ISSN:0976-0172

Journal of Bioscience And Technology
www.jbstonline.com

Prospective Soccer Injuries and Their Prediction using Machine Learning

[1] T. Bala Nandini, [2] P. Bhavana,

[1]Assistant Professor, Megha Institute of Engineering & Technology for Women, Ghatkesar.
[2] MCA Student, Megha Institute of Engineering & Technology for Women, Ghatkesar.

Article Info

Received: 30-04-2025 Revised: 16-06-2025 Accepted: 28-06-2025

Abstract

Players and clubs alike deal with the enormous issue of injuries in professional soccer. A player's career might take a nosedive or terminate abruptly due to a serious injury. When important players are out with injuries, clubs have to adjust their game plans to make do with the players they have, which may have a devastating effect on their performance, particularly in high-pressure competitions. Finding a way to train a Machine Learning model that can accurately predict injuries in professional soccer using a big dataset is the main objective of this project. Machine learning algorithms have traditionally relied on short datasets to address this issue, which may lead to inconsistent and unpredictable outcomes. I will build a bigger dataset using publicly accessible data to establish that it is possible to design an accurate injury prediction program. Minutes played, ages, appearances, and injury status will all be part of the data set used in this research, which will encompass a number of years. The study's findings support the feasibility of implementing a Machine Learning-based injury prediction tool for use in professional soccer, namely in the prevention and management of non-contact injuries. Additionally, as teams obtain more precise data collecting technologies, the quality of these Machine Learning tools will rise.

Keywords

Soccer-Related Injuries, AI, Large-Scale Data, Injury Forecasting, and Injury Avoidance

I. INTRODUCTION

The study of how to program computers to mimic human intelligence and perform new tasks is known as machine learning (ML). They are useful in many fields, including sports science, which touches every facet of society. The use of ML to decipher the often-chaotic sports industry has garnered a lot of attention as of late. Predicting player performance, injury risk, and future talent are just a few of the many ML applications that have been found so far [1]. In instance, ML has shown significant potential as a practical predictor for musculoskeletal injuries in top-tier soccer players [2]. Prior research has investigated a wide variety of injury prediction parameters and ML approaches. The correlation between a sportsman's blood type and injuries was investigated by Rossi et al. [3]. In a separate study, the researchers employed "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]. Youth soccer players are more likely to sustain injuries if they have a history of injuries, according to study by Mandorino et al. [5]. Also, utilizing information like age, substitutions, playing time, and appearances, Satvedi et al. [6] zeroed down on publicly available EPL data. The subject of injury prediction in soccer still has several prominent concerns because to the relative novelty of ML in

ISSN:0976-0172

Journal of Bioscience And Technology
www.jbstonline.com

sports science. The injury class was severely underrepresented in several of the datasets utilized in these investigations [7]. Thus, algorithms have a hard time seeing trends in the injury class and making reliable predictions. Rossi et al. successfully overcame this obstacle by using the "ADASYN" oversampling method [8]. It used to be difficult to get data that was useful for soccer injury prediction models, but new technology has made it much simpler to gather and utilize correct data. In particular, "systems and devices that collect and provide position tracking data" have increased in number [9]. This makes it far simpler for teams and clubs to gather massive amounts of reliable data, which they can then use to their advantage. The advent of enormous data-collecting technologies and interest in injury prediction in sports have certainly contributed to the underexplored potential of big data in this area. It is not uncommon for athletes to suffer devastating injuries that terminate their careers. The cost of an athlete's recovery, which may have a significant influence on their season-long performance, is another expense that clubs and teams must bear. Some injuries, like those sustained in a game, are impossible to foresee, but those brought on by exhaustion or abuse of the body are easier to quantify and anticipate. There is a dearth of publicly accessible data, which hinders study in this area. This is due to the fact that several clubs may be reluctant to provide the personally identifiable information that would be used for research purposes. So, this project's objective is to come up with a fresh way to showcase the capabilities of a machine learning algorithm that was trained using just publicly accessible data in a bigger dataset. The field of public health is greatly affected by the findings of this study. Both the athletes and their coaches will benefit from this system, as the former will have an easier time keeping their players safe during practices and games while the latter will have an easier time keeping their players healthy overall. A manager may use an injury predictor to save his players from potentially devastating harm during practice and games.

II. MATERIALS AND METHODS (PROCESS)

Injuries in professional soccer may be difficult to forecast, but this study proposes that ML systems can make sense of massive datasets. Another goal of this research is to determine what variables have the greatest impact on athletes' risk of injury. Section A. Synopsis Web scraping refers to the process of extracting targeted data from websites using an algorithm or script. Python modules Requests and BeautifulSoup4 were used for data scraping. Requests is an application programming interface (API) that allows Python programs to make HTTP requests. Using a GET request, the website's raw HTML may be retrieved using Requests. In order to format data and effectively parse HTML, BeautifulSoup4 is used. In this situation, we used BeautifulSoup4 to discover all the data we needed to go forward with the scraping process. Pandas is a Python module that facilitates the use of machine learning methods by transforming data into a format that is conducive to their processing. The data that was scraped was formatted using Pandas and stored as JSON files. The English Premier League (EPL) Player statistics public Kaggle dataset was utilized to get the non-scraped player data (Table I contains the parameters used in this work) [10]. The Kaggle dataset was used to derive the "Appearances" parameter. The parameters "Age," "Height," "Minutes," "Recovering," and "Injured" were constructed using data extracted from Transfermarkt.com. [11]. When the "Height" parameter was empty, the median height of the values that were filled in was used as a replacement.

