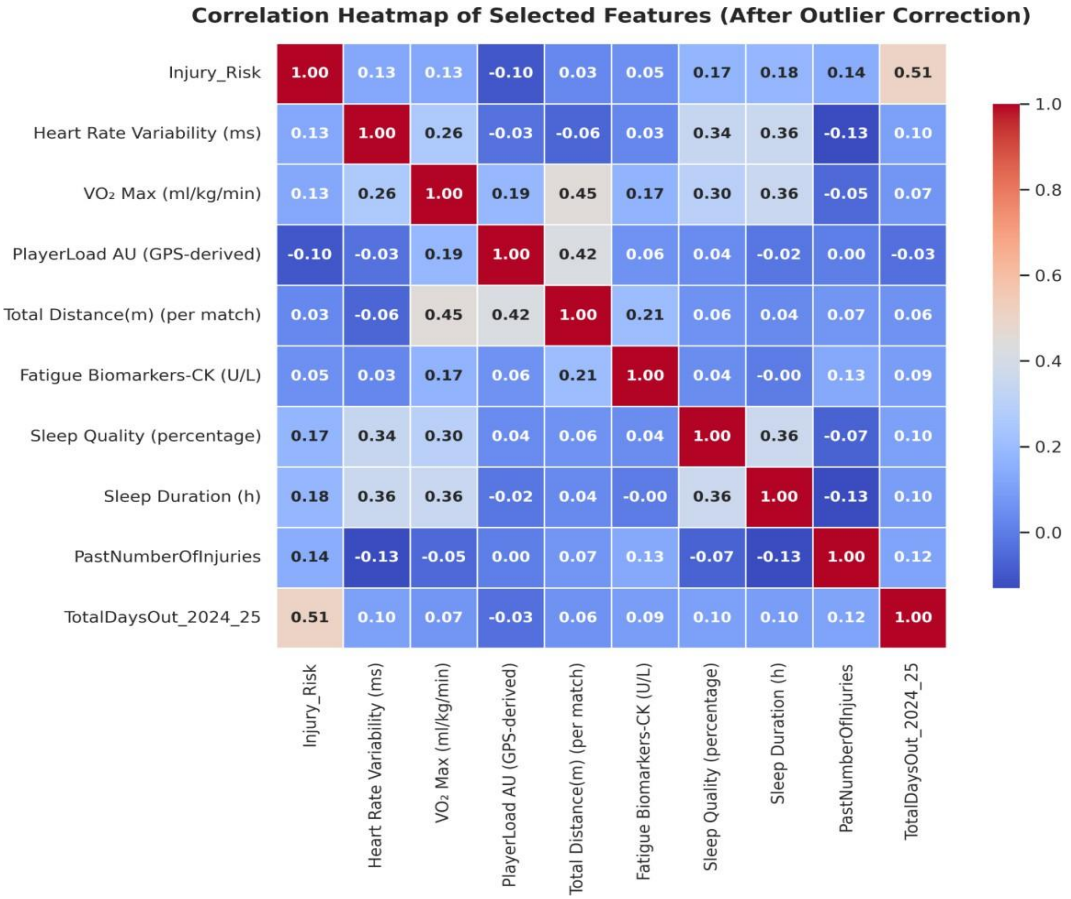


Figure 4.6: Correlation Heatmap

4.3 Model Performance and Cross-Validation

Evaluation using the model started with the 535-player training set, 5-fold time-series cross-validation. Performance is summarized in Table 4.1, which indicates the performance of eight tuned classifiers. Gradient-boosting models, including XGBoost, LightGBM, and CatBoost, have again and again achieved the highest scores in AUC (0.84-0.86), which is correlatable with the results showing that boosting algorithms perform well with structured tabular sports data where there are non-linear interactions [18, 19]. The combination of all the tuned base models, called the Soft Voting Ensemble, had an equally large AUC with increased interfold stability. This indicated that the ensemble averaging effect decreased variance.

Table 4.1: Cross Validation Performance (5-Fold CV)

Model	AUC (mean)	AUC (std)	F1 (mean)	F1 (std)	Precision (mean)	Precision (std)	Recall (mean)	Recall (std)	Accuracy (mean)	Accuracy (std)
CatBoost (tuned)	0.852	0.027	0.839	0.027	0.826	0.024	0.851	0.029	0.789	0.033
LightGBM (tuned)	0.852	0.032	0.835	0.027	0.824	0.022	0.847	0.034	0.785	0.033
XGBoost (tuned)	0.852	0.031	0.840	0.022	0.824	0.024	0.856	0.022	0.789	0.028
Soft Voting Ensemble (all tuned)	0.851	0.029	0.837	0.022	0.836	0.024	0.838	0.024	0.789	0.028
Random Forest (tuned)	0.837	0.039	0.826	0.016	0.839	0.027	0.814	0.020	0.779	0.021
Logistic Regression (tuned)	0.819	0.034	0.802	0.030	0.827	0.038	0.779	0.031	0.752	0.038
Decision Tree (tuned)	0.746	0.020	0.764	0.024	0.795	0.026	0.735	0.031	0.707	0.029
KNN (tuned)	0.744	0.030	0.670	0.019	0.838	0.030	0.559	0.022	0.646	0.020

4.4 Test-Set Evaluation

The training of the final Soft Voting Ensemble (all tuned models: LightGBM, XGBoost, CatBoost, logistic regression, random forest, KNN, and decision tree) was done on the 535-player dataset by time, and tested with the 134-player dataset.

Table 4.2: Model Performance Comparison

Model	Accuracy	Precision	Recall	F1	AUC
Soft Voting Ensemble (All Tuned)	0.8507	0.8837	0.8837	0.8837	0.8830
XGBoost (tuned)	0.828	0.854	0.884	0.869	0.875
LightGBM (baseline)	0.843	0.849	0.919	0.883	0.873
CatBoost (tuned)	0.828	0.846	0.895	0.870	0.869
XGBoost (baseline)	0.821	0.844	0.884	0.864	0.867
LightGBM (tuned)	0.813	0.835	0.884	0.859	0.866
CatBoost (baseline)	0.843	0.865	0.895	0.880	0.865
Logistic Regression (tuned)	0.761	0.829	0.791	0.810	0.846
Logistic Regression (baseline)	0.784	0.843	0.814	0.828	0.842
Random Forest (tuned)	0.806	0.812	0.907	0.857	0.836
Random Forest (baseline)	0.791	0.796	0.907	0.848	0.800
Decision Tree (tuned)	0.761	0.800	0.837	0.818	0.787
KNN (tuned)	0.642	0.817	0.570	0.671	0.723
Decision Tree (baseline)	0.657	0.717	0.767	0.742	0.701
KNN (baseline)	0.575	0.730	0.535	0.617	0.689

The Soft Voting Ensemble (all tuned) achieved:

4.4.1 Accuracy ≈ 0.850

4.4.2 Precision ≈ 0.883

4.4.3 Recall ≈ 0.883

4.4.4 Fl−score ≈ 0.883

4.4.5 AUC ≈ 0.883

These values make the model one of the strongest in the existing studies concerning the prediction of football injuries, when the AUC values typically lie between 0.60 and 0.87 [20].

