Multi-Season Injury Prediction in Football Using Machine Learning and Open Data

Abdalrahman Siddig Ali Bashir

Final International University
June 2026
Girne, North Cyprus

Abdalrahman Siddig Ali Bashir
Multi-Season Injury Prediction in Football Using Machine Learning and Open Data
MS, 2026

Multi-Season Injury Prediction in Football Using Machine Learning and Open Data

by

Abdalrahman Siddig Ali Bashir

A thesis submitted to the Institute of Graduate Studies
in partial fulfillment of the requirements for the Degree of
Master of Science
in
Software Engineering

Final International University
June 2026
Girne, North Cyprus

FINAL INTERNATIONAL UNIVERSITY
INSTITUTE OF GRADUATE STUDIES

APPROVAL

Title: Multi-Season Injury Prediction in Football Using Machine Learning and Open Data

Approval of the Examining Committee:

Asst. Prof. Dr. Ibrahim Adeshola (Chair)

_____ _____

Asst. Prof. Dr. Payam Zarbakhsh

_____ _____

Asst. Prof. Dr. Ghazaal Sheikhi (Supervisor)

_____ _____

Approval of the Institute of Graduate Studies:

_____

Asst. Prof. Dr. Seniha Besim
Director

Abdalrahman Siddig Ali Bashir

abdalrahman@bashir.edu.tr

ORCID ID: 0009-0006-8864-3298

©Abdalrahman Siddig Ali Bashir 2026

“But over every possessor of knowledge is one [more] knowing.”

— Qur’an 12:76 (Sūrah Yūsuf)

Dedicated to the pursuit of knowledge,
to perseverance in the face of hardship,
and to those who remain steadfast when circumstances
would make surrender seem easier than endurance.

To every soul that taught the meaning of resilience,
every community that preserved hope amidst adversity,
and every individual who chose dignity, patience,
and faith over despair.

To my parents,

whose sacrifices were the first lessons I ever learned,
long before any classroom taught me anything worth knowing.

Everything I have built stands on ground they prepared,
often at a cost I am still learning to understand.
This work is theirs before it is mine.

And to my wife,

my constant through every distance and difficulty,
whose patience made this work possible.

ETHICAL DECLARATION

I, ABDALRAHMAN SIDDIG ALI BASHIR, hereby, declare that I am the sole author of this thesis and it is my original work. I declare that I have followed ethical standards in collecting and analyzing the data and accurately reported the findings in this thesis. I have also properly credited and cited all the sources included in this work.

Abdalrahman Siddig Ali Bashir

ACKNOWLEDGMENTS

First and foremost, I want to thank my supervisor, **Assist. Prof. Dr. Ghazaal Sheikhi**, whose guidance, patience, and continued support made this research what it is, and whose feedback not only shaped the direction of this thesis but pushed me to grow as a researcher in ways I did not expect when I started.

I am also grateful to the faculty members and staff of the Department of Software Engineering at **Final International University**, who created an environment where learning and research could genuinely thrive throughout my master's studies.

This work would not have been possible without the open data and tools that the research community shares so freely, and I want to acknowledge **Understat** and **Transfermarkt** in particular, since their publicly available football statistics formed the entire empirical foundation of this thesis, alongside the developers behind PyTorch, scikit-learn, Optuna, SHAP, and the many other open-source projects that make rigorous scientific research accessible and reproducible for everyone. In a very real sense, this thesis is a product of open science, and I hope it gives something back to that community.

Finally, and most importantly, none of this would have happened without my parents, whose belief in me, sacrifice, and encouragement carried me further than they probably realise, and without my wife, whose patience and steadfast support kept me going through every difficult stretch of this journey. To them, and to everyone who stood beside me along the way, I am deeply grateful.

ABSTRACT

Every season, thousands of footballers at the elite level get injured, and the cost of those injuries is enormous, with a review covering the five seasons up to 2024/25 recording 22,596 injuries across Europe's top five men's leagues at a combined financial toll of more than £2.97 billion in lost income and wasted resources. Researchers and clubs have tried to get ahead of this problem using wearable monitoring systems and machine learning, yet most published studies run into the same two obstacles, namely that they draw on small single-club datasets of only 25 to 40 players over one or two seasons which makes it hard to generalise findings, and that the wearable hardware they depend on is expensive with data that is proprietary and out of reach for most clubs and researchers. This thesis takes a different route by relying entirely on publicly available match statistics, pulling data from Understat and Transfermarkt to build a dataset of 47,477 player-match records covering 1,017 English Premier League players across five seasons from 2020/21 to 2024/25, and since injury risk does not come from a single match but builds up gradually over time each player's appearances are treated as a time series, which allows the model to pick up on workload accumulation, fatigue patterns, and the toll of repeated match exposure. A Long Short-Term Memory (LSTM) network with an attention mechanism handles this sequential task and is compared against Decision Tree and Random Forest baselines, with the prediction framed as a binary classification problem where, given a player's recent run of matches up to a specific game-week, the model determines whether they will be injured within the next 30 days. Across the five seasons the models achieve AUC values ranging from 0.44 to 0.75, with the LSTM with attention reaching the best AUC in three of the five seasons while the Random Forest achieves the single highest result at 0.746, and a three-tier explainability framework combining SHAP, attention weights, and Integrated Gradients consistently points to cumulative workload and recent performance load as the dominant injury- risk signals across all model families.

Keywords: Football injury prediction, LSTM, sequential data, explainable machine learning

ÖZ

Futbol sakatlıkları, elit seviyedeki performans için en önemli risklerden birini temsil etmekte olup, finansal maliyeti milyarlarca sterline ulaşmaktadır. 2024/25 sezonuna kadar olan beş sezonu kapsayan bir inceleme, Avrupa'nın en iyi beş erkek liginde 22.596 sakatlık kaydetmiş ve bunun oyunculara ödenen maaşlar açısından toplamda 2,97 milyar sterlinden fazla gelir ve takım kaynağı kaybına yol açtığını ortaya koymuştur. Bu sakatlıkları tahmin etmek için giyilebilir izleme sistemleri ve makine öğrenmesi kullanılmıştır; ancak yayınlanan çoğu çalışma, gerçek hayatta kullanımı zorlaştıran iki ortak sınırlamaya sahiptir: bir veya iki sezon boyunca tek bir kulüpten yalnızca 25 ila 40 oyuncudan oluşan küçük veri kümelerine dayanırlar ki bu da genellenebilirliği kısıtlar; ayrıca bağımlı oldukları giyilebilir donanım pahalıdır ve verileri özel mülkiyettir, bu da onu çoğu kulüp ve araştırmacı için erişilemez kılar. Bu tez, giyilebilir cihaz verileri yerine kamuya açık maç istatistiklerini kullanan alternatif bir yaklaşım önermektedir. Veriler Understat ve Transfermarkt'tan toplanmış ve beş sezon boyunca (2020/21–2024/25) 1.017 İngiltere Premier Ligi oyuncusunu kapsayan 47.477 oyuncu-maç kaydından oluşan bir veri kümesinde birleştirilmiştir. Sakatlık riski tek bir maçtan kaynaklanmak yerine zamanla biriktiği için, her oyuncunun maçları bir zaman serisi olarak ele alınmış ve böylece modelin iş yükü birikimini, yorgunluk örüntülerini ve tekrarlanan maç maruziyetini yakalaması sağlanmıştır. Bu sıralı görev için dikkat (attention) mekanizmasına sahip bir Uzun Kısa-Süreli Bellek (LSTM) ağı kullanılmış ve Karar Ağacı ile Rastgele Orman temel modelleriyle karşılaştırılmıştır. Görev, ikili sınıflandırma olarak tanımlanmıştır: belirli bir oyun haftasına kadar olan bir maç dizisi verildiğinde, model oyuncunun sonraki 30 gün içinde sakatlanıp sakatlanmayacağını tahmin eder. Beş sezon genelinde modeller 0,44 ile 0,75 arasında AUC değerleri elde etmiştir; dikkat mekanizmalı LSTM beş sezonun üçünde en iyi AUC değerine ulaşırken, Rastgele Orman tek başına en yüksek AUC değerini (0,746) elde etmiştir. Üç katmanlı bir açıklanabilirlik çerçevesi (SHAP, dikkat ağırlıkları ve Integrated Gradients), tüm model aileleri genelinde birikimli iş yükü ve yakın dönem performans yükünü baskın sakatlık riski göstergeleri olarak tutarlı biçimde belirlemiştir.

Anahtar Kelimeler: Futbol sakatlanma tahmini, LSTM, sıralı veri, açıklanabilir makine öğrenimi

TABLE OF CONTENTS

ETHICAL DECLARATION vi
ACKNOWLEDGMENTS vii
ABSTRACT viii
ÖZ ix
LIST OF TABLES xii
LIST OF FIGURES xiii
LIST OF APPENDICES xiv
LIST OF ABBREVIATIONS xv
CHAPTER 1 1
INTRODUCTION 1
CHAPTER 2 5
LITERATURE REVIEW 5
2.1 Injury Burden and Workload Monitoring 5
2.1.1 Epidemiology and Consequences 5
2.1.2 Wearable Workload Monitoring and Its Barriers 7
2.2 Machine Learning for Injury Prediction 8
2.2.1 Tree-Based Methods and GPS-Derived Data 8
2.2.2 Methodological Critiques and Public-Data Approaches 8
2.3 Sequential Models and Explainability 10
2.3.1 Recurrent Models and Attention 10
2.3.2 Explainability Methods 11
CHAPTER 3 13
METHODOLOGY 13
3.1 Methodology 13
3.1.1 Research Design Overview 13
3.1.2 Data Curation 13
3.1.3 Feature Engineering 17
3.1.4 Sequence Construction, Splitting, and Scaling 18
3.1.5 Sequence Modelling 20
3.1.6 Attention Mechanism 22
3.1.7 Integrated Gradients 24
3.1.8 SHAP Explainability 25
3.2 Experimental Setup 26
3.2.1 Baseline Models 27
3.2.2 Hyper-parameter Tuning 28
3.2.3 Evaluation Metrics 31
CHAPTER 4 34

RESULTS 34
4.1 Results and Discussion 34
4.1.1 Overall Model Comparison 34
4.1.2 Feature Importance and Cross-Model SHAP Analysis 37
4.1.3 Attention Weight Analysis 41
4.1.4 Integrated Gradients Analysis 42
4.2 Practical Implications 44
4.2.1 What a Club Could Do with This System 44
4.2.2 Interpreting AUC in the Context of Actionability 44
4.2.3 Limitations of the Performance Envelope 45
CHAPTER 5 47
CONCLUSION AND IMPLICATIONS 47
5.1 Summary of Findings 47
5.2 Contributions 49
5.3 Limitations 50
5.4 Future Work 53
REFERENCES 56
APPENDICES 61

LIST OF TABLES

Table 1 Summary of The Unified Player-match Dataset. 15
Table 2 Summary of The 23 Engineered Features Used for Injury Prediction. . . 17
Table 3 LSTM + Attention Hyperparameter Search Space for Bayesian Optimization Via Optuna. 30
Table 4 Decision Tree Hyperparameter Search Space for Bayesian Optimization Via Optuna. 30
Table 5 Random Forest Hyperparameter Search Space for Bayesian Optimization via Optuna. 30
Table 6 Best LSTM + Attention Hyperparameters Found by Optuna for Each EPL Season. 31
Table 7 Best Decision Tree Hyperparameters Found by Optuna for Each EPL season. 31
Table 8 Best Random Forest Hyperparameters Found by Optuna for Each EPL Season. 31
Table 9 Season-level Baseline (Leave-one-season-out) Versus Match-level Results From Test Set Performance Of All Models Across Five EPL Seasons, Averaged Across The Five EPL Seasons. 35
Table 10 Test Set Performance Of All Models Across Five EPL Seasons. Best Value Per Season is Shown in **bold**. 36
Table 11 Cross-model SHAP Feature Importance Across the Three Model Families, Averaged Over the Five EPL Seasons (2020/21–2024/25). . . 39

LIST OF FIGURES

Figure 1 End-to-end methodology pipeline, from public data acquisition through feature engineering, sequence construction, model training and chronological evaluation, to multi-method explainability. Section references beneath each stage indicate where it is described in the text. . 13
Figure 2 Distribution of Injury Events Across Five Premier League Seasons (2020/21–2024/25). Total Injury Events Recorded In the Unified Dataset: 5,504. 16
Figure 3 Average Days Missed Per Injury (bars) and Overall Injury Rate Percentage (line) Per Season Across The Premier League (2020/21–2024/25). 16
Figure 4 LSTM With Attention Architecture For Football Injury Prediction. . . . 22
Figure 5 SHAP Feature Importance For The LSTM + Attention Model (Season 2023/24). `match_number`, `xG_roll5` And `cumulative_minutes` Are The Top-ranked Features. 38
Figure 6 Mean|SHAP| Of The Five Universally Important Features, Averaged Over Seasons 2020/21–2024/25, For The Decision Tree, Random Forest, and LSTM+Attention Models. The same Five Features Dominate Across All Three Model Families Despite the TreeExplainer/KernelExplainer Asymmetry. 41
Figure 7 Mean attention weights across match positions for Season 2021/22 (lookback = 6). The model assigns 45.8% of its attention to the most recent match, showing a strong recency bias in injury prediction. 42
Figure 8 Integrated Gradients attribution heatmap for a single high-risk sequence (Season 2023/24, lookback = 10). 43

LIST OF APPENDICES

APPENDIX A Similarity Report 61
APPENDIX B Statement Of Artificial Intelligence Usage 68

LIST OF ABBREVIATIONS

ACWR	Acute-to-Chronic Workload Ratio
ANN	Artificial Neural Network
AUC	Area Under the Curve
DNN	Deep Neural Network
DT	Decision Tree
EI	Expected Improvement
EPL	English Premier League
FN	False Negative
FP	False Positive
FPR	False Positive Rate
GPS	Global Positioning System
GRU	Gated Recurrent Unit
IG	Integrated Gradients
IOC	International Olympic Committee
LSTM	Long Short-Term Memory
MDI	Mean Decrease in Impurity
OOB	Out-of-Bag
RF	Random Forest
RNN	Recurrent Neural Network
ROC	Receiver Operating Characteristic
SHAP	SHapley Additive exPlanations
TN	True Negative
TP	True Positive
TPE	Tree-structured Parzen Estimator
xA	Expected Assists
xG	Expected Goals
XGBoost	Extreme Gradient Boosting

CHAPTER 1
INTRODUCTION

Football injuries represent one of the most significant risks to performance at the elite level, and their financial toll runs into the billions. A review of the five seasons up to 2024/25 recorded 22,596 injuries across Europe's top five men's leagues, at a combined salary cost of £2.97 billion in wages paid to sidelined players (Howden Sport and Entertainment, 2025). Premier League clubs accounted for nearly a quarter of these injuries, absorbing roughly £240 million per season in wages paid to injured players (Howden Sport and Entertainment, 2025). The most direct remedy is adequate recovery, yet physiological research indicates that players require up to 96 hours to recover fully from a competitive match, a window the modern calendar rarely allows (Den Hollander et al., 2024). The expansion of the Champions League and the Club World Cup means players now regularly compete in successive matches with fewer than five days of rest, a burden compounded by long-distance travel that further erodes limited recovery time (Den Hollander et al., 2024).

