https://doi.org/10.1007/s00167-022-07082-4

Pre-injury performance is most important for predicting the level of match participation after Achilles tendon ruptures in elite soccer players: a study using a machine learning classifier

Pedro Diniz[1,2,3,4] · Mariana Abreu[2,5] · Diogo Lacerda[1] · António Martins[1,4] · Hélder Pereira[6,7,8] · Frederico Castelo Ferreira[2,3] · Gino MMJ Kerkhoffs[9,10,11] · Ana Fred[2,5]

Received: 11 April 2022 / Accepted: 18 July 2022

Abstract

Purpose Achilles tendon ruptures (ATR) are career-threatening injuries in elite soccer players due to the decreased sports performance they commonly inflict. This study presents an exploratory data analysis of match participation before and after ATRs and an evaluation of the performance of a machine learning (ML) model based on pre-injury features to predict whether a player will return to a previous level of match participation.

Methods The website *transfermarkt.com* was mined, between January and March of 2021, for relevant entries regarding soccer players who suffered an ATR while playing in first or second leagues. The difference between average minutes played per match (MPM) 1 year before injury and between 1 and 2 years after the injury was used to identify patterns in match participation after injury. Clustering analysis was performed using *k*-means clustering. Predictions of post-injury match participation were made using the XGBoost classification algorithm. The performance of this model was evaluated using the area under the receiver operating characteristic curve (AUROC) and Brier score loss (BSL).

Results Two hundred and nine players were included in the study. Data from 32,853 matches was analysed. Exploratory data analysis revealed that forwards, midfielders and defenders increased match participation during the first year after injury, with goalkeepers still improving at 2 years. Players were grouped into four clusters regarding the difference between MPMs 1 year before injury and between 1 and 2 years after the injury. These groups ranged between a severe decrease ($n=34$; -59 ± 13 MPM), moderate decrease ($n=75$; -25 ± 8 MPM), maintenance ($n=70$; 0 ± 8 MPM), or increase ($n=30$; 32 ± 13 MPM). Regarding the predictive model, the average AUROC after cross-validation was 0.81 ± 0.10, and the BSL was 0.12, with the most important features relating to pre-injury match participation.

Conclusion Most players take 1 year to reach peak match participation after an ATR. Good performance was attained using a ML classifier to predict the level of match participation following an ATR, with features related to pre-injury match participation displaying the highest importance.

Level of evidence I.

Keywords Achilles tendon · General sports trauma · Football (soccer) · Epidemiology · Statistics · Machine learning

Abbreviations

AT	Achilles tendon
ATR	Achilles tendon rupture
AUROC	Area under the receiver operating characteristic curve
ML	Machine learning
RTP	Return to play
ΔMPM	Difference between average minutes played per match during Year 1 and Year − 1

Introduction

Achilles tendon ruptures (ATR) are career-threatening injuries in elite soccer players. Unfortunately, despite a relatively high return to play (RTP) rate, 96%, according to Grassi et al. [11], 18% of players will not return to the same level of competition within two seasons following

✉ Pedro Diniz
pedro.diniz@scml.pt

Extended author information available on the last page of the article

Published online: 09 August 2022

injury [34]. Furthermore, previous research has also shown that soccer players suffering from these injuries have their careers shortened, on average, by two seasons compared to matched controls [30].

Several studies reporting outcomes of ATRs in elite athletes are based on publicly available information [11, 13, 15, 24, 26, 30, 33, 34]. In soccer, one notable source is *transfermarkt.com* [10, 11, 16, 21, 34, 36], which has been considered accurate, regarding injury denomination and location, in 89% of cases [8, 16]. Although primarily aimed at aggregating player market values and transfer fees, it includes other valuable data for sports analytics, such as match results, player performance indicators (namely goals, assists, and fouls), and injury history. This database is publicly available and maintained by *transfermarkt.com* and its user community [32].

Artificial intelligence is a field that studies artificial agents that can mimic or surpass human-level intelligent tasks and has become increasingly popular in the past decade [7]. Machine Learning (ML) is a subset of artificial intelligence related to “advanced statistical techniques that use computer algorithms to model complex relationships between variables”, with these computer algorithms *learning* automatically from *experience*, i.e. data, without direct human intervention [20]. These algorithms rely on data analysis models to uncover hidden patterns and other meaningful insights from large datasets [28]. Among these algorithms, one can find both unsupervised and supervised learning methods [1]. Unsupervised learning is used when “labels” are unavailable [1], i.e. individual instances in the dataset are not categorized. These algorithms can organize individual instances according to naturally emerging patterns in the dataset, detect anomalous patterns and perform dimensionality reduction [1, 7]. Supervised learning is used when data are “labeled”, i.e. the algorithm is fed training data where individual instances—observations—and corresponding output values, obtained with human intervention, are known [1]. Regression and classification problems are the two main categories into which supervised learning can be divided [1, 7].

Despite recent advances in the characterization of consequences of ATRs for elite soccer players [10, 11, 30, 34], both an evaluation of how match participation evolves after injury and a set of prognostic tools to gauge the likelihood of return to the same level of play are still missing in the literature. In addition, previous studies of elite soccer players treated for ATRs have also been limited by their reduced number of cases under consideration [10, 30, 34], by being restricted to a single league [10, 30], or by missing performance measures besides the return to play at the same competitive level [11].

This study has a double objective. Firstly, an exploratory data analysis aims to inform athletes and staff how match participation evolves after ATRs. Secondly, it evaluates the performance of an ML model based on pre-injury features to predict whether a player would return to a similar level of match participation, together with a study of the most relevant features for this task.

Materials and methods

Player screening and selection

The website *transfermarkt.com* (Transfermarkt, Hamburg, Germany) was mined, between January and March of 2021, for relevant entries regarding soccer players who suffered an ATR while playing in first or second leagues.

A customized web scraper was developed using Scrapy [22]. Player screening and selection were carried out using the following scheme: fir. Firstly, a list of all first and second leagues across the world was manually compiled; secondly, team rosters for each team in each league, since season 2007/2008, were extracted to a list; finally, the injury data of each player in the list were retrieved. The resulting injury data were filtered for entries containing the string “Achilles tendon rupture” or “Achilles” combined with more than 90 days of absence. Another group of players with absence times of more than 90 days was built from the following strings: “calf”, “leg”, and “ankle tendon”.

Each entry was then evaluated independently by two researchers. Only players with club reports, press releases, or interviews mentioning a complete ATR were eligible for inclusion. A minimum follow-up of 24 months was also required. Due to the COVID-19 pandemic and ensuing match calendar rearranging, only injuries occurring before 31st March^{t} of 2018 were included. Players that suffered partial or focal tears of the Achilles, and players that suffered an ATR while playing for teams not in first or second leagues, or were unaffiliated with any team at the moment of injury, were excluded. Disagreements were settled by discussion with a third researcher on a case-by-case basis.

