https://doi.org/10.21289/KSIC.2024.27.2.257

Comparative Evaluation of Machine Learning Models for Predicting Soccer Injury Types

Davronbek Malikov[1], Jaeho Kim[2*], Jung Kyu Park[3*]

〈Abstract〉

Soccer is type of sport that carries a high risk of injury. Injury is not only cause in the unlucky soccer carrier and also team performance as well as financial effects can be worse since soccer is a team-based game. The duration of recovery from a soccer injury typically relies on its type and severity. Therefore, we conduct this research in order to predict the probability of players injury type using machine learning technologies in this paper. Furthermore, we compare different machine learning models to find the best fit model. This paper utilizes various supervised classification machine learning models, including Decision Tree, Random Forest, K-Nearest Neighbors (KNN), and Naive Bayes. Moreover, based on our finding the KNN and Decision models achieved the highest accuracy rates at 70%, surpassing other models. The Random Forest model followed closely with an accuracy score of 62%. Among the evaluated models, the Naive Bayes model demonstrated the lowest accuracy at 56%. We gathered information about 54 professional soccer players who are playing in the top five European leagues based on their career history. We gathered information about 54 professional soccer players who are playing in the top five European leagues based on their career history.

Keywords : Soccer, Data Analysis, Soccer Injury Type, Classification Machine Learning Models

1 First author, Department of AI Convergence Engineering, Gyeongsang National University, Ph.D. Student
E-mail: davronbekmalikov96@gmail.com

2* Corresponding Author, Department of AI Convergence Engineering & Department of Software Engineering, Gyeongsang National University, Professor
E-mail: jaeho.kim@gnu.ac.kr

3* Corresponding Author, Department of Computer Engineering, Changshin University, Professor
E-mail: jkpark@cs.ac.kr

1. Introduction

Football, also known as soccer, is a sport that has captured the hearts and minds of people all around the world. Due to its physical nature, soccer can be challenging to participate in and as a team all players are under pressure of different types of injuries ranging from head to toe since it requires from players multifunctioning such as: running, sprinting, jogging, etc. [1]. The type of injury can vary and may be either common or uncommon, depending on the specific type of injury [2].

Machine learning is a component of artificial intelligence that can be used in the field of sports medicine [17]. By using machine learning models, it is feasible to predict and avoid injuries by taking into account various risk factors such as the player's past injury record, the number of games played, the significance of the competition, etc. [3]. Numerous academic papers and articles have explored the utilization of machine learning models within the field of sports science, particularly in soccer, to address various injury-related challenges. These include tasks such as detecting injuries, assessing injury risks, monitoring performance, predicting injuries, and guiding rehabilitation processes [1 ,4-6].

For example, a multi-dimensional approach combining GPS data and machine learning has been developed for injury forecasting in soccer. This approach strikes a balance between accuracy and interpretability [8]. Moreover, in that approach compares the performance of Decision Tree (DT) with Random Forest (RF), Logistic Regression (LR), Autoencoder-Stacked WaveNet Regression (ASWR), Multi-Stacked WaveNet Regression (MSWR), forecasters, as well as four baseline methods [8]. In a soccer injury prediction study that utilized GPS technology and wearable devices a combination of rule-based and fuzzy rule-based approaches, along with the XGBoost algorithm as a machine learning baseline, was implemented [6].Numerous academic articles examining the prediction of soccer injuries have predominantly emphasized the collection of data via GPS technology or similar electronic devices[14,15,16]. Nevertheless, challenges arise in acquiring consistent data using GPS technology for soccer teams due to limitations, permission from soccer teams, time-intensive procedures, and potentially prohibitive cost of such devices, rendering them unsuitable for all teams. Furthermore, there is a need for further research to focus on predicting specific types of injuries.

One of the main objectives of this paper is to showcase the potential of collecting data through professional soccer websites that data readily available on the web as an alternative to using GPS technology or other mechanical devices. Additionally, this paper introduces a method for predicting specific types of soccer injuries among players. In order to achieve this goal, we evaluate and compare different classification machine learning models in our

experiment. The results of our study provide valuable insights into the key factors that contribute to injury occurrences. This information can be advantageous not only for the future careers of soccer players but also for coaches, scouts, team directors, and even fans who aspire to witness their favorite players perform without interruptions.