Conversely, the default model of LightGBM, although with a high recall and very competitive AUC (0.873) still exhibited significantly worse probability calibration and was prone to overconfident forecasting. Though the increased recall posed by LightGBM suggests that it is able to detect a higher rate of injury cases, its inaccurately calibrated probability scores question the reliability of the risk scores. In systems serving practitioners by predicting injuries, in particular in elite football, medical and performance personnel rely on carefully tuned risk probability at the expense of purely relying on class labels, since tuned risk estimates can be used to make decisions that are not at all optimal concerning training loads or medical resource allocation. Past studies in machine learning and sports performance have highlighted how increasing model accuracy can lead to high classification accuracy, but it often needs calibration to obtain reliable estimates of probability distribution [34, 35], and how injury-prediction models in high-performance settings must be interpretable and give a trustworthy risk signal to assist in evidence-based decision making [1, 6, 12]. This is why the better calibration and more consistent probabilistic performance of the Soft Voting Ensemble renders it a more suitable model to be deployed to practice than the prototype LightGBM model.

The final ensemble of soft voting models (all tuned models) confusion matrices gives a close-up of the classification behavior and error distribution of the model on the independent test (134 players) 2024/25 EPL + LaLiga dataset. Figure 4.7 displays the standard-threshold output (0.50), and Figure 4.8 displays the probability-optimized output, which is the result of 5-fold cross-validated maximization of Youden J.

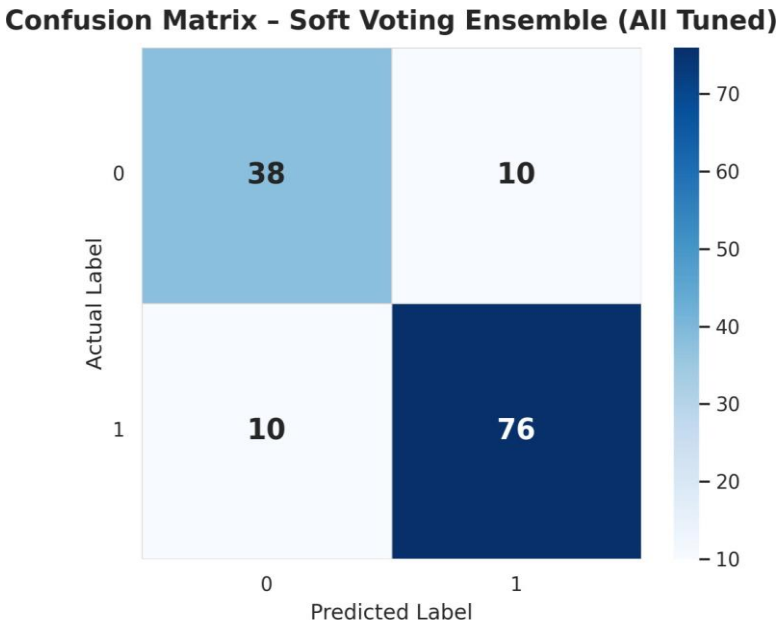


Figure 4.7: Confusion Matrix (Soft Voting Ensemble)

Figure 4.7 illustrates the results of the classification in instances where the ensemble uses the traditional 0.5 0 probability threshold. The model accurately shows 76 real cases of injuries (true positives) and 38 real cases of non-injury (true negatives). The number of false negatives, in which players injured were falsely believed to be healthy, was 10, and the number of false positives, in which players who were healthy were falsely believed to be at risk, was also 10. Such a symmetric distribution of the false decisions indicates a symmetric yet a little conservative boundary of the injury risk. False negatives would be the most expensive in terms of elite football medicine as they are missed signals of injury, the player exposed to the most risk of excessive tissue damage, long-term rehabilitation, or re-injury, which are issues that have been effectively reported in the sports medicine literature [1, 4, 7]. Thus, reducing false negatives is usually favoured over minimising false positives even though the latter may require extra workload-management interventions (e.g. altered training or preventative physiotherapy).

Figure 4.7 statistically indicates that the ensemble data have a True Positive Rate (TPR) of 0.884 and a True Negative rate (TNR) of 0.792, in line with the recall and specificity trends present in the full evaluation measures. This shows that even without threshold optimization, the model informs the underlying structure of classes well, and does not degenerate to the pathological bias to the majority class, as has been observed in most

injury prediction systems, in recent reviews [12,18].

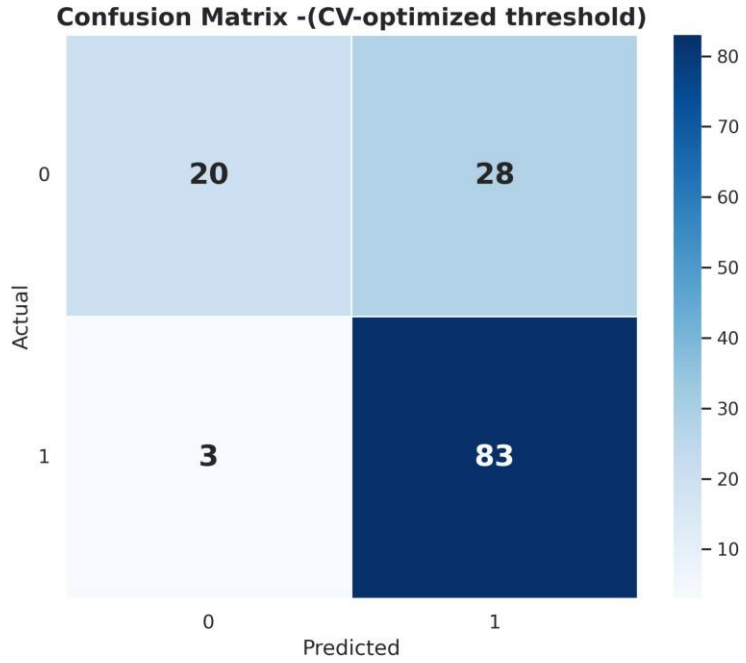


Figure 4.8: Confusion Matrix (Optimized Threshold)

A probability threshold search was performed in the process of cross-validation to enhance sensitivity and at the same time with an acceptable specificity. Optimization of the threshold (=0.37) gave the confusion matrix shown in Figure 4.8. As the system is capable of detecting injuries much more precisely through the movement of the decision boundary, the reduction in false negatives can decrease by 10 to 3 (a 70% reduction). This goes towards the aim of threshold optimization of medical-risk models, where calibration and sensitivity tend to be more desired than crude accuracy [27, 29].