Data extraction and dataset handling

The *transfermarkt.com* website was also scraped for the following items: date of birth, height, preferred foot, playing position, club transfers (including projected market values and transfer fees), whether the player had played for the national team (at any time during the player’s career), date of clearance for unrestricted practice, and match participation data (as minutes on the playing field; for the season of injury, the preceding season, and the two seasons following injury). Specific match participation data included: minutes played, whether the player was in the starting team, whether the player did not play but sat on the bench, and the reason for not playing (medical injuries, coach choice, or other). Data were anonymized, pooled into a database, inspected, and formatted for consistency. In cases where players sustained bilateral ruptures, the first rupture was considered the index event.

Dealing with missing data

Missing data regarding minutes played per match were imputed using spline interpolation. In addition, missing values regarding categorical features related to match participation (reason for player absence from the playing field and whether the player was in the starting eleven) were imputed using backfilling. Of note, less than 0.01% of matches had missing information.

Feature engineering

The following features were computed from the available data: age at rupture, relative market value (obtained from the division of the player's market value by the squad total market value), whether a re-rupture or a contra-lateral rupture happened, whether there were other preceding or following Achilles Tendon (AT) problems, date of the first official match participation following rupture, whether the player retired, changed clubs or was left without club within the 2 years following injury, minutes and matches played in the 24- (Year − 2) and 12-months preceding (Year − 1), and 12- (Year 0) and 24-months (Year 1) after injury. In addition, to account for discrepancies in playtime available, players' data related to match participation was averaged by the number of matches played by the team in 30-, 90-, 120-, 180- and 360-day intervals.

Additional feature engineering was then performed, leading to the creation of the following features: the player's market value multiplied by the average minutes played per match in Year − 1, the market value of the team multiplied by the player's average minutes played per match in Year − 1, the difference in minutes played per match in Year − 1 and Year − 2, how many days had elapsed since the player joined the team when the injury happened and the number of months elapsed since the beginning of the season when the injury occurred.

Machine learning model development and calibration

Unsupervised and supervised machine learning models were trained and evaluated using the Python *SciKit-Learn* library on the *Google Colab* platform [2, 25].

The difference between average minutes played per match during Year 1 and Year − 1 (ΔMPM) was used to survey patterns in match participation after injury. Clustering analysis was performed using *k*-means clustering [1]. The optimal number of clusters was determined using the silhouette score [27], which varies between − 1 and + 1, and evaluates how similar data points are to their clusters compared to other clusters. A value of 0 represents overlapping clusters, and negative values signify that data points have been assigned to the wrong cluster. The silhouette score is frequently used to assess clustering quality, in the absence of a standard method in the research community [27]. Cluster stability was evaluated by repeatedly randomly dividing the main dataset into training and test datasets (number of repeats: 100; train/test split: 50/50) and measuring the similarity of the resulting clustering with the Adjusted Rand Index and Fowlkes–Mallows scores, using the main dataset cluster labels as ground truth.

The post-injury match participation level was predicted using the XGBoost classification algorithm [6, 12]. Continuous variables were scaled with standardization. Feature selection was performed using forward selection, in which the model is started with no features, and features are added sequentially and kept if results are improved. Model outputs were subjected to cross-validation using a ten *k*-fold strategy [19]. In a stratified ten *k*-fold cross-validation, 90% of the dataset is used to train, and 10% is used to evaluate the model. The procedure is repeated ten times, each with a different train/test split until the entire dataset has been used as the test set. The model was evaluated using the area under the receiver operating characteristic curve (AUROC) and Brier score loss. A representation of the machine learning processing pipeline can be found in Fig. 1.

Statistical analysis

Statistical analysis was performed using Python libraries *Statsmodels* and *SciPy*. Except otherwise specified, values are presented as means and standard deviation. Groups were compared using Student's *t*-test, Kruskal–Wallis, or one-way ANOVA (depending on the number of groups and whether data followed a normal distribution). The assumption of normality was tested using the Shapiro–Wilk test. The Pearson's correlation coefficient was used to explore potential correlations between variables. Statistical significance was set at $p < 0.05$. Sample size calculation was not performed for this study.

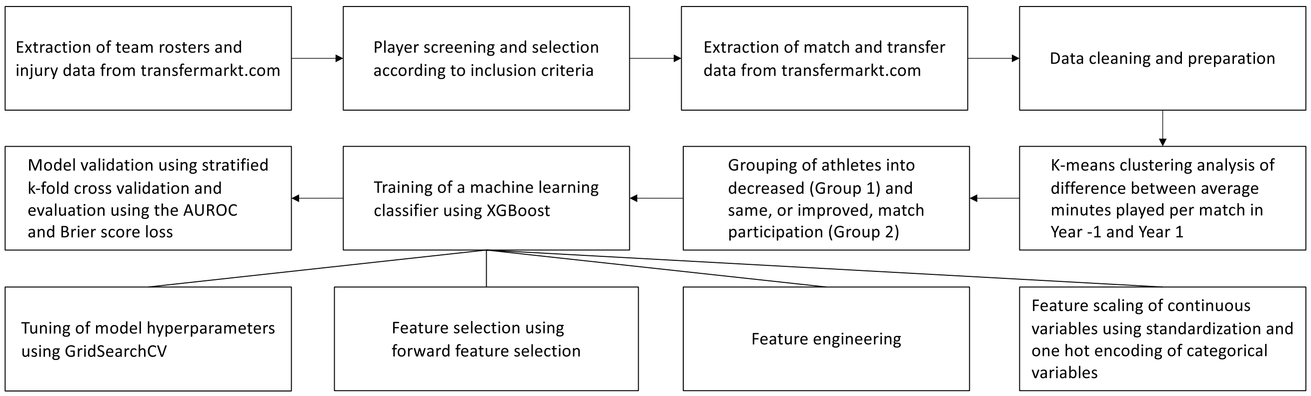
Extraction of team rosters and injury data from transfermarkt.com
Player screening and selection according to inclusion criteria
Extraction of match and transfer data from transfermarkt.com
Data cleaning and preparation
Model validation using stratified k-fold cross validation and evaluation using the AUROC and Brier score loss
Training of a machine learning classifier using XGBoost
Grouping of athletes into decreased (Group 1) and same, or improved, match participation (Group 2)
K-means clustering analysis of difference between average minutes played per match in Year -1 and Year 1
Tuning of model hyperparameters using GridSearchCV
Feature selection using forward feature selection
Feature engineering
Feature scaling of continuous variables using standardization and one hot encoding of categorical variables

Fig. 1 Machine learning processing pipeline. *AUROC* area under the receiver operating characteristic curve

Results

The scraping process retrieved 748 entries. After applying exclusion criteria, 209 players were selected for analysis. Detailed information regarding the screening and selection process, with exclusion criteria, can be found in Fig. 2.

Player demographics and baseline characteristics

Data related to player demographics and baseline characteristics can be found in Table 1. The mean age at rupture was 28.2 ± 4.0 years (range 20–40).

Return to competition and career changes

Players were cleared for unrestricted practice after a mean of 223 ± 129 days (range 92–1553). The first post-injury match was played after a mean of 287 ± 136 days (range 106–825).