2. Data for Predicting Soccer Injury

In this section, we present the data used in this paper. Additionally, we provide information on the soccer injury types and features.

2.1 Data Introduction

Data is a crucial component in every machine learning model. The effectiveness and accuracy of the model heavily rely on the quality and quantity of the data utilized for training. Machine learning algorithms learn through the data they receive, which emphasizes the importance of proper data selection. This paper focuses on the analysis of 54 professional soccer players who have been playing in the highest-ranking European football leagues, including the English Premier League, Spanish La Liga, Italian Serie A, German Bundesliga, French League 1. The Top 5 European football leagues, often regarded as the most competitive globally, attract premier talent and host intense competitions like the UEFA Champions League. Analyzing players with experience in these prestigious events can provide valuable insights into their abilities at the highest level. To delve into this aspect, we have identified and chosen 54 players currently active in the Top 5 European leagues in our research. The dataset we utilized comprised of soccer players who have a significant impact on game outcomes. There are 54 players included in the dataset, and they occupy various positions in the game. The dataset consists of 54 players, with 20 players identified as center-forwards, 10 as left-wingers, 5 as right-wingers, 3 as second-strikers, 6 as attacking midfielders, and the remaining 10 as central midfielders. we provide information on the soccer injury types and features.

Additionally, in order to acquire information on the career and injury records of the players, we made use of a specialized website for soccer player's information [9]. The website is a useful resource to gain significant knowledge about professional soccer players, including their career statistics, transfer history, market value, and injury records.

2.2 Data Set of soccer injury

A soccer injury is defined as any physical ailment that resulted in a player being taken out of a game, sitting out a game, or being sufficiently impaired to require medical attention [9]. A soccer injury is defined as any physical

ailment that resulted in a player being taken out of a game, sitting out a game, or being sufficiently impaired to require medical attention [10]. Moreover, contact injuries and non-contact injuries are the two main types of soccer injuries. Contact injuries are caused by collisions with objects or other players, whereas non-contact injuries result from factors such as repetitive strain, poor technique, or sudden changes in movement. The primary objective of this paper is to concentrate on the four predominant injury types that are commonly experienced by professional soccer players. These four types of injuries include thigh injury, ankle injury, muscle injury, and knee injury. Our research paper utilized the career and injury history data of soccer players from our dataset.

Table 1 shows an example of dataset that depicts the career and injury history of professional soccer players, as examined in our study. Our dataset consists of a total of 12 columns. The "Player" column displays the names of the players in accordance with the season of their career, which is indicated in the "Season" column. The other columns are as follows: "Total" indicates the total number of games played in a particular season, while "Minutes" represent the total number of minutes played across all the games in that season. Our dataset encompasses FIFA, UEFA, and team games in the top five leagues due to their significant influence on the occurrence of soccer injuries among players. FIFA games, including renowned tournaments like the World Cup, Continental Cup, and Olympic Games, hold significant importance for several reasons.

These reasons include the opportunity for players to represent their country, evoke national pride, enjoy fan support, and engage in high-level competition. Moreover, in top leagues, matches typically take place on weekends, while UEFA games are scheduled for Tuesdays and Wednesdays. This arrangement results in players having to play at least 2 or 3 games per week. Consequently, the UEFA games can increase the risk of soccer players getting injured due to the added workload and potential overexertion. Furthermore, considering the factors mentioned above matches in FIFA, UEFA, and the top five leagues are deemed important and can have an impact on the likelihood of getting injured. The number of

Table 1. Example of soccer player dataset

Player	Season	Total	Minutes	FIFA	CHL/EL	Important	Past4	Past3	Past2	Past1	Injury
G. Bale	2006-2007	46	3948	3	0	39	0	0	0	0	0
G. Bale	2015-2016	38	2941	7	8	37	1	0	2	2	2
C.Immobile	2020-2021	56	4349	15	5	46	0	0	2	0	0
C.Immobile	2021-2022	41	3433	1	7	39	0	1	0	1	0
E. Hazard	2017-2018	63	4661	11	8	48	2	2	0	1	0

Note. 0:Non-injury, 1: Ankle-injury, 2: Muscle-injury

games played in these competitions is reflected in the "Important" column of the table. The "Injury" column indicates the type of injury, while the "Past4" to "Past1" columns indicate the type of injuries the player had in the past.