Some elite clubs deploy wearable technology to manage this risk. Systems from vendors such as Catapult and STATSports measure physical workload—distance covered, sprinting, high-speed running, and acceleration and deceleration—and derive indicators such as the acute-to-chronic workload ratio. When a player's workload rises sharply relative to their established baseline, injury risk increases, with some studies reporting a rise of up to 15% (Mahmood, 2025). Because clubs collect large volumes of such data, machine-learning models including Random Forest and XGBoost have been applied to predict injuries, with reported AUC values approaching 0.90, while tools such as SHAP help explain which workload factors drive each prediction (Leckey et al., 2025). Most published work in this area nonetheless carries limitations that hinder practical deployment: studies typically draw on small samples of 25 to 40 players at a single club over one or two seasons, so a model that performs well in that setting tends

to generalise poorly elsewhere (Martins et al., 2025). Cost compounds the problem, as wearable systems can exceed €2,400 per unit, placing them beyond the reach of smaller clubs and locking the underlying data away from most researchers (Tierney et al., 2026).

This study addresses these barriers by working entirely with publicly available match statistics that are free, accessible, and require no proprietary hardware, drawn from Understat and Transfermarkt. It is important to note that such statistics serve as *proxies* for physical workload rather than direct measurements: public match data record discrete on-ball events and minutes played, but omit the continuous locomotor load that GPS and inertial systems capture—total distance, high-speed-running distance, sprint counts, accelerations and decelerations, and metabolic-power estimates. The present approach therefore trades the physiological precision of wearable data for scale and accessibility, a trade-off examined throughout this thesis. Rather than treating each match in isolation, the study models each player's appearances as a sequence unfolding over time, making it possible to track how workload accumulates and how sustained exposure to a dense fixture schedule affects a player. This is precisely the problem the LSTM is built for: as a model designed for sequential data, it carries information forward from earlier matches and detects patterns that emerge only over longer stretches, so that a player enduring a punishing run of fixtures can be identified in advance in a way that static models cannot. The resulting dataset spans five seasons from 2020/21 to 2024/25 and comprises 47,477 player-match records pairing injury histories with match-performance statistics. The prediction task is framed as binary classification: given a player's recent sequence of matches up to a given game-week, the model determines whether they will be injured within the next 30 days.

Research Questions

To structure this investigation the thesis addresses the following four research questions where each one targets a distinct gap in the existing literature:

- **RQ1 (Public-data sufficiency):** Can freely available match statistics produce injury predictions that remain consistent across multiple seasons, and how does their predictive performance compare to the levels reported for GPS-based systems?

- **RQ2 (Temporal-modelling advantage):** The second asks whether treating a player's match history as an ordered sequence through an LSTM with attention makes a real difference in practice, or whether simpler tree- based models like the Decision Tree and Random Forest are good enough to compete.

- **RQ3 (Granularity value):** The third looks at whether it is more useful to predict injury risk on a match-by-match basis through a rolling 30-day window or to frame the whole thing at the season level, and which of the two actually serves the practical needs of in-season risk monitoring.

- **RQ4 (Explainability consistency):** Do independent explainability methods—SHAP across all model families, attention weights, and Integrated Gradients—converge on a consistent set of injury-risk features, and is that ordering stable across model classes?

This study makes four main contributions:

1. **A scalable public-data benchmark:** demonstrating that injury prediction is feasible using only publicly available match data, without wearable sensors or proprietary club datasets;

2. **A chronologically evaluated LSTM with attention:** applying an LSTM with an attention mechanism to capture temporal injury-risk patterns across a player's match history under leak-safe, per-player chronological splits;

3. **A principled match-level versus season-level comparison:** establishing the value of in-season temporal granularity over a season-level baseline; and

4. **A multi-method explainability framework:** identifying, through SHAP, attention weights, and Integrated Gradients applied across model families, the match statistics most predictive of injury.

CHAPTER 2
LITERATURE REVIEW

This chapter surveys the body of knowledge that contextualises and motivates the present work. The review is organised into three thematic areas: the epidemiology, economic burden, and workload basis of football injuries (Section 2.1); machine learning approaches to injury prediction and their reliance on wearable data (Section 2.2); and sequential models together with the explainability methods used to interpret them (Section 2.3).

2.1 Injury Burden and Workload Monitoring

2.1.1 Epidemiology and Consequences

Football is the world's most widely played sport, and because of the physical demands of the professional game, tight schedules and congested fixtures, musculoskeletal injuries are a common occurrence rather than a rare event. The most clear picture of this load has been produced by observation throughout Europe's top leagues. Ekstrand et al. (2011) conducted one of the foundational studies of its kind, reporting that muscle injuries alone account for approximately **37%** of all time-loss injuries in professional football and that each club can expect an average of 17 muscle injuries per season, with lower-limb injuries dominating the distribution, since the hamstrings, adductors, quadriceps, and calf muscles collectively represent the majority of cases across all positions. Beyond muscle injuries, systematic reviews have identified a broader epidemiological landscape that includes ankle sprains, knee ligament injuries, and bone stress reactions (Jones et al., 2019), and Jones et al. (2019) synthesised evidence from youth football specifically, finding that injury incidence increases sharply in the transition from sub-elite to elite academies, a period characterised by a rapid increase in training load relative to the player's chronic

baseline, wich is the dose–response relationship between load and injury serves as a central hypothesis of the predictive modelling literature reviewed in subsequent sections.

The financial and competitive dimensions of these injuries have been growing rapidly, and a detailed industry analysis conducted by Howden Sport and Entertainment (2025) reviewed five seasons of data up to 2024/25 for Europe's top five men's leagues, recording 22,596 injuries at a combined cost of £2.97 billion (€3.45 billion) in salary paid to unavailable players. Dallmeyer et al. (2025) added to this picture with an econometric analysis of German Bundesliga data, showing that a team losing four to five additional players to injury relative to the league median falls on average by one position in the final table, while the absence of a single outfield player for every matchday costs the club approximately one league point across the season, and Courty (2022) argued further that from a game-theoretic perspective clubs systematically under-invest in injury prevention because the primary beneficiaries of reduced injury risk, namely player health and longevity, are shared between the club and the player, whereas the cost of monitoring infrastructure is covered entirely by the club. The modern congested fixture calendar amplifies all of these effects, as Den Hollander et al. (2024) showed that injuries are significantly more prevalent during periods of fixture congestion, with injured players exhibiting higher match workloads and shorter inter-match recovery windows in the weeks preceding the injury compared to matched non-injured controls across five elite European leagues, and the expansion of the UEFA Champions League group stage from 32 to 36 clubs from 2024/25 alongside the introduction of an expanded FIFA Club World Cup have since further compressed recovery windows for clubs competing on multiple fronts, suggesting that the congestion problem identified in the literature will intensify rather than abate.

2.1.2 Wearable Workload Monitoring and Its Barriers

Professional football performance science monitors athlete workload along two complementary dimensions, namely external load referring to the physical work performed and internal load referring to the physiological stress it induces, of which the internal response is most proximally related to tissue damage (Impellizzeri et al., 2019) and Rago et al. (2020) conducted a systematic review of methodologies for collecting external load via microtechnology-equipped GPS units in professional football, reporting that 10,Hz GPS devices with integrated accelerometers, magnetometers, and heart rate sensors have become the de facto standard at the elite level enabling the continuous capture of total distance covered, high-speed running distance, sprint distance, and mechanical load derived from accelerometer impulse data, the most influential conceptual framework derived from wearable data is the acute-to-chronic workload ratio (ACWR), which compares workload accumulated in the most recent week against a longer rolling baseline operationalising the intuition that injury risk rises when training increases abruptly relative to the athlete's established level (Gabbett, 2016), and a subsequent International Olympic Committee consensus statement reinforced that rapid spikes in load and inadequate recovery are among the most consistently supported modifiable risk factors for injury (Soligard et al., 2016), despite their scientific validity, GPS-based monitoring systems impose significant financial and logistical barriers, since commercially available systems from providers such as Catapult Sports and STATSports cost approximately €2,400 per unit making full-squad instrumentation a substantial capital investment that is prohibitive for clubs outside the top divisions (Tierney et al., 2026), and beyond cost, proprietary data access poses a structural barrier to independent research as clubs that do collect GPS and heart rate data are generally reluctant to share it given that it constitutes a competitive asset, which means most published studies are limited to single-club, single-season samples of 25 to 40 players (Martins et al., 2025; Rossi et al., 2018), and Catterall et al. (2026) explicitly note that data-sharing limitations are particularly acute in women's and elite continental football where even basic injury surveillance is

incomplete, which is precisely what motivates approaches such as the present thesis that rely exclusively on publicly available sources.

2.2 Machine Learning for Injury Prediction

2.2.1 Tree-Based Methods and GPS-Derived Data

When machine learning was first applied to sports injury prediction, researchers went straight to GPS-derived wearable data as their primary source, and Rossi et al. (2018)trained random forest and gradient boosting models on 72 GPS features collected daily from 26 professional footballers at a single Italian Serie A club over two seasons, achieving an AUC of up to 0.92 for the following-day injury prediction task, though that impressive figure needs to be taken with some caution since both the training and test sets came from the same club and the same players, which means the model essentially learned each player's individual injury patterns rather than anything that could generalise more broadly across the sport. Saberisani et al. (2025) reported a similarly strong AUC of 0.89 using XGBoost on GPS data from 34 Iranian professional players, with SHAP analysis pointing to ACWR-derived metrics and high-speed running distance as the features that mattered most, and while Piłka et al. (2023) took things a step further by extending the GPS-based approach to a multi-club dataset, their reported 81% accuracy came with the honest acknowledgement that accuracy is actually a misleading metric under the severe class imbalance that characterises injury datasets, where injured samples typically make up fewer than 15% of all observations.

2.2.2 Methodological Critiques and Public-Data Approaches

A series of systematic reviews have identified persistent methodological weaknesses in the injury prediction literature, and Bullock et al. (2022) synthesised 44 prediction models across sports, finding that fewer than a third reported calibration statistics, fewer than a quarter performed any form of temporal validation as opposed to random-split cross-validation, and almost none were externally validated on data

from a different club or league, leading them to conclude that published AUC values substantially overestimate the performance that would be observed in deployment. Majumdar et al. (2022) reviewed 27 studies of machine learning in football injury prediction specifically and identified three recurring limitations, namely small single-club datasets that limit statistical power and prevent generalisation, inadequate handling of class imbalance where injured cases typically represent 8 to 15% of observations, and the use of random rather than chronological train-test splits which introduces temporal leakage by allowing future information to influence training, and Leckey et al. (2025) confirmed these findings in a broader scoping review across multiple sports, additionally noting that the explanatory framework is often absent since models are evaluated on held-out accuracy but the features that drive predictions are rarely interrogated. Martins et al. (2025) provided one of the few longitudinal multi- season analyses in the literature, following 35 professional footballers at a single club for four seasons, and despite covering multiple seasons the study remained limited to a single club, with model performance degrading meaningfully when the most recent season was held out for testing, confirming Bullock et al. (2022)'s warning about within-club overfitting.

A smaller but growing sub-literature has explored whether injury-relevant signals can be extracted from public data sources, thereby circumventing the wearable data barrier altogether, and Chang et al. (2024) used match statistics from WhoScored.com to train XGBoost models on English Premier League data, identifying squad affiliation, age, and cumulative workload as the strongest predictors, though while their approach demonstrated that public data contains meaningful signal the study was limited to a single season and framed the task as season- level binary classification rather than in-season rolling forecasting. Catterall et al. (2026) highlight that even for studies using public data the lack of sub-match physical metrics remains a ceiling on predictive performance, and that more sophisticated temporal modelling may partially compensate for this feature poverty.

2.3 Sequential Models and Explainability

2.3.1 Recurrent Models and Attention

The majority of machine learning studies reviewed above treat each player appearance as an independent observation, ignoring the temporal dependence between successive matches that is fundamental to the workload accumulation hypothesis, and since injury risk is a function of cumulative load over weeks or months a model that considers only the most recent match or an aggregated season total discards the sequential information that is most diagnostically relevant, which is precisely what motivates the use of recurrent architectures that explicitly model the ordered sequence of match appearances as a time series. The long short-term memory (LSTM) network, introduced by Hochreiter and Schmidhuber (1997), was designed precisely to overcome the vanishing gradient problem that prevents standard recurrent neural networks from retaining information across long sequences, making it a natural choice for match-event sequences in which a congested fixture period from several weeks prior may be the primary causal antecedent of a current-week injury.

A standard LSTM compresses the full input sequence into a single final hidden state, which implicitly assigns higher importance to recent time steps, and the attention mechanism, originally proposed for neural machine translation by Bahdanau et al. (2014), addresses this limitation by computing a learned importance weight for every hidden state in the sequence and aggregating them into a context vector that replaces the final hidden state as input to the classifier, with Wang et al. (2016) having adapted this approach to text classification and demonstrated that attention weights provide human-interpretable evidence about which parts of an input sequence drive the model's output. In the sports analytics domain the attention mechanism offers a dual advantage in that it potentially improves predictive performance by allowing the model to focus on the most informative match in the lookback window irrespective of its temporal position while also providing a first-level explainability signal by making the model's temporal focus observable, and Zhao et al. (2026) recently

demonstrated this advantage in a comparative study of injury prediction pipelines for female varsity soccer, finding that a Hybrid Transformer with attention-based temporal interpretability achieved the best performance with an AUC of 0.647 ahead of a standard LSTM at 0.611, XGBoost at 0.556, and SVM at 0.497 under leave-one- subject-out cross-validation, while the attention weights highlighted specific movement phases such as the landing and push-off phases of jump and hop tasks that differentiated injured from non-injured athletes, though their study used rich biomechanical motion-capture data rather than public match statistics.

2.3.2 Explainability Methods

Predicting sports injuries sits at the intersection of athletic performance management and clinical decision-making, since interventions such as reduced training load, squad rotation, or focused rehabilitation are usually triggered by a prediction that a player is at increased risk, and practitioners therefore need to be certain that the prediction is grounded in meaningful physiological signals rather than erroneous correlations. The SHapley Additive exPlanations (SHAP) framework, introduced by Lundberg and Lee (2017), grounds feature attribution in cooperative game theory by defining the contribution of each feature as its average marginal effect across all possible feature subsets, and Lundberg et al. (2020) extended this framework to tree-based models with the `TreeExplainer` algorithm, which computes exact Shapley values in polynomial time and has become the de facto standard for interpreting gradient boosted trees and random forests in the sports science literature, with Leckey et al. (2025) reviewing the use of SHAP in sport injury prediction and finding that across 27 studies cumulative training load, recovery time, and acute workload spikes were the features most consistently assigned high SHAP values regardless of the specific model architecture.