TABLE I. PARAMETER DESCRIPTIONS

Parameter	Description
Name	Name of the athlete
Age	Age of athlete during the respective season
Minutes	Total time the athlete played in matches during the season
Appearances	Number of times the athlete played in matches during the season
Recovering	1 if the athlete was injured in the previous season, 0 otherwise
Injured	1 if the athlete was injured in the respective season, 0 otherwise
Height	Height of athlete in meters
Position	Position the athlete plays on the field
Assists	Number of times the player assisted a goal during the season
Passes	Number of completed passes by the athlete during the season
Yellow Cards	Number of yellow cards received by the athlete during the season
Red Cards	Number of red cards received by the athlete during the season
Fouls	Number of fouls committed by the athlete during the season
Tackles	Number of successful tackles made by the athlete during the season

ISSN:0976-0172

Journal of Bioscience And Technology
www.jbstonline.com

The Jupyter Notebook Integrated Development Environment was used to write the data gathering procedure. It facilitates the execution and debugging of Python programs. A Google Colaboratory Notebook was used for model training because of its integrated GPUs, which greatly enhance the training process. Our models were created using a classification algorithm. B. Scraping: To scrape, one had to follow a series of stages that built upon one another. With each new season comes a shakeup in the Premier League as some clubs drop down to the lower divisions and others rise to the top. The first stage, then, was to compile a roster of all Premier League clubs for every season. This was accomplished by scraping the Premier League team links from Transfermarkt's season-by-season team pages. To see the connection structure, refer to Figure 1.

Figure 1: Each link contains the team's name and the year the season started. The link shows the team "Manchester City" during the 2015-2016 season.

A dictionary was created using the player's name as the key, using the links for each team and season to scrape player links from that team (Figure 2).

Figure 2: Each player is matched with their respective subpage on the Transfermarkt website.

The injury subpages of each player were scraped for their seasonal injuries. To ensure the validity of the trial, only qualitative reports of musculoskeletal injuries were included in the data collection. The number of approved injuries dropped from 207 to 151 after the screening process. Afterwards, we gave a true/false value to each Premier League player's season: if the player had a musculoskeletal ailment that season, the value would be true; otherwise, it would be false (Figure 3).

Figure 3: Each player is matched to a dictionary where the keys are the seasons they played in the Premier League and the values are whether they were injured during the respective season.

After each season, we extracted the player's age, height, and minutes played. If a player's height was unavailable, we averaged their height and used that.

C. Cleaning the Data Processing and merging the data with the Kaggle dataset were necessary steps after data collection to finish the dataset. The process began with importing the Kaggle dataset into Python in the form of a Pandas DataFrame. The additional data from Transfermarkt might be more easily combined into a single structure because of this. After that, every element in the Pandas DataFrame had the seasonal player parameters appended to it, which included "age," "height," "recovering," "injured," and more. A JSON file containing the finished dataset was saved in preparation for its input into Google Colaboratory. Part D.: The Project Base A total of 20,069 English Premier League players from the 2015–2016 to the 2021–2022 seasons were included in the dataset. We added a record for every player that appeared in a given season to the dataset, and we viewed each season independently. Each season's classification is based on whether or not the athlete had an injury while playing. Musculoskeletal injuries were the only ones retained after sorting through Transfermarkt's extensive database of unusual ailments. The data does not allow us to tell whether the injury happened during practice or a game. For that reason, the injuries parameter will include both in-game and out-of-game injuries. The project dataset, which has 3262 items in total, was finished after filtering. A total of 1,457 submissions had injuries, while 1,805 entries were injury-free. This research conducted a thorough investigation of the dataset that was produced. A link between individual factors was shown using variable analysis (See Figure 4). The "Injured" component was most strongly correlated with the following four variables: "Age," "Appearances," "Minutes," and "Recovering." The categorization model makes use of these four parameters.