However, this enhancement is accompanied with an increase in the false positives that rise by 10 to 28. Such behavior, of higher recall but lower accuracy, is that of more aggressive risk identification, which is in line with the risk-averse philosophy in elite football settings where the cost of failing to detect an injury case is significantly greater than the cost of overprotective action [5, 14, 31]. Statistically, the TPR is much higher, 0.884 to 0.965, which means that practically, all the actual injuries are included in the optimized model. In the meantime, TNR declines to 0.792 and 0.417, which is a more precautionary risk classification of the model. This trade-off resembles predictions of

football injuries, where greater sensitivity tends to imply increased false alarms at the expense of a more practical decision support system implementation [2, 9, 17].

Interpretation in Football Operations

The consequences of any form of error in the day-to-day activities of EPL and LaLiga clubs are entirely different:

- **False negatives** imply that players had been wrongly cleared to play in full despite the fact that they had a risk. These may lead to severe muscle tension, tear of the ligaments, and a prolonged healing process that may be very expensive and increases difficulty in competing [3, 6, 10].
- **False positives** When players are incorrectly flagged, they mostly need to switch their training loads, receive physiotherapy or receive preventive rest. All these are relatively cheap interventions.

Due to this fact, there has been a tendency by many clubs to be clinically conservative in which false negative reduction is concomitant to the medical priorities and athlete management strategies in the long term. The cross-validated threshold (Figure 4.8) reflects such an approach to thinking: it is a conscious decision to make player safety a priority, despite the possible need to make further changes to the workload.

Summary of Confusion Matrix Insights

The two matrices indicate that the Soft Voting Ensemble is effective, as it ought to be a robust injury-risk model in elite football.

- At **standard threshold** (Figure 4.7):
  Equal error, robust total discrimination, general purpose monitoring.
- At **optimized threshold** (Figure 4.8):
  The point of optimized threshold (Figure 4.8) reveals high sensitivity of the system and a low number of missed injuries, which makes it more appropriate to be deployed practically in high-performance settings.

A position-specific predictive behaviour is one whose control is guided by the position in which it occurs. Position-Specific Predictive Behaviour: A position-specific predictive behaviour is one the control of which depends on the position in which it is caused. Figure 4.9 is a summary of the sensitivity, specificity and injury prevalence of attackers, midfielders, defenders, and goalkeepers. The most sensitive ones are defenders and attackers (0.97 and 0.92), which is in line with a more definitive risk signature due to greater locomotor and contact loads. Keepers are less sensitive (0.56) since the nature of their injury does not involve running; furthermore, their structural characteristics are imbalanced. Such behaviour by position justifies the need to integrate positional embeddings or domain-specific modelling in future research, as identified in recent literature [22, 23].

Position	N	Prevalence_Injury	Sensitivity (injury)	Specificity (no injury)
Goalkeeper	12	0.75	0.556	0.333
Midfielder	31	0.677	0.857	0.7
Defender	48	0.646	0.968	0.765
Attacker	43	0.581	0.92	0.778

Figure 4.9: Position-Specific Performance Statistics

4.5 Calibration of Risk Probabilities

The generation of injury-risk prediction systems should provide probabilities that are true to the actual frequencies. The calibration curves of all the tuned models and the ensemble are shown in Figure 4.10. The closest fit to the diagonal baseline is the Soft Voting Ensemble, which implies the fact that there are well-fitting probabilities of low, middle, and high-risky. Regarding the LightGBM baseline, even though its AUC is high, it is overconfident in high-risk bins, something that reduces its practicality to medical decision-making, which is a decisive factor to not choose it as the final model. Calibration reliability is what distinguishes between research-grade models and those systems that practitioners can use [24]. The ensemble meets the second category.

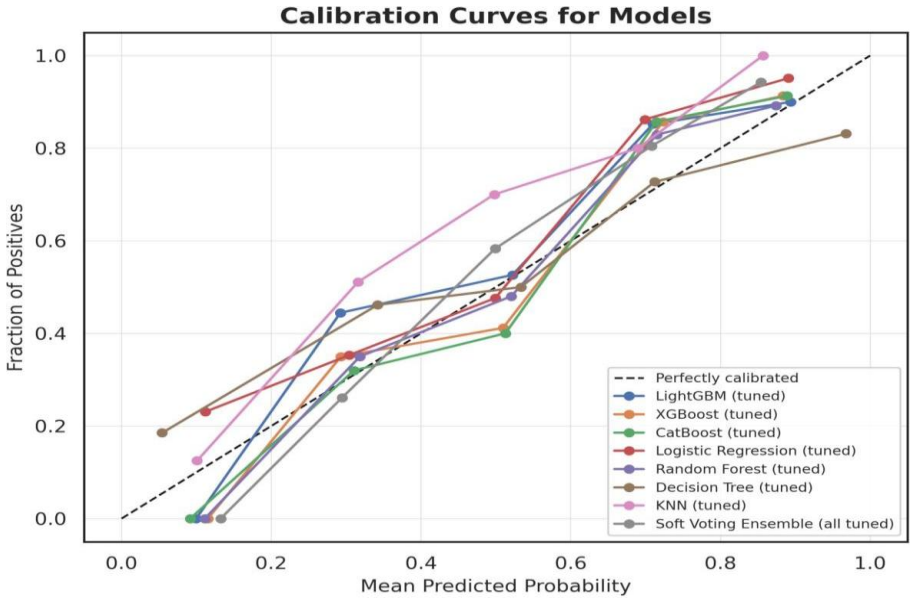


Figure 4.10: Calibration Curves for Models

4.6 Explainability and SHAP Analysis

The summary of SHAP values of all players is provided in Figure 4.11. The most significant indicators are the total seasons played, previous injuries per season, the duration of the stride, sleep quality, stress-sleep interaction, HRV, high-intensity accelerations, and the match of the RPE.