For neural network architectures, the `DeepExplainer` variant of SHAP relies on a gradient-based backward pass through the network. PyTorch's cuDNN LSTM backend does not support the second-order backward pass this requires, which

limits its applicability to the recurrent architecture used here. An alternative that is provably consistent for differentiable models is Integrated Gradients (IG), proposed by Sundararajan et al. (2017), which accumulates the gradient of the model output with respect to the input along a straight-line interpolation from a reference baseline to the actual input, producing an attribution for every individual cell in the input matrix. Sun et al. (2024) applied IG to an LSTM model for highway infrastructure degradation prediction and found that its completeness property made IG attributions substantially more consistent across random seeds than saliency maps, which motivated its adoption in the present study as a complement to SHAP.

A growing body of work argues that no single explainability method provides a complete picture of model behaviour, and that triangulating across multiple methods—each with distinct theoretical assumptions—provides stronger evidence for substantive conclusions (Molnar, 2022). Li et al. (2025) illustrate the value of combining complementary sources of information in a multimodal injury-prediction model that fuses CT-imaging features (extracted with an attention-based Swin-UNet) with biomechanical time-series features (encoded by an LSTM), reporting that the fused representation outperformed either modality alone. Their work motivates the broader principle adopted here—that combining multiple, methodologically distinct views of a model (or its inputs) yields more trustworthy conclusions than any single view. In the context of this thesis, this principle is operationalised through a three-tier explainability framework: SHAP values are computed for all three model families to enable cross-model feature-importance comparison (Section 4.1.2); attention weights characterise the LSTM's temporal focus (Section 4.1.3); and Integrated Gradients provide a fine-grained feature-by-time-step attribution matrix for the LSTM (Section 4.1.4).

CHAPTER 3

METHODOLOGY

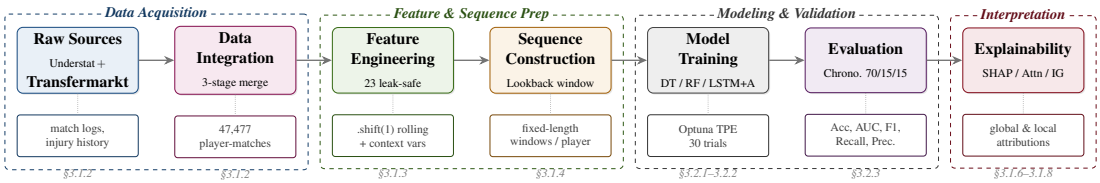
3.1 Methodology

3.1.1 Research Design Overview

The methodology follows a seven-stage pipeline, organised into four conceptual phases, that uses two public web sources to introduce an evaluated, interpretable injury prediction system. The stages are illustrated in Figure 1 and detailed in the sections that follow.

Figure 1

End-to-end methodology pipeline, from public data acquisition through feature engineering, sequence construction, model training and chronological evaluation, to multi-method explainability. Section references beneath each stage indicate where it is described in the text.



3.1.2 Data Curation

This study used two main types of data; the first was per-match player performance statistics that used as workload indicators, the second was historical injury records, used to construct the prediction labels, however, no private club data or wearable sensor data were used. All datasets were collected from public sources that permit non-commercial academic research.

Match Statistics — Understat. Match performance data were collected from Understat (Understat, 2024) which is a public football analytics website that provides detailed player match records for Premier League games, where for each player appearance, Understat provides minutes played, goals, assists, shots, expected goals

(xG), expected assists (xA), and key passes. These data were collected for five Premier League seasons, from 2020/21 to 2024/25 (380 matches per season $\times$ 5 seasons $=$ 1,900 matches), producing 55,174 raw player-match records.

Appearance Records and Injury Labels — Transfermarkt. Player appearance data and injury histories were collected from Transfermarkt (Transfermarkt GmbH & Co. KG, 2024), a public football database that provides match appearances and injury information, where appearance data include minutes played, goals, assists, and disciplinary events across many competitions. The injury records comprise 143,195 historical entries for more than 93,000 players worldwide where each record includes injury type, start date, return date, days missed, and matches missed.

Data Integration Pipeline. Bringing the two datasets together required a three-step integration pipeline designed to match as many records as possible while keeping the injury labels accurate.

Stage 1 — Scope Filtering. The first step was simply narrowing things down, since Transfermarkt's appearance records were filtered to include only English Premier League matches (competition ID: `GB1`) from the 2020/21 to 2024/25 seasons, which left 47,805 player-match records covering 1,125 players, and the injury records were filtered in the same way to keep only those linked to these players.

Stage 2 — Injury Label Construction. With the scope defined, binary injury labels were created using a 30-day lookahead window, so that for each player-match record the player's injury history was checked and if an injury start date fell within 30 days after the match date the record was labelled injured ($y = 1$), and otherwise not injured ($y = 0$), and this window was chosen because it captures injuries that may be linked to recent workload accumulation while reducing the chance of picking up injuries caused by unrelated events later in the season (Den Hollander et al., 2024), and after labelling the dataset contained 11.6% injured and 88.4% not-injured cases, a rate consistent with the reported Premier League injury incidence of approximately 8 to 15% (Ekstrand et al., 2011).

Stage 3 — Cross-Source Name Matching. The trickiest part of the pipeline

was matching player records across Understat and Transfermarkt, since both platforms identify players mainly by name rather than by any shared ID, and a two-level string-matching method was therefore used to handle this, where Tier 1 applied exact matching on player name, season, and match date and successfully matched 76.5% of records, while Tier 2 applied Unicode normalisation and HTML decoding to reconcile differences in spelling and special characters, for example converting *Tomáš Souček* $\rightarrow$ *tomas soucek*), and recovered a further 9.5% of records, leaving a remaining 14.0% that could not be matched primarily because of missing player coverage between the two sources rather than any data errors, and these unmatched records were removed from the final dataset and are acknowledged as a limitation in Chapter 4.2.3).

Final Dataset. The unified dataset that came out of this pipeline comprises 47,477 player-match records across 1,017 unique EPL players and five complete seasons, with 23 features per record and a binary injury label, and the 20 raw parquet columns stored on disk consist of 9 raw match statistics together with identifier and metadata fields covering season, match and player identifiers, team and venue information, playing position, and the injury labels, with feature engineering in Section 3.1.3 then expanding those 9 raw statistics into the full 23 model features by adding 9 rolling-average counterparts and 5 contextual variables. Table 1 summarises the dataset characteristics, while Figures 2 and 3 illustrate the seasonal distribution of injury events and average severity across the dataset.

Table 1
Summary of The Unified Player-match Dataset.

Attribute	**Value**
Seasons covered	2020/21 to 2024/25
Total player-match records	47,477
Unique players	1,017
Raw parquet columns	20
Engineered model features	23
Match statistics source	Understat (Understat, 2024)
Injury labels source	Transfermarkt (Transfermarkt GmbH & Co. KG, 2024)

Figure 2

Distribution of Injury Events Across Five Premier League Seasons (2020/21–2024/25). Total Injury Events Recorded In the Unified Dataset: 5,504.

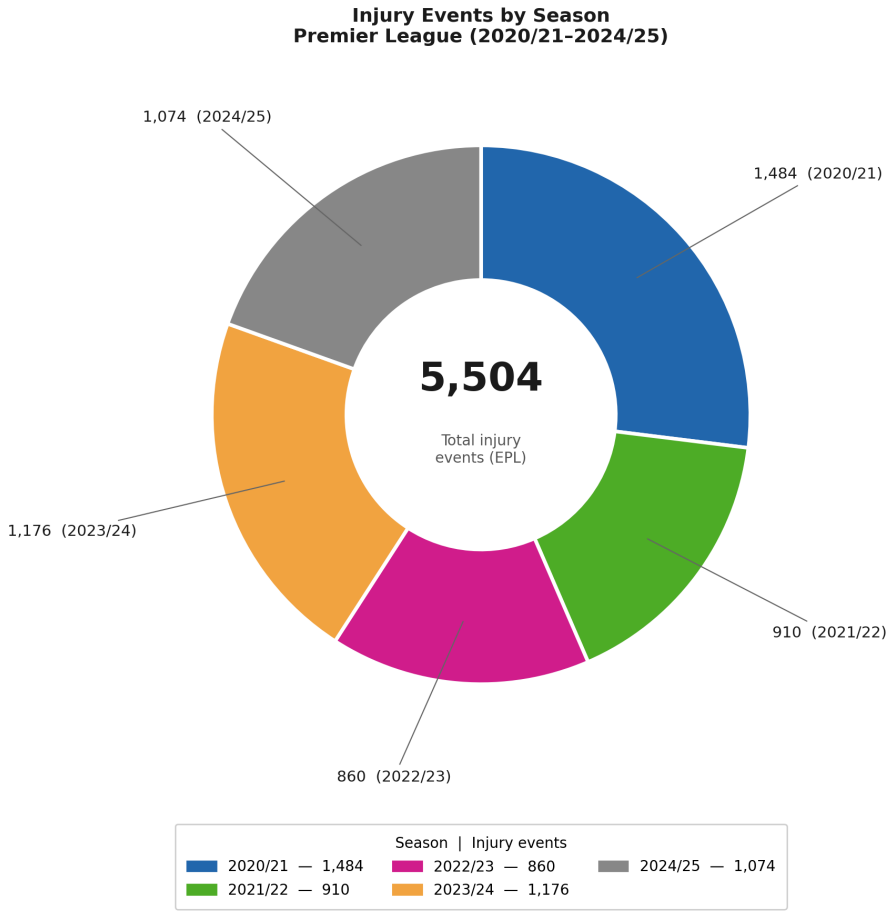
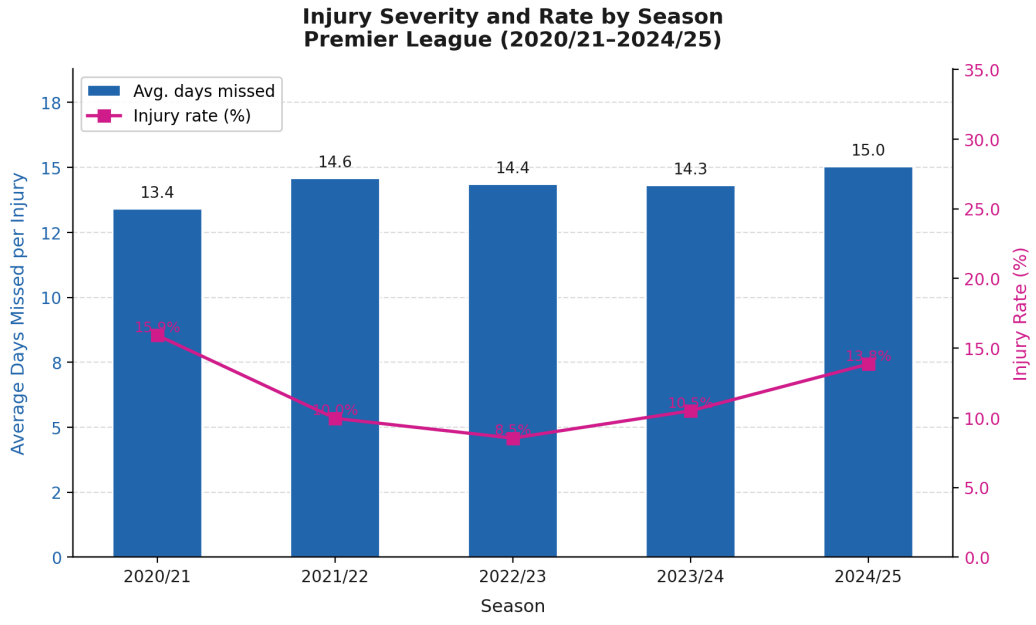


Figure 3

Average Days Missed Per Injury (bars) and Overall Injury Rate Percentage (line) Per Season Across The Premier League (2020/21–2024/25).



3.1.3 Feature Engineering

From the unified dataset described in Section 3.1.2, a total of 23 features were engineered and organized into three categories: 9 raw match statistics, 9 rolling averages derived from those statistics, and 5 contextual features. Table 2 provides the complete list.

Table 2
Summary of The 23 Engineered Features Used for Injury Prediction.

Feature	Category	Description
goals	Raw	Goals scored in the match
assists	Raw	Assists recorded in the match
minutes	Raw	Minutes played in the match
shots	Raw	Total shots attempted
key_passes	Raw	Key passes completed
xG	Raw	Expected goals
xA	Raw	Expected assists
red_cards	Raw	Red cards received
yellow_cards	Raw	Yellow cards received
goals_roll5	Rolling	5-match rolling mean of goals
assists_roll5	Rolling	5-match rolling mean of assists
minutes_roll5	Rolling	5-match rolling mean of minutes
shots_roll5	Rolling	5-match rolling mean of shots
key_passes_roll5	Rolling	5-match rolling mean of key passes
xG_roll5	Rolling	5-match rolling mean of xG
xA_roll5	Rolling	5-match rolling mean of xA
red_cards_roll5	Rolling	5-match rolling mean of red cards
yellow_cards_roll5	Rolling	5-match rolling mean of yellow cards
cumulative_minutes	Contextual	Total minutes played up to but not including the current match
days_rest	Contextual	Days since the player's previous match appearance
match_number	Contextual	Sequential count of appearances within the season
is_home	Contextual	Binary flag for home venue
position_code	Contextual	Numerically encoded playing position

All rolling averages were computed with a window size of 5 matches and shifted by one match to exclude the current match from its own calculation, to prevent

data leakage. Contextual features such as cumulative_minutes and match_number were similarly computed using only prior appearances. The injury label uses a forward-looking 30-day window that does not overlap with any input feature's time reference.

3.1.4 Sequence Construction, Splitting, and Scaling

Per-player sequence construction: For each player, matches were ordered chronologically and converted into fixed-length input sequences using a sliding lookback window. For a lookback length L, the L matches immediately preceding match i form the input sequence and the binary label of match i (injured within the next 30 days) is the target. Players with fewer than the minimum required number of matches in a season were excluded, since they could not yield a complete sequence. The lookback length L was treated as a tunable hyperparameter (Section 3.2.2).

Chronological per-player train/validation/test split: To prevent temporal leakage, the data were split chronologically *within each player* rather than randomly across the pooled dataset. For every player, the generated sequences were ordered by match date and divided into the earliest 70% (training), the next 15% (validation), and the final 15% (test). This guarantees that, for any given player, no sequence in the training set occurs later in time than a sequence used for validation or testing, so the model is never trained on information that post-dates the data it is evaluated on. The same split ratios and procedure were applied identically to the tree-based models and to the LSTM to keep the comparison fair.

Feature scaling: Because the LSTM is sensitive to input magnitude, all features were scaled to the $[0, 1]$ range using a `MinMaxScaler`. To avoid leakage, the scaler was fitted on the training partition only and then applied unchanged to the validation and test partitions. Scaling parameters were therefore never derived from data the model was evaluated on.

