Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 225 (2023) 4139–4147

Procedia Computer Science

www.elsevier.com/locate/procedia

27th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2023)

Integrating Machine Learning for Football Injury Prediction: A Concept for an Intelligent System

Marcin Prys[a]*, Łukasz Rosiński[b, c], Rafał Buryta[b, c], Łukasz Radzimiński[d], Przemysław Różewski[a], Izabela Rejer[a]

[a] *Faculty of Computer Science and Information Technology, West Pomeranian University of Technology in Szczecin, Żołnierska 49, 71-210 Szczecin, Poland*
[b] *Pogoń Szczecin, Footbal Club, Karłowicza 28, 71-102 Szczecin, Poland*
[c]*Institute of Physical Education, University of Szczecin, Piastów 40b, 71-064 Szczecin, Poland*
[d] *Department of Physiology, Gdansk University of Physical Education and Sport, K. Górskiego 1, 80-336 Gdansk, Poland*

Abstract

In recent years, sports analytics has gained significant attention, particularly in the area of using machine learning techniques to analyze athlete performance and injury risks. This paper focuses on the use of machine learning for predicting injuries in football players and proposes a concept for an intelligent system that can provide early warnings of potential injuries. To achieve this, the proposed intelligent system will gather input from various sources such as athlete performance data, muscle strength assessments, and DNA markers. Using this information, the system will employ machine learning algorithms to identify patterns and predict the probability of injury occurrence. The intelligent system will offer coaches and medical staff actionable insights to help them take preventative measures and decrease the likelihood of injuries.

Keywords: injury prediction; machine learning; football analytics; sport science

1. Introduction

Sports analytics has experienced a surge in popularity in recent years, as researchers have started to explore the benefits of incorporating machine learning techniques into various sports-related areas. One such domain of interest

* Corresponding author. Marcin Prys
E-mail address: mprys@zut.edu.pl

10.1016/j.procs.2023.10.410

is football injury prediction, as injuries not only pose a threat to the athlete's health but also impact the team's overall performance and financial prospects. In this article, we introduce a concept for an intelligent system capable of predicting football injuries and providing early warnings to mitigate their consequences. Injuries often necessitate extensive rehabilitation, which can be both costly and time-consuming; therefore, an intelligent system that can predict and warn of potential injuries is crucial for minimizing their impact.

Although sports analytics is an expanding field, the research on football injury prediction remains in its infancy. Several studies [1] [2] [3] [4] [5] [6] [7] [8] have illustrated the potential of machine learning techniques in sports analytics. However, the implementation of these methods in predicting football injuries is not without challenges.

As mentioned, some studies have reported positive results in football injury prediction using machine learning techniques; however, others have pointed out the limitations of these algorithms, particularly in clinical settings. One significant concern is the "black box" nature of these models [9], which can hinder validation and implementation due to the lack of reporting transparency. Black box models are prone to overfitting, especially when their full equations, code, hyperparameters, or algorithms are not provided. While machine learning methods have the potential for accurate predictions, they often result in overfitted models with poor performance in prospective validation.

In a recent systematic review, 57% of the studies did not report the full model or updated model, which raises concerns about their external validation quality [10].

Despite the employment of multiple machine learning algorithms in predicting football injuries, several challenges persist. Issues such as data quality, data availability, and the choice of suitable algorithms can influence the accuracy of the predictions. In this article, we present a concept for an intelligent system designed to address these challenges and offer effective injury prediction capabilities. We will discuss both the potential of machine learning methods in injury prediction and their inherent limitations, acknowledging the nascent nature of this research area and its promising yet visible constraints.

2. Literature review

2.1. Sports Science perspective. Non-contact injury, external and internal factors

Non-contact injuries are a serious concern in football, and understanding the factors that contribute to these injuries is crucial for prevention. Non-contact injuries occur when an injury is sustained without any direct impact from another player or object. Studies have shown that non-contact injuries account for a significant proportion of all football injuries [11] [12]. The contributing factors to non-contact injuries can be classified into external and internal risk factors.

Research has identified several external risk factors for non-contact injuries in football. High training loads have been shown to increase the risk of injury due to fatigue, overuse, and inadequate recovery time [13] [14] [15]. A study [16] found that muscle fatigue, poor balance, and previous injuries were significant risk factors for non-contact injuries in male football players. Another study [17] identified poor neuromuscular control, reduced flexibility, and poor balance as important risk factors for non-contact injuries in female football players.

