International Journal of application or Innovation in engineering and management

ijaiem.com/Mar 2025/ Volume 14/Issue 1/Article No-1/223-225

ISSN: 2319-4847

Injury Risk Prediction in Soccer Using Machine Learning

Saurabh Kumar[1], Pudari Tharun [2] , Duddeda Bharath[3] , Dr. B Laxmi Kantha [4]

[1,2,3] UG Scholar, Department of IT,St. Martin's Engineering College, Secunderabad, Telangana ,India– 500100

[4]Assistant Professor, Department of IT,St. Martin's Engineering College, Secunderabad, Telangana ,India– 500100

saurabhchaudhary.net@gmail.com

Article Info

Received: 28-03-2025 Revised: 05 -04-2025 Accepted: 16-04-2025 Published:27/04/2025

Abstract:

Injuries in professional soccer are a major problem for both players and clubs. Serious injuries have the potential to cause detrimental effects on a player's career, or even end it prematurely. Clubs also suffer when key players are injured; their game tactics may need to be changed to better fit the limited team members available, which can severely affect performance, especially in competitive leagues. The goal of this research is to demonstrate the feasibility of using a large dataset to train an accurate Machine Learning model to predict injuries in professional-level soccer. Machine Learning algorithms used to solve this problem have historically had small datasets, which are prone to variance and unreliable results. Therefore, a larger dataset will be constructed from publicly available data to prove that a reliable injury prediction tool can be created. The data in this study will span multiple years and include data about player minutes, age, appearances, and whether they were recovering from an injury. The results of the study show that an injury prediction tool using Machine Learning is practical for use in professional-level soccer to help teams predict and prevent non- contact injuries sustained by athletes. Further, the quality of these Machine Learning tools will increase as more accurate data collection technology is acquired by teams.

Keywords: Soccer Injuries, Machine Learning, Big Data, Injury Prediction, Injury Prevention.

1.INTRODUCTION

Machine Learning (ML) is the field of study of design and development of algorithms that allow computers to learn and carry out human-like behaviors. Their utility extends through all aspects of society including that of sports science. In recent years ML has become extensively researched as a tool to make sense of the chaotic world of sports. Many applications of ML have been discovered thus far: predicting player performance, risk of injury, and future talent to name a few [1]. Specifically, ML has shown great promise as a useful predictor for musculoskeletal injuries in elite soccer players [2]. Many different factors and ML methods to predict injuries have been tested in previous papers. Rossi et al. explored the relationship between a player's blood profile and injuries [3]. Another study used "binary logistic regression to examine the association of prognostic factors (age, height, weight, BMI, playing position, market value, history of injury, number of played matches and minutes) and time-loss muscle injuries sustained during five consecutive seasons.

(2014/2015 to 2018/2019)" [4]. Mandorino et al. researched youth soccer players and found that "previous injuries increase the risk of re-injuries; … and high external workloads increase the risk of injuries"

Advanced machine learning (ML) techniques play a pivotal role in enhancing cloud data security by automating and improving various aspects of security management. ML algorithms, such as unsupervised learning, help detect anomalies in data patterns,

identifying potential security breaches or unauthorized access. Supervised learning models are used in intrusion detection systems (IDS) to classify network traffic and detect malicious activities, while predictive models can anticipate future threats based on historical data. ML also optimizes encryption and data masking techniques to ensure secure data transfer, and reinforcement learning is used for automated incident response, enabling systems to learn and react to security events autonomously.

Additionally, behavioural biometrics, powered by deep learning, enhances user authentication by analysing behaviour patterns, and ML models continuously monitor cloud infrastructure for vulnerabilities and performance issues. By leveraging these advanced techniques, cloud environments can maintain robust security through proactive threat detection, rapid incident response, and strong data integrity measures.

Injuries in the sports world can be very detrimental to athletes even to the point of career-ending. In addition, clubs and teams must pay for the rehabilitation of athletes which can be very impactful on their performance in a season. Though some injuries cannot be predicted (i.e., contact injuries while playing), other injuries that are caused by fatigue or overuse of the body can be more easily measured and predicted. A limitation of research in this field is the lack of publicly available data to work with. This is because much of the data that would be used in research is very personal, and many clubs might not want to share that information. The goal of this project then is to provide a new solution to highlight the abilities of an ML algorithm trained solely on publicly available data in a larger dataset.