Class imbalance handling: The 11.6% injury prevalence was addressed differently for the two model families, according to each one's conventions. The

Decision Tree and Random Forest used scikit-learn's `class_weight='balanced'`, which reweights samples inversely to class frequency. The LSTM was trained with a weighted binary cross-entropy loss (`BCEWithLogitsLoss`) in which the positive class was assigned a weight equal to the ratio of negative to positive training examples, so that misclassifying an injury contributes proportionally more to the loss than misclassifying a healthy observation. No synthetic oversampling (e.g. SMOTE) was applied, so that the temporal structure of each player's real match sequence was preserved.

Summary of temporal-leakage safeguards: Four concrete mechanisms jointly prevent any future information from reaching the model at training or evaluation time:

1. **Lagged rolling features.** Every rolling average is computed with a one-match shift (`.shift(1)`) applied before the rolling window, so a match's own statistics never enter its own feature vector (Section 3.1.3).

2. **Prior-only contextual features.** Cumulative and count-based features (`cumulative_minutes`, `match_number`) are accumulated strictly over appearances preceding the current match.

3. **Chronological per-player splitting.** Sequences are ordered by match date within each player and partitioned 70/15/15, guaranteeing that no training sequence post-dates any validation or test sequence for that player (Section 3.1.4).

4. **Train-only scaler fitting.** The `MinMaxScaler` is fitted on the training partition alone and applied unchanged to validation and test, so no distributional information from evaluation data influences the inputs.

The forward-looking 30-day label is disjoint in time from every input feature, so target construction introduces no overlap with the predictor window.

3.1.5 Sequence Modelling

A Recurrent Neural Network (RNN) is a type of neural network designed to work with sequential data, meaning the order of the data matters. It maintains a memory of previous inputs using the hidden state from the previous time step to influence the current prediction and when training backpropagation through time is used, but the issue is that gradients can become extremely small; for example, if the network needs to retain information from many matches earlier in the sequence it often fails, which is known as the vanishing gradient problem.

The LSTM is a special type of RNN, introduced in 1997, that allows the network to retain important information over long sequences, avoid vanishing gradients, and learn long-range dependencies better than standard RNNs. It achieves this by introducing a separate memory line, the cell state C_t, which changes slowly so that information can flow across many time steps without disappearing. The cell state is regulated by three gates, each using a sigmoid activation function (σ) that outputs values between 0 and 1 and acts as a soft switch controlling the flow of information.

The forget gate f_t determines what proportion of the previous cell state C_{t-1} to retain:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{1}$$

The input gate i_t controls what new information is written into the cell state, together with a candidate cell state $\tilde{C}_t$:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{2}$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{3}$$

The cell state is then updated as:

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \tag{4}$$

Finally, the output gate o_t determines what portion of the cell state is exposed as the hidden state h_t:

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{5}$$

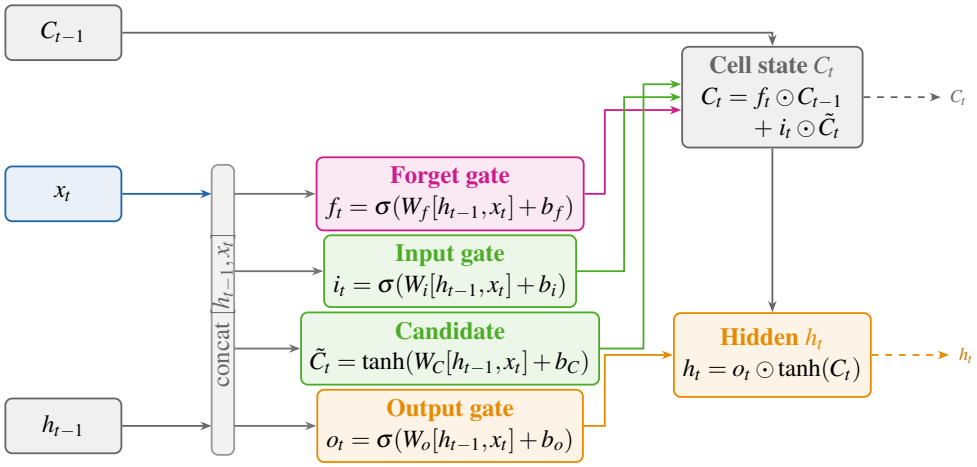
$$h_t = o_t \odot \tanh(C_t) \tag{6}$$

where σ denotes the sigmoid activation function, tanh is the hyperbolic tangent, $\odot$ denotes element-wise multiplication, and W_*, b_* are learned weight matrices and bias vectors.

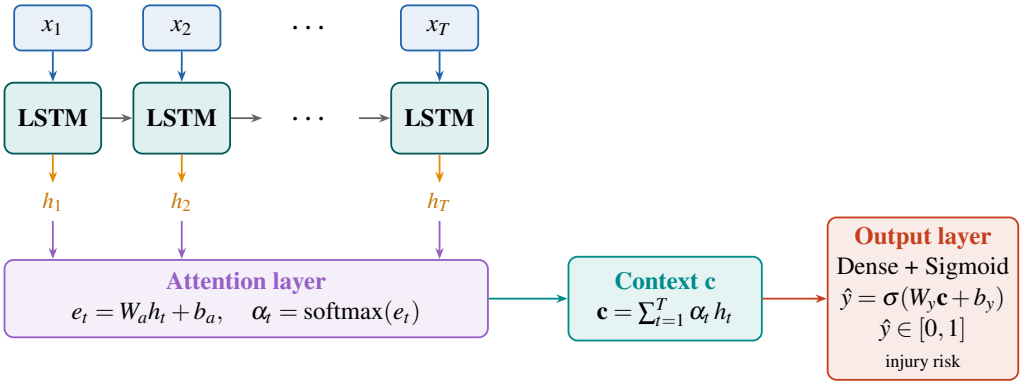
The LSTM is well suited to football injury prediction because injury risk depends on history and time rather than on a single isolated match. A player's fitness and injury risk evolve over the season: whether a player is injured in a given week depends on what happened in preceding weeks, such as match density and accumulated fatigue. These factors build up over time, much like a time series. The LSTM can retain relevant information from previous weeks in its memory and, instead of treating each week in isolation, model the player's longer-term workload and recovery pattern, which makes it a strong and logical choice for predicting injuries.

Figure 4

LSTM With Attention Architecture For Football Injury Prediction.



(a) Inside a single LSTM cell at time step t



(b) LSTM unrolled across the lookback window with attention pooling

Note: Internals of a single LSTM cell: the previous hidden state h_{t-1} and the current match features x_t are concatenated and routed through four gates (forget, input, candidate, output) that update the long-term memory C_t and emit the short-term hidden state h_t. **(b)** The cell is unrolled across the lookback window of T prior matches; an attention layer learns importance weights α_t for every hidden state, and the resulting context vector $\mathbf{c}$ is fed to a dense sigmoid layer that outputs the injury probability $\hat{y}$.

3.1.6 Attention Mechanism

In a traditional Long Short-Term Memory (LSTM) architecture, only the final hidden state (h_T) is passed to the classifier, where it acts as a shorthand for the entire time series. This creates a temporal bias toward recent observations, which is a significant limitation for injury prediction, because the physiological mechanisms leading to injury do not always correlate with immediate fatigue alone. For example,

a workload increment several matches earlier (for instance, during a congested December schedule) may be more predictive than the most recent match in January when the schedule is normal.

To address this, instead of relying on the final hidden state alone, an attention layer is added that computes a learned importance score for every time step in the sequence. This idea was originally proposed for machine translation by **bahdanau2015neural<empty citation>** and later adapted to classification tasks by Wang et al. (2016). The layer assigns a relevance score to each hidden state produced by the LSTM:

$$e_t = \mathbf{W}_a \mathbf{h}_t + b_a, \quad t = 1, 2, \ldots, T \tag{7}$$

where e_t is the attention score for time step t; $\mathbf{h}_t \in \mathbb{R}^d$ is the LSTM hidden state at time step t; $\mathbf{W}_a \in \mathbb{R}^{1\times d}$ is a learnable weight matrix; $b_a \in \mathbb{R}$ is a bias term; d is the hidden size of the LSTM; and T is the total number of time steps in the lookback window.

For these raw scores to be interpreted as a probability distribution over time steps, they are normalised using the **softmax** function, which ensures all weights are positive and sum to one. The normalised weights can then be interpreted as the relative importance the model assigns to each match in the input window.

$$\alpha_t = \frac{\exp(e_t)}{\sum_{k=1}^{T} \exp(e_k)}, \quad t = 1, 2, \ldots, T \tag{8}$$

where α_t is the normalized attention weight for time step t, satisfying $\alpha_t \geq 0$ and $\sum_{t=1}^{T} \alpha_t = 1$.

After normalisation, the context vector is formed as a weighted combination of all hidden states. This context vector replaces the final hidden state as the input to the fully connected classification layer, so that information from any time step can contribute to the prediction in proportion to its learned relevance, rather than being compressed into a single final state.

$$\mathbf{c} = \sum_{t=1}^{T} \alpha_t \mathbf{h}_t \tag{9}$$

where $\mathbf{c} \in \mathbb{R}^d$ is the context vector that aggregates information across all T time steps, weighted by the learned attention distribution α_t.

The context vector $\mathbf{c}$ is then passed to the fully connected layer for classification, enabling the model to focus on the most relevant matches within the lookback window.

3.1.7 Integrated Gradients

In Section 3.1.6 we can see which time steps the model prioritises, but within each time step there are multiple features, and attention does not distinguish between them. Similarly, the SHAP method (Section 3.1.8) measures each feature's total impact, but averages this contribution over the whole input window. To address this, Integrated Gradients (IG) is introduced, an attribution method proposed by Sundararajan et al. (2017) that assigns an importance score to every individual cell in the input matrix, so that each feature at each time step receives its own attribution value. This provides the most detailed level of explainability in the framework, complementing the temporal view from attention and the feature-level view from SHAP.

The method works as follows. Starting from a baseline input (typically all zeros, representing the absence of information), a straight-line path is constructed from the baseline to the actual input sequence, and the gradients of the model output are accumulated along this path.

$$\mathrm{IG}_i(\mathbf{x}) = (x_i - x_i') \times \int_{\alpha=0}^{1} \frac{\partial F(\mathbf{x}' + \alpha(\mathbf{x} - \mathbf{x}'))}{\partial x_i} \, d\alpha \tag{10}$$

where $\mathbf{x}$ is the input sequence; $\mathbf{x}'$ is the baseline input (a zero vector representing the absence of information); x_i and x_i' are the i-th elements of $\mathbf{x}$ and $\mathbf{x}'$ respectively; F is the model output function (injury probability); and $\alpha \in [0,1]$ parameterizes the straight-line path from the baseline to the input. Since the integral in Eq. 10 cannot be computed analytically, it is approximated using a Riemann summation with m discrete

steps:

$$\mathrm{IG}_i(\mathbf{x}) \approx (x_i - x_i') \times \frac{1}{m} \sum_{k=1}^{m} \frac{\partial F\left(\mathbf{x}' + \frac{k}{m}(\mathbf{x} - \mathbf{x}')\right)}{\partial x_i} \tag{11}$$

where m is the number of interpolation steps used to approximate the integral. In this study, $m = 50$ steps were used following the recommendation of (Sundararajan et al., 2017).

The main reason for using Integrated Gradients over simpler gradient-based methods is that it satisfies two axioms. The first is sensitivity: if changing a feature changes the model prediction, that feature receives a nonzero attribution. The second is completeness, expressed as:

$$\sum_{i=1}^{n} \mathrm{IG}_i(\mathbf{x}) = F(\mathbf{x}) - F(\mathbf{x}') \tag{12}$$

where n is the total number of input elements, $F(\mathbf{x})$ is the model prediction for the actual input, and $F(\mathbf{x}')$ is the model prediction for the baseline.

This property ensures that the attributions perfectly match the difference between the model's output and the baseline. It just means nothing gets lost or made up along the way.

In this study, Integrated Gradients were computed for the sequences with the highest predicted injury probability in each season. The output is an attribution matrix that has the shape lookback $\times$ 23, where each cell represents how much a specific feature at a specific match position contributed to the injury prediction. Then, by taking the mean absolute attribution across sequences, a feature-level ranking was obtained and compared against the SHAP-based ranking to assess consistency across the two explainability methods.

3.1.8 SHAP Explainability

SHAP (**SH**apley **A**dditive ex**P**lanations) is a concept coming from cooperative game theory. Shapley values quantify the average contribution of a feature to a model's prediction (Lundberg & Lee, 2017). Rather than retraining the model on

every feature subset—which would be computationally infeasible—SHAP estimates each feature's contribution by comparing the model's output when the feature is present in the conditioning set against its output when the feature is marginalised out, $f(x_{S\cup\{i\}}) - f(x_S)$, where x_S denotes the values of the features in subset S. Because the effect of a feature depends on which other features are present, this difference is evaluated across feature subsets and combined into a weighted average. The Shapley value is this weighted average of marginal contributions:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!} \left[f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_S(x_S) \right] \tag{13}$$

where F represents the complete collection of input features, S is a subset of F not containing feature i, $|S|$ is the cardinality of subset S, $|F|$ is the overall count of features, $f_S(x_S)$ is the model's expected output when only the features in subset S are present, and ϕ_i is the Shapley value representing the contribution of feature i to the prediction.

In this research, SHAP values for the LSTM model were computed using the **KernelExplainer** rather than the **DeepExplainer**. `DeepExplainer` relies on a gradient-based backward pass through the network, and PyTorch's cuDNN LSTM backend does not support the second-order (double) backward pass that this requires, which prevents reliable attribution for the recurrent architecture used here. `KernelExplainer`, by contrast, is a model-agnostic approach that treats the model as a black-box function and approximates Shapley values by evaluating predictions over perturbed (masked) input subsets, requiring no gradient access.

3.2 Experimental Setup

Having defined the data, features, sequence construction, and explainability methods, this section specifies the models compared in the study, the procedure used to tune them, and the metrics by which they are evaluated.

3.2.1 Baseline Models

Decision Tree: A Decision Tree is a model in machine learning used for supervised learning tasks (learns from labeled data) that works by dividing the data once more and again into smaller subsets through yes/no choices (Binary splits), by constructing a tree where each internal node asks a question about one feature (e.g. "is age > 25?"), and each branch is the answer (yes/no) and the final prediction is given by each leaf node (injured/not injured) (Breiman et al., 1984). At every node the tree chooses the split that gives best separation between classes to increase information gain, as evaluated by Gini impurity for classification tasks or minimizing mean squared error for regression:

$$\text{Gini}(t) = 1 - \sum_{k=1}^{K} p_k^2 \tag{14}$$

where p_k is the proportion of class k samples at node t.