Weather conditions, such as rain and wet playing surfaces, can have a significant impact on injury risk in football players. Research has shown that playing football on wet artificial turf or natural grass increases the risk of lower extremity injuries, particularly ankle sprains and knee ligament tears [18]. The increased risk of injury is likely due to reduced traction on the wet surface, leading to greater instability during movement. Cold exposure can increase muscle glycolysis and lactate accumulation suggesting a lower muscle efficiency and/or an effect of a lower perfusion in cold muscle [19]

Intrinsic factors such as muscle imbalances and joint laxity may also contribute to injury risk. [20] found that female football players with a combination of neuromuscular and biomechanical risk factors had a significantly higher risk of ACL injuries. Similarly, a study [21] found that players with weaker hip muscles were more likely to suffer from groin injuries.

Psychological factors can also contribute to non-contact injuries in football. Stress, anxiety, and low mood can all affect an athlete's mental state, leading to reduced concentration, motivation, and confidence. This can increase the risk of injury due to reduced focus and increased risk-taking behavior. A study [22] found that psychological distress was associated with an increased risk of injury in professional football players.

Research has also focused on identifying genetic and other intrinsic factors that may predispose football players to injury. Several studies have investigated the potential relationship between genetic markers and injury risk, with some suggesting that specific variations in genes related to muscle structure and function may increase the likelihood of certain types of injuries. For example, a study [23] found that football players with a specific variant of the COL5A1 gene had a higher risk of suffering from Achilles tendon injuries.

Creatine kinase (CK) has been studied as a biomarker for muscle damage, and some research has suggested that elevated CK levels may be associated with increased injury risk [24]. CK is an enzyme that plays a key role in energy production within muscle cells, and its presence in the bloodstream can indicate muscle damage or stress. In a study [25], athletes with higher CK levels were found to have a greater risk of injury, particularly muscle strains and ligament injuries. This suggests that monitoring CK levels could potentially help identify athletes at higher risk of injury, allowing for targeted interventions and prevention strategies.

In summary, internal injury risk factors include intrinsic factors such as muscle imbalances, joint laxity, psychological factors, genetic predispositions, and elevated CK levels. These factors, along with external factors such as training load, weather conditions, and playing surfaces, should be considered when developing a comprehensive approach to injury prevention in football. Understanding the complex interplay between these factors is crucial for the design of an intelligent system that can predict and prevent football injuries. Further research is needed to confirm the associations between these factors and injury risk, as well as to determine their practical implications for injury prevention and management.

2.2. Computer Science perspective. Machine Learning in football injury predictions

Machine learning focuses on developing algorithms and models, which allow computers to learn from data and make better decisions based on experience. It has been applied in a wide range of fields, including healthcare, finance, marketing, and sports, for tasks such as image classification, disease detection, financial forecasting, customer retention analysis, and injury prediction. The advancements in machine learning have significantly transformed the way data is analyzed and processed across various industries

The aim is to compare several classic studies frequently cited, especially in systemic analysis, and also include other articles that we believe are significant in this field. The focus of this comparison will be on the integration of machine learning in predicting football injuries. The comparative parameters have been selected in accordance with the literature on sports science and medicine, mentioned previously in the paper. In our comparison, we will refer to the article "Machine Learning for Understanding and Predicting Injuries in Football” [26], which provides an extensive review of the most popular articles on injury prediction in football and presents a comparison of different algorithms and their effectiveness. This comprehensive review helps to identify the most effective machine learning techniques for predicting football injuries, allowing researchers and practitioners to make informed decisions when designing and implementing injury prevention strategies.