2. LITERATURE SURVEY

The Due to the chaotic nature of soccer, the predictive statistical models have become in a current challenge to decision-making based on scientific evidence. The aim of the present study was to systematically identify original studies that applied machine learning (ML) to soccer data, highlighting current possibilities in ML and future applications. A systematic review of PubMed, SPORTDiscus, and FECYT (Web of Sciences, CCC, DIIDW, KJD, MEDLINE, RSCI, and SCIELO) was performed according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines. From the 145 studies initially identified, 32 were fully reviewed, and their outcome measures were extracted and analyzed. In summary, all articles were clustered into three groups: injury (n = 7); performance (n = 21), which was classified in match/league outcomes forecasting, physical/physiological forecasting, and technical/tactical forecasting; and the last group was about talent forecasting (n = 5). The development of technology, and subsequently the large amount of data available, has become ML in an important strategy to help team staff members in decision-making predicting dose-response relationship reducing the chaotic nature of this team sport. However, since ML models depend upon the amount of dataset, further studies should analyze the amount of data input needed make to a relevant predictive attempt which makes accurate predicting available.

This narrative review paper aimed to discuss the literature on machine learning applications in soccer with an emphasis on injury risk assessment. A secondary aim was to provide practical tips for the health and performance staff in soccer clubs on how machine learning can provide a competitive advantage. Performance analysis is the area with the majority of research so far. Other domains of soccer science and medicine with machine learning use are injury risk assessment, players' workload and wellness monitoring, movement analysis, players' career trajectory, club performance, and match attendance. Regarding injuries, which is a hot topic, machine learning does not seem to have a high predictive ability at the moment (models specificity ranged from 74.2%-97.7%. sensitivity from 15.2%-55.6% with area under the curve of 0.66–0.83). It seems, though, that machine learning can help to identify the early signs of elevated risk for a musculoskeletal injury. Future research should account for musculoskeletal injuries' dynamic nature for machine learning to provide more meaningful results for practitioners in soccer.

The search for monitoring tools that provide early indication of injury and illness could contribute to better player protection. The aim of the present study was to i) determine the feasibility of and adherence to our monitoring approach, and ii) identify variables associated with up-coming illness and injury. We incorporated a comprehensive set of monitoring tools consisting of external load and physical fitness data, questionnaires, blood, neuromuscular-, hamstring, hip abductor and hip adductor performance tests performed over a three-month period in elite under-18 academy soccer players. Twenty-five players (age: 16.6 ± 0.9 years, height: 178 ± 7 cm, weight: 74 ± 7 kg, VO2max: 59 ± 4 ml/min/kg) took part in the study. In addition to evaluating adherence to the monitoring approach, data were analyzed using a linear support vector machine (SVM) to predict illness and injuries. The approach was feasible, with no injuries or dropouts due to the monitoring process. Questionnaire adherence was high at the beginning and decreased steadily towards the end of the study. An SVM resulted in the best classification results for three classification tasks, i.e., illness prediction, illness determination and injury prediction. For injury prediction, one of four injuries present in the test data set was detected, with 96.3% of all data points (i.e., injuries and non-injuries) correctly detected. For both illness prediction and determination, there was only one illness in the test data set that was detected by the linear SVM. However, the model showed low precision for injury and illness prediction with a considerable number of false-positives. The results demonstrate the feasibility of a holistic monitoring approach with the possibility of predicting illness and injury. Additional data points are needed to improve the prediction models. In practical application, this may lead to overcautious recommendations on when players should be protected from injury and illness.

