World Academy of Science, Engineering and Technology
International Journal of Sport and Health Sciences
Vol:16, No:3, 2022

Injury Prediction for Soccer Players Using Machine Learning

Amiel Satvedi, Richard Pyne

***Abstract*—**Injuries in professional sports occur on a regular basis. Some may be minor while others can cause huge impact on a player's career and earning potential. In soccer, there is a high risk of players picking up injuries during game time. This research work seeks to help soccer players reduce the risk of getting injured by predicting the likelihood of injury while playing in the near future and then providing recommendations for intervention. The injury prediction tool will use a soccer player's number of minutes played on the field, number of appearances, distance covered and performance data for the current and previous seasons as variables to conduct statistical analysis and provide injury predictive results using a machine learning linear regression model.

***Keywords*—**Injury predictor, soccer injury prevention, machine learning in soccer, big data in soccer.

I. Introduction

THE main purpose for conducting this research work is to create an application for predicting the likelihood that a soccer player will obtain an injury while playing. Machine Learning regression models will be used to determine when a soccer player will likely get injured while playing in the near future by training and testing the application with statistical data extracted from sports analytics websites and articles. Injury prediction will be performed by analyzing the relationships among variables such as number of matches played during the current season; number of minutes played during current season; player position; distance covered each game; and distance covered during a season. The resulting prediction analysis will then be used to inform the player's coach of the likelihood that the player will be injured so that the player's coach can take the necessary intervention measures to prevent the injury from occurring.

II. Related Literature

Injuries are one of the biggest challenges players and team managers face in professional sports. Some of these injuries can easily be treated, while others can be critical, leaving players injured for months or sometimes ending a player's career. This is especially evident in the game of soccer where most of the top-class players are playing league matches and international matches for up to nine months in a year. Heuer and Rubner describe statistical analysis techniques used for decision making by coaches and managers. These techniques use historical data to find patterns in data sources that help to better predict future injuries [13].

The English Premier League (EPL) data sets will be used as test cases for the experiments conducted in this research work [22]-[24]. The EPL is one of the most competitive leagues in the world and is filled with high intensity games [11]. For this reason, it would be very useful to have a mechanism for coaches and team scouts to assist them in assessing the injury risk profile of players. This injury risk profile can also inform team administrators of the players that add the best value to their team and thereby increase the team's chance of success in the EPL [1].

A. Machine Learning in Sports

Previous research was conducted using linear regression model for injury prediction on players in Soccer and American Football where machine learning was involved [20], [5]. Linear regression is one of the most relevant algorithm techniques in the field of statistics that is used in machine learning. Machine Learning is a form of programming that transforms input retrieved from multiple data sources to output on the basis of experiences and changes in the environment that the program has learnt over time [2]. Machine learning is vital for automatically learning and enhancing previous systems without performing much programming effort. Machine learning can be very useful when performing tasks that are very complex and difficult for humans to do within a reasonable timespan [3], [7].

In a research based on decision tree and neural networks, unstructured complex data were studied that contained different scenarios for problem-solving, this unique scenario was known as game playing. The analysis of historical data was used to assist in finding different patterns and reduced uncertainty [18]. Decision tree building is a predictive modeling approach used in machine learning that is simple to used yet can be very powerful. Neural networks are a class of machine learning algorithms that can be used to determining complex mappings from input sources to the output spaces [10]. In a typical problem-solving scenario, the primary task is to reduce the complexity by trimming down a set of items required into smaller set for a classified result. When analyzing the performance of a player, one key point that needs to be extracted is past data that match the variables being observed. Furthermore, a key approach for an efficient prediction system would be to involve an expert or knowledge-based method [10]. In this method, a program would be provided with a set of rules to describe various

A. Satvedi is a recent graduate of the Faculty of Applied Science & Technology, Sheridan College, 1430 Trafalgar Road, Oakville, On, L6H 2L1, Canada (e-mail: satvedi@sheridancollege.ca).