ISSN:0976-0172

Journal of Bioscience And Technology
www.jbstonline.com

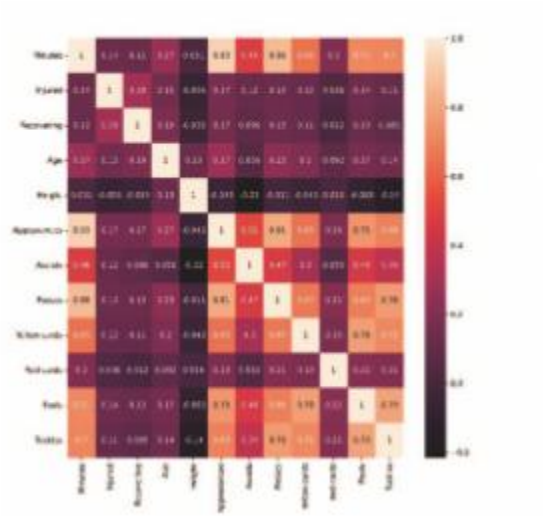

Figure 4: Feature correlation generated through Matplotlib

In order to keep the validation size appropriate and avoid overfitting, the training and test data were divided 80/20. When assessing and validating models, we employed f1 measures in addition to AUC, accuracy, precision, and recall. Additionally, K-fold cross-validations were documented using 10 data divisions. Decision tree (DT), random forest (RF), support vector machine (SVM), and XGBoost (XGB) were the four models that received the training data.

III. RESULTS AND DISCUSSION

In order to test the dataset, four different approaches were used. Among these techniques are Decision Tree, Support Vector Machine (SVM), Random Forest, and Extreme Gradient Boosting (XGBoost). In Table II, you can see the outcomes of all the models.

TABLE II. EVALUATION METRICS FOR PREDICTIVE MODELS

Model	AUC	Accuracy (%)	Precision	Recall	F1	K-fold
XGBoost	0.7602	71.21	65.08	72.44	68.56	66.12
Decision Tree	0.6363	64.32	59.29	58.25	58.76	59.69
Random Forest	0.6826	68.76	64.56	64.11	64.34	61.87
SVM	0.6838	69.68	62.31	61.83	62.07	65.48

Table II shows that XGBoost produced the greatest results and guided the study to novel and intriguing conclusions. Figure 5 shows the identified feature importances, which were obtained via additional examination into the XGBoost model.

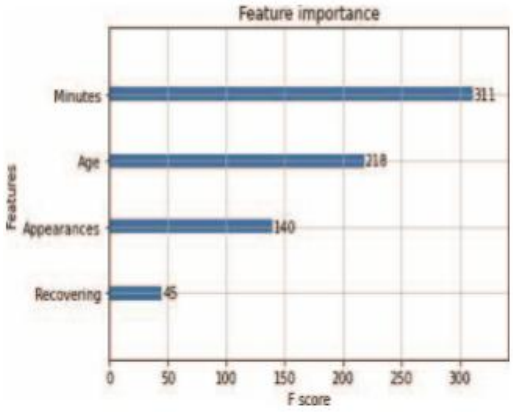

Figure 5: Relative feature importance of the 4 parameters used to train the XGBoost model

Based on the findings from the four models, we know what causes soccer injuries. Some possible reasons for how each metric influences an athlete's risk of injury are as follows. Athletes are more likely to get injuries due to wear and exhaustion the longer they play, hence minutes play a crucial role. Older athletes are more likely to sustain injuries, which is another major factor in injury risk. The more games a player participates in, the more probable it is that their body will be stressed, therefore appearances are just like minutes played in this regard. The recuperating class, which had the strongest link with the injury class when generating predictions, is surprisingly the least important in the model. Despite only having four parameters, the XGBoost model was able to successfully discover a pattern in the vast dataset. Its recall score of 72.44 percent indicates, in particular, that it is quite good at predicting real injuries. A false positive injury prediction is preferable than a false negative one, thus this is very promising. This work

ISSN:0976-0172

Journal of Bioscience And Technology
www.jbstonline.com

is groundbreaking since it is the first to show the power of a machine learning algorithm when fed a massive dataset including 3262 items. As more and more data is collected by teams during training and matches, dataset sizes will grow due to new innovations in data gathering technologies. The findings of the model demonstrate the significance of evaluating the practicality and utility of a possible injury prediction tool in a real-world context. The inclusion of both contact and non-contact injuries is acknowledged as a limitation of the data in this investigation. Because of this, very unexpected contact injury outliers are introduced. Also, we only have Premier League statistics, and as other soccer leagues use different tactics, the variables that contribute to injuries may vary significantly from one league to the next. In order to address such problems and discover new answers, our initiative will continue to conduct research.

IV. CONCLUSION

Injuries sustained by soccer players and teams are a major concern, and this article presents the findings of original research demonstrating the prediction ability of several ML techniques to this issue. According to the latest research, out of all the methodologies tested, the XGBoost model performed the best in terms of injury classification for players. It successfully parsed massive amounts of data and produced useful output. Therefore, it is shown that an algorithm for injury prediction may be developed with a decent level of accuracy, which would be beneficial for both athletes and their management. Additional publically unavailable data needs to be included into the dataset in order to enhance prediction capabilities in the ongoing investigation of this project. More player attributes, such body mass index (BMI), weight, bone density, etc., as well as training-specific data, might be included of this supplementary data set. Because player monitoring technology is improving, teams will have easy access to this data, which they can utilize in prediction models to potentially save lives, careers, and a lot of money.

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. Retrieved July 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

ISSN:0976-0172

Journal of Bioscience And Technology
www.jbstonline.com

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/englishpremier- leagueepl-player-statistics

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