The tree keeps splitting until a stopping rule is met, such as reaching the maximum depth, having too few samples in a leaf, or achieving no further improvement in purity. Decision Trees are valued for their interpretability and their ability to capture non-linear relationships without feature scaling. However, a single Decision Tree can be unstable: small changes in the training data can produce a very different tree, a property known as high variance. For this reason, Decision Trees are often used as base learners inside the ensemble methods described below.

Random Forest: Random Forest (RF), presented by Breiman (2001), is a technique based on bagging that generates a substantial quantity of decision trees $\{T_1, T_2, \ldots, T_B\}$, each trained on a bootstrap sample of the training data with an additional layer of feature randomization. At each split within a tree only a random selection of m features (typically $m = \sqrt{p}$ for classification, where p is the total number of features) is taken as candidates for the split. The final prediction combines the outputs from each tree:

$$\hat{y} = \frac{1}{B} \sum_{b=1}^{B} T_b(x) \quad \text{(regression)} \tag{15}$$

$$\hat{y} = \text{majority vote}\{T_b(x)\}_{b=1}^{B} \quad \text{(classification)} \tag{16}$$

Bootstrap sampling helps reduce variance (trees are trained on slightly different data) and feature randomization makes trees less similar to each other (less correlated), thus the final model becomes more stable and reliable than one decision tree.

Random Forest is also naturally resistant to overfitting as B increases. It additionally provides a built-in evaluation method, the Out-of-Bag (OOB) error: because each tree is trained on a bootstrap sample, the points left out of that sample serve as a test set, and the error on them gives a good estimate of generalisation performance. For injury prediction, Random Forest is useful because it can quantify feature importance, identifying which variables (workload, performance metrics, and so on) most influence injury risk. This is conventionally done using Mean Decrease in Impurity (MDI); however, MDI can assign unfairly high importance to features with many unique values. Because of this bias, SHAP explanations (Section 3.1.8) are used as an additional method to interpret the model.

3.2.2 Hyper-parameter Tuning

Hyperparameter optimisation was conducted for all models using Bayesian optimisation, an iterative strategy designed for the global optimisation of expensive-to-evaluate objective functions (Frazier, 2018). The process operates through two components: a surrogate model and an acquisition function. First, the surrogate model constructs a probabilistic estimate of the objective across the hyperparameter search space, providing both a predicted value and a measure of uncertainty. Second, the acquisition function uses this surrogate to decide which configuration to evaluate next, balancing exploration of high-uncertainty regions against exploitation of regions where high performance is predicted. This study uses

the Expected Improvement (EI) criterion, defined as:

$$EI_n(\mathbf{x}) := \mathbb{E}_n\left[\left[f(\mathbf{x}) - f_n^*\right]^+\right] \tag{17}$$

Bayesian optimisation was implemented using the Optuna framework, which employs a Tree-structured Parzen Estimator (TPE) as its surrogate model. Unlike Gaussian-process approaches that model the objective directly as $P(y \mid \mathbf{x})$, TPE models the distribution of hyperparameter configurations conditioned on the objective value, $P(\mathbf{x} \mid y)$, which makes it more efficient for conditional and mixed search spaces. Optimisation was run for 30 trials per model per season, using validation AUC as the objective. During optimisation, each trial was evaluated on the single held-out chronological validation fold defined in Section 3.1.4 (the middle 15% of each player's sequences) rather than by k-fold cross-validation; a fixed chronological fold was used deliberately to preserve the temporal ordering of matches and avoid the leakage that random cross-validation folds would introduce. AUC was chosen over accuracy because, given the 11.6% injury prevalence, a model predicting only the majority class would achieve misleadingly high accuracy while identifying no injury cases.

For the LSTM with attention, the optimization included a two-stage search: the first stage optimized the network architecture and training hyperparameters (Table 3), and the second stage searched for the optimal lookback window length and minimum match threshold using the best hyperparameters from the first stage. For the Decision Tree and Random Forest models, Bayesian optimization was applied over the search spaces shown in Tables 4 and 5. Both tree models used `class_weight="balanced"` in all trials, which automatically adjusts sample weights inversely proportional to class frequencies to address the class imbalance.

Table 3
LSTM + Attention Hyperparameter Search Space for Bayesian Optimization Via Optuna.

Hyperparameter	Range	Scale
`hidden_size`	32–128 (step 32)	Linear
`num_layers`	1–3	Linear
`dropout`	0.1–0.5 (step 0.1)	Linear
`learning_rate`	1×10^{-4} – 1×10^{-3}	Logarithmic
`batch_size`	{64, 128}	Categorical
`LOOKBACK`	3–10	Linear
`MIN_MATCHES`	LOOKBACK+1 – LOOKBACK+5	Linear

Table 4
Decision Tree Hyperparameter Search Space for Bayesian Optimization Via Optuna.

Hyperparameter	Range	Scale
`max_depth`	3–20	Linear
`min_samples_leaf`	5–50	Linear
`min_samples_split`	2–30	Linear
`criterion`	{gini, entropy}	Categorical

Table 5
Random Forest Hyperparameter Search Space for Bayesian Optimization via Optuna.

Hyperparameter	Range	Scale
`n_estimators`	50–500 (step 50)	Linear
`max_depth`	5–30	Linear
`min_samples_leaf`	5–30	Linear
`min_samples_split`	2–20	Linear
`max_features`	{sqrt, log2, 0.3–0.7}	Categorical / Linear

Tables 6, 7, and 8 report the best hyperparameters found by Optuna for each model across all five seasons.

Table 6
Best LSTM + Attention Hyperparameters Found by Optuna for Each EPL Season.

Season	hidden_size	num_layers	dropout	learning_rate	batch_size	look_back
2020/21	128	2	0.3	0.000235	128	10
2021/22	64	2	0.3	0.000382	64	6
2022/23	64	1	0.4	0.000138	128	3
2023/24	64	2	0.4	0.000290	64	10
2024/25	128	2	0.5	0.000880	64	10

Table 7
Best Decision Tree Hyperparameters Found by Optuna for Each EPL season.

Season	max_depth	min_samples_leaf	min_samples_split	criterion
2020/21	9	44	16	entropy
2021/22	16	33	18	entropy
2022/23	5	6	20	entropy
2023/24	4	38	14	entropy
2024/25	6	25	30	entropy

Table 8
Best Random Forest Hyperparameters Found by Optuna for Each EPL Season.

Season	n_estimators	max_depth	min_samples_leaf	min_samples_split	max_features
2020/21	300	21	12	9	sqrt
2021/22	200	10	18	14	log2
2022/23	350	5	15	14	sqrt
2023/24	200	24	11	5	sqrt
2024/25	200	23	5	17	0.5[a]

Note: Continuous fraction sampled by Optuna's float-valued `max_features`.

3.2.3 Evaluation Metrics

In binary classification the model assigns each sample to one of two classes, in this study either injured or healthy. By comparing the predicted labels with the true labels, four outcomes are possible: True Positive (TP) where the model correctly predicts an injury, False Positive (FP) where the model predicts an injury but the player is healthy, True Negative (TN) where the model correctly predicts a healthy player, and False Negative (FN) where the model predicts healthy but the player is actually injured. All evaluation metrics used in this study are derived from these four quantities.

Accuracy assesses the ratio of all predictions that are accurate:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{18}$$

However, accuracy may be misleading when the classes are imbalanced. In this dataset, 88.4% of records belong to the healthy class, so a model that always predicts healthy would achieve 88.4% accuracy while detecting zero injuries. Additional metrics that focus specifically on minority-class performance are therefore needed.

Precision evaluates the proportion of predicted injuries that are true injuries:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{19}$$

Recall (also called sensitivity) assesses the number of actual injuries accurately recognized by the model:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{20}$$

In the context of injury prediction, recall is particularly important because a missed injury (FN) can have serious consequences for the player's health, whereas a false alarm (FP) only results in unnecessary caution.

F1-Score is the harmonic mean of precision and recall, providing a single metric that balances both:

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{21}$$

F1-score ranges from 0 to 1 and is especially useful when class distribution is skewed, as it penalises models that sacrifice one metric for the other.

AUC-ROC (Area Under the Receiver Operating Characteristic Curve) evaluates how well the model can differentiate between classes at various classification thresholds. The ROC curve graphs the True Positive Rate (recall) against the False Positive Rate (FPR $= FP/(FP+TN)$) at every threshold. AUC summarises this curve

into a single value: a score of 0.5 indicates random guessing and 1.0 indicates perfect separation. Because AUC is threshold-independent, it measures how well the model ranks injured players above healthy ones regardless of the decision cutoff chosen, making it the most robust metric for comparing models on imbalanced datasets.

CHAPTER 4

RESULTS

4.1 Results and Discussion

4.1.1 Overall Model Comparison

To verify that the per-match temporal granularity adopted in this thesis is operationally justified, a season-level baseline was run that re-frames the task as a single binary prediction per `(player, season)` pair: "did this player suffer any injury at any point during this season?". Features were aggregated to one row per player-season: each raw match statistic contributed its season total and its per-match average (18 features), supplemented by the number of matches played, the number of home appearances, the average days of rest between matches, and the modal playing position. The same Decision Tree and Random Forest model classes used in the match-level experiments were trained with fixed, conservative hyper-parameters and `class_weight='balanced'`, and evaluated under a leave-one-season-out protocol. Table 9 reports the comparison.

The season-level baseline achieves a higher mean test AUC (0.681 for the Decision Tree and 0.698 for the Random Forest) than any of the match-level models, but this comparison is not directly meaningful: the season-level task has a positive-class prevalence of 45.9% (virtually balanced) against the 11.6% prevalence of the match-level task, and the season-level model issues a single end-of-season prediction rather than a rolling 30-day forecast. The season-level formulation therefore solves an easier statistical problem but offers no in-season early-warning value: by the time the player-season features are fully observed, the season is already over. More importantly, the season-level formulation collapses exactly the signals that the cross-model SHAP analysis identifies as universally important—`cumulative_minutes`, `match_number`, and the recent rolling means of

`xG` and `minutes` (Table 11)—because those features are intrinsically defined by their per-match trajectory rather than by their season-end total. This positions the two formulations as complementary rather than competing: match-level prediction targets the in-season risk-monitoring scenario this study addresses, whereas season-level prediction is better suited to coarse pre-season risk screening at the squad level.

Table 9

Season-level Baseline (Leave-one-season-out) Versus Match-level Results From Test Set Performance Of All Models Across Five EPL Seasons, Averaged Across The Five EPL Seasons.

Model / formulation	AUC	F1	Recall	Positive rate
Decision Tree, season-level (this baseline)	0.681	0.609	0.678	45.9%
Random Forest, season-level (this baseline)	0.698	0.588	0.638	45.9%
Decision Tree, match-level	0.570	0.203	0.409	11.6%
Random Forest, match-level	0.617	0.130	0.105	11.6%
LSTM + Attention, match-level	0.607	0.147	0.148	11.6%

Note: The two formulations target different tasks with very different positive-class prevalence and are reported here for transparency rather than as competing predictors of the same quantity (see text).

Table 10
*Test Set Performance Of All Models Across Five EPL Seasons. Best Value Per Season is Shown in **bold**.*

Season	Model	Acc	AUC	F1	Rec.	Prec.
2020/21	DT	0.801	0.563	**0.164**	0.209	0.135
	RF	**0.893**	0.557	0.086	0.054	**0.211**
	LSTM	0.803	0.563	0.123	**0.294**	0.078
	LSTM+A	0.865	**0.577**	0.099	0.083	0.123
2021/22	DT	0.741	0.601	**0.228**	**0.375**	0.163
	RF	0.872	0.640	0.159	0.119	**0.241**
	LSTM	**0.887**	0.618	0.048	0.077	0.035
	LSTM+A	0.859	**0.654**	0.217	0.205	0.230
2022/23	DT	0.656	0.521	0.186	0.349	0.126
	RF	0.796	0.600	0.202	0.229	0.180
	LSTM	0.728	**0.609**	0.113	**0.403**	0.065
	LSTM+A	**0.803**	0.600	**0.227**	0.270	**0.195**
2023/24	DT	0.741	0.669	**0.267**	**0.466**	0.187
	RF	0.897	**0.746**	0.134	0.079	**0.455**
	LSTM	0.867	0.443	0.064	0.119	0.043
	LSTM+A	**0.903**	0.656	0.096	0.067	0.167
2024/25	DT	0.326	0.495	**0.171**	**0.647**	0.098
	RF	**0.877**	0.540	0.071	0.044	**0.188**
	LSTM	0.598	0.518	0.100	0.397	0.057
	LSTM+A	0.771	**0.549**	0.098	0.116	0.085
	Winner	RF / LSTM+A (2/5)	LSTM+A (3/5)	DT (4/5)	DT (3/5)	RF (4/5)

Note: In F1, Recall and Precision Calculation, Injured is Considered The Positive Class.

The per-metric winners in Table 10 require careful interpretation, because the metrics answer different operational questions. The Decision Tree attains the highest F1 in four of the five seasons, but F1 is computed at a single fixed decision threshold (0.5) and therefore reflects performance at one arbitrary operating point. For a screening tool whose threshold is set by a club's medical staff according to their preferred sensitivity–specificity trade-off (Section 4.2), the threshold-independent ranking quality captured by AUC is the more decision-relevant metric: it measures how well a model orders players by risk regardless of where the cutoff is placed. On AUC, the LSTM with attention is the best model in three of the five seasons and is

competitive in the remainder, which is why it is treated as the primary model of this thesis. The high accuracy figures should be read with the same caution: with an 11.6% injury rate, a model that predicts "not injured" for every player already scores 88.4% accuracy, so the accuracy column largely reflects the class imbalance rather than discriminative skill, and the low F1 and recall values of the higher-accuracy models confirm that they achieve their accuracy partly by under-predicting the minority (injured) class. This is also why model selection during optimisation (Section 3.2.2) targeted validation AUC rather than accuracy or F1.

4.1.2 Feature Importance and Cross-Model SHAP Analysis

This section complements the metric-based comparison of Section 4.1.1 with a unified SHAP analysis applied to all three model families. Tree-based models (DT, RF) admit an exact polynomial-time SHAP estimator via `TreeExplainer` (Lundberg et al., 2020), whereas the LSTM+Attention network is non-tree and is therefore explained with the model-agnostic `KernelExplainer` (Lundberg & Lee, 2017); this methodological asymmetry is acknowledged when interpreting absolute magnitudes, but does not affect the *rank* information on which the comparative discussion below relies. Per-season rankings were computed from the held-out test fold of each chronological split (Section 3.2.1) and aggregated into Table 11, with a single-season illustration of the LSTM+Attention model shown in Figure 5.