Fourteen players (6.7%) did not play any match after the AT injury and subsequently retired, with five more players retiring within 2 years after injury, for a total of 19 (9.1%). Three other players (1.4%) had their contracts expire and were left without a club sometime in the 2 years after injury. Thirteen players (6.2%) changed clubs within the 2 years following injury, with nine changing to teams playing below second league (4.3%).

Re-ruptures and other Achilles tendon issues

Ten players sustained re-ruptures (4.8%). These re-ruptures occurred after a mean of 621 ± 532 days after the index injury (153–1634). Six players (2.9%) sustained contralateral ruptures at some point in their careers. Eight players (3.8%) had a recording of previous AT problems, and 16 players (7.7%) had another time-loss injury (other than re-rupture or contra-lateral ATR) related to AT problems after the index injury.

Exploratory analysis of match participation data

Data from 32,853 matches were analysed. The average minutes played per match was 48 ± 25 in Year − 2, 46 ± 24 in Year − 1, 11 ± 13 in Year 0, and 32 ± 25 in Year 1. Players were in the squad in 64.1 ± 26.2% of games in Year − 2, 62.6 ± 25.0% in Year − 1, 17.9 ± 18.1% in Year 0, and 47.0 ± 29.4% in Year 1. Players were in the starting eleven in 53.5 ± 28.0% of games in Year − 2, 51.6 ± 27.0% in Year − 1, 12.0 ± 15.1% in Year 0, and 35.3 ± 28.7% in Year 1. These differences were statistically significant ($p<0.001$) for all comparisons except between Year − 2 and Year − 1. A plot of average minutes played per match throughout the study time frame, computed in 30-day intervals for each playing position, can be seen in Fig. 3.

The Pearson's correlation coefficient showed a small inverse correlation between days until clearance for unrestricted practice and the ΔMPM ($r=-0.2$; 95% confidence interval − 0.33 to − 0.07; $p<0.01$). A very small positive correlation was found between days elapsed since injury until first match played and the ΔMPM ($r=0.13$; 95% confidence interval − 0.01 to 0.26; n.s.). After removal of outliers (those with values above 500 days; $n=18$), this correlation was 0.2 (95% confidence interval 0.06–0.33; $p<0.01$). Finally, a small positive correlation was also found between the number of days from clearance for unrestricted practice to first match played and the ΔMPM ($r=0.24$; 95% confidence interval 0.11–0.36; $p<0.001$).

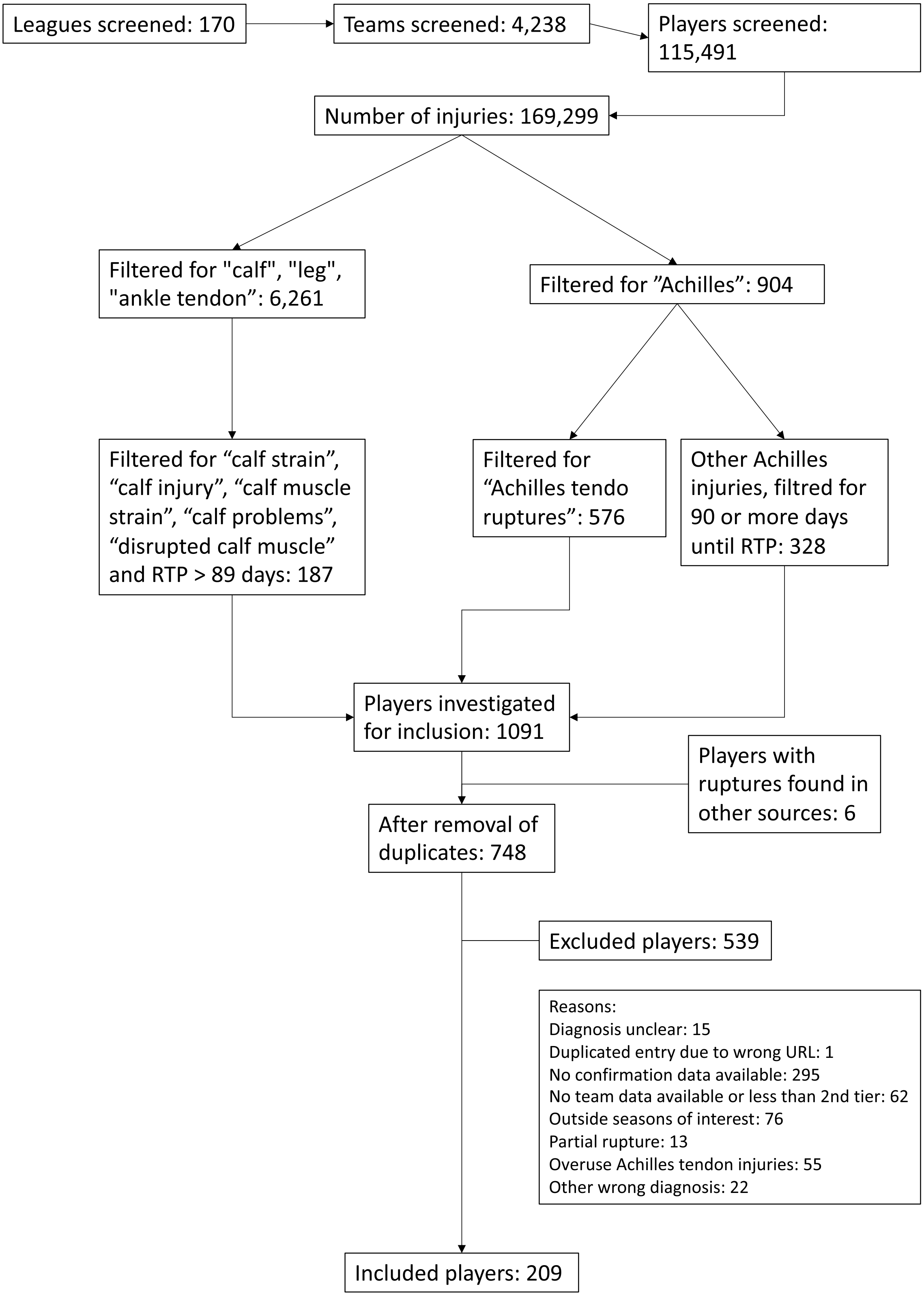
Leagues screened: 170
Teams screened: 4,238
Players screened: 115,491
Number of injuries: 169,299
Filtered for "calf", "leg", "ankle tendon”: 6,261
Filtered for ”Achilles”: 904
Filtered for “calf strain”, “calf injury”, “calf muscle strain”, “calf problems”, “disrupted calf muscle” and RTP > 89 days: 187
Filtered for “Achilles tendo ruptures”: 576
Other Achilles injuries, filtred for 90 or more days until RTP: 328
Players investigated for inclusion: 1091
Players with ruptures found in other sources: 6
After removal of duplicates: 748
Excluded players: 539
Reasons:
Diagnosis unclear: 15
Duplicated entry due to wrong URL: 1
No confirmation data available: 295
No team data available or less than 2nd tier: 62
Outside seasons of interest: 76
Partial rupture: 13
Overuse Achilles tendon injuries: 55
Other wrong diagnosis: 22
Included players: 209