Table 1. Machine Learning in football injury prediction

Author	No. of players	No. of injuries	Age group	Injury type	Type of features	Dataset time span	Machine learning algorithms	Accuracy (%)	Precision (%)	AUC	Recall (%)	F1–score (%)	Specificity (%)
Rossi et al. [1]	26	21	20–30	Every non-contact	External load, Injury history, Player profile	23 weeks	Decision tree	–	50	0.76	80	64	–
Naglah et al. [2]	21	36	N/A	Every non-contact	External load, Heart Rate	16 months	SVM	83.50	–	–	–	–	–
López-Valenciano et al. [3]	132	32	N/A	Lower leg muscle	Player profile, Psychology, Injury history	Pre-season +	SmoteBoost	–	–	0.75	65.90	–	79.10
Ayala et al. [4]	96	18	N/A	Hamstring strain	Player profile, Psychology, Injury history	1 Season	SmoteBoost	–	–	0.84	77.80	–	83.80
Rommers et al. [5]	734	368	10–15	Acute and overuse	Match activity, Player profile	Pre-season + 1 season	XGBoost	–	85	–	85	85	–
Oliver et al. [6]	400	99	10–18	Non-contact lower leg	Player profile, Screening	Pre-season + 1 season	Decision tree	–	–	0.66	55.60	–	74.20
Vallance et al. [7]	40	142	23.6–35.2	Every non-contact	External load, Player profile, Injury history	Pre-season + 1 season	Random forest, XGBoost	95.5, 97	92.2, 97	0.92, 0.97	94.5, 97	–, –	–, –
Kampakis [8]	N/A	N/A	N/A	Not specified	External load, Player profile, Injury history, Match details, Surface conditions	Pre-season + 1 season	Supervised principal components analysis	88.80	55	–	33	–	–

Currently, implementations mainly rely on a few types of data, for example, either workload or screening. It is much less common to find examples where different sources are combined to build a more complex data set, including not only workload or muscle strength but also psychological questionnaire studies. No studies have been found to incorporate a wide range of data, including weather and DNA markers.

One of the major challenges is that data is often derived from a single club, over a limited time period, where players come and go, resulting in a missed opportunity for observation or the lack of access to a more complete injury and workload history. The limitation of sample size and data types restricts the predictive capabilities of the models. Therefore, the development of more comprehensive, unbiased models is necessary to improve injury prediction accuracy and reduce the risk of injury.

It seems that the current approach to applying ML algorithms does not take into account the critical conclusion from sports science that the rhythm of loading and recovery is crucial. In the literature, the workload is usually considered as the sum of minutes played from the beginning of the season or the study. As a result, the algorithm is not informed about how the workload and recovery were distributed over time. That is, it cannot see whether the effort was accumulated over several days without recovery or the opposite.

The ML algorithms used in injury prediction suffer from typical problems of probabilistic sample classification. This is due to the limited number of samples and the lack of external validation of the adopted methods, which hinder their generalizability. Despite reported prediction efficacy levels of > 75% (AUC, Accuracy) in the literature, there is a lot of justified criticism. Currently, the algorithms used are consistent with state-of-the-art approaches in ML, such as Bayesian algorithms, decision trees, and XGBoost, among other commonly used methods. However, there are no reported attempts in the literature to propose new, more innovative approaches to the data.

Considering the above, and reviewing the literature, it is reasonable to agree with most authors' conclusions that there are significant limitations in the application and usefulness of injury prediction and machine learning for coaching staff at this stage of development. Often, the reason for this is the too general nature of the predictions, for example, after muscle screening, the prediction is for the entire season, similarly, genes alone without the context of workload do not provide useful value. Furthermore, it is worth noting that there are no methods in the literature that could provide the coach with the probability of injury in the next game based on recent activity, observations, and injury history. This is another limitation of current injury prediction models.

3. System analysis in the case of Pogoń Szczecin

This study focuses on analyzing data from Pogoń Szczecin, a professional football club in the Polish top-flight league with a rich tradition, and extensive experience in injury monitoring and prevention. The motivation behind this study is to enhance decision-making processes by leveraging the latest scientific knowledge in injury prediction. By examining Pogoń Szczecin's injury prevention strategies and utilizing their vast dataset, we aim to identify injury patterns and develop a more accurate prediction model. The ultimate objective is to reduce the incidence of injuries and improve the team's overall performance.

As a result of the data analysis, a comprehensive set of relevant parameters has been identified and compiled for the development of an intelligent injury prediction system.

- Player Profile: This includes basic information such as age, height, weight, and body mass index (BMI). These factors may influence an athlete's injury risk, as well as their ability to recover from injuries and respond to various training loads.
- Activity Data: Pogoń Szczecin collects data from Tracab during matches and Catapult during training sessions and matches. These systems provide detailed insights into player movements, physical exertion, and workload, which can be used to identify potential overuse injuries and monitor recovery progress.
- Screening Data: The club conducts regular screening assessments, including the NordBord hamstring strength test and functional movement assessments. These screenings help identify muscle imbalances, flexibility issues, and other potential risk factors for injury.