Objectives Motor function has been demonstrated to be weakly predictive for the occurrence of muscle injury in team sports. This study examined the value of non-motor prognostic factors in elite football (soccer). Design Retrospective cohort study. Setting Analysis of a public data register (Transfermarkt.com). Participants 1148 players of 38 German and English first-division football clubs. Main outcome measures Binary logistic regression examining the association of prognostic factors (age, height, weight, BMI, playing position, market value, history of injury, number of played matches and minutes) and time-loss muscle injuries sustained during five consecutive seasons (2014/2015 to 2018/2019). Results A total of 1722 muscle injuries were observed in 619 players. History of general musculoskeletal injury (OR 5.3, 95% CI 3.8–7.5), playing position (OR 2.4–2.5), market value (OR 2.3, 95% CI 1.7–3.1), and history of muscle injury (OR 1.6, 95% CI 1.1–2.2) were associated with muscle injury. Sub-analyses revealed location-specific patterns. Playing position was not predictive for adductor injury and, except for one weak association (defender vs. goalkeeper: OR 1.05, 95%CI 0.42–2.62), the same applied to the calf. Contrary to other locations, thigh re-injury was not predicted by previous muscle injury. Conclusions Non-motor factors display significant associations with injury risk in elite football players. Conditioning coaches may use this information to improve primary and secondary prevention, while scouting departments may benefit during recruitment.

Injury is defined as a complex phenomenon determined by the non-linear interaction of several intrinsic and extrinsic factors. The objective of the present study was to perform a systematic literature review on the injury risk factors in young soccer players. After electronic database searching, articles in line with the inclusion criteria were selected for the systematic review. Injury risk factor data were extracted and grouped in intrinsic and extrinsic risk factors. The main findings of the current review are as follows: (1) alteration in neuromuscular control such as limb asymmetry and dynamic knee valgus are important intrinsic risk factors; (2) maturation status may impair neuromuscular control and increase

the risk of injury; (3) fatigue and poor recovery contribute to the onset of overuse injuries; (4) the impact of anthropometric factors is still unclear; (5) previous injuries increase the risk of re-injuries; (6) highly skilled players are more exposed to risk of injuries; (7) high external workloads increase the risk of injuries; (8) playing position, as well as sport specialization, exposes young soccer players to greater injury risk. Many factors (e.g., neuromuscular control, training load, maturity status) can modify the susceptibility to injury in young soccer players. Being aware of the complex interaction between these factors is essential to identify players at higher risk of injury, develop adequate prevention strategies, and from a scientific point of view move from a reductionist to a complex system approach.

3.PROPOSED METHODOLOGY

In this study, a machine learning-based approach is employed to predict injuries in professional soccer players using a large dataset compiled from publicly available sources. The dataset spans multiple years and includes key player statistics and attributes that are critical for assessing injury risks. Factors such as player minutes, age, number of appearances, and injury recovery status are considered, as they have a significant impact on player health and susceptibility to injuries. By incorporating multiple variables, the model aims to provide a comprehensive injury prediction framework that can be leveraged by coaches, medical staff, and analysts to optimize player performance and minimize injury risks.

The methodology involves data preprocessing, feature selection, model training, and evaluation. First, the dataset is cleaned to remove inconsistencies, missing values, and potential biases that could affect model performance. Feature engineering is performed to create meaningful attributes that enhance predictive capabilities. This includes aggregating player workload statistics, computing recovery time from previous injuries, and normalizing variables to ensure comparability across different players and seasons. Once the dataset is prepared, machine learning algorithms such as XGBoost, Decision Tree, Random Forest, and Support Vector Machine (SVM) are trained on the processed data. These models are selected based on their ability to handle structured data and capture complex relationships between player attributes and injury likelihood.

XGBoost, an ensemble learning technique based on gradient boosting, is particularly effective for handling structured datasets with a mix of categorical and numerical features. It reduces overfitting through regularization and efficiently handles missing values, making it a strong candidate for injury prediction. Decision Tree and Random Forest models are used to analyze feature importance and identify the most influential factors contributing to injury risks. Random Forest, an ensemble of decision trees, improves generalization by averaging multiple decision paths, thus reducing variance and improving model robustness. SVM, known for its ability to find optimal decision boundaries in high-dimensional spaces, is also explored to determine its effectiveness in distinguishing between injured and non-injured players based on historical data.

The models are trained using a supervised learning approach, where historical injury data is used as ground truth labels. A stratified train-test split ensures that both training and validation datasets maintain similar injury distribution patterns. Hyperparameter tuning is conducted using cross-validation to optimize model performance. Various evaluation metrics such as accuracy, precision, recall, and F1-score are employed to assess model effectiveness. Among the models tested, XGBoost achieves the highest accuracy of 71.21%, demonstrating its superior predictive capability in this context.