To quantify the degree of agreement between the three model families, the pairwise rank correlation between their mean|SHAP| feature rankings (the columns of Table 11) was computed. The rankings are strongly and significantly correlated across all three pairs: the mean pairwise Spearman correlation is $\rho = 0.905$ (Decision Tree vs. Random Forest $\rho = 0.866$; Decision Tree vs. LSTM $\rho = 0.858$; Random Forest vs. LSTM $\rho = 0.993$; all $p < 0.001$), with a corresponding mean Kendall's $\tau = 0.775$. This confirms quantitatively that the feature-importance ordering is largely *model-independent*: the same features are prioritised whether the explanation is derived from an exact tree estimator or from a model-agnostic approximation of the

recurrent network.

Figure 5

SHAP Feature Importance For The LSTM + Attention Model (Season 2023/24). `match_number`*,* `xG_roll5` *And* `cumulative_minutes` *Are The Top-ranked Features.*

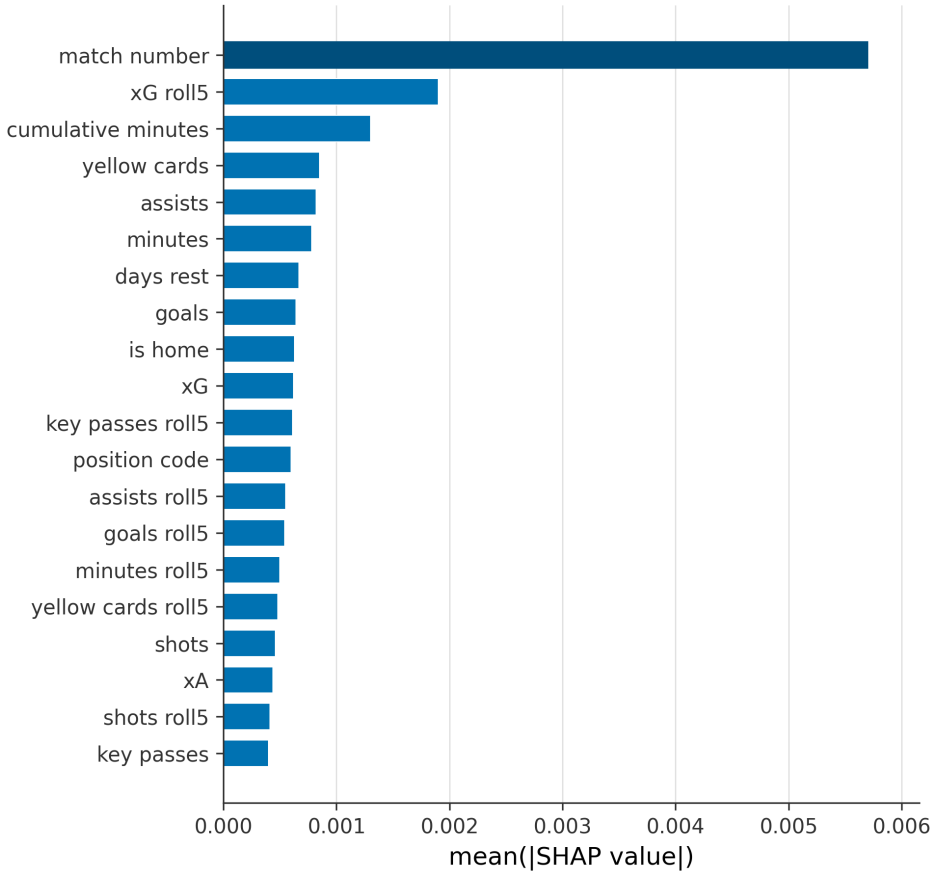


Table 11

Cross-model SHAP Feature Importance Across the Three Model Families, Averaged Over the Five EPL Seasons (2020/21–2024/25).

	Average rank (1–23)			Seasons in top-10 / 5		
Feature	**DT**	**RF**	**LSTM**	**DT**	**RF**	**LSTM**
Universally important (top-10 across all three models)						
`cumulative_minutes`	5.6	1.6	1.5	4	5	5
`xG_roll5`	2.4	3.8	4.0	5	5	5
`match_number`	9.6	1.4	1.5	3	5	5
`minutes_roll5`	5.6	4.6	4.0	5	5	5
`xA_roll5`	7.6	7.2	6.0	5	5	5
`shots_roll5`	8.4	6.8	9.0	4	5	4
`minutes`	8.2	9.0	8.0	3	3	4
`days_rest`	12.8	8.2	7.5	3	3	5
Model-specific (importance varies between families)						
`position_code`	11.0	9.2	10.0	3	3	3
`key_passes_roll5`	11.0	9.8	11.0	2	3	1
`yellow_cards_roll5`	10.0	10.6	12.0	4	2	0
`xG` (raw)	11.0	10.4	11.0	1	3	1
`xA` (raw)	8.8	11.6	13.0	2	1	0
`goals_roll5`	9.6	12.8	14.0	4	2	0
Consistently uninformative (bottom-third in all models)						
`assists_roll5`	16.4	15.6	16.0	0	0	0
`key_passes`	13.2	16.6	17.0	2	0	0
`shots` (raw)	16.2	15.4	14.0	0	0	0
`yellow_cards`	16.2	18.0	18.0	0	0	0
`is_home`	19.4	18.2	18.0	0	0	0
`goals`	15.8	20.0	21.0	0	0	0
`assists`	16.2	20.2	21.0	0	0	0
`red_cards_roll5`	21.4	22.0	22.0	0	0	0
`red_cards`	19.6	23.0	23.0	0	0	0

Note: The "Top-10" columns count how many of the five seasons placed the feature in that model's top 10. Decision-tree and random-forest ranks are derived from exact `TreeExplainer` values; LSTM+Attention ranks are derived from `KernelExplainer` approximations on the same test fold. Rank 1 corresponds To The Feature With The Highest Mean|SHAP| In The Corresponding Per-season Ranking.

Universally important features: Eight features consistently appear in the top-10 of all three models across at least three of the five seasons: `cumulative_minutes`, `xG_roll5`, `match_number`, `minutes_roll5`, `xA_roll5`, `shots_roll5`, `minutes`, and `days_rest`. The four highest-ranked features

(cumulative_minutes, xG_roll5, match_number, minutes_roll5) achieve a mean rank of 4.7 or better in every model and are placed in the top-10 by 23 of the 25 (model, season) combinations. This convergence across two methodologically very different paradigms—axis-aligned greedy splitting (DT/RF) and attention-weighted recurrent encoding (LSTM+Attention)—is the strongest evidence in this work that cumulative workload (cumulative_minutes, match_number) and recent attacking output (xG_roll5, minutes_roll5) are intrinsic injury-risk signals in match-event data rather than artefacts of any one model class. This finding is consistent with the wearable literature on accumulated training load (Gabbett, 2016; Soligard et al., 2016), even though the present study uses only public match-event statistics.

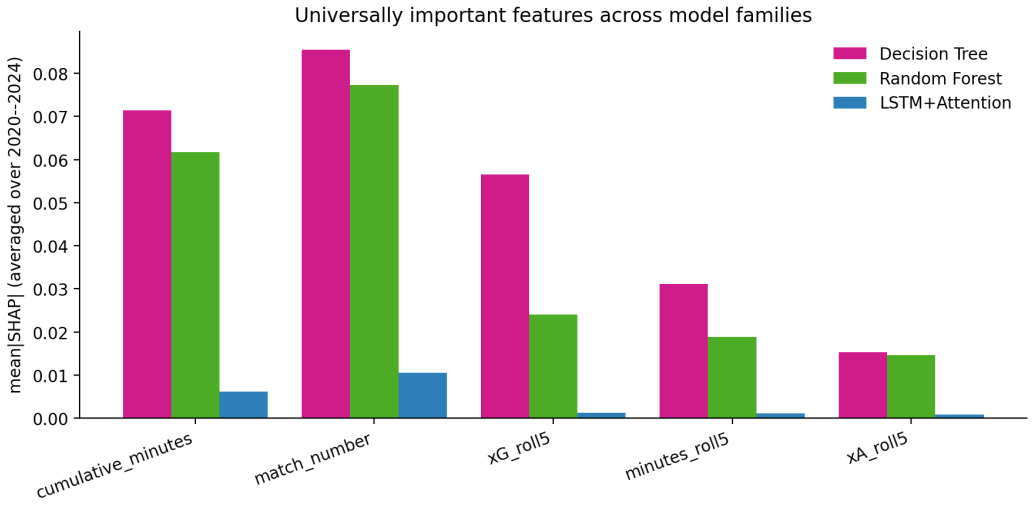
Model-specific patterns: A second band of features—yellow_cards_roll5, goals_roll5, and key_passes_roll5—reaches the top-10 in DT and RF but not in LSTM (Table 11). This divergence is interpreted as a property of the model classes themselves: tree splits can exploit a 5-match rolling mean directly as a single static covariate, whereas the LSTM already integrates short-horizon temporal structure through its lookback window (Table 6, $T \in \{3, \ldots, 10\}$), making the explicit rolling means partially redundant. Conversely, the LSTM gives higher importance to days_rest (mean rank 7.5 vs. 12.8 for DT) because it can combine days_rest with the temporal position of each match in the sequence, an interaction that tree models must approximate through deeper splits and that class_weight='balanced' alone cannot recover under the 11.6% injury base rate.

Irrelevant features and methodological implications: Eight features—red_cards, red_cards_roll5, is_home, yellow_cards (raw), goals, assists, shots (raw), and key_passes—are ranked in the bottom third by all three models in every season (top-10 count $= 0$ in Table 11). The unanimous demotion of raw goals, assists, and shots in favour of their _roll5 counterparts confirms that *recent form rather than one-off match outputs* is what drives injury risk in the model's eyes—a result that survives both the greedy-tree and the sequential-attention inductive biases. The near-zero importance of red_cards reflects its extreme sparsity in the dataset rather

than a substantive finding about disciplinary behaviour, and future work should either drop the feature or replace it with a season-level aggregate. The cross-model SHAP agreement reported here therefore not only strengthens the substantive conclusions of the study—workload and recent form dominate—but also functions as a sanity check on the feature engineering itself: features that survive SHAP scrutiny in three independent model classes are unlikely to be dataset-specific artefacts, and features that fail in all three are strong candidates for pruning in subsequent work, following the same multi-explainer discipline used in recent infrastructure-prediction studies (Sun et al., 2024).

Figure 6

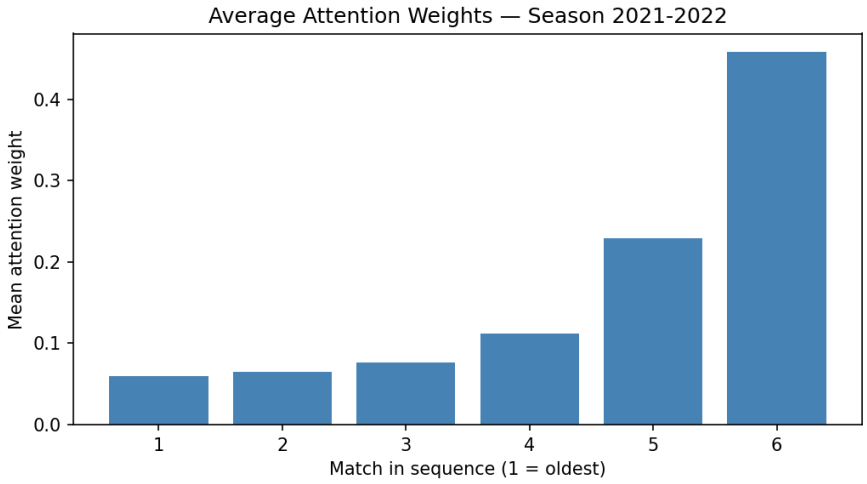
Mean|SHAP| Of The Five Universally Important Features, Averaged Over Seasons 2020/21–2024/25, For The Decision Tree, Random Forest, and LSTM+Attention Models. The same Five Features Dominate Across All Three Model Families Despite the TreeExplainer/KernelExplainer Asymmetry.



4.1.3 Attention Weight Analysis

The attention layer (Section 3.1.6) assigns each match in the lookback window a learned weight α_t that quantifies how much that match contributes to the final injury prediction. Aggregating these weights across all test sequences reveals which positions in the sequence the model relies on most. Figure 7 shows the mean attention distribution for the 2021/22 season, for which the optimal lookback window was six matches.

Figure 7
Mean attention weights across match positions for Season 2021/22 (lookback = 6). The model assigns 45.8% of its attention to the most recent match, showing a strong recency bias in injury prediction.



The distribution exhibits a pronounced recency bias: the model concentrates 45.8% of its total attention on the most recent match in the window, with the remaining weight distributed across earlier positions. Read together with the cross-model SHAP analysis of Section 4.1.2, this produces a coherent picture of how the model reasons about injury risk. The SHAP results identify *cumulative* signals—`cumulative_minutes` and `match_number`—as the dominant features, while the attention mechanism shows that the model uses the *most recent* match as the trigger point at which that accumulated risk is read out. In other words, long-term workload accumulation establishes the underlying risk level, and the latest match acts as the moment of assessment, which is consistent with the physiological intuition that injuries materialise when an already-fatigued player is exposed to one further load.

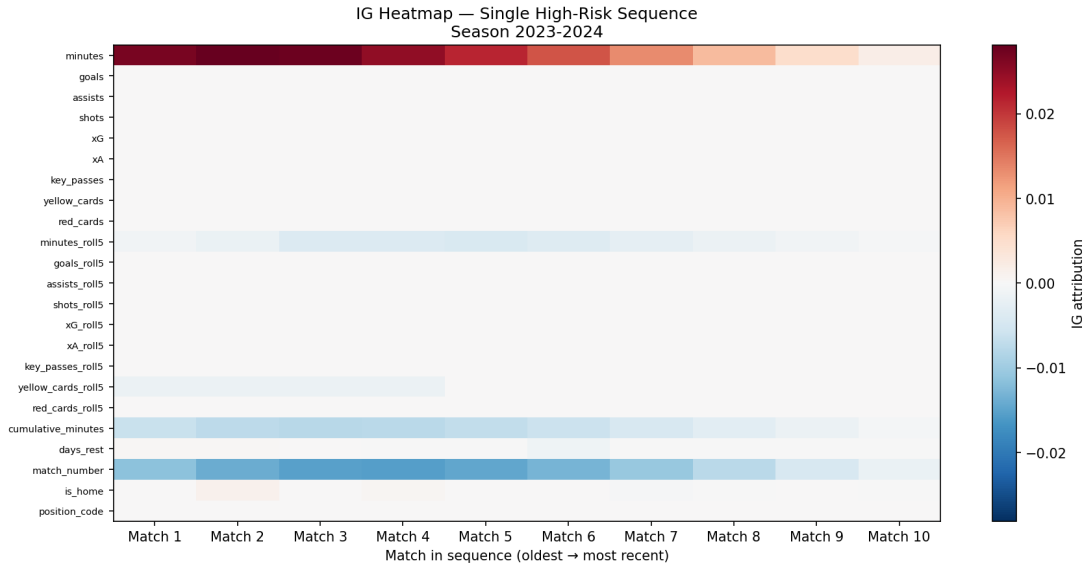
4.1.4 Integrated Gradients Analysis

While the attention analysis (Section 4.1.3) reveals which *matches* the model prioritises, and the SHAP analysis (Section 4.1.2) ranks features by their overall contribution, neither resolves the interaction between the two: which specific feature at which specific match position drives an individual prediction. Integrated Gradients (Section 3.1.7) fills this gap by assigning an attribution value to every cell of the

lookback $\times$ 23 input matrix. Figure 8 visualises this attribution for a single high-risk sequence from the 2023/24 season.