Fig. 2 Player screening and selection flowchart, with exclusion criteria

Table 1 Player demographics and baseline characteristics

	Forwards	Midfielders	Defenders	Goalkeepers	Total
N (%)	55 (26.3)	43 (20.6)	95 (45.5)	16 (7.7)	209 (100)
Age, years	27.6 ± 3.7	28.6 ± 4.1	28.5 ± 4.0	28.6 ± 4.7	28.3 ± 4.0
Height, cm	181 ± 7	179 ± 7	184 ± 6	186 ± 5	182.2 ± 6.5
Preferred foot (%)					
Right	38 (18.2)	36 (17.2)	67 (32.1)	14 (6.7)	155 (74.2)
Left	14 (6.7)	5 (2.4)	26 (12.4)	2 (1.0)	47 (22.5)
Both	3 (1.3)	2 (1.0)	2 (1.0)	0 (0)	7 (3.3)
League (%)					
First	43 (20.6)	32 (15.3)	82 (39.2)	10 (4.8)	167 (79.9)
Second	12 (5.7)	11 (5.3)	13 (6.2)	6 (2.9)	42 (20.1)
National team (%)					
Yes	35 (16.8)	20 (9.5)	57 (27.3)	9 (4.3)	121 (57.9)
No	20 (9.5)	23 (11.0)	38 (18.2)	7 (3.4)	88 (42.1)
World region (%)					
Europe	48 (22.9)	33 (15.7)	69 (33.0)	12 (5.7)	162 (77.3)
America	6 (2.9)	6 (2.9)	20 (9.5)	1 (0.5)	33 (15.8)
Africa	1 (0.5)	2 (1.0)	6 (2.9)	1 (0.5)	10 (4.9)
Asia/Australasia	0 (0)	2 (1.0)	0 (0)	2 (1.0)	4 (2.0)

Player demographics and baseline characteristics. Values are represented as means and standard deviations or percentages of total values

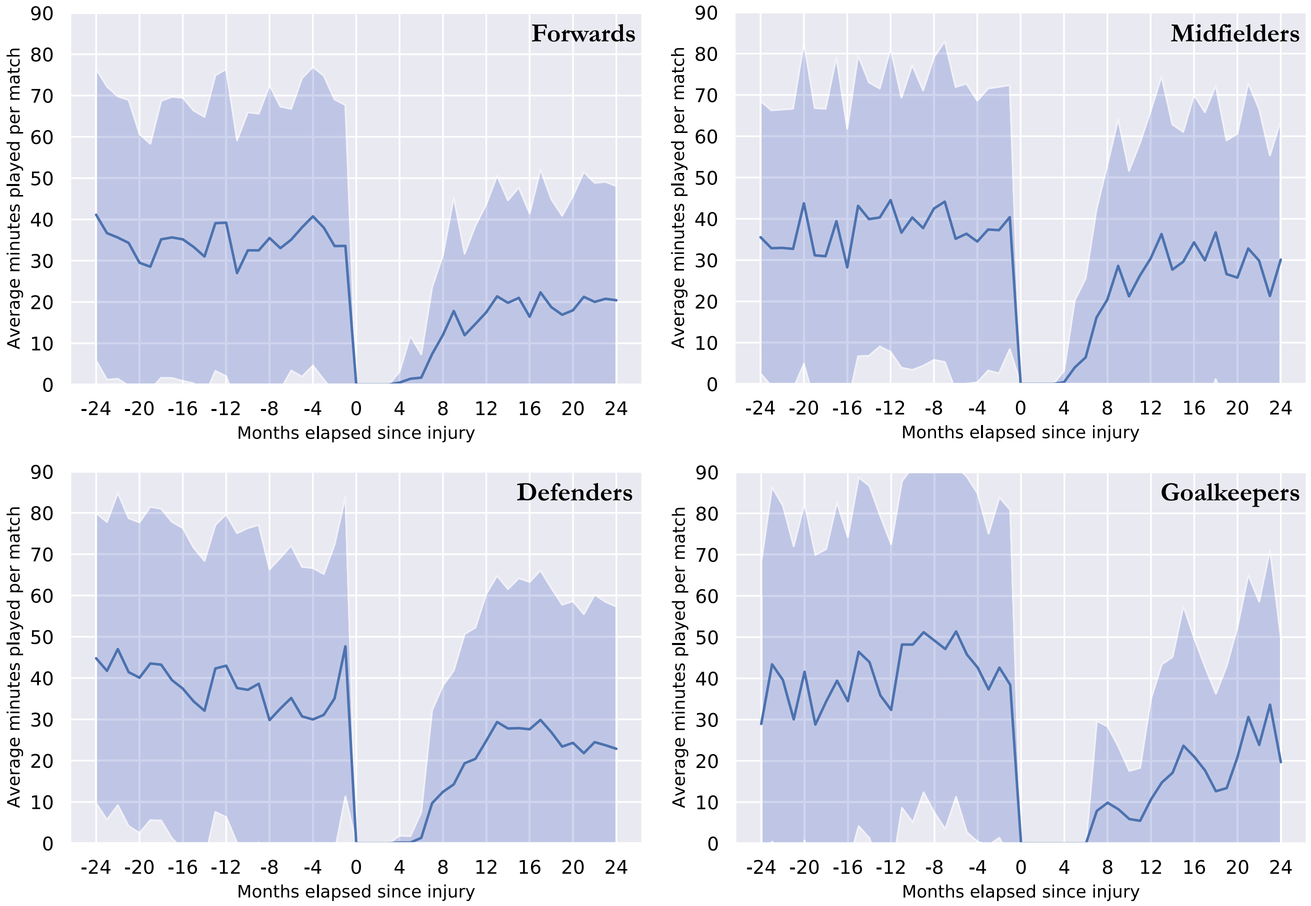

Fig. 3 Plot of average minutes played per match (*y*-axis) for all players included throughout the study time frame and computed in 30-day intervals (*x*-axis) per playing position. Shaded areas correspond to standard deviation