Furthermore, we used various machine learning models, which we compared in order to identify the optimal fit for predicting different types of soccer injuries. Our objective is to enhance our comprehension of the elements that contribute to such injuries, with the ultimate goal of devising more efficacious tactics for preventing injuries among soccer players.

3. Machine Learning models and Comparative Evaluation

This section presents a summary of the machine learning models and their categories that are applicable in the prediction of soccer injuries. Moreover, our study involved the comparison of numerous machine learning models to determine the most suitable one for predicting different types of soccer injuries.

3.1 Machine Learning Models

Machine learning models are algorithms that can automatically learn from data without being explicitly programmed. These models use statistical techniques to identify patterns in data, and with sufficient training, can make accurate predictions or decisions on new data. Machine learning can be classified into different types depending on the nature of the problem being addressed, the type of data available, and the learning process involved. These types of machine learning encompass supervised learning, unsupervised learning, semi-supervised learning, reinforcement learning, and deep learning.

In our study, we employed a supervised machine learning model. This is because supervised machine learning models are trained using labeled data, which indicates that the input data is provided with corresponding output values. However, supervised machine learning can be divided into two main categories: classification and regression. Classification is used to predict which category or class a new data point belongs to, based on its features. Regression, on the other hand, is used to predict a numerical value, such as a price or a score, based on the input features. However, this type of model is used when the output variable is continuous and can not be classified into categories. Therefore, we utilized supervised classification machine learning models in our study due to the characteristics of our dataset and the objectives of our research.

We employ a range of machine learning models, including Decision Tree, Random Forest, K-Nearest Neighbors (KNN), and Naive Bayes since our main focus revolves

around the classification of injury types, necessitating the utilization of specialized machine learning models designed for classification tasks. The ultimate objective is to pinpoint the most appropriate model for our dataset, placing a significant emphasis on achieving a high level of accuracy to enhance confidence in the model's predictive capabilities. Morover, each of these classification models possesses its own set of strengths and weaknesses, and the selection of which one to utilize hinges upon the characteristics of the data and the specific problem being addressed. Decision trees employ a hierarchical structure to represent decisions and their potential outcomes. Conversely, random forests are ensemble methods that combine multiple decision trees to enhance accuracy and mitigate over-fitting. K-Nearest Neighbors (KNN) is a non-parametric technique that assigns data points to classes based on the classes of their nearest neighbors in the feature space. KNN can be applied for both classification and regression tasks. Naive Bayes is a probabilistic model that leverages Bayes' theorem to estimate the probability of a sample belonging to a particular class. Furthermore, thorough and systematic comparison process empowers us to confidently select models that demonstrate superior performance in classifying injury types within the specific context of our research.

3.2 Comparative Evaluation

Classification is one of the fundamental problems in machine learning, and it involves assigning a class or category to a given data point. In this context, we used various classification machine learning models to classify data points based on their features.

As part of our study, we employed a data splitting approach to effectively assess the performance of our machine learning model. In order to achieve this, we utilized the `train_test_split` function from the `sklearn. model_selection `module. By specifying a test size of 30%, we ensured that a significant portion of our dataset was dedicated to evaluating the model's generalization capabilities.

Consequently, 70% of the data was allocated for training the model. This larger portion allowed the model to learn patterns and relationships within the data, facilitating its ability to make accurate predictions or classifications when exposed to new, unseen instances.

By dividing our data into separate training and testing sets, we were able to train the model on the training data, fine-tuning its parameters and optimizing its performance. The testing data, on the other hand, served as an independent benchmark to evaluate how well the model generalized to new, unseen data. The precise separation of data into training and testing subsets plays a vital role in objectively assessing the model's performance, identifying over-fitting, and validating its capability to handle real-life situations. It guarantees that the model is dependable and proficient when confronted

with unfamiliar data outside of its training set. Our study involved a comparison of the aforementioned models, utilizing two different approaches. The first approach involved applying three distinct types of injuries, whereas the second approach utilized four different types of injuries. We conducted this comparison to assess the performance of each model under different injury scenarios and to determine which model is better suited for each type of approaches.