Figure 8

Integrated Gradients attribution heatmap for a single high-risk sequence (Season 2023/24, lookback = 10).



Note: Red indicates features that increase injury probability; blue indicates features that decrease it. `minutes` shows strong positive attribution across all time steps, `match_number` negative attribution.

What the heatmap shows lines up well with what the other two explainability methods already suggested, since the `minutes` feature carries strong positive attribution at nearly every match position, which confirms that sustained playing time is being read by the model as a cumulative risk-increasing signal rather than something tied to any single appearance, and in contrast `match_number` carries negative attribution throughout, suggesting that later positions in the season are, all else equal, associated with slightly lower predicted risk for this particular sequence, which is a plausible reflection of survivorship bias in that players who are still available deep into a season tend to be those who have simply handled the workload better than others. Taken as a whole, the fact that SHAP, attention weights, and Integrated Gradients all converge on the same small set of cumulative workload features is not a minor detail, since agreement across three methodologically independent attribution

techniques is itself a credibility statement, confirming that the risk factors the model has identified are stable properties of the data rather than artefacts of whichever explanation method happened to be applied.

4.2 Practical Implications

4.2.1 What a Club Could Do with This System

The framework developed in this thesis runs entirely on public data, which means that any club or analyst with access to Understat and Transfermarkt could realistically deploy it without investing in specialised hardware or negotiating proprietary data agreements, and the day-to-day workflow is straightforward enough that at the start of each matchweek the analyst simply downloads the most recent player appearance records, updates the 23-feature sequence for each active squad member, and lets the trained LSTM+Attention model generate a 30-day injury probability $\hat{y}$ for each player, with those exceeding a threshold calibrated by the club's medical staff to their preferred balance between sensitivity and specificity being flagged for additional assessment, targeted recovery sessions, or more conservative substitution timing, beyond the workflow itself, perhaps the most significant practical advantage of this approach is that it requires no hardware, no data-sharing agreements, and no proprietary subscriptions, making it just as relevant for academies, lower-division clubs, and independent researchers as it is for top-flight clubs that happen to lack GPS infrastructure, since anyone who can access two freely available websites can in principle run the same pipeline.

4.2.2 Interpreting AUC in the Context of Actionability

The AUC values achieved in this study, ranging from 0.44 to 0.75 with most models falling between 0.55 and 0.66, are substantially lower than those reported by single-club GPS-based studies (0.89–0.92; Rossi et al. (2018) and Saberisani et al. (2025)) though they are broadly comparable to what the public-data literature has consistently shown, and Chang et al. (2024) achieved similar AUC levels using

WhoScored statistics on a single season, suggesting that this performance ceiling has been observed independently across multiple studies. For clinical decision support, an AUC in the upper part of this range at around 0.65 to 0.75 is generally considered to offer actionable discrimination, since it means that in a direct comparison between a player who will be injured and one who will not the model ranks the at- risk player higher 65 to 75% of the time, which is meaningfully better than the 50% expected from random selection, and the more important question in any case is not whether this matches what GPS-based systems can do but whether it is better than the informal judgment currently applied by medical staff at clubs without sensor access, since for that segment of the football ecosystem which represents the majority of professional clubs worldwide a public-data model at AUC 0.65 may already provide more structured decision support than anything currently available to them.

4.2.3 Limitations of the Performance Envelope

The comparison with GPS-based approaches makes clear that there is a genuine cost to working with public data, since GPS-based workload metrics including total distance, high-speed running distance, sprint counts, and ACWR are far more direct physiological drivers of injury than the match-event statistics used here, and their absence likely accounts for most of the performance gap, and Zhao et al. (2026) found that an attention-based transformer outperformed SVM, XGBoost, and a standard LSTM for injury prediction on biomechanical motion-capture data, indicating that attention-based temporal modelling can improve discrimination over both classical and recurrent baselines, though in their study this was achieved with rich motion-capture features rather than public match data, which is an important distinction since it means the gap to wearable-data performance is unlikely to be fully closed by architectural sophistication alone and will probably require richer input features to bridge, which is precisely why the multimodal fusion direction explored by Li et al. (2025), combining complementary modalities such as medical imaging and biomechanical sensor data, represents one of the more promising avenues for future work.

The survival modelling approach of Catterall et al. (2026) represents another important alternative framing: rather than predicting binary injury/no-injury within a fixed 30-day window, a survival model predicts the time-to-next-injury distribution, which provides richer information for load management decisions and is more robust to the label-construction choices (window length, announcement-date vs tissue-level timing) that affect binary classification labels in crowd-sourced datasets.

CHAPTER 5
CONCLUSION AND IMPLICATIONS

5.1 Summary of Findings

This thesis set out to determine whether publicly available match-event statistics, collected without any wearable sensors or proprietary club data, can support statistically meaningful and interpretable injury prediction at the player-match level across multiple seasons of elite football, and the four research questions posed in Chapter 1 are answered explicitly below.

RQ1 — Public-data sufficiency. Short answer is yes, but with important caveats Using 47,477 player-match records from 1,017 EPL players across five seasons 2020/21 to 2024/25, AUC values ranged from 0.44 to 0.75 across all models and seasons, with most results falling between 0.55 and 0.66, and The LSTM with attention won the AUC ranking in three of the five seasons, reaching peak of 0.654 in 2021/22, while the Random Forest achieved the highest single-season AUC in the study 0.746 in 2023/24, confirming that ensemble baselines remain competitive once class imbalance is addressed with `class_weight='balanced'`. Performance was not uniformly above chance, however, since in 2023/24 the standalone LSTM without attention scored an AUC of 0.443 falling below the 0.50 random baseline as discussed in Section 4.1.1), and taken together these results indicate that public match-event data does contain a genuine but modest statistical signal for injury risk, with the typical AUC of roughly 0.55 to 0.66 and the best-case of 0.75 falling substantially below the 0.89 to 0.92 values reported by GPS-based single-club studies, wich reflects the genuine information gap between event-based and sensor-based load metrics.

RQ2 — Temporal modelling advantage. This was partially supported, as the LSTM with attention led on AUC in three of five seasons and remained competitive in the remaining two, suggesting that sequential temporal modelling does confer a

discriminative advantage over static tree-based baselines, though that advantage is not universal since the standalone LSTM achieved a higher AUC than LSTM with attention in 2022/23 at 0.609 versus 0.600, and the Random Forest outperformed both LSTM variants in 2023/24 at 0.746 versus 0.656, and the temporal advantage appears largest when the lookback window is sufficiently long, specifically where (LOOKBACK $\geq$ 6) to capture multi-match accumulation patterns, and weakest when Optuna selects a short lookback of LOOKBACK = 3 in 2022/23, which limits the LSTM's sequential advantage over tree models that can exploit the same information through rolling-mean features.

RQ3 — Granularity value. The match-level and season-level formulations proved to be complementary rather than directly comparable, since they are not solving the same prediction problem, and season-level classification achieves a higher mean AUC of 0.681 to 0.698 primarily because its positive-class prevalence sits at 45.9%, making it a much easier statistical problem, and also because it can only be evaluated retrospectively once the season is complete, whereas match- level prediction operating at 11.6% prevalence generates rolling forecasts that are available from the earliest gameweeks of the season, enabling real- time load management decisions, and crucially it is the only formulation that preserves the sequential workload signals including `cumulative_minutes`, `match_number`, and rolling performance means that the cross-model SHAP analysis identifies as the primary injury risk indicators.

RQ4 — Explainability consistency. The answer here is yes, and with high confidence, since across all three model families and all five seasons SHAP analysis consistently identifies cumulative workload features such as (`cumulative_minutes` and `match_number`) alongside recent performance load including (`minutes_roll5`, `xG_roll5`, and `xA_roll5`) as the universally top-ranked features, and the LSTM attention weight analysis confirms a strong but non-exclusive recency focus with 45.8% of attention mass concentrated on the most recent match while significant weight of approximately 40% is still retained by the two to three preceding appearances, and Integrated Gradients at the instance level further confirms the SHAP

narrative since `minutes` receives the strongest positive attribution across all time steps for the highest-risk sequences while `match_number` receives consistent negative attribution that reflects the model's implicit capture of season-progression dynamics, meaning that the three-method framework converges on a single workload-based causal story that is grounded in established sports physiology.

5.2 Contributions

This thesis makes four original contributions to the sport field of injury prediction where each of which was previewed in Chapter and are substantiated in detail below with implications that extend beyond the specific dataset studied.

1. **Scalable public-data benchmark.** Rather than relying on the single-club, single-season datasets that dominate the literature, this work assembles a rigorously constructed five- season, 47,477-record EPL injury prediction dataset entirely from two public sources through a documented three-stage pipeline, and because it spans 1,017 players across five consecutive seasons and is fully reproducible from freely available data, it offers a reference point against which future methodological improvements can be measured without running into the access barriers that have historically made progress in this field so difficult.

2. **Chronologically evaluated LSTM with attention for public-data injury prediction.** An LSTM network augmented with an additive attention layer is applied to per-player match sequences under strict chronological per-player train/validation/test splits that prevent temporal leakage. This is, to the author's knowledge, the largest-scale chronologically validated application of sequential neural networks to EPL injury prediction using exclusively public data. The architecture achieves the best AUC in three of five seasons and demonstrates that match-sequence modelling extracts information that static tree models cannot fully replicate.

3. **Methodologically principled match-level versus season-level framing.** By

evaluating both temporal granularities on the same data with the same models, this thesis provides the first head-to-head comparison that controls for model class and feature set. The analysis demonstrates that the apparent superiority of season-level prediction on AUC is an artefact of its higher class prevalence rather than a genuine signal advantage, and establishes the conceptual case for match-level rolling prediction as the operationally relevant formulation for in-season risk management.

4. **Three-tier explainability framework.** By combining SHAP across three model families, attention weight visualisation, and Integrated Gradients, the thesis produces a multi-method explainability picture that triangulates from population- to instance-level attribution. The convergence of all three methods on cumulative workload features provides stronger evidence for the workload–injury relationship than any single method could establish alone, and demonstrates that the relationship identified by the wearable literature is also recoverable from public event data.

5.3 Limitations

Several limitations of this work should be acknowledged explicitly, as they define the boundary conditions under which the findings are valid and motivate the future directions proposed in Section 5.4.

Absence of GPS-derived workload metrics: The primary limitation is the absence of the physical load metrics that wearable-based studies use as direct injury precursors: total distance covered, high-speed running distance, sprint counts, acceleration and deceleration loads, and ACWR (Impellizzeri et al., 2019). The present features are indirect proxies for physical workload at best: `minutes` approximates exposure time but not intensity, and `xG_roll5` reflects attacking involvement rather than physical exertion. The AUC ceiling of approximately 0.65 to 0.75 reflects this feature poverty, and the gap to sensor-based models cannot be fully closed by architectural improvements alone, which means the study deliberately traded feature

depth for accessibility and scale but this trade- off has a measurable cost in predictive power.

Crowd-sourced injury label quality: Injury labels were derived from Transfermarkt, a crowd-sourced platform whose injury records are updated based on publicly announced injury news rather than direct clinical reporting, and this introduces two forms of noise. First, minor injuries such as muscle tightness and minor knocks that do not lead to match absences are systematically under-reported, potentially omitting injuries that occurred during the 30-day lookahead window, and since injuries frequently develop incrementally with a minor issue becoming more serious under continued workload, these silent injuries may be the most relevant to early-warning systems, meaning their absence in the labels biases the model towards capturing only the most severe cases. Second, Transfermarkt records the date on which an injury is *publicly announced* rather than the date of actual tissue damage, which may lag by days or weeks particularly for muscle injuries that are often managed conservatively before disclosure, and this annotation-time versus occurrence-time discrepancy is especially problematic for the LSTM since it treats the match sequence as a precisely ordered temporal signal, meaning the 30-day lookahead window may misalign the causal workload antecedents with the injury label.

Static position encoding: The `position_code` feature is a static, numerically encoded playing position that does not capture the actual positional demands of each individual match, and modern football involves frequent tactical repurposing such as a winger playing as a false nine, a central midfielder deployed as a box-to-box runner, or a fullback used as a wing-back in a 3-5-2 formation, all of which carry substantially different physical load profiles that the static code cannot encode, and since position-specific injury risk profiles are well- established in the literature (Thema et al., 2025) a dynamic position label derived from tracking data would likely improve both prediction performance and the specificity of the feature importance findings.

No inter-match training load: The dataset contains no information about training sessions, gym work, or player wellness between matches, and all contextual

features such as `days_rest` and `is_home` are derived from the match schedule rather than from actual inter-match activities, meaning that a player with a three-day rest window before their next match may have conducted a high- intensity training session two days prior or may have undergone complete rest, and since the model cannot distinguish between these scenarios this represents a fundamental constraint of any public-data approach, given that the inter- match period is the primary opportunity for recovery interventions and its absence from the feature set limits what any model built on public data alone can realistically capture.

14% record loss from name matching: The cross-source matching pipeline removed 14.0% of Understat records that could not be matched to Transfermarkt entries. While post-hoc inspection confirmed that these failures were primarily attributable to absent player coverage rather than data quality errors, the removed records are not missing at random: they likely over-represent players with non-Latin name characters (who are disproportionately affected by normalisation failures at Tier 2) and short EPL careers (who appear in fewer seasons and therefore are less likely to be covered consistently by both platforms). If these player groups have systematically different injury risk profiles, the dataset may be subject to selection bias.

Single-run estimates and absence of significance testing: Each model was trained once per season and the reported metrics are therefore single point estimates rather than averages over repeated runs. Because the LSTM is stochastic—its weights are randomly initialised and dropout is applied during training—its results are subject to run-to-run variation that was not quantified here. The differences between models on AUC are in several seasons small (for example, 0.654 versus 0.640 in 2021/22), and without repeated runs with different random seeds it cannot be established that these differences are statistically significant rather than the product of initialisation noise which means that the per-metric "winner" counts in Table 10 should be read as indicative rather than as statistically confirmed rankings, and repeating each experiment across multiple seeds while reporting mean and standard deviation

together with paired significance tests between models would place the comparative claims on a firmer footing, representing a clear methodological improvement for future work.