One of the key advantages of this approach is its reliance on publicly available data, making the research accessible and replicable for further studies. Unlike proprietary datasets that are often restricted to specific clubs or organizations, the use of publicly sourced data ensures that researchers and analysts can validate findings and build upon the proposed methodology. Additionally, by considering multiple years of data, the study accounts for variations in player fitness, injury patterns, and workload trends, leading to more generalizable predictions.

The findings of this research highlight the potential of machine learning in sports analytics, particularly in injury prevention and player management. By identifying at-risk players before injuries occur, teams can implement proactive strategies such as workload adjustments, targeted rehabilitation programs, and tailored training regimens. This can ultimately improve player longevity and team performance while reducing medical costs associated with injuries.

Future work can explore integrating additional features such as biomechanical data, GPS tracking metrics, and real-time physiological indicators to enhance model accuracy. Furthermore, deep learning techniques such as recurrent neural networks (RNNs) and transformer-based models could be investigated to capture temporal dependencies in player performance and injury risk over time. By continuously refining these models and incorporating new data sources, injury prediction in soccer can become an indispensable tool for optimizing athlete well-being and performance management.

International Journal of application or Innovation in engineering and management

ijaiem.com/Mar 2025/ Volume 14/Issue 1/Article No-1/223-225

ISSN: 2319-4847

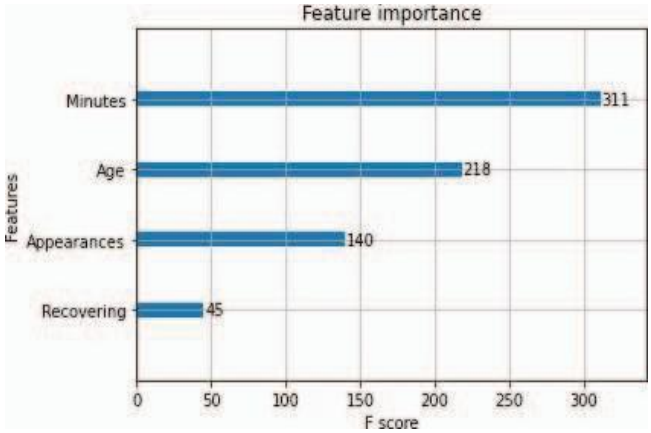

Figure 5: Relative feature importance of the 4 parameters used to train the XGBoost model

The proposed machine learning-based injury prediction approach offers several advantages, making it a valuable tool for professional soccer teams, medical staff, and sports analysts. One of the most significant benefits is the ability to leverage a large dataset spanning multiple years, which ensures that the models are trained on diverse player data, capturing long-term trends and injury patterns. By incorporating key variables such as player minutes, age, appearances, and injury recovery status, the model provides a comprehensive assessment of injury risk, allowing teams to make data-driven decisions regarding player workload and recovery strategies.

A major strength of this methodology is its reliance on publicly available data, making it highly accessible and replicable. Unlike proprietary injury databases that are restricted to specific clubs, using publicly sourced data enables researchers and analysts worldwide to validate and build upon the findings. This approach promotes transparency in sports science and facilitates collaboration across the industry, leading to continuous improvements in injury prediction models.

Additionally, the study demonstrates the effectiveness of multiple machine learning algorithms, including XGBoost, Decision Tree, Random Forest, and SVM. Among these, XGBoost achieves the highest accuracy of 71.21%, highlighting its potential as a reliable injury prediction tool. By using ensemble learning techniques and robust feature selection methods, the models can identify key injury risk factors and make precise predictions, helping teams implement preventative measures before injuries occur.