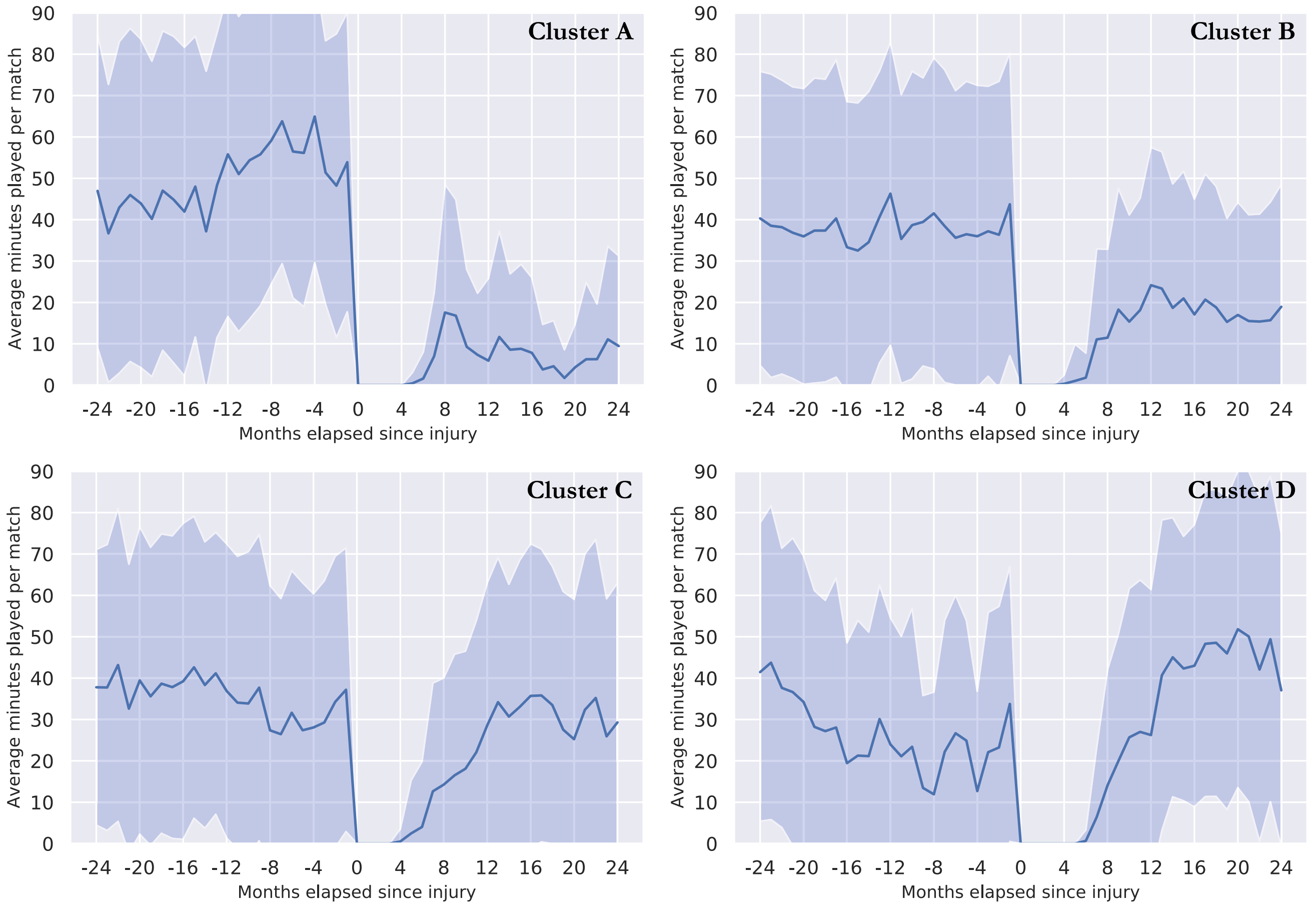

Fig. 4 Plot of average minutes played per match (*y*-axis) throughout the study time frame and computed in 30-day intervals (*x*-axis) for each cluster. Shaded areas correspond to standard deviation

Clustering analysis

The optimal number of clusters was four. The silhouette score was 0.55. The Adjusted Rand Index and Fowlkes–Mallows scores were 0.84 and 0.88, respectively. A plot of average minutes played per match for each cluster, computed in 30-day intervals, can be found in Fig. 4. The main characteristics of clusters and respective statistical comparisons can be found in Table 2.

Prediction of post-injury match participation

Players were divided into two groups based on whether they suffered a decrease in match participation while comparing average minutes per match in Year 1 and Year − 1. Players were assigned to Group 1 if they showed a decrease larger than 15 min played per match, and this difference was more than 20% of the value in Year − 1 (decreased match participation, $n = 103$). Otherwise, they were assigned to Group 2 (maintenance or improvement of match participation, $n = 106$). These designations were used as classification labels to train a ML classification algorithm. A list of included features and relative feature importance can be found in Table 3. After cross-validation, the average model AUROC was 0.81 ± 0.10, and the Brier score loss was 0.12.

Discussion

The most important findings of this study were: most players gradually increased match participation during the first year after injury, with goalkeepers still improving after 2 years; and the ML classifier displayed good performance predicting whether a player would return to a similar, or even improved, level of match participation, with the most important features being related with pre-injury performance.

Plateauing of post-injury match participation occurred approximately 1 year after injury for forwards, midfielders, and defenders. Goalkeepers kept increasing playing time throughout the 2 years following injury, albeit at a slower

Table 2 Main characteristics of clusters and statistical comparisons

	Cluster A	Cluster B	Cluster C	Cluster D	*p*
N (%)	34 (16.2)	75 (35.9)	70 (33.5)	30 (14.4)	–
Age, years	29.9 ± 4.5	27.8 ± 3.8	28.6 ± 3.7	26.9 ± 3.9	**0.02**
Height, cm	184 ± 7	182 ± 6	181 ± 7	183 ± 6	(n.s.)
Position (%)					
Forward	11 (32.3)	22 (29.3)	19 (27.1)	3 (10.0)	(n.s.)
Midfielder	2 (5.9)	18 (24.0)	14 (20.00)	9 (30.0)	
Defender	15 (44.1)	30 (40.0)	34 (48.6)	16 (53.3)	
Goalkeeper	6 (17.7)	5 (6.7)	3 (4.3)	2 (6.7)	
Preferred foot (%)					
Right	27 (79.4)	50 (66.7)	54 (77.1)	24 (80.0)	(n.s.)
Left	5 (14.7)	23 (30.7)	14 (20.0)	5 (16.7)	
Both	2 (5.9)	2 (2.6)	2 (2.9)	1 (3.3)	
League (%)					
First	26 (76.5)	60 (80.0)	55 (78.6)	26 (86.7)	(n.s.)
Second	8 (23.5)	15 (20.0)	15 (21.4)	4 (13.3)	
National team (%)					
Yes	21 (61.8)	42 (56.0)	41 (58.6)	17 (56.7)	**0.02**
No	13 (38.2)	33 (44.0)	29 (41.4)	13 (43.3)	
World region (%)					
Europe	25 (73.5)	59 (78.7)	54 (77.1)	24 (80.0)	(n.s.)
America	6 (17.7)	11 (14.7)	12 (17.2)	4 (13.3)	
Africa	2 (5.9)	3 (4.0)	3 (4.3)	2 (6.7)	
Asia/Australasia	1 (2.9)	2 (2.6)	1 (1.4)	0 (0.0)	
Market value (Euros)	1.2 ± 1.3 Mil	2.2 ± 4.0 Mil	1.7 ± 2.1 Mil	2.6 ± 3.8 Mil	(n.s.)
Time since joining the team (days)	1060 ± 1288	655 ± 769	658 ± 613	441 ± 549	**0.01**
Time between season start and injury (months)	6 ± 4	5 ± 4	5 ± 3	5 ± 4	(n.s.)
Previous AT injuries (%)					
Yes	1 (2.9)	6 (8.0)	0 (0.0)	1 (3.3)	(n.s.)
No	33 (97.1)	69 (92.0)	70 (100.0)	29 (96.7)	
Number of previous injuries (*N*)	2.3 ± 1.5	3.0 ± 2.6	2.5 ± 2.3	2.8 ± 2.7	(n.s.)
Time until unrestricted practice (days)	280 ± 244	215 ± 118	202 ± 73	207 ± 55	(n.s.)