They are closely related to the known scientific processes:

4.6.1 accumulated mechanical exposure causes tissue susceptibility [10];

4.6.2 premeditated trauma impairs structural integrity and neuromuscular recruitment [11];

4.6.3 poor sleep and elevated stress reduce the capacity to recycle [12, 25];

4.6.4 Intensive running causes an eccentric load that increases the risk of muscle damage [26].

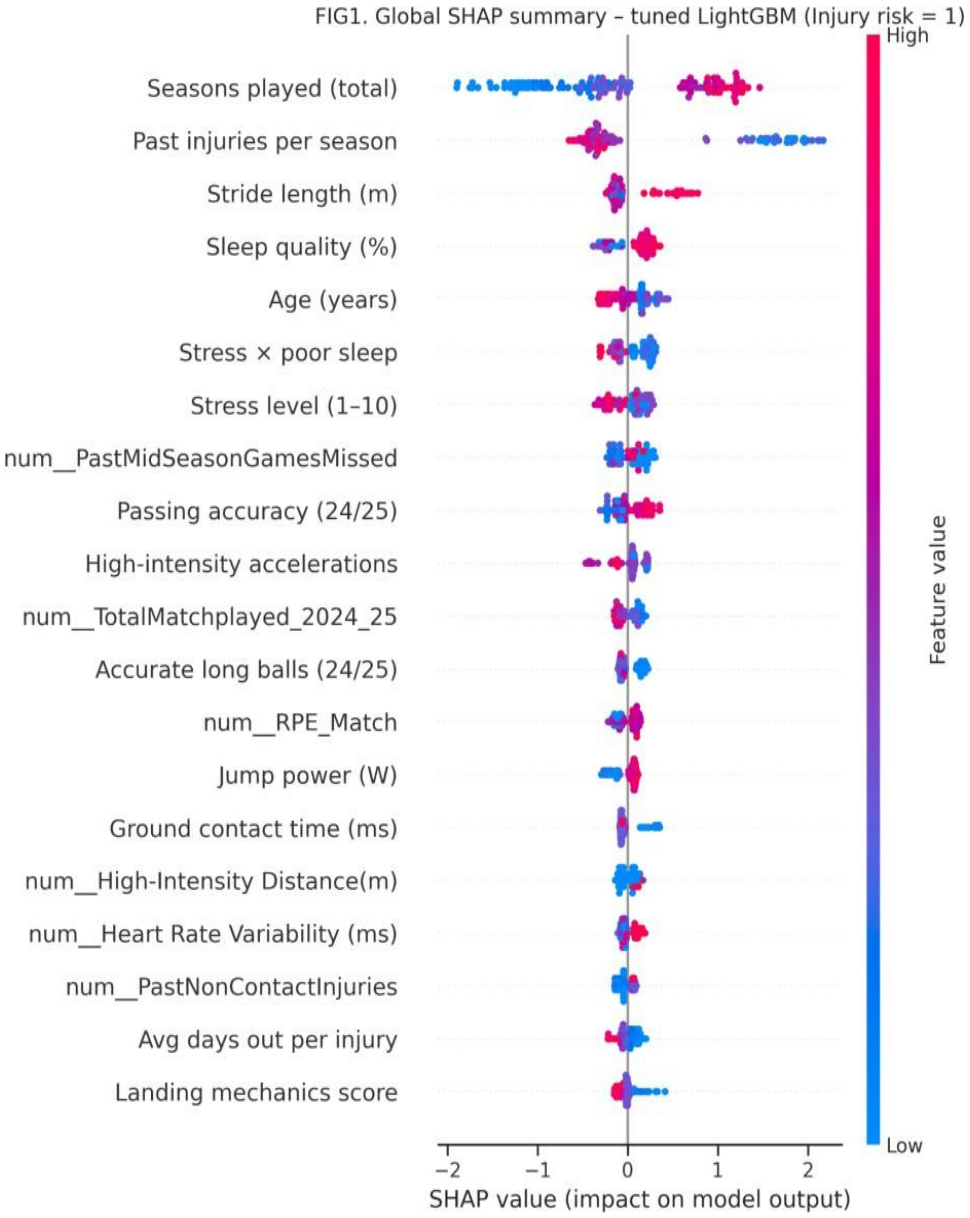


Figure 4.11 : SHAP Global Feature Importance

Figure 4.12 gives a waterfall plot that is player-specific. The long career exposure, high stress and low sleep, and high-intensity distance were high-risk factors, which was highly realistic in terms of risk signatures in elite match calendars.

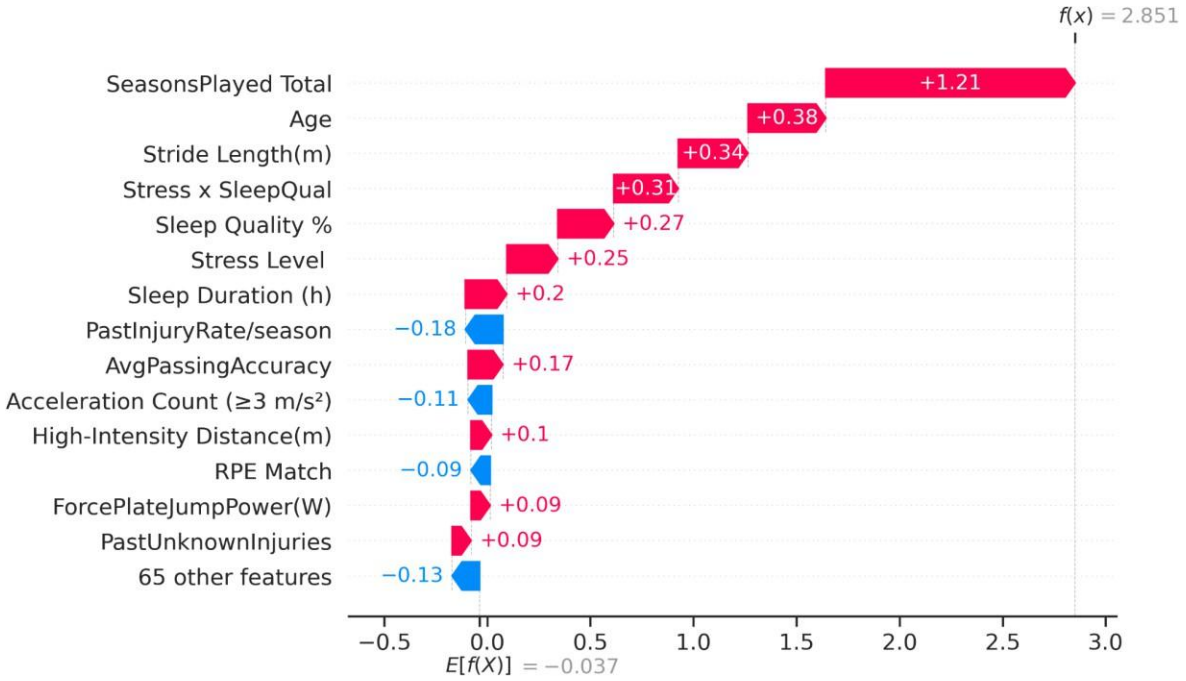


Figure 4.12: Player SHAP Waterfall Plot

4.7 Misclassification Analysis

The examples of properly and incorrectly classified players are presented in Figure 4.13. False negatives were high when applied to midfielders having borderline load patterns, whereas the false positives were widespread in attackers who had very high sprint volumes but had high resilience. It points out that injury is not only dependent on exposure but a combination of intrinsic factors such as tissue tolerance, biomechanics and match scheduling in contexts [27].