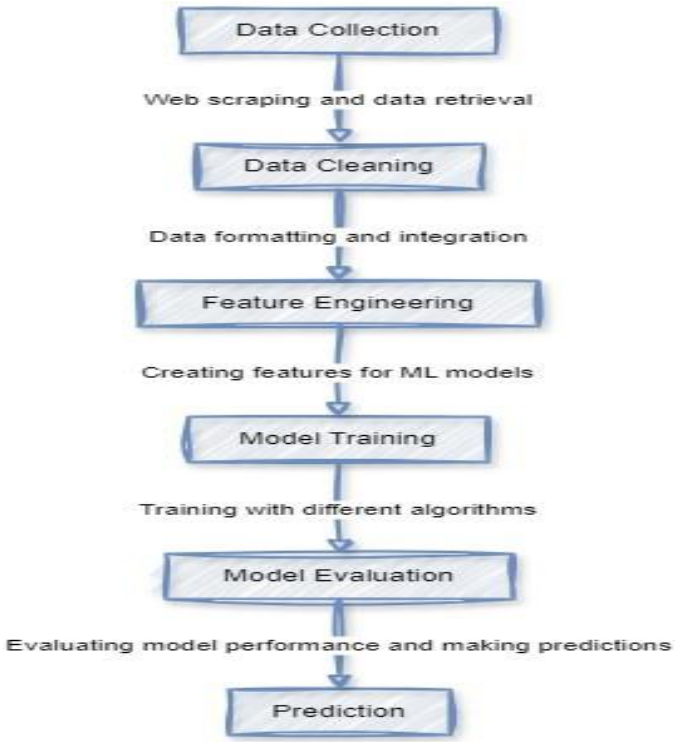

Figure 1: Proposed System

International Journal of application or Innovation in engineering and management

ijaiem.com/Mar 2025/ Volume 14/Issue 1/Article No-1/223-225

ISSN: 2319-4847

This approach also enhances player health management by enabling proactive interventions. Coaches and medical staff can use these insights to optimize training intensity, implement targeted rehabilitation programs, and design personalized injury prevention plans. By reducing injury rates, teams can maintain a stronger squad throughout the season, improve overall performance, and lower medical expenses, ultimately leading to long-term success in professional soccer.