Table 2 (continued)

	Cluster A	Cluster B	Cluster C	Cluster D	*p*
Time until first match (days)	242 ± 199	269 ± 137	271 ± 158	315 ± 119	**0.05**
Average minutes played per match					
Year − 2	57 ± 28	47 ± 22	49 ± 26	40 ± 23	(n.s.)
Year − 1	69 ± 14	48 ± 19	40 ± 25	27 ± 17	**< 0.01**
Year 0	7 ± 9	10 ± 12	13 ± 15	12 ± 15	(n.s.)
Year 1	10 ± 12	23 ± 19	40 ± 25	59 ± 19	**< 0.01**
Delta minutes played per match Year 1 and Year − 1	− 59 ± 13	− 25 ± 8	0 ± 8	32 ± 13	**< 0.01**
Re-rupture (%)					
Yes	2 (5.9)	6 (8.0)	2 (2.9)	0 (0.0)	(n.s.)
No	32 (94.1)	69 (92.0)	68 (97.1)	30 (100.0)	
Bilateral rupture (%)					
Yes	1 (2.9)	2 (2.7)	3 (4.3)	0 (0.0)	(n.s.)
No	33 (97.1)	73 (97.3)	67 (95.7)	30 (100.0)	
Other AT problems afterwards (%)					
Yes	2 (5.9)	5 (6.7)	6 (8.6)	3 (10.0)	(n.s.)
No	32 (94.1)	70 (93.3)	64 (91.4)	27 (90.0)	
Changed club within 2 years (%)					
Yes	1 (2.9)	7 (9.3)	5 (7.1)	2 (6.7)	(n.s.)
No	33 (97.1)	68 (90.7)	65 (92.9)	28 (93.3)	
Left without club within 2 years (%)					
Yes	1 (2.9)	2 (2.7)	0 (0.0)	0 (0.0)	(n.s.)
No	33 (97.1)	73 (97.3)	70 (100.0)	30 (100.0)	
Retired within 2 years (%)					
Yes	9 (26.5)	6 (8.0)	4 (5.7)	0 (0.0)	**< 0.01**
No	25 (73.5)	69 (92.0)	66 (94.3)	30 (100.0)	

Comparison between clusters of match participation patterns. Values are represented as means and standard deviations or percentages of total values. *AT* Achilles tendon. Clusters A, B, C and D relate to severe decrease, moderate decrease, maintenance or improvement of match participation

Bold font indicates statistical significance rate. Of note, previous research has shown that outcomes after ATRs improve for at least 1 year after injury [3, 4], possibly due to a need to adapt to biomechanical changes in the lower limb resulting from tendon elongation [9]. Another critical aspect to consider is that psychological factors may be involved [29, 35], in which players need to regain confidence in their abilities and overcome the fear of re-injury.

Differences in match participation between Year − 1 and Year 1 were the subject of clustering analysis. A silhouette score of 0.55 was found for the optimum number of clusters, which denotes moderate cluster separability. In addition, good clustering stability was found through the Adjusted Rand Index and Fowlkes–Mallows scores, meaning that these clusters were relatively consistent, even when only subsections of the dataset were randomly evaluated.

Younger age has been previously recognized as a favourable prognostic factor after ATRs in soccer players [10]. However, this point is controversial since other studies have not found statistically significant differences regarding age in players with favourable versus unfavourable outcomes in soccer [34], American football [24], basketball [15], and baseball [26]. In this study, the average age was lower in clusters C and D (maintenance or improvement of match participation) than clusters A and B (decreased match participation).

The number of days the player has been with the team at the time of injury is a previously unrecognized prognostic factor in ATRs. In this study, it was found that players in Cluster A were with the team for a significantly longer time (1060.1 ± 1287.6 days) compared with the remaining cohort ($p < 0.01$). The longer time with the team (or since the last market transfer) may signal a different career context for these players. For example, their contracts may be near expiration, and prospects of joining another team are dim. Coincidentally, players in this cluster also retired within 2 years in a statistically significant higher proportion than the remaining cohort (26.5% versus 5.7%; $p < 0.01$).

Players in cluster D took a significantly longer time before playing their first official match compared with the remaining cohort (315 days ± 119 versus 264 days ± 159; $p < 0.05$), despite similar time intervals from injury to unrestricted practice (207 days ± 55 versus 222 days ± 140; $p = 0.72$). Therefore, it can be speculated that by allowing these players more time to recover, they made their comeback at a higher performance level—closer to the full recovery potential—which would be perceived as a superior recovery from injury, encouraging increased match participation. In addition, players in Cluster A (those with the most significant decrease in average minutes played per match in Year 1 compared to Year − 1) showed the shortest time until first match

Table 3 Features included in the predictive model and their importance

	Feature importance
Base features	
Days elapsed since joining the team	0.02
International level player?	0.02
Playing position	0.02
First or second league	0.01
Months elapsed since the beginning of the season when the injury occurred	0.01
Player market value	0.01
Engineered features	
Matches in which player was in the starting eleven divided by number of matches available, averaged in 30-day intervals, in Year − 1	0.23
Minutes player per match, averaged in 30-day intervals, in Year − 1	0.23
Matches in which player did not play because of medical issues divided by the number of matches available, averaged in 30-day intervals, in Year − 1	0.15
Matches sat on bench divided by the number of matches available, averaged in 30-day intervals, in Year − 1	0.12
Matches in the player's team won divided by the number of matches available, averaged in 180-day intervals, in Year − 1 and Year − 2	0.07
Player market value times minutes played per match in Year − 1	0.04
Average minutes played per match in Year − 1 divided by the same variable in Year − 2	0.03
Team market value times days elapsed since player joined the team	0.02
Team market value times minutes played per match in Year − 1	0.02

Features included in the predictive model. Engineered features result from combining continuous variables or mathematical operations between two other features. Feature importance relates to the relative contribution of that feature to the model, where higher values imply a higher impact on model performance played. However, statistical correlations between days until unrestricted practice or first match played and the ΔMPM were small (albeit statistically significant). Further research is required to determine how a delayed return to competition may relate to improved outcomes after ATRs.

A ML classifier was trained, with an AUROC of 0.81 ± 0.10 after cross-validation, through careful feature engineering and selection, which translates as good discriminating performance [23]. The model's performance was also evaluated regarding output probabilities using the Brier score loss, as it was deemed helpful for players and staff to gauge these against their individual beliefs and experiences. It should be noted that only pre-injury features were used to train the model, and no data regarding treatments was available. Of note, since features related to pre-injury match participation showed the highest feature importance, it can be inferred that the future level of match participation is related to the sporting context at the time of injury, directly or indirectly (e.g. a tendency for early RTP in high-performing players which may reflect negatively in match participation afterward).