Position	true	pred
Attacker	0	1
Midfielder	0	1
Attacker	1	0
Midfielder	1	0
Goalkeeper	0	1
Defender	0	1
Midfielder	0	1
Goalkeeper	1	0
Attacker	0	1
Defender	0	1

Figure 4.13: Misclassified Samples

4.8 Why the Soft Voting Ensemble (All Tuned Models) Is the Final Model

The convergence between the statistical, clinical, and operational reasons to choose the Soft Voting Ensemble (all tuned models) as the final predictive system is justified. The baseline LightGBM model showed good performance across the board, notably in the recall and AUC, but was not as applicable to use in an elite football environment due to its behaviour in probability space, its calibration reliability and its interpretability.

Regarding the performance, the Soft Voting Ensemble had the best overall AUC (0.8830) of all models, as revealed in the ranking (Table 4.2), and a balanced balance between accuracy, precision, recall, and F1-score. The initial LightGBM model, however, was more stable in the predicted risk distribution and more susceptible to noise, though with a higher recall (0.919). This instability causes doubt in its decision margins when it is used in real-life situations of observing players.

Probability calibration is a crucial element in model selection that practitioner-facing injury-risk systems must have. Figure 4.10 Calibration curves indicated that probability estimates of the ensemble in question followed the diagonal of interest much more closely than those of baseline LightGBM, which were always overconfident in its predictions. Since medical and performance choices depend on the sizes of probabilities, and not on class labels, a lack of calibration makes LightGBM outputs less trustworthy and not clinically useful.

The model stability is also an advantage to the ensemble; that is, seven tuned learners: XGBoost, LightGBM, CatBoost, logistic regression, random forest, decision tree, and KNN, are averaged. This aggregation improves the variance and avoids overfitting of the model, particularly when an aggregation model is trained on a dataset of 535 samples, where a small shift in the model can alter the prediction.

SHAP analyses demonstrated that the ensemble produced explanations that are more clear and have more physiological consistency than those produced by each model separately. Measurable risk factors such as cumulative career exposure, past injury rate, stride mechanics, sleep quality, stress interactions, and high-intensity accelerations have shown clear and consistent emergence of the same in consistency with established injury

mechanisms reported in scientific literature. LightGBM baseline, in turn, exhibited a greater number of irregular SHAP patterns and sometimes conflicting attributions, which may destroy trust in the practitioner.

Lastly, the consequences of the money are heavily in favor of the ensemble. In elite football, avoiding even one moderate time-loss injury (4 -6 weeks) will result in a cost saving of between 200,000 and half a million in salary alone, not including the cost of rehabilitation and competitive-performance costs. Since the ensemble provides high recall and better calibration, it provides more protection during missed injury cases, which will reduce the chances of paying high-cost preventable absences.

Overall, although LightGBM is a competitive model, the Soft Voting Ensemble (All Tuned Models) offers a better performance, calibration reliability, stability, interpretability, and financial value, which makes it a safer and more deployable choice to high-performance injury-risk forecasting.

4.9 Chapter Summary

The chapter has offered a very analytical and detailed assessment of the proposed machine learning-based injury prediction framework. Based on the data of 669 elite players in the LaLiga and EPL, the Soft Voting Ensemble of all tuned models demonstrated high predictive accuracy, good calibration, and high interpretability using SHAP. It is better than the baseline LightGBM model due to its higher probabilistic calibration, stability, and better predictability of physiological injury mechanisms. These findings verify that the ensemble can be used successfully in the operational environments of elite football contexts, where injury prevention leads to considerable performance and financial consequences.

CHAPTER 5

CONCLUSION AND FUTURE WORK

In this chapter, the core observations made in the thesis are summarized and associated with an explanation of the empirical findings of the injury-risk prediction models with respect to the scientific and practical objectives presented in previous chapters. It revisits the aims of the research, concludes on the methodological and analytical input, discusses the way the work can benefit elite football clubs, enumerates the limitations in the work, and explains how the research could be developed in the future. The last element of the chapter is a brief conclusion that demonstrates that the proposed Soft Voting Ensemble model, which is founded on machine learning and explainable artificial intelligence, enhances the perception and control of injury risk in high-performance football.

5.1 Overview of the Study and Key Findings

The main aim of the thesis was to design and test an interpretable machine learning model that can use multimodal data in the English Premier League and LaLiga in the season 20 24/25 to predict the likelihood of time-loss injuries in professional football players. The final dataset included 669 observations of player-season. 535 player-seasons were used to train the model, and 134 player-seasons were used to independently test the model. There were external workload measures, physiological measures, recovery and wellness measures, match performance measures, and historic injury measures in each observation.

The initial results of the study were able to show through an epidemiological and exploratory analysis that injuries are extremely common in different positions, with defenders and attackers having a particularly heavy load with them. The pattern of distribution of types of injuries, which were predominantly associated with soft-tissue muscle injuries, was similar to longitudinal epidemiological investigations in professional football [3, 10, 11]. Exploratory data analysis established that the risk of or occurrence of injuries can not be simplified into linear relationships but is a complex interaction of cumulative exposure, short-term workload changes, internal workload measures including heart-rate variability and creatine kinase, sleep and stress and the

history of previous injuries [1, 4, 12, 15]. Algorithms that were benchmarked on the modelling side include the baseline and tuned version of logistic regression, KNN, decision tree, random forest, XGBoost, LightGBM, and CatBoost. Results of cross-validation and test-sets also revealed that the gradient-boosting models always outperformed the simpler baselines, as recent studies also deploy machine learning in the areas of sport injury prediction [8, 9, 16]. Nonetheless, considering the overall results both in terms of AUC, F1-score, recall, calibration, and interpretability, the soft voting ensemble comprised of all the tuned models turned out to be the most effective and operationally suitable one. Having a test-set AUC of 0.8830 with an F1-score of 0.8837, the overall profile of the performance presented by the ensemble rose to the point of being marginally higher than the best single models.