R. Pyne, Ph. D., is a professor with the Faculty of Applied Science & Technology, Sheridan College, 1430 Trafalgar Road, Oakville, On, L6H 2L1, Canada (phone: 1-905-459-7533; fax: 1-905-874-4331, e-mail: richard.pyne@sheridancollege.ca).

situations that can be applied to a knowledge base. As the knowledge base of the subject grows and more rules are set in place the capabilities of this system can be enhanced.

Additionally, [5] discusses efficient machine learning application that uses skeletal training. Signals from a video camera are captured to perform real-time tracking at a low cost. This sort of tracking adapts the technique of decision tree combined with training data. The depth images pick up poses of different body parts that are labeled. When this information gets transferred into a machine learning application for the training phase, these parameters are added to the decision tree along with the depth of the tree. In recent times, the abundance of digitized data has grown all across the world and with the help of cloud computing many datasets are hosted at data centers affordably. The extent to which data sources are available has broadened the impact of machine learning applications.

The research work being undertaken will require a statistical analysis in order to understand which key variables will be needed to perform arithmetic functions for prediction analysis. Previous work has been conducted for soccer teams using regression analysis to predict the final outcome for individual teams in soccer matches [13], [4]. Variables such as goal difference in the past and half-of the current seasons were also used to provide a prediction for the final number of goals at the end of the season. At the start of this process, a linear regression analysis was performed by separating dependent and independent variables keeping in mind the number of matches played, team strength and observations of past and present data [13]. Calculations and predictions of these sorts can easily be transferred to our current research on injury prediction for soccer players. A linear regression analysis can be used to model a relationship between a dependent and an independent variable. Additionally, there are various techniques available to automate the training, testing and validation of linear regression models using machine learning. Furthermore, [21] presented the use of neural networks to manipulate data from sensors. They recognize patterns that are numerical and vector-based and translate them to real-world data. Creative analysis using a neural network in any sport can assist in understanding various aspects of an individual player's performance. One type of network discussed in [21] was called dynamically controlled network. Neurons were trained to build clusters of input data that match items without the use of additional training information [21]. A single neuron can learn and deliver data constantly for statistical purpose when the network is used after training is complete. An autonomous neural network model can then be developed by analyzing each controlled neuron [21]. We observe that data of various types can be trained in different phases based on previous player data collected that show changes in motion, the kilometers traveled every season as well as injuries obtained in previous games. This type of quantitative information can help to understand the input data as well as show different patterns that were not visible due to the enormous amounts of the data set being investigated.

Finally, other researches [15], [17] done in the sport of baseball for high school players who obtained serious injuries concluded that the chances of an injury occurring are increased during the later stages of the game. One of the biggest reasons provided was the increase intensity of the game when it comes to running, pitching, hitting, and the general increase in competitiveness between the teams when the game was on the line. As a result, soccer coaches and managers tend to substitute star players out of the game between the 75th to 80th minute marks of the game if the game is already decided. This way they can reduce the risk of injury. Therefore, as it relates to injury, another important category of player attribute is when a player is substituted in, or out of the game [19]. As such, we will need to use Big Data tools for categorizing and sorting information properly.

B. Big Data

Big data traditionally means a massive volume of data [12]. Historical player data will be used to predict a likelihood of future injury to a player. In a study done on the 2014/2015 German Bundesliga Professional Soccer League, it was determined that data cleaning and data integration of Big Data from multiple data sources were highly important when performing analyses. Data manipulations were performed using conventional data analysis, with calculations based solely on positional information collected from that particular season [20].

Big Data can be found in weblogs, GPS systems, text documents, spreadsheets and databases on the internet as well as in various other places. For this research, the three Big Data characteristics we will focus on are volume, variety and velocity [12]. With respect to the game of soccer, volume can be the set of games that a player plays in a season, number of passes completed, assists given and goals scored. Variety refers the different types of data sources that are available to use such as player data for different positions, health reports and medical data of previous injuries as well as the player performance metrics. These data sources may contain data that are both quantitative and qualitative. Velocity would be the real-time data that would be collected while players are playing. All three concepts are very important for analyzing data in professional sports [16].