- Injury History: Detailed records of each player's injury history, including the type of injury, when it occurred, and the duration of recovery time. This information can help identify patterns and potential risk factors for future injuries.
- Psychological Data: Pogoń Szczecin uses morning wellness questionnaires and post-training questionnaires to assess players' psychological well-being. These questionnaires help monitor stress levels, mood, and overall mental health, which can impact injury risk and recovery.
- Creatine Kinase and Fatigue Markers: The club monitors creatine kinase levels and other fatigue markers to assess muscle damage and stress. This information can help identify athletes at higher risk of injury and inform appropriate recovery strategies.
- DNA Data: Pogoń Szczecin has access to 14 genetic markers for some players, which can potentially provide insights into individual injury risks and susceptibilities. This information can help tailor injury prevention strategies based on each player's unique genetic profile.

By compiling and analyzing this comprehensive dataset, Pogoń Szczecin can develop a more accurate and effective injury prediction model, ultimately helping to reduce the incidence of injuries and improve the team's overall performance.

In addition to the data used by Pogoń Szczecin, weather conditions will also be considered. Thanks to the publicly available database from the Institute of Meteorology and Water Management, it is possible to obtain a complete history of temperatures and precipitation amounts for the locations where matches and training sessions took place. Incorporating weather data into the injury prediction model can help identify potential correlations between specific weather conditions and injury risk. For example, wet playing surfaces may increase the likelihood of lower extremity injuries, while cold temperatures could contribute to muscle strains.

4. System Concept

4.1. System assumptions

All data layers must be considered: This includes daily activities, training sessions, matches, questionnaires, muscle strength measurements, genetic factors, and weather conditions. Incorporating all these variables allows for a more accurate and comprehensive injury risk assessment.

System utility and sequences: Injury risk classification should be responsive to the predicted future activity sequence of a given player. For example, if a player is expected to play in a match the following day, their probability of sustaining an injury in the subsequent days or sequence steps is relatively higher than if their planned activity was a training session. As research on UEFA Champions League injuries has shown [27], the number of injuries sustained during matches is significantly higher than those sustained during training sessions.

Due to the presence of independent variables, the system should be based on multiple classifiers. This approach allows for a more robust and accurate injury prediction model, capable of accounting for the complex interplay between various factors contributing to injury risk. By leveraging multiple classifiers, the system can better adapt to each player's unique profile and provide personalized injury prevention strategies.