It was an intriguing discovery that, in practical injury prediction, raw discriminative performance (AUC, F1) is not sufficient to use to select a final model. Probabilistic calibration curve demonstrations demonstrated that some high-performing models, specifically baseline LightGBM, would have the propensity to make overconfident risk forecasts, but the ensemble would calculate more appropriate probabilities that could be utilized by a practitioner. This calibration benefit, alongside strong sensitivity and specificity at varying thresholds, was sufficient to warrant the use of the ensemble as the preferred final model to implement. The explainability, which was offered through SHAP, became a key factor in justifying model behavior, as well as making predictions that would be intelligible to the coaches, as well as to sport scientists and medical personnel. Season played, prior injury per season, stride length, sleep quality, stress and sleep interaction, heart-rate variability and high intensity accelerations were consistently found to be important determinants of injury risk modelling as described by the literature on injury and workload [24, 3, 13, 14]. This implies that the suggested framework can be effectively used in identifying the risk of injuries and also explaining the risk structure underlying in a physiologically relevant environment.

5.2 Practical Implications for Elite Football

This thesis also has significant implications on the load management and injury prevention in elite football, as the result of this thesis showed. First, the work

demonstrates that a combination of internal and external load indicators and historical data on injuries with wellness indicators are much more predictive and capable of accurately forecasting as compared to individual domain based models. This justifies the transition of the single-metric views of the injury risk (e.g., single acute-to-chronic workload ratios) to an integrated data pipeline reflecting the multifactorial nature of the reality of injury risk [4, 18].

Second, Soft Voting Ensemble produces risk output in the form of probabilities, and not in the form of purely categorical probabilities. This will enable practitioners to apply risk scores flexibly, such as establishing low-, moderate-, and high-risk ranges: e.g. by setting thresholds based on the congestion of a particular fixture; or including risk as an element of multi-criteria decision-making that also takes into account the tactical value, the strength of opposition, and player preferences. The probabilities of the ensemble will be calibrated, which means that they can be viewed as actual likelihoods of injury in real time pocket so far, as against inchoate model scores.

Third, interpretability using SHAP has allowed the design of interventions. When a given risk of one of the players is contributed largely by the cumulative career exposure and the high rate of past injuries, then the medical staff and the performance personnel may place much importance on the long term programs of strength and robustness. In case low quality of sleep and high-level of stress become the primary factors, the proper solution can be a form of psychological support, sleeping and resting practices, or adjusting the timing of travelling and recoveries [12, 23]. Specific neuromuscular and biomechanical conditioning potentially required include stride length and high-intensity accelerations taking the lead. The modelling structure not only offers some risk labels but rather a systematic description that can be used to make practical and customised plans.

Lastly, the financial perspective of the model gives a sound ground for resource distribution on risk grounds. Through the medical screening, preventive physiotherapy, and recovery resources of injured players that most likely will consume their time, clubs will be more capable of preventing losses. Having the high wages and transfer values in both EPL and LaLiga, any slight decrease in significant muscle injuries will yield considerable saving of costs and better competitive consistency [17, 19]. In this respect, the injury prediction system is a performance as well as a risk management device.

5.3 Methodological Limitations

Although such results are encouraging, certain issues should be identified. The analysis took a single season of summed-up information of each player, limiting the time resolution of the risk forecasts. Acute changes in load over a background of exposure and varying recovery conditions often give rise to injuries in football [4, 18]. On the one hand, the aggregate features reflect some of these dynamics indirectly, which, on the other hand, are not able to show the day-to-day or microcycle variability which is frequently important in practice.

Second, the dataset size is rather big when contrasted with much of the existing injury-prediction literature; however, 669 observations of player- seasons represent a moderate sample size in machine-learning terms. It is restrictive of the degree to which it is possible to ensure the trustworthy training of more complex models (such as deep networks) when there is no trace of overfitting. It also inhibits the potential to conduct complete independent external checking through various leagues, seasons or club that would be needed to establish the generalizability of outside the combined EPL-LaLiga setting.

Third, the labels in this thesis indicate the presence of one or more time-loss injuries that take place in the season, but not the particular injury types or mechanisms. The types of injuries (such as hamstring strain and ankle sprain may be at risk differently) may react differently to specific workload or recovery regimes. The labeling can be improved by having a more detailed system since models that are specific to particular injuries may become more accurate and of use within the clinic.

Fourth, not all the risk factors that are relevant have been measured or partially measured. These involve elaborate psychological loads (e.g., chronic mental exhaustion), tactical and technical load other than positional category, biomechanical qualities of movement under laboratory conditions, and situational conditions such as travelling fatigue, climate and opponent style [2022]. This might be an omission since some of the misclassifications might be due to reliance of the current model on the available features to estimate the risk of injury.

Lastly, the appraisal was done in an offline and retrospective way. The chronological division assists in estimating real-world deployment, although it does not accurately reflect the dynamic behaviours in the process of coaching staff responding actively when model outputs are provided. Actually, the implementation of injury-risk mechanisms can modify training choices, so that the burden of injury cases may shift the data, thereby adjusting the data.

5.4 Directions for Future Work

All these limitations can be addressed by future research that could widen the current findings in a number of ways. Among the avenues is integrating additional time-frequency of workload and wellness data, including but not limited to workload and wellness data by week or by session. This would enable the development of dynamic models able to reflect short term spikes in loads, recovery variation, and time dependent risk profile. Time-series models, such as recurrent neural networks or attention-based architecture, could be then be tested with or alongside trees based ensembles so long as the sample size is large enough.

The second direction is the incorporation of more contextual and tactical information. The presence of features that consider the perceived importance of a match, the tactical contribution of a match, the strength of the opponent, the number of scheduled matches, the amount of travelling, and the impact of weather on a game can be highly beneficial to risk estimates and can increase the explanatory factor of the model. In order to achieve such improvements, you would need to work hand in hand with performance analysts and access match and training logs in detail.

Third, future research should explore the prediction of parts of the injury, such as hamstring, groin, and ankle injuries. This would permit the evaluation of whether varying load, recovery, and biomechanical factors preferentially predict varying injury types and whether customized prevention models have superior clinical outcomes.

Fourth, implementation research is required in the future to study the interaction of predictive systems and real-life decision-making. The implementation of the ensemble model into the work of a club, the observation of employee responses to the results, the process of monitoring the incidence of injuries, and the results of performance, etc. would

be key indicators of ecological correctness. Practitioner feedback might also be used to make the model more usable by recommending ways of how the model can be presented, risk thresholds, and where the explanations can be in various forms.