Additionally, there are generally three different types of Big Data, namely, structured, unstructured and semi-structured [12]. Structured data are data that are more organized data and describe the relationship present between all contents available within the data space. The information can easily be analyzed as there is the presence of a data pattern which can facilitate data analyzing and SQL queries. An unstructured data do not contain any sort of pattern, making it a challenge to find information quickly. Finally, semi-structured data are in the middle between structured and unstructured where only a portion of the data is well-described [8]. Both semi-structured and unstructured data will need to be cleaned, organized and integrated before meaningful automated analysis can be performed.

C. Analyzing Big Data in Professional Soccer

In professional soccer, there has been a wide variety of changes in the style of play, training methods, and fitness regimen as players get older. One noticeable change in the EPL is the evolution of match performance statistics and metrics compiled on players [6]. There was an increase in intensity and the amount of distance traveled in the premier league between the years 2006-2013. These data were collected using camera tracking system [6]. This system then generated a data set of each player's technical and physical performance. This player information was stored in public domain data files that contained information of players who completed at least 90 minutes each game [6]. These players were then categorized into separate sections based on the positions they played, such as, defenders, central midfield, wide midfield, attackers etc. These data files on player statistics and metrics have been organized by seasons and have been made available to the general public for consumption, analysis and forecasting purposes [6].

In recent years, player and match analyses have concentrated on finding the strengths and weaknesses of a team and the presumed readiness of a team for upcoming matches. These types of analysis have allowed managers to focus on specific areas to improve, such as fitness; style of play; pace of play; team formation and field management. A set of selected items and indicators can be observed for a specific player or group of players (for example, all the defenders) to determine adjustments that are needed to be ready for an upcoming game. This is generally referred to as a game plan, and is crucial to the success of a team [18]. Some of the basic analysis that has been discussed [18] includes patterns and shapes made by attacking players when searching for goal scoring opportunities and player movement off-the-ball to provide more passing options [9]. This analysis was done by the Spanish Professional Soccer league for the 2008-2009 season [18]. The information gathered from all teams was broken down into four different groups. Statistics gathered and analyzed were goals scored, offensive efficiency, defensive efficiency and game context, such as home-game, away-game and venue played. One of the interesting results obtained from the analysis showed that the home advantage played a huge role in the final result of a game. Teams that were winning more games were more effective at their offensive sets, team organization on the field of play, pace, and field management which allowed them to create more goal scoring situations. Furthermore, the team at the top of the table was averaging a higher number of assists compared to the ones in the bottom. Reference [18] demonstrates that performing statistical analysis can provide helpful information that goes toward team success. This analysis helps managers and coaches to understand the game better and also gives data driven evidence of improvement that need to be made to have a good game plan for each game.

Importantly, the right tool to support Big Data manipulation is critical in performing strong and in-depth data analysis. Large scale data need to be managed in the right manner, as it comes from multiple sources that need to be extracted, cleaned and integrated properly to get more accurate results. In the case of serious injury prediction related to accidents and crashes dealing with cars, [14] divided a data set into multiple groups with samples of accidents that occurred during a 12-year period. One group consisted of data that contained crashes that occurred with one or two passengers and whether or not there was involvement of heavy vehicles. Other groups consisted of accidents with new vehicles; accidents with old; and severity of the injuries. This kind of categorization assisted in formulating insurance premiums for drivers and their various types of automobiles. Reference [14] demonstrated that informed categorization of data can aid significantly in the data analysis process. Therefore, we will adopt a similar strategy of categorization of our data set as part of our data analysis. The data sources we will use are available on team and league website for the general public to access. Since this research will involve Big Data from varying sources, it is best to use a Non-relational Structured Query Language (NoSQL) based application such as SAS or Tableau to establish multiple categories according to player attributes when manipulating the data. We have chosen to use Tableau as the data analysis application. We will use the Jupyter Notebook Integrated Development Environment (IDE) which allows the execution of Python code, data cleaning, data integration, and gives access to the Python machine learning libraries that will determine the injury predictor value. We will then use the Tableau data analytic workbench for data visualization that will enable us to establish our injury prediction indicators and risk profile categories.