4.2. System architecture

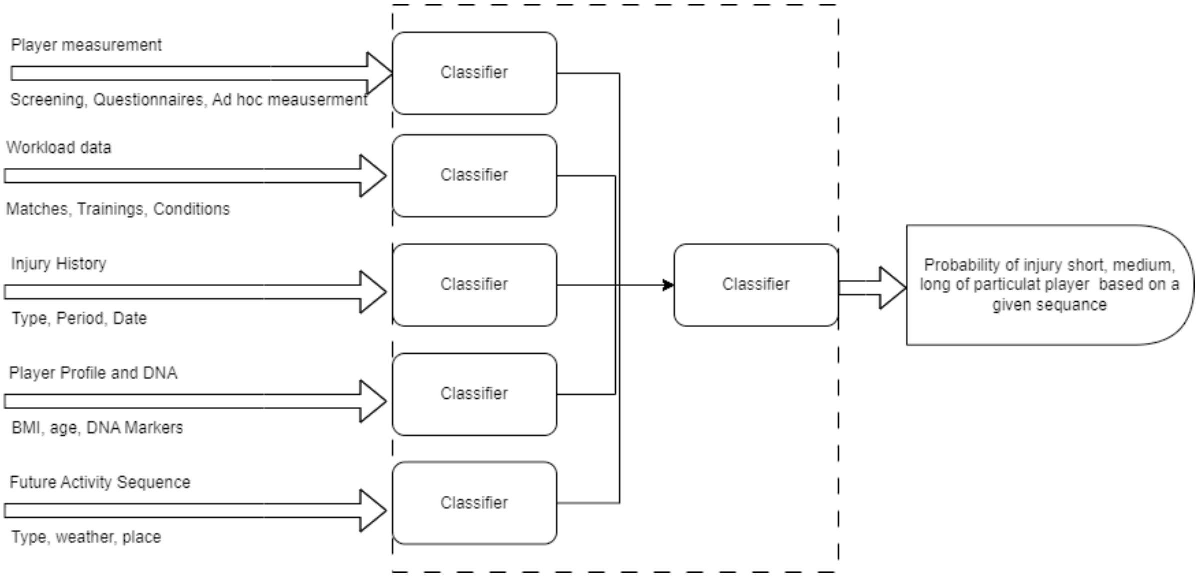

Fig. 1. System architecture

5. Conclusions and future work

Taking into account the guidelines of the project, there is a great diversity of inputs in the system. This means that there is a variation in the tempo, quantity, and importance of features. To achieve the system's objectives, which involve considering all available data, the challenge is to manage this diversity. Some data sources are frequent, such as training and survey data, while others are updated infrequently, such as muscle strength measurements before the season. Therefore, feature engineering and the way the data is processed by machine learning algorithms will be crucial, including which features to include and to what extent.

Moreover, the limited amount of fragmented data in all studies is a significant challenge. With of varying observation lengths, the dynamic nature of the team's roster leads to a loss of data continuity or limited history. Considering the need for guidelines for clinical predictive models, we may have statistically less certainty as a result. However, it still provides an opportunity for relative progress and decision support, albeit with less certainty from the outset.

Innovations and novelty in this area should focus on a broader range of data sources, encompassing the full spectrum of types from workload and surveys to DNA. Another area of focus is the problem's sequence, meaning the data and information about the player should define the workload distribution over time, ideally in a sequential form.

Due to the high variety of data sources, the first step should be to prepare an approach to data structuring and organization, including their description, completeness, and final usefulness.

After preparing and processing the data, comparative studies will be possible regarding the use of machine learning algorithms and the answer to the question of whether the achieved results with fuller data are better.

The next step will be to investigate a sequential approach to the problem of injury prediction and compare it to the previously obtained results. The ultimate goal is to achieve readiness for external validation, taking into account the condition of the next activity.

In addition, it will be essential to continue expanding the range of data sources, including a wider spectrum of types, from workload and surveys to DNA, and to approach the problem of player information as a sequence, which would define the distribution of load over time in the most explicit way.

In conclusion, future work should focus on refining data collection and organization, exploring sequential approaches to injury prediction, and expanding the scope of data sources to achieve better accuracy and usefulness for coaching staff in preventing player injuries.

References

[1] Rossi A, Pappalardo L, Cintia P (2018) “Effective injury forecasting in football with GPS training data and machine learning”, *PLoS One*. **13** (**7**):1–15.

[2] Naglah A, Khalifa F, Mahmoud A, (2018) “Athlete-customized injury prediction using training load statistical records and machine learning.” *IEEE Int Symp Signal Process Inf Technol* (ISSPIT) 459–464.

[3] López-Valenciano A, Ayala F, Puerta Jos M, (2018) “A preventive model for muscle Injuries: a novel approach based on learning algorithms.” *Med Sci Sports Exerc*. **50** (**5**):915–27.

[4] Ayala F, López-Valenciano A, Gámez Martín JA, (2019) “A preventive model for hamstring injuries in professional football: learning algorithms.” *Int J Sports Med*. **40** (**5**):344–53

[5] Rommers N, Rössler R, Verhagen E, (2020) “A machine learning approach to assess injury risk in elite youth football players.” *Med Sci Sport Exerc*. **52** (**8**):1745–51

[6] Oliver JL, Ayala F, De Ste Croix MBA, et al. (2020) “Using machine learning to improve our understanding of injury risk and prediction in elite male youth football players.” *J Sci Med Sport.* **23** (**11**):1044–1048.

[7] Vallance E, Sutton-Charani N, Imoussaten A, et al. (2020) “Combining internal- and external-training-loads to predict non-contact injuries in football.” Appl Sci. **10** (**15**):5261

[8] Kampakis S. (2016) “Predictive modelling of football injuries.” Available from: http://arxiv.org/abs/1609.07480.

[9] Bullock GS, Hughes T, Arundale AH, Ward P, Collins GS, Kluzek S. (2022) “Black Box Prediction Methods in Sports Medicine Deserve a Red Card for Reckless Practice: A Change of Tactics is Needed to Advance Athlete Care.” *Sports Med.* **52** (**8**):1729-1735

[10] Collins GS, de Groot JA, Dutton S, Omar O, Shanyinde M, Tajar A, et al. (2014) “External validation of multivariable prediction models: a systematic review of methodological conduct and reporting”. *BMC Med Res Methodol.* **14** (**1**):40.