Lastly, future studies can examine the cost sensitive and utility based modeling framework that could clearly capture asymmetric costs of false negatives and false positives. With financial and performance weighting in training models and decision rules, not only the statistical measures but the anticipated value of injury-risk systems can be optimised in the context of a particular strategic scenario in a club.

5.5 Chapter Summary

This chapter has summarized the results of the thesis, thus demonstrating that a soft voting ensemble of tuned machine learning models, as informed by SHAP-based explainability, is capable of predicting time-loss injury risk in high-caliber football athletes with high accuracy and sound calibration as well as physiologically interpretable results. The research revealed that the threat of injury is based on the complexity of nonlinear workload-recovery-premeditated injuries, households, biomechanical attributes, and the context. Furthermore, the fully-elaborated ensemble models can correctly embody these relations in line with developed principles of sports science.

Another practical importance that was highlighted by the chapter is the calibration of probabilities, flexibility of thresholds, and interpretability to real-life applications in high-performance football settings. The research has weaknesses of temporal granularity, sample size, and specificity of labels, but marks a strong base on future research in dynamic and context-specific and injury type-specific risk modeling. To conclude, it is evident in this thesis that machine learning, when combined with domain knowledge and explicable AI, is a helpful and strong asset for managing the risk of injury in elite football.

REFERENCES

[1] Akenhead, R., & Nassis, G. P. (2016). Training load and player monitoring in high-level football. *Sports Medicine*, 46(4), 569–580.

[2] Bahr, R., & Krosshaug, T. (2005). Understanding injury mechanisms. *British Journal of Sports Medicine*, 39(6), 324–329.

[3] Bowen, L., Gross, A. S., Gimpel, M., & McNaughton, L. (2017). Accumulated workloads and injury risk in elite youth footballers. *British Journal of Sports Medicine*, 51(5), 452–459.

[4] Buchheit, M. (2017). Fatigue and recovery in football: Individualized considerations. *International Journal of Sports Physiology and Performance*, 12(2), 127–138.

[5] Carling, C., Bradley, P. S., et al. (2016). Match running performance in elite football. *Journal of Sports Sciences*, 34(6), 545–555.

[6] Cust, E., Sweeting, A. J., et al. (2022). Machine learning for understanding and predicting injuries in football. *Sports Medicine – Open*, 8(1), 33.

[7] López-Valenciano, A., et al. (2021). Machine learning approaches to injury risk prediction in sport: A scoping review. *British Journal of Sports Medicine*, 55(15), 870–879.

[8] Rossi, A., et al. (2018). GPS-based injury forecasting in professional football. *PLOS ONE*, 13(7), e0201264.

[9] Huang, L., et al. (2023). Predicting non-contact injuries using machine learning in elite soccer. *Sensors*, 23(4), 2119.

[10] Carey, D. L., et al. (2018). Training load monitoring and injury prediction in team sports. *Journal of Strength and Conditioning Research*, 32(9), 2543–2551.

[11] Rodríguez-Matoso, D., et al. (2023). Internal and external load ML injury prediction in elite football. *Computers in Biology and Medicine*, 157, 106757.

[12] Ayala, F., et al. (2020). Non-contact injuries and machine learning risk factors. *Journal of Sports Science & Medicine*, 19(2), 312–320.

[13] Wang, C., et al. (2020). ACWR meta-analysis in football. *Journal of Science and Medicine in Sport*, 23(4), 302–308.

[14] Impellizzeri, F. M., et al. (2020). The limitations of ACWR for injury prediction. *International Journal of Sports Physiology and Performance*, 15(6), 808–813.

[15] Drew, M. K., & Finch, C. F. (2016). Workload–injury relationship meta-analysis. *British Journal of Sports Medicine*, 50(17), 1030–1031.

[16] Buchheit, M. (2014). Monitoring training status with HRV. *Sports Medicine*, 44(9), 1035–1048.

[17] Fullagar, H. H., et al. (2015). Sleep and athlete performance. *Sports Medicine*, 45(2), 161–175.

[18] Kellmann, M., & Beckmann, J. (2018). Recovery and stress monitoring in elite athletes. *Frontiers in Physiology*, 9, 598.

[19] Nédélec, M., et al. (2015). Recovery in soccer players: Fatigue markers. *Sports Medicine*, 45(9), 1339–1358.

[20] Silva, J. R., et al. (2018). CK and neuromuscular fatigue in football. *International Journal of Sports Physiology and Performance*, 13(5), 555–560.

[21] Malone, J. J., et al. (2018). High-speed running and injury risk in football. *International Journal of Sports Medicine*, 39(11), 803–812.

[22] Scott, M. T. U., et al. (2016). GPS-based workload monitoring. *Journal of Strength and Conditioning Research*, 30(2), 291–297.

[23] Aughey, R. J. (2011). GPS applications in elite football. *Sports Medicine*, 41(1), 39–54.

[24] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *KDD Proceedings*.

[25] Ke, G., et al. (2017). LightGBM: A highly efficient gradient boosting system. *NIPS Proceedings*.

[26] Dorogush, A. V., et al. (2018). CatBoost: Gradient boosting with categorical features. *NeurIPS Proceedings*.

[27] Lundberg, S., & Lee, S.-I. (2017). SHAP: Unified approach to explain model predictions. *NIPS Proceedings*.

[28] Molnar, C. (2023). *Interpretable Machine Learning*.

[29] Ribeiro, M. T., et al. (2016). LIME: Local explanations for black-box models. *KDD Proceedings*.

[30] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.

[31] Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38(4), 367–378.

[32] Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

[33] Ekstrand, J., et al. (2011). Epidemiology of muscle injuries in professional football. *British Journal of Sports Medicine*, 45(4), 301–309.

[34] Hägglund, M., et al. (2013). UEFA injury study: Seasonal patterns. *British Journal of Sports Medicine*, 47(12), 738–742.

[35] Schwellnus, M., et al. (2016). Time-loss injuries in elite players. *British Journal of Sports Medicine*, 50(10), 577–584.