III. Experimental Framework and Preliminary Results

During our experiment, the predictive model will be obtained through the use of a linear regression machine learning algorithm from the scikit-learn machine learning library available in Python. The linear regression algorithm uses a linear equation of the form:

$$y = mx + c$$

where y is the predicted value (dependent variable), m represents the slope of the regression line, x is the input to be supplied (independent variable) and c in the y-intercept, that is, the point at which the regression line crosses the y-axis. Since we will be observing the relationships of several independent variables with a single dependent variable (Distance_per_season from Fig. 1), then a multiple linear regression algorithm will be used to generate the coefficients (the m-value) for each player attribute (input, independent variable). A split, train, test and validate process will be used to setup our predictive model. Splitting of the data will be done by filtering the dataset to focus on only the independent input variables that are needed to be used to obtain the predicted value. The variables in Fig. 1 will be observed.

The split data sets are used for the purpose of training and testing the linear model. After the completion of the multiple linear regression algorithm, a set of coefficients (m-values) and the y-intercept (c-values) will be determined. In Fig. 1, we observed the character of the data set to be used for the purpose of training and testing the linear regression model. The focus variables used include the dependent variable “distance covered all season” and the independent variables “number of appearances”, “number of minutes played”, “players age”, “number of times subbed on” and “the number of times subbed off”. During the next step, all the independent variables will be used to split the data for training and testing, where the training set will contain 80% of the data and the test set will contain 20% of the data as seen in Fig. 2.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 253 entries, 0 to 252
Data columns (total 6 columns):
Distance_per_season    253 non-null float64
Appearance             253 non-null int64
Minutes_Played         253 non-null int64
Age                    253 non-null int64
Subbed_on              253 non-null int64
Subbed_off             253 non-null int64
dtypes: float64(1), int64(5)
```

Fig. 1 Data set characteristic

```
Training prediction variable includes : 202 rows
Training the independent variable includes : 202 rows
Testing prediction variable includes : 51 rows
Testing independent variable includes : 51 rows
```

Fig. 2 Testing and Training data

After running the machine learning regression algorithm, a set of coefficients and the intercept values are obtained. These values, as seen in Fig. 3, will be used for analyzing and visualizing the data in the Tableau workbench.

```
Intercept : 76.1863185868

   coeficients focus Variables
0     0.171873        Appearance
1     0.095864    Minutes_Played
2    -2.493532               Age
3     0.088562         Subbed_on
4     2.864500        Subbed_off
```

Fig. 3 Intercept and coefficient for prediction

Using the values in Fig. 3, the next step was to use the 80% of training-data-set to train the linear regression model to determine the training score. When this was done, the training score obtained was 77%, see Fig. 4.

```
Training score: 0.77
```

Fig. 4 Training score using 80% of the data set

At this point we used the 20% of testing-data-set to test the linear regression model to determine the test score. This test score will measure how well our linear model can be used for prediction. When this was done, the test score obtained was 76%, see Fig. 5.

20% Regression Results			
==			
Dep. Variable:	Distance_per_season	R-squared:	0.766

Fig. 5 Testing score using 20% of the data-set

```
Cross Validation score: [ 0.64987097  0.75823383  0.30704045  0.83932281  0.64866024  0.78894233
  0.83822462  0.69183908  0.93426389  0.819153   ]

Average 10-Fold Score: 0.7275551225624345
```

Fig. 6 K-Fold Cross Validation score