Fig. 1 depicts the representation of the first approach and second approaches, which included three and four different types of injury categories. The graph shows the variation of machine learning models with respect to their accuracy scores. In the first approach, the K-Nearest Neighbors (KNN) model performed the best on our dataset, achieving an accuracy score of 70%. The second-best model was the Decision Tree model, which achieved an accuracy score of 69%. The Random Forest model achieved an accuracy score of 66%, whereas the Naive Bayes model had the lowest accuracy score of 58% compared to the other models.

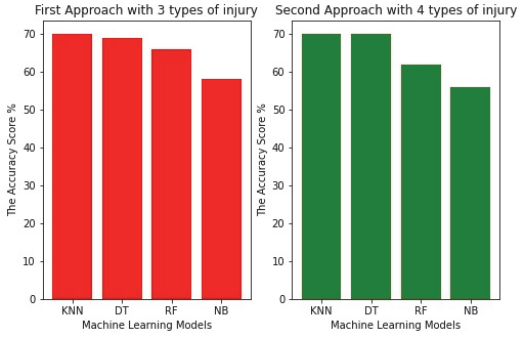

Fig. 1 Representation of first and second approaches

In the second approach it is evident that the second approach involved four different injury categories and illustrates the disparities in accuracy scores among the various machine learning models. The KNN and Decision models obtained the highest accuracy scores of 70%, followed by the Random Forest model with an accuracy score of 62%. Among the models, the Naive Bayes model had the lowest accuracy score of 56%.

4. Experimental Evaluation

In the Comparative Evaluation subsection, four different machine learning models were compared using two different approaches. It was determined that KNN was the most suitable machine learning model for our dataset in both approaches.

In this section, the KNN model is utilized to determine the probability of each class. The model is specifically designed for 3 classes, as it has been found to have a higher accuracy compared to when it is used with 4 classes. In our study, we implemented a novel approach to predict injury types for individual players. We leveraged the historical injury data available to public and organized the players into groups based on their injury history. By analyzing the frequency and severity of specific injury types, we aimed to identify patterns and tendencies within the dataset.

For instance, if a player had experienced a higher number of ankle injuries compared to other players, our analysis indicated that there was a higher likelihood of that player sustaining an ankle injury in the upcoming season. This information enabled us to make predictions regarding the potential injury risks faced by individual players. Finally, the decision is made based on the probabilities of each class and the groupings of players.

4.1 Probability of Classes

Initially, we utilize three categories of injury and leverage the scikit-learn library, which is accessible in Python, to determine the probability of each injury type. Moreover, we analyzed three categories of injuries: non-injury, ankle injury, and muscle in-jury. The probability values for each category are shown in Fig. 2. We observed that the class with the highest probability among the others

Injury Type	Probability
non-injury	0.6
ankle injury	0.2
muscle injury	0.2

Fig. 2 Probability of injury type

Class	Precision	Recall	F1-Score	Support
non-injury	0.74	0.89	0.81	171
ankle injury	0.25	0.09	0.13	33
muscle injury	0.40	0.24	0.30	34

Fig. 3 Probability of injury type

is non-injury, which has a probability of 60%. Moreover, ankle and muscle injury have the same probability, which is 20% respectively.

In addition to displaying the probability of each injury type, we have also included a classification report in Fig. 3. The scikit-learn library, which is a widely used Python library, was utilized to generate a classification report that would enable us to evaluate the effectiveness of our model.

This report provided us with a comprehensive overview of the model's performance by presenting a variety of metrics such as precision, recall, and F1-score for each injury category. By examining these metrics, we were able to identify areas where the model performed well and where it fell short. This information can be used to refine and improve the model's performance in future iterations. The use of scikit-learn library for generating the classification report made the evaluation process efficient and streamlined, enabling us to gain insights into the model's performance with ease.

Moreover, precision was used to evaluate the true positive rate among the total predicted positives for each injury category. The precision score for "non-injury" was 0.74, indicating that only 74% of predicted non-injury cases were actually correct. Recall, or sensitivity, was also computed for each category to measure the proportion of true positives among all actual positives. For "non-injury," the recall score was 0.89, which

means the model could only identify 89% of actual non-injury cases. The F1-score, a comprehensive measure of the model's accuracy that combines precision and recall, was also calculated and resulted in 81%. Finally, the support column listed the number of instances of each injury type in the dataset, revealing insights into the dataset's composition. It was found that the number of "non-injury" instances was 171. These metrics are crucial for evaluating the model's performance and determining its ability to accurately detect various types of injuries.