5.4 Future Work

The limitations identified above and the new questions raised by the present findings, suggest several concrete directions for future research:

1. **Transformer and attention-based architectures.** Zhao et al. (2026) demonstrated that a Hybrid Transformer with multi-head attention outperformed standard LSTM in an analogous injury prediction setting, achieving higher AUC with improved temporal interpretability through the attention distribution. Applying transformer-based sequence models to the present five-season public dataset would provide a direct comparison point and could close part of the performance gap to GPS-based studies, particularly if the multi-head structure allows the model to simultaneously attend to recency (recent workload) and history (seasonal load accumulation).

2. **Survival modelling for time-to-injury forecasting.** The present binary classification formulation treats all injuries within a 30-day window equally, regardless of whether the injury occurred one day or 29 days after the match. Catterall et al. (2026) applied the DeepHit competing-risks survival model to elite female football data and showed that time-to-injury predictions provide richer and more calibrated information for load management decisions than binary labels. Adapting the present public-data pipeline to a survival modelling framework would also address the label-boundary sensitivity issue identified in the limitations, since the outcome would be the waiting time to injury rather than a window-specific binary indicator.

3. **Multimodal data fusion.** Li et al. (2025) combined medical imaging features with biomechanical time-series data using attention-based Swin-UNet

and LSTM architectures, demonstrating that multimodal fusion substantially improved prediction performance over any single modality. A natural extension of the present work would integrate publicly available contextual signals—opponent quality ratings, travel distance and duration, weather conditions, or player contract status as a proxy for motivation, with the match-event features used here, although these additional signals fall short of GPS sensor data because they are accessible and may add orthogonal predictive information not captured by workload statistics alone.

4. **Cross-league and cross-season generalisation testing.** The present study is intentionally scoped to the EPL to control for between-league confounders, and a direct test of the model's generalisability would involve training on EPL seasons and evaluating on La Liga or Bundesliga data, where Understat and Transfermarkt provide equivalent coverage, since such a study would reveal whether the workload features identified as universally important in Table 11transfer across playing styles and competitive environments and would provide a stronger assessment of the dataset's replicability as a public benchmark.

5. **Player-specific recalibration.** The present models are trained on pooled player data and apply the same injury probability function to all players, yet injury susceptibility varies substantially between individuals due to age, position, prior injury history, and physiological recovery capacity, and a hierarchical modelling approach that trains a population-level model and fine-tunes player-specific layers from individual historical sequences could improve both discrimination and calibration by allowing the model to learn idiosyncratic risk profiles without requiring the large within-player sample sizes that individual models would need to train from scratch.

6. **Label quality improvement.** The noise introduced by Transfermarkt's announcement-date labels and incomplete minor-injury coverage could be

partially addressed by combining Transfermarkt records with additional public sources such as official club injury bulletins, press conference transcripts, or the CIES Football Observatory match absence records, and a data fusion pipeline that triangulates across multiple crowd-sourced and official records would produce more accurately timed and more complete injury labels, thereby reducing one of the most fundamental sources of error in the present training set.

REFERENCES

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

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

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Wadsworth; Brooks/Cole.
ISBN: 978-0412048418.

Bullock, G. S., Mylott, J., Hughes, T., Nicholson, K. F., Riley, R. D., & Collins, G. S. (2022). Just how confident can we be in predicting sports injuries? A systematic review of the methodological conduct and performance of existing musculoskeletal injury prediction models in sport. *Sports Medicine*, *52*(10), 2469–2482. https://doi.org/10.1007/s40279-022-01698-9

Catterall, V., Midoglu, C., & Lynch, S. (2026). Time-to-injury forecasting in elite female football: A DeepHit survival approach [DeepHit survival model on the public SoccerMon dataset (37 players, two seasons); concordance index 0.762; SHAP interpretability]. *arXiv preprint arXiv:2601.19479*.

Chang, V., Sajeev, S., Xu, Q. A., Tan, M., & Wang, H. (2024). Football analytics: Assessing the correlation between workload, injury and performance of football players in the english premier league. *Applied Sciences*, *14*(16), 7217. https://doi.org/10.3390/app14167217

Courty, P. (2022). *An economic approach to sports injury policies* (Working Paper). University of Victoria, Department of Economics.

Dallmeyer, S., Steinfeldt, H., Hübers, T., Pietzonka, M., & Breuer, C. (2025). Monetising misfortune: The financial consequences of injuries in professional football teams. *BMJ Open Sport & Exercise Medicine*, *11*(2).

Den Hollander, S., Kerkhoffs, G., & Gouttebarge, V. (2024). The impact of match workload and international travel on injuries in professional men's football. *Sports*, *12*(8), 212. https://doi.org/10.3390/sports12080212

Ekstrand, J., Hägglund, M., & Waldén, M. (2011). Epidemiology of muscle injuries in professional football (soccer). *The American Journal of Sports Medicine*, *39*(6), 1226–1232.

Frazier, P. I. (2018). A tutorial on bayesian optimization. *arXiv preprint arXiv:1807.02811*.

Gabbett, T. J. (2016). The training-injury prevention paradox: Should athletes be training smarter and harder? *British Journal of Sports Medicine*, *50*(5), 273–280.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Howden Sport and Entertainment. (2025, December). *Men's European football injury index 2024/25* (tech. rep.). Howden Group Holdings.

Impellizzeri, F. M., Marcora, S. M., & Coutts, A. J. (2019). Internal and external training load: 15 years on. *International Journal of Sports Physiology and Performance*, *14*(2), 270–273. https://doi.org/10.1123/ijspp.2018-0935

Jones, S., Almousa, S., Gibb, A., Allamby, N., Mullen, R., Andersen, T. E., & Williams, M. (2019). Injury incidence, prevalence and severity in high-level male youth football: A systematic review. *Sports Medicine*, *49*, 1879–1899. https://doi.org/10.1007/s40279-019-01170-3

Leckey, C., Van Dyk, N., Doherty, C., Lawlor, A., & Delahunt, E. (2025). Machine learning approaches to injury risk prediction in sport: A scoping review with evidence synthesis. *British Journal of Sports Medicine*, *59*(8), 491–500. https://doi.org/10.1136/bjsports-2024-108576

Li, S., Hou, Z., Amjad, K., & Mushtaq, H. (2025). Multimodal fusion of medical imaging and biomechanical data using attention-based Swin-UNet and LSTM for sports injury prediction. *Frontiers in Physiology*, *16*. https://doi.org/10.3389/fphys.2025.1687895

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, *2*(1), 56–67. https://doi.org/10.1038/s42256-019-0138-9

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *30*.

Mahmood, A. (2025). The role of AI and wearables in injury prediction and performance optimization. *Premier Journal of Artificial Intelligence*, *4*, 100019.

Majumdar, A., Bakirov, R., Hodges, D., Scott, S., & Rees, T. (2022). Machine learning for understanding and predicting injuries in football. *Sports Medicine – Open*, *8*, 73. https://doi.org/10.1186/s40798-022-00465-4

Martins, F., Sarmento, H., Gouveia, É. R., Saveca, P., & Przednowek, K. (2025). Machine learning-based prediction of muscle injury risk in professional football: A four-year longitudinal study. *Journal of Clinical Medicine*, *14*(22), 8039.

Molnar, C. (2022). *Interpretable machine learning: A guide for making black box models explainable* (2nd) [https://christophm.github.io/interpretable-ml-book/]. Independently published.
ISBN: 979-8411463330.

Piłka, T., Grzelak, B., Sadurska, A., Górecki, T., & Dyczkowski, K. (2023). Predicting injuries in football based on data collected from GPS-based wearable sensors. *Sensors*, *23*(3), 1227. https://doi.org/10.3390/s23031227

Rago, V., Brito, J., Figueiredo, P., Costa, J., Barreira, D., Krustrup, P., & Rebelo, A. (2020). Methods to collect and interpret external training load using microtechnology incorporating GPS in professional football: A systematic review. *Research in Sports Medicine*, *28*(3), 437–458. https://doi.org/10.1080/15438627.2019.1686703

Rossi, A., Pappalardo, L., Cintia, P., Iaia, F. M., Fernàndez, J., & Medina, D. (2018). Effective injury forecasting in soccer with GPS training data and machine learning. *PLoS ONE*, *13*(7), e0201264. https://doi.org/10.1371/journal.pone.0201264

Saberisani, R., Barati, A. H., Zarei, M., Santos, P., Gorouhi, A., Ardigò, L. P., & Nobari, H. (2025). Prediction of football injuries using GPS-based data in Iranian professional football players: A machine learning approach. *Frontiers in Sports and Active Living*, *7*, 1425180. https://doi.org/10.3389/fspor.2025.1425180

Soligard, T., Schwellnus, M., Alonso, J.-M., Bahr, R., Clarsen, B., Dijkstra, H. P., Gabbett, T., Gleeson, M., Hägglund, M., Hutchinson, M. R., et al. (2016). How much is too much? (Part 1) International Olympic Committee consensus statement on load in sport and risk of injury. *British Journal of Sports Medicine*, *50*(17), 1030–1041.

Sun, X., Wang, H., & Mei, S. (2024). Explainable highway performance degradation prediction model based on LSTM. *Advanced Engineering Informatics*, *61*, 102539. https://doi.org/10.1016/j.aei.2024.102539

Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. *In proceeding book of the 34th International Conference on Machine Learning (ICML)*, *70*, 3319–3328.

Thema, M. T., Jacobs, S., van den Berg, L., Strauss, A., & Mahlangu, M. P. (2025). The role of playing position in soccer injury characteristics: Evidence from sub-elite athletes. *Frontiers in Sports and Active Living*, *7*, 1542300. https://doi.org/10.3389/fspor.2025.1542300

Tierney, P., Clarke, N., & Roberts, S. (2026). Use and application of wearable technology in football further education settings in the UK. *Sport, Education and Society*, *31*(2), 296–309.

Transfermarkt GmbH & Co. KG. (2024). Transfermarkt — football statistics, player profiles and injury records [Accessed: 2024-XX-XX].

Understat. (2024). Understat — football statistics and expected goals [Accessed: 2024-XX-XX].

Wang, Y., Huang, M., Zhu, X., & Zhao, L. (2016). Attention-based LSTM for aspect-level sentiment classification. *In proceeding book of the 2016 Conference on Empirical Methods in Natural Language Processing*, 606–615 ISBN: 978-1-945626-25-8.

Zhao, X., Chan, V., & Graham, R. B. (2026). From classical models to attention-based transformers: A comparative study of injury prediction pipelines in female varsity soccer [Compares SVM, XGBoost, LSTM and a Hybrid Transformer with attention-based temporal interpretability on 25 athletes using leave-one-subject-out CV]. *Journal of Biomechanics*. https://doi.org/10.1016/j.jbiomech.2026.112730

APPENDICES

APPENDIX A: SIMILARITY REPORT

10% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Quoted Text
- Small Matches (less than 10 words)

Match Groups

47 Not Cited or Quoted 9%
Matches with neither in-text citation nor quotation marks

3 Missing Quotations 1%
Matches that are still very similar to source material

0 Missing Citation 0%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

8% Internet sources
4% Publications
6% Submitted works (Student Papers)

Integrity Flags

1 Integrity Flag for Review

Replaced Characters
13 suspect characters on 5 pages
Letters are swapped with similar characters from another alphabet.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Match Groups

- **47** Not Cited or Quoted 9%
  Matches with neither in-text citation nor quotation marks
- **3** Missing Quotations 1%
  Matches that are still very similar to source material
- **0** Missing Citation 0%
  Matches that have quotation marks, but no in-text citation
- **0** Cited and Quoted 0%
  Matches with in-text citation present, but no quotation marks

Top Sources

8% Internet sources
4% Publications
6% Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Internet	
www.final.edu.tr		<1%
2	Student papers	
The Scientific & Technological Research Council of Turkey (TUBITAK)		<1%
3	Internet	
docslib.org		<1%
4	Student papers	
Associatie K.U.Leuven		<1%
5	Internet	
www.ecologysci.com		<1%
6	Internet	
dspace.iiti.ac.in		<1%
7	Publication	
Su, Juntao. "Explainable and Robust AI for Medical Applications.", The George Wa...		<1%
8	Student papers	
Strathmore University (Main Account)		<1%
9	Internet	
research.chalmers.se		<1%
10	Internet	
arxiv.org		<1%

11	Internet	
core.ac.uk		<1%
12	Internet	
www.mdpi.com		<1%
13	Student papers	
University of Bristol		<1%
14	Student papers	
University of New South Wales		<1%
15	Student papers	
University of Liberal Arts Bangladesh		<1%
16	Internet	
careerskills.interviewgemini.com		<1%
17	Student papers	
Liverpool John Moores University		<1%
18	Student papers	
Munster Technological University (MTU)		<1%
19	Publication	
Zhao, Haoran. "Search for Charged Higgs Bosons in Multi-Lepton and Dijet Final S...		<1%
20	Student papers	
Dublin Business School		<1%
21	Student papers	
Asia Pacific University College of Technology and Innovation (UCTI)		<1%
22	Internet	
brill.com		<1%
23	Student papers	
Georgia Southern University		<1%
24	Internet	
epub.ub.uni-muenchen.de		<1%

25 Internet
erepo.uef.fi <1%

26 Internet
medium.com <1%

27 Internet
repository.sharif.edu <1%

28 Internet
www.itspy.cz <1%

29 Internet
www.matscisci.com <1%

30 Internet
www.tutorialspoint.com <1%

31 Publication
Herbers, Cara. "Leveraging Machine Learning Tools to Develop Objective, Interpr... <1%

32 Internet
dspace.bracu.ac.bd <1%

33 Internet
www.astesj.com <1%

34 Publication
Hossein Mirzaee, Serveh Kamrava. "Estimation of internal states in a Li-ion batter... <1%

35 Publication
Shashank Kumar Singh, Amrita Chaturvedi. "An efficient multi-modal sensors fea... <1%

36 Internet
mediatum.ub.tum.de <1%

37 Internet
repository.ubn.ru.nl <1%

38 Internet
vdoc.pub <1%

39 Internet

web.ist.utl.pt <1%

40 Internet

www.preprints.org <1%

41 Internet

www.repositorio.ufal.br <1%

Multi-Season Injury Prediction in Football Using Machine Learning and Open Data

Abdalrahman Siddig Ali Bashir

Multi-Season Injury Prediction in Football Using Machine Learning and Open Data

MS 2026

Abdalrahman Siddig Ali Bashir

Final International University
June 2026
Girne, TRNC

APPENDIX B: STATEMENT OF ARTIFICIAL INTELLIGENCE USAGE

FINAL INTERNATIONAL UNIVERSITY
STATEMENT OF ARTIFICIAL INTELLIGENCE USAGE

To the Institute of Graduate Education of Final International University,

I hereby declare that the subject matter, methodology, analyses, interpretations, and conclusions of this research are entirely my own work. I accept full academic, ethical, and legal responsibility for this study. Furthermore, I declare and undertake that I have not used artificial intelligence at any stage of the aforementioned processes.

Sincerely,

Full Name: **ABDALRAHMAN SIDDIG ALI BASHIR**

Student ID Number: **2410140001**

Signature:

Date: **07/24/2026**