```
InjPred

(76.1863185868 +
[Appearance] * 0.171873 +
[Minutes Played] * 0.095864 +
[Age] * -2.493532+
[Subbed_on]* 0.088562 +
[Subbed_off]* 2.864500)
```

Fig. 7 Injury Predictor Computation

An important step in the process is to validate the 76% test score of the linear model by using a k-fold cross validation process. This validation allows for the generalizing of the model when it is provided with new data. This is done by dividing the data into k equal parts and then making each part contribute to the testing set. The final score will be computed as a mean average of all the scores for each of the kth-part of the data used for testing. From Fig. 6, the mean average score of the k-fold validation (where k = 10 in this experiment) was 72%; the median average lies between 75 and 78% while the mode average lies within the range of 64 and 83% as shown in Fig. 6. As such, we can consider the 76% test score for the linear model to be valid since it falls within the average ranges of the validation scores.

The injury predictor value is then obtained by the computation in Fig. 7 derived from the multiple linear regression machine learning algorithm.

The computation in Fig. 7 adds the intercept value to the coefficients multiplied by the several independent variables used to determine the prediction value. At this point, we have the option to determine the prediction values for all players; a set of players by position played; or an individual player. Fig. 8 displays the list of players for the forward (attacking) position for the 2018-2019 EPL season. This function can be performed by choosing the appropriate <Position> checkbox from the right panel. It is important to note that when a player (or group of players) is selected their actual coordinate value is displayed by a green-dot while the predicted value is displayed using a degree of light to dark red-dot.

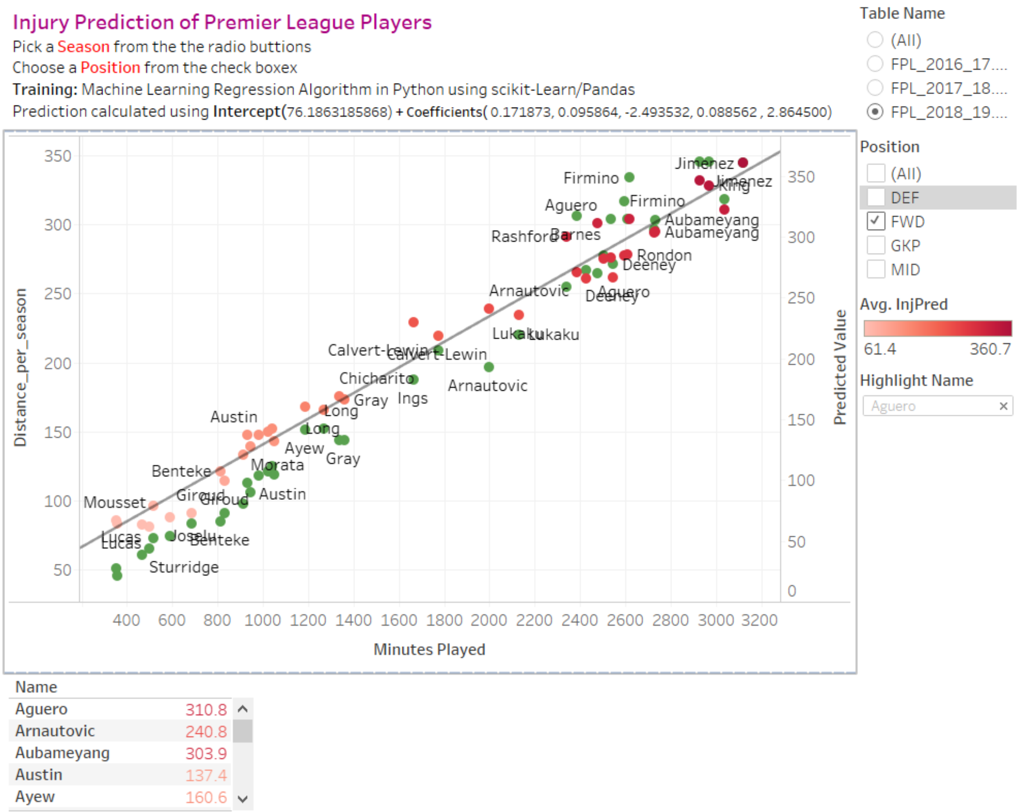

Fig. 8 Forward Position Players (