4.2 Probability of Classes

In the above subsection, we have successfully calculated the probability for each class, using the KNN machine learning model, which we determined to be the most appropriate for our dataset.

In the following subsection, we will utilize the probability information obtained to identify players who may be susceptible to specific types of injuries in our dataset of 54 professional soccer players. To accomplish this, we will divide the players into three groups based on the three classes previously established and we consider the number of cases that players got injury with that specific type. Our model will be used to select the top five players who have already sustained injuries with those specific types. The top five players who have already suffered injuries with those specific types will be prioritized, as they have the highest risk of being injured again.

Player	Injury_Type	Count
P.Aubameyang	non-injury	20
C.Immobile	non-injury	18
O.Giroud	non-injury	17
G.Bale	non-injury	16
T.Muller	non-injury	15

Fig. 4 Top-5 non-injury players

Fig. 4 presents the top five players in our dataset who are the least likely to sustain injuries. Additionally, the figure reveal that the count column represents the number of instances where players did not suffer any injuries over the course of their career. Based on the probability information obtained, all five players listed in the figure have a 60% chance of not getting injured according to our model. Moreover, it is worth mentioning that based on the count column data, P. Aubameyang has the lowest probability of getting injured with 20 non-injury cases, whereas among the top five players with the least likelihood of sustaining injuries, T. Muller has the highest probability with only 15 non-injury instances.

By utilizing the method that we employed, we generated a group for class 2, which comprises players who have sustained ankle injuries.

Fig. 5 displays information on the top five players who have experienced the highest number of ankle injuries in our dataset, in comparison to other players. The data in Fig.

Player	Injury_Type	Count
M.Reus	Ankle Injury	7
J.Felix	Ankle Injury	7
P.Pogba	Ankle Injury	5
H.Kane	Ankle Injury	5
G.Bale	Ankle Injury	5

Fig. 5 Top-5 ankle injury players

Player	Injury_Type	Count
M.Reus	Muscle Injury	13
E.Hazard	Muscle Injury	9
K.Benzema	Muscle Injury	8
E.Cavani	Muscle Injury	8
P.Dybala	Muscle Injury	7

Fig. 6 Top-5 muscle injury players

2, combined with the probability data obtained earlier, suggests that M. Reus and J. Felix are the players with the highest chance of suffering an ankle injury among those listed with 7 ankle injury cases, as there have been seven instances of ankle injuries and a 20% likelihood for both players. Furthermore, our model predicts that P. Pogba, H. Kane, and G. Bale have a reduced likelihood of experiencing an ankle in-jury compared to the other players, as there are only five instances of ankle injury associated with them.

Finally, we employed the same methodology to create a group of players who have suffered from muscle injuries. Fig. 6 shows the top five players who have experienced the highest number of muscle injuries, when compared to the other players in our dataset.

Moreover, based on Fig. 6 and the probability information gathered earlier, it can be inferred that M. Reus is the player most prone to muscle injuries, with a probability of 20% and 13 instances of muscle injuries compared to other players. Conversely, P. Dybala is the player with the least probability of sustaining a muscle injury among the top five players mentioned in the table, as there are only seven instances of muscle injuries associated with him.

5. Conclusion

The main goal of this paper is to ascertain the chances of different types of injuries that professional soccer players may encounter during the forthcoming season. Ultimately, we managed to forecast the likelihood of each of the three-injury categories, with non-injury having a 60% probability, while ankle and muscle injuries both having a 20% probability. Furthermore, our findings revealed that P. Aubameyang had the lowest probability of sustaining an injury compared to other players in our dataset, with a 60% chance of not getting injured. On the other hand, M. Reus and J. Felix had the highest likelihood of experiencing an ankle injury among all soccer players in our dataset, with a 20% chance of injury. Additionally, our results indicated that M. Reus had the highest probability of suffering a muscle injury, with

a 20% chance of getting injured.