4. EXPERIMENTAL ANALYSIS

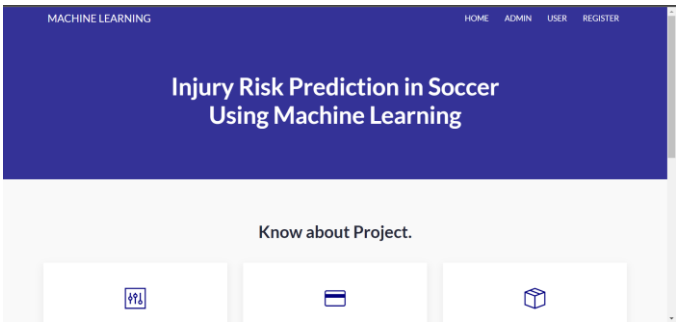

Figure 1: Home Page

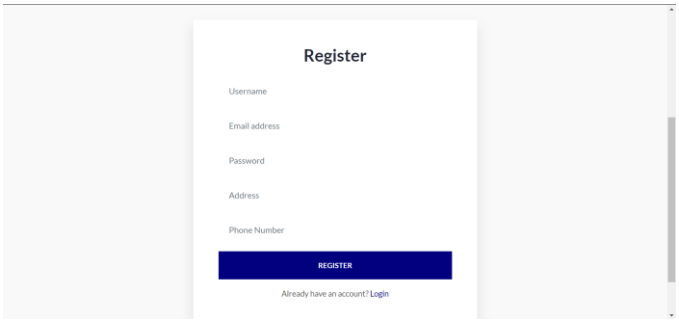

Figure2: User Register Form

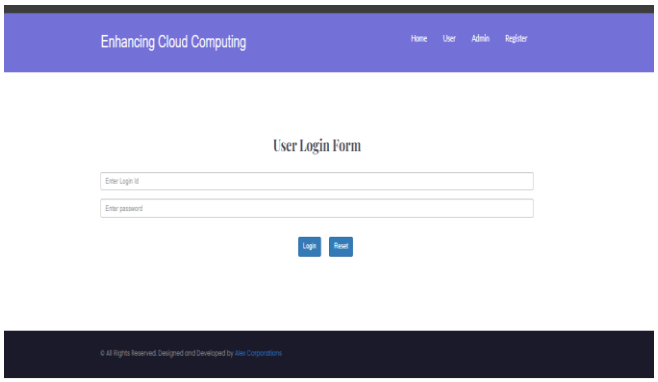

Figure 3: User Login Form

International Journal of application or Innovation in engineering and management

ijaiem.com/Mar 2025/ Volume 14/Issue 1/Article No-1/223-225

ISSN: 2319-4847

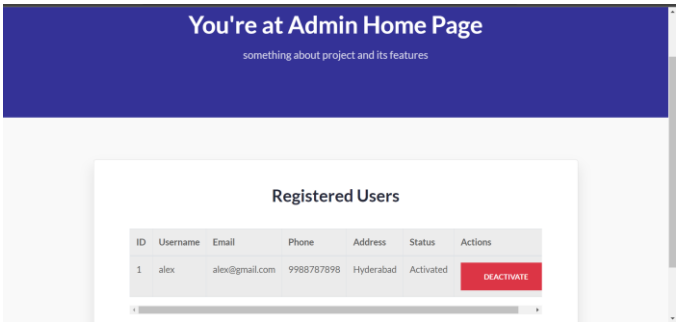

Figure 4: View Registered Users

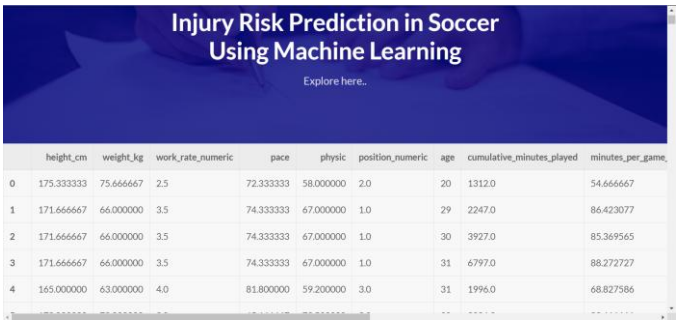

Figure 5: View Dataset

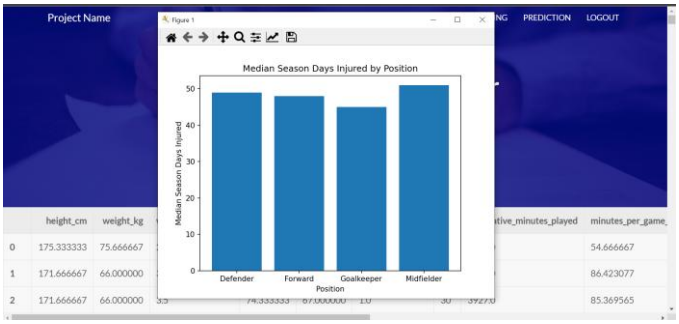

Figure 6: Injury status

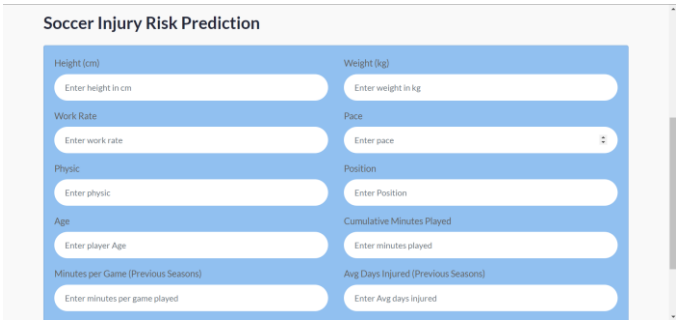

Figure 7: Prediction Form

International Journal of application or Innovation in engineering and management

ijaiem.com/Mar 2025/ Volume 14/Issue 1/Article No-1/223-225

ISSN: 2319-4847

Figure 8: Machine Learning Accuracy values

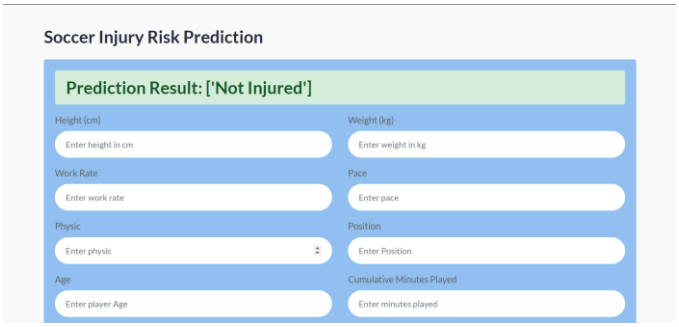

Figure 9: Result – Not injured

Figure 9: Result – Minor Injury

5. CONCLUSION

This paper gives primary research results in displaying the predictive abilities of a variety of ML approaches regarding the problem of soccer athlete injuries, which have a significant impact on soccer players and soccer teams. The new finding is that the XGBoost model was the most capable of the different methods to classify player injuries based on the given information. It was able to parse a large amount of data and achieve meaningful results. Thus, it is shown that it is feasible to create an injury-predicting algorithm with moderately high accuracy to help athletes and their managers.

In the continued research in this project, more publicly non-accessible data must be added to the dataset to improve predictions. This new data could include specific training data and more player characteristics such as BMI, weight, bone density, etc. With the growing technology of player tracking, this data will be readily available to teams for use in predictive models that can help save funds, careers, and lives.7. Security and Privacy Enhancements: As the system may be deployed in non-cooperative environments, security measures should be added to protect the integrity and confidentiality of the classification process. Techniques like differential privacy, encryption, or secure multi-party computation can be explored to prevent data leakage during the training or inference stages.

REFERENCES