The use of ML algorithms to predict sports injuries is a current trend in research [14, 17, 31], but practitioners should remain cautious regarding their use despite recent advances. There are ethical implications to consider [5], such as inadvertently hindering a player's career through a wrongfully attributed worse prognosis. Model results may also be overly optimistic, either due to *overfitting* (when the model is fitted too close to a particular set of data and becomes unable to make good predictions in a generalized environment) or accidental *data leakage* (when information contained in the test set is wrongfully fed to the model during training). Nevertheless, the increasing accessibility and ease of use of ML tools and development frameworks offer an excellent opportunity to improve the care of musculoskeletal injuries, though researchers and clinicians should stay vigilant about its shortcomings.

The main limitation of this study is the inability to confirm the diagnosis. However, all included cases were manually double-checked using other sources by two researchers independently to avoid the inclusion of misclassified injuries. Other limitations are the unknown measurement accuracy of match participation data found on *transfermarkt.com*, the unavailability of treatment data, and the lack of a strictu sensu measure of player performance.

This study can guide the objectives and expectations of athletes and staff regarding how match participation evolves after an ATR, noting that it takes approximately 1 year to reach its peak (except for goalkeepers, who may keep improving for at least 2 years). In addition, the cluster of players with improved match performance showed a statistically significant increase in the number of days until first match played compared to the remaining cohort. Also, a small but statistically significant positive correlation was found between time until first match played and the ΔMPM. Finally, recent research has shown improved outcomes in patients undergoing *slowed-down* rehabilitation programs [18]. Thus, it may make sense to prioritize recovery of lower limb strength and sport-specific skills over an early return to competition.

Conclusion

Exploratory data analysis revealed that forwards, midfielders and defenders increased match participation during the first year after injury, with goalkeepers still improving at 2 years. Good performance was attained using a ML classifier to predict the level of match participation following an ATR, with features related to pre-injury match participation displaying the highest importance.

Acknowledgements This work is financed by national funds from FCT—Fundação para a Ciência e a Tecnologia, I.P., in the scope of the Project UIDB/04565/2020 and UIDP/04565/2020 of the Research Unit Institute for Bioengineering and Biosciences—iBB and the project LA/P/0140/2020 of the Associate Laboratory Institute for Health and Bioeconomy—i4HB.

Author contributions PD, FCF, GK, and AF designed the study; PD wrote the computer code to perform data extraction from a publicly available source; PD and DL screened and selected athletes for inclusion in the study; PD prepared the data, performed exploratory data analysis, and developed the predictive model with support from MA and AF; PD drafted the manuscript with input from MA, AM, and HP; FCF, GK and AF revised the final manuscript.

Funding None.

Declarations

Conflict of interest The authors declare that they have no competing interests.

Ethical approval Not applicable.

Informed consent Not applicable.

References

1. Badillo S, Banfai B, Birzele F, Davydov II, Hutchinson L, Kam-Thong T, Siebourg-Polster J, Steiert B, Zhang JD (2020) An introduction to machine learning. Clin Pharmacol Ther 107:871–885
2. Bisong E (2019) Building machine learning and deep learning models on google cloud platform: a comprehensive guide for beginners. Apress, Ottawa
3. Carmont MR, Knutsson SB, Brorsson A, Karlsson J, Nilsson-Helander K (2022) The release of adhesions improves outcome