We plan to address these limitations in future work to improve our ability to predict a wider range of injury types beyond the three types considered in our study. By doing so, we can enhance the usefulness and applicability of our research for injury prevention and management in soccer players. For example, we used data from only one professional soccer website, which may have introduced biases and limited the generalizability of our findings. Therefore, future research could incorporate.

Acknowledgements

This work was supported in part by the National Research Foundation of Korea (NRF) grant number (No. 2021R1F1A1063524 & 2021R1F1A1052129) and the research grant of the Gyeongsang National University in 2023.

References

[1] P.Wong and Y Hong, "Soccer Injury in the lower extremities," Br J. Sports Med. vol.39, pp.473-482. (2005).

[2] SOCCER INTER-ACTION S.L., <https://soccerinteraction.com/the-most-serious-injuries-in-football> viewed 12 December (2023).

[3] H. Sigurdson and H. Chan, "Machine Learning applications to Sports Injury," Proc. of the 9th International Conference on Sport Sciences Research and Technology Support (icSPORTS 2021), pp.157-168, (2021).

[4] Catapult. Provider of Wearable Technology for the Management of Athletes, <https://www.catapultsports.com/.> viewed 12 December (2023).

[5] A. Rossi, L. Pappalardo and P. Cintia, "A Narrative Review for a Machine Learning Application in Sports: An Example Based on Injury Forecasting in Soccer," Sports vol.10, no.1:5. (2022).

[6] A. Lyubovsky, Z. Liu, A. Watso, S. Kuehn, E. Korem and G. Zhou, "A Pain Free Nociceptor: Predicting Football Injuries with Machine Learning," Smart Health, vol.24, pp.10026, (2022).

[7] T. Piłka, B. Grzelak, A. Sadurska, T. Górecki and K. Dyczkowski, "Predicting Injuries in Football Based on Data Collected from GPS-Based Wearable Sensors," Sensors vol.23, no.1227, (2023)

[8] A. Rossi, L. Pappalardo, P. Cintia, F.M. Iaia, J. Fernàndez and D. Medina, "Effective injury forecasting in soccer with GPS training data and machine learning," PLoS One. vol.13(7), no. e0201264, (2018).

[9] Transfermarkt GmbH & Co. <https://www.transfermarkt.com/> viewed 12 December (2023).

[10] W.B. Kibler, "Injuries in adolescent and preadolescent soccer players," Med SciSports Exerc vol.25 no.1330–2, (1993)

[11] T.J. Gabbett, "The training-injury prevention paradox: Should athletes be training smarter and harder?," Br. J. Sports Med. vol.50, pp.273–280, (2016)

[12] B. Zhang, "Tactical Decision System of Table Tennis Match based on C4.5 Decision Tree," 13th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA), pp.632-635, (2021).

[13] F. Nurwanto, I. Ardiyanto and S. Wibirama, "Light sport exercise detection based on smartwatch and smartphone using k-Nearest Neighbor and Dynamic Time Warping algorithm," 8th International Conference on Information

Technology and Electrical Engineering (ICITEE), pp.1-5, (2016).

[14] T. O. D. Beéck, A. Jaspers, M. S. Brink, W. G. P. Frencken, F. Staes, J. J. Davis and W. F. Helsen, “Predicting Future Perceived Wellness in Professional Soccer: The Role of Preceding Load and Wellness,” Int J Sports Physiol Perform. vol.14, no.8, pp.1074-1080, (2019).

[15] A. Jaspers, T. O. D. Beéck, M. S. Brink, W. Frencken and F. C. Groningen, “Relationship between the external and internal training load in professional soccer: what can we learn from machine learning?,” Int J Sports Physiol Perform. vol.13, no.5, pp.625-630, (2017).

[16] P. G. Campbell, I. B. Stewart, A. C. Sirotic, C. Drovandi, B. H. Foy and G. M. Minett, “Analysing the predictive capacity and dose-response of wellness in load monitoring,” J Sports Sc. vol.39, no.12, pp.1339-1347, (2021).

[17] R. K. Martin, A. Pareek, A. J. Krych, H. M. Kremers and L. Engebretsen, “Machine learning in sports medicine: need for improvement,” Journal of ISAKOS, vol.6, no.1, pp.1-2, (2021).

(Manuscript received January 23, 2024;
revised March 06, 2024; accepted March 13, 2024)