[11] Hägglund M, Waldén M, Ekstrand J. (2006) “Previous injury as a risk factor for injury in elite football: a prospective study over two consecutive seasons.” *Br J Sports Med*. **40** (**9**):767-72.

[12] Wong-On M, Turmo-Garuz A, Arriaza R, Gonzalez de Suso JM, Til-Perez L, Yanguas-Leite X, Diaz-Cueli D, Gasol-Santa X. (2017) “Injuries of the obturator muscles in professional soccer players.” *Knee Surg Sports Traumatol Arthrosc*. **26** (**7**):1936-1942.

[13] Lathlean, T. J., Gastin, P. B., Newstead, S. V., & Finch, C. F. (2020). “Player Wellness (Soreness and Stress) and Injury in Elite Junior Australian Football Players Over 1 Season” *International Journal of Sports Physiology and Performance* 15(10), 1422-1429

[14] Larruskain J, Lekue JA, Diaz N, Odriozola A, Gil SM. (2018) “A comparison of injuries in elite male and female football players: A five-season prospective study.” *Scand J Med Sci Sports*. **28** (**1**):237-245

[15] Williams S, Robertson C, Starling L, McKay C, West S, Brown J, Stokes K. (2022) “Injuries in Elite Men's Rugby Union: An Updated (2012-2020) Meta-Analysis of 11,620 Match and Training Injuries.” *Sports Med*. **52** (**5**):1127-1140

[16] 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

[17] Mandorino M, Figueiredo AJ, Gjaka M, Tessitore A. (2023) “Injury incidence and risk factors in youth soccer players: a systematic literature review. Part II: Intrinsic and extrinsic risk factors.” *Biol Sport*. **40** (**1**):27-49.

[18] Gould HP, Lostetter SJ, Samuelson ER, Guyton GP. (2022) “Lower Extremity Injury Rates on Artificial Turf Versus Natural Grass Playing Surfaces: A Systematic Review.” *Am J Sports Med*. **20**:695

[19] Blomstrand E, Bergh U, Essén-Gustavsson B, Ekblom B. (1984) “Influence of low muscle temperature on muscle metabolism during intense dynamic exercise.” *Acta Physiol Scand*. **120**:229-36..

[20] Sugimoto D, Myer GD, Barber Foss KD, Pepin MJ, Micheli LJ, Hewett TE. (2016) “Critical components of neuromuscular training to reduce ACL injury risk in female athletes: meta-regression analysis.” *Br J Sports Med***.** **50** (**20**):1259-1266

[21] Markovic G, Šarabon N, Pausic J, Hadžić V. (2020) “Adductor Muscles Strength and Strength Asymmetry as Risk Factors for Groin Injuries among Professional Soccer Players: A Prospective Study.” *Int J Environ Res Public Health*. **17** (**14**):4946

[22] Ivarsson A, Johnson U. (2010) “Psychological factors as predictors of injuries among senior soccer players. A prospective study.” *J Sports Sci Med.* **9** (**2**):347-52

[23] Posthumus M, September AV, O'Cuinneagain D, van der Merwe W, Schwellnus MP, Collins M. (2009) “The COL5A1 gene is associated with increased risk of anterior cruciate ligament ruptures in female participants.” *Am J Sports Med*. **37** (**11**):2234-40

[24] Baird MF, Graham SM, Baker JS, Bickerstaff GF. (2012) “Creatine-kinase- and exercise-related muscle damage implications for muscle performance and recovery.” *J Nutr Metab*. 2012:960363.

[25] Fatouros IG, Chatzinikolaou A, Douroudos II, Nikolaidis MG, Kyparos A, Margonis K, Michailidis Y, Vantarakis A, Taxildaris K, Katrabasas I, Mandalidis D, Kouretas D, Jamurtas AZ. (2010) “Time-course of changes in oxidative stress and antioxidant status responses following a soccer game.” *J Strength Cond Res*. **24** (**12**):3278-86

[26] Majumdar A, Bakirov R, Hodges D, Scott S, Rees T. (2022) “Machine Learning for Understanding and Predicting Injuries in Football.” *Sports Med Open*. **8** (**1**):73

[27] UEFA Champions League injuries study Season 2018/2019, 2019; Available from: https://www.uefa.com/MultimediaFiles/Download/uefaorg/Medical/02/61/67/86/2616786_DOWNLOAD.pdf