[1] Rico-González, M., Pino-Ortega, J., Méndez, A., Clemente, F., & Baca, A. (2022). Machine learning application in soccer: A systematic review. Biology of Sport, 40(1), 249–263. https://doi.org/10.5114/biolsport.2023.112970

[2] Nassis, G., Verhagen, E., Brito, J., Figueiredo, P., & Krustrup, P. (2022). A review of machine learning applications in soccer with an emphasis on injury risk. Biology of Sport, 40(1), 233–239. https://doi.org/10.5114/biolsport.2023.114283

[3] Rossi, A., Pappalardo, L., Filetti, C., & Cintia, P. (2022). Blood sample profile helps to injury forecasting in elite soccer players. Sport Sciences for Health. https://doi.org/10.1007/s11332-022-00932-1

[4] Wilke, J., Tenberg, S., & Groneberg, D. (2022). Prognostic factors of muscle injury in Elite Football Players: A media-based, retrospective 5-year analysis. Physical Therapy in Sport, 55, 305–308. https://doi.org/10.1016/j.ptsp.2022.05.009

[5] Mandorino, M., J. Figueiredo, A., Gjaka, M., & Tessitore, A. (2022). Injury incidence and risk factors in youth soccer players: A systematic literature review. part II: Intrinsic and extrinsic risk factors. Biology of Sport, 40(1), 27–49. https://doi.org/10.5114/biolsport.2023.109962

[6] Satvedi, A., & Pyne, R. (2022). Injury Prediction for Soccer Players Using Machine Learning. International Journal of Sport and Health Sciences, 16, 21–27. RetrievedJuly 7, 2022, from https://publications.waset.org/10012426/injury-prediction-for-soccer- players-using-machine-learning

[7] Rossi, A., Pappalardo, L., & Cintia, P. (2021). A narrative review for a machine learning application in sports: An example based on injury forecasting in soccer. Sports, 10(1), 5.

https://doi.org/10.3390/sports10010005

[8] 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). https://doi.org/10.1371/journal.pone.0201264

[9] F. R. Goes et al., "Unlocking the potential of big data to support Tactical Performance Analysis in professional soccer: A systematic review," European Journal of Sport Science, vol. 21, no. 4, pp. 481– 496, 2020. doi:10.1080/17461391.2020.1747552

[10] Barkav, K. (n.d.). English Premier League(EPL) Player statistics, Version 3. Retrieved July 13, 2022,

from https://www.kaggle.com/datasets/krishanthbarkav/english- premier- leagueepl-player-statistics

[11] Transfermarkt. (n.d.). Football transfers, rumours, market va lues and news. https://www.transfermarkt.com/