following minimally invasive repair of Achilles tendon rupture. Knee Surg Sports Traumatol Arthrosc 30:1109–1117
4. Carmont MR, Silbernagel KG, Edge A, Mei-Dan O, Karlsson J, Maffulli N (2013) Functional outcome of percutaneous achilles repair: improvements in achilles tendon total rupture score during the first year. Orthop J Sports Med. https://doi.org/10.1177/2325967113494584
5. Char DS, Abràmoff MD, Feudtner C (2020) Identifying ethical considerations for machine learning healthcare applications. Am J Bioeth 20:7–17
6. Chen T, Guestrin C (2016) XGBoost: a scalable tree boosting system. In: Krishnapuram B, Shah M (eds) KDD '16: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, San Francisco, August 2016. Association for Computing Machinery, New York, pp 785–794
7. Choi RY, Coyner AS, Kalpathy-Cramer J, Chiang MF, Campbell JP (2020) Introduction to machine learning, neural networks, and deep learning. Transl Vis Sci Technol. https://doi.org/10.1167/tvst.9.2.14
8. Della Villa F, Hägglund M, Della Villa S, Ekstrand J, Waldén M (2021) High rate of second ACL injury following ACL reconstruction in male professional footballers: an updated longitudinal analysis from 118 players in the UEFA Elite Club Injury Study. Br J Sports Med 55:1350–1357
9. Diniz P, Pacheco J, Guerra-Pinto F, Pereira H, Ferreira FC, Kerkhoffs G (2020) Achilles tendon elongation after acute rupture: is it a problem? A systematic review. Knee Surg Sports Traumatol Arthrosc 28:4011–4030
10. Grassi A, Macchiarola L, Filippini M, Lucidi GA, Della Villa F, Zaffagnini S (2020) Epidemiology of anterior cruciate ligament injury in italian first division soccer players. Sports Health 12:279–288
11. Grassi A, Rossi G, D'Hooghe P, Aujla R, Mosca M, Samuelsson K, Zaffagnini S (2020) Eighty-two per cent of male professional football (soccer) players return to play at the previous level two seasons after Achilles tendon rupture treated with surgical repair. Br J Sports Med 54:480–486
12. Hinterwimmer F, Lazic I, Langer S, Suren C, Charitou F, Hirschmann MT, Matziolis G, Seidl F, Pohlig F, Rueckert D, Burgkart R, von Eisenhart-Rothe R (2022) Prediction of complications and surgery duration in primary TKA with high accuracy using machine learning with arthroplasty-specific data. Knee Surg Sports Traumatol Arthrosc. https://doi.org/10.1007/s00167-022-06957-w
13. Jack RA, Sochacki KR, Gardner SS, McCulloch PC, Lintner DM, Cosculluela PE, Varner KE, Harris JD (2017) Performance and return to sport after Achilles tendon repair in national football league players. Foot Ankle Int 38:1092–1099
14. Karnuta JM, Luu BC, Haeberle HS, Saluan PM, Frangiamore SJ, Stearns KL, Farrow LD, Nwachukwu BU, Verma NN, Makhni EC, Schickendantz MS, Ramkumar PN (2020) Machine learning outperforms regression analysis to predict next-season major league baseball player injuries: epidemiology and validation of 13,982 player-years from performance and injury profile trends, 2000–2017. Orthop J Sports Med. https://doi.org/10.1177/2325967120963046
15. Lemme NJ, Li NY, Kleiner JE, Tan S, DeFroda SF, Owens BD (2019) Epidemiology and video analysis of Achilles tendon ruptures in the National Basketball Association. Am J Sports Med 47:2360–2366
16. Leventer L, Eek F, Hofstetter S, Lames M (2016) Injury patterns among elite football players: a media-based analysis over 6 seasons with emphasis on playing position. Int J Sports Med 37:898–908
17. Lövdal SS, Den Hartigh RJR, Azzopardi G (2021) Injury prediction in competitive runners with machine learning. Int J Sports Physiol Perform. https://doi.org/10.1123/ijspp.2020-0518
18. Maffulli N, Gougoulias N, Maffulli GD, Oliva F, Migliorini F (2022) Slowed-down rehabilitation following percutaneous repair of Achilles tendon rupture. Foot Ankle Int 43:244–252
19. Marcot BG, Hanea AM (2021) What is an optimal value of k in k-fold cross-validation in discrete Bayesian network analysis? Comput Stat 36:2009–2031
20. Martin RK, Ley C, Pareek A, Groll A, Tischer T, Seil R (2022) Artificial intelligence and machine learning: an introduction for orthopaedic surgeons. Knee Surg Sports Traumatol Arthrosc 30:361–364
21. Mazza D, Viglietta E, Monaco E, Iorio R, Marzilli F, Princi G, Massafra C, Ferretti A (2022) Impact of anterior cruciate ligament injury on European professional soccer players. Orthop J Sports Med. https://doi.org/10.1177/23259671221076865
22. Nigam H, Biswas P (2021) From web scraping to web crawling. In: Choudhary A, Agrawal AP, Logeswaran R, Unhelkar B (eds) Applications of artificial intelligence and machine learning. Lecture notes in electrical engineering, vol 778. Springer, Singapore, pp 97–112
23. Olsson S, Akbarian E, Lind A, Razavian AS, Gordon M (2021) Automating classification of osteoarthritis according to Kellgren–Lawrence in the knee using deep learning in an unfiltered adult population. BMC Musculoskelet Disord. https://doi.org/10.1186/s12891-021-04722-7
24. Parekh SG, Wray WH, Brimmo O, Sennett BJ, Wapner KL (2009) Epidemiology and outcomes of Achilles tendon ruptures in the National Football League. Foot Ankle Spec 2:283–286
25. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay É (2011) Scikit-learn: machine learning in python. J Mach Learn Res 12:2825–2830
26. Saltzman BM, Tetreault MW, Bohl DD, Tetreault D, Lee S, Bach BR (2017) Analysis of player statistics in major league baseball players before and after Achilles tendon repair. HSS J 13:108–118
27. Shahapure KR, Nicholas C (2020) Cluster quality analysis using Silhouette Score. In: Geoffrey IW, Zhongfei Z, Vincent TS, Graham W, Michalis V, Longbing C (eds) IEEE DSAA'2020: The 7th IEEE international conference on data science and advanced analytics, Sydney, October 2020. Institute of Electrical and Electronics Engineers (IEEE), New York, pp 747–748
28. Shamshirband S, Fathi M, Dehzangi A, Chronopoulos AT, Alinejad-Rokny H (2020) A review on deep learning approaches in healthcare systems: taxonomies, challenges, and open issues. J Biomed Inform. https://doi.org/10.1016/j.jbi.2020.103627
29. Slagers AJ, van Veen E, Zwerver J, Geertzen JHB, Reininga IHF, van den Akker-Scheek I (2021) Psychological factors during rehabilitation of patients with Achilles or patellar tendinopathy: a cross-sectional study. Phys Ther Sport 50:145–152
30. Sochacki KR, Jack RA II, Hirase T, McCulloch PC, Lintner DM, Varner KE, Cosculluela PE, Harris JD (2019) There is a high return to sport rate but with reduced career lengths after Achilles tendon repair in Major League Soccer players. J ISAKOS 4:15–20
31. Taborri J, Molinaro L, Santospagnuolo A, Vetrano M, Vulpiani MC, Rossi S (2021) A machine-learning approach to measure the anterior cruciate ligament injury risk in female basketball players. Sensors (Basel). https://doi.org/10.3390/s21093141
32. Transfermarkt.com (2021) Transfermarkt.com: Football Transfers, Rumours, Market Values, News and Statistics. http://www.transfermarkt.com. Accessed 31 Mar 2021
33. Trofa DP, Miller JC, Jang ES, Woode DR, Greisberg JK, Vosseller JT (2017) Professional athletes' return to play and performance

after operative repair of an Achilles tendon rupture. Am J Sports Med 45:2864–2871

34. Trofa DP, Noback PC, Caldwell J-ME, Miller JC, Greisberg JK, Ahmad CS, Vosseller JT (2018) Professional soccer players' return to play and performance after operative repair of Achilles tendon rupture. Orthop J Sports Med. https://doi.org/10.1177/2325967118810772
35. Turner J, Malliaras P, Goulis J, Mc Auliffe S (2020) "It's disappointing and it's pretty frustrating, because it feels like it's something that will never go away." A qualitative study exploring individuals' beliefs and experiences of Achilles tendinopathy. PLoS One. https://doi.org/10.1371/journal.pone.0233459
36. Wilke J, Tenberg S, Groneberg D (2022) Prognostic factors of muscle injury in elite football players: a media-based, retrospective 5-year analysis. Phys Ther Sport 55:305–308

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Pedro Diniz[1,2,3,4] · Mariana Abreu[2,5] · Diogo Lacerda[1] · António Martins[1,4] · Hélder Pereira[6,7,8] · Frederico Castelo Ferreira[2,3] · Gino MMJ Kerkhoffs[9,10,11] · Ana Fred[2,5]

[1] Department of Orthopaedic Surgery, Hospital de Sant'Ana, Rua de Benguela, 501, 2775-028 Parede, Portugal

[2] Department of Bioengineering and iBB, Institute for Bioengineering and Biosciences, Instituto Superior Técnico, Universidade de Lisboa, Lisbon, Portugal

[3] Associate Laboratory i4HB, Institute for Health and Bioeconomy, Instituto Superior Técnico, Universidade de Lisboa, Lisbon, Portugal

[4] Fisiogaspar, Lisbon, Portugal

[5] Instituto de Telecomunicações, Lisbon, Portugal

[6] Orthopaedic Department, Centro Hospitalar Póvoa de Varzim, Vila do Conde, Portugal

[7] Ripoll y De Prado Sports Clinic: FIFA Medical Centre of Excellence, Murcia-Madrid, Spain

[8] University of Minho ICVS/3B's-PT Government Associate Laboratory, Braga/Guimarães, Portugal

[9] Department of Orthopaedic Surgery, Amsterdam Movement Sciences, Amsterdam University Medical Centers, Amsterdam, The Netherlands

[10] Academic Center for Evidence Based Sports Medicine (ACES), Amsterdam, The Netherlands

[11] Amsterdam Collaboration for Health and Safety in Sports (ACHSS), Amsterdam, The Netherlands