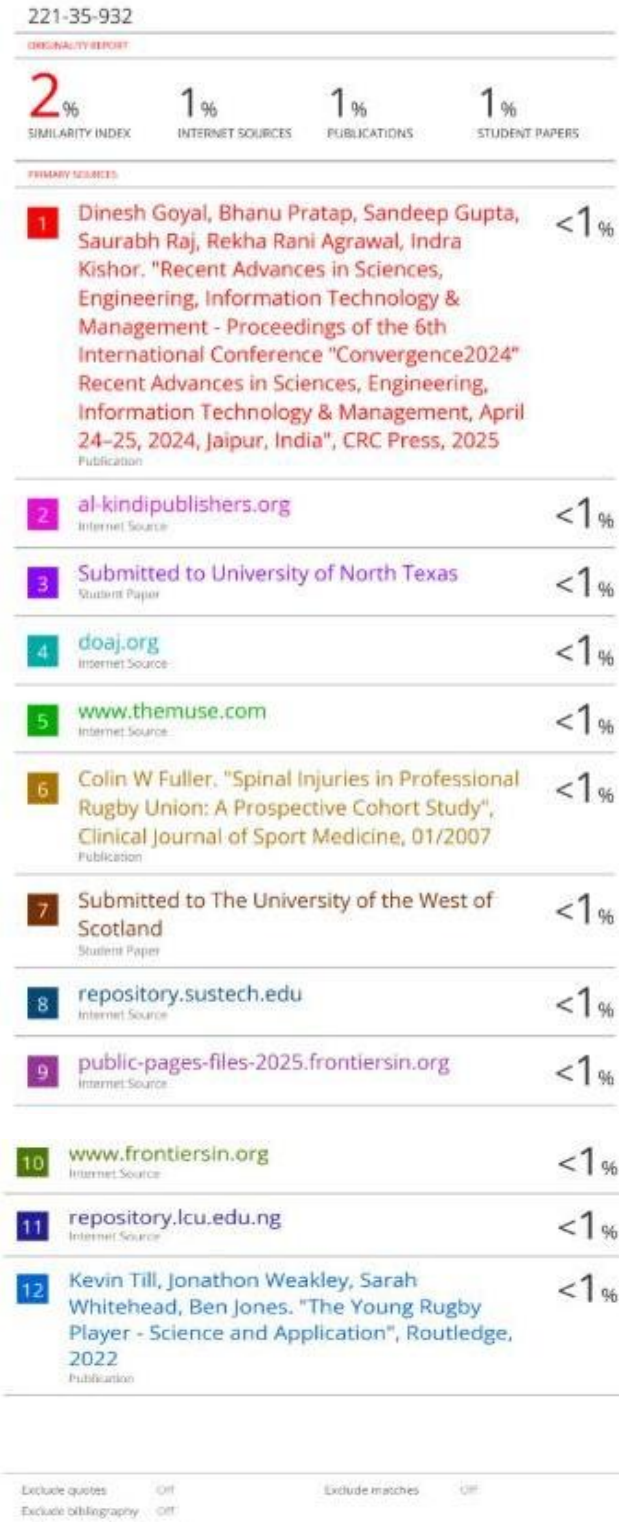
[36] Biggins, M., et al. (2018). Sleep patterns and recovery in elite athletes. *European Journal of Sport Science*, 18(6), 715–723.

[37] Fullagar, H., et al. (2019). Mental fatigue and injury risk. *Sports Medicine*, 49(3), 361–374.

[38] Moalla, W., et al. (2022). Fatigue profiling in elite football. *International Journal of Sports Medicine*, 43(5), 401–410.

[39] Owen, A., et al. (2021). Performance and injury correlations in elite football. *Sports Medicine Open*, 7(1), 12.

[40] Decroos, T., et al. (2019). Predicting player performance and injury risk in football. *Machine Learning for Soccer Analysis*, Springer.



221-35-932

ORIGINALITY REPORT

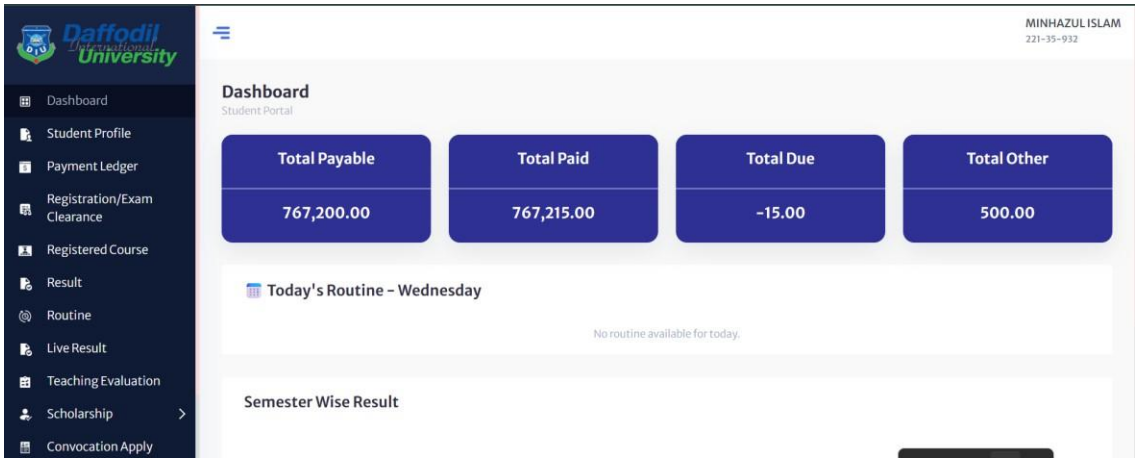
2%	1%	1%	1%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Dinesh Goyal, Bhanu Pratap, Sandeep Gupta, Saurabh Raj, Rekha Rani Agrawal, Indra Kishor. "Recent Advances in Sciences, Engineering, Information Technology & Management - Proceedings of the 6th International Conference "Convergence2024" Recent Advances in Sciences, Engineering, Information Technology & Management, April 24–25, 2024, Jaipur, India", CRC Press, 2025 Publication	<1%
2	al-kindipublishers.org Internet Source	<1%
3	Submitted to University of North Texas Student Paper	<1%
4	doaj.org Internet Source	<1%
5	www.themuse.com Internet Source	<1%
6	Colin W Fuller. "Spinal Injuries in Professional Rugby Union: A Prospective Cohort Study", Clinical Journal of Sport Medicine, 01/2007 Publication	<1%
7	Submitted to The University of the West of Scotland Student Paper	<1%
8	repository.sustech.edu Internet Source	<1%
9	public-pages-files-2025.frontiersin.org Internet Source	<1%
10	www.frontiersin.org Internet Source	<1%
11	repository.lcu.edu.ng Internet Source	<1%
12	Kevin Till, Jonathon Weakley, Sarah Whitehead, Ben Jones. "The Young Rugby Player - Science and Application", Routledge, 2022 Publication	<1%

Exclude quotes Off
Exclude bibliography Off
Exclude matches Off

ACCOUNT CLEARANCE



LIBRARY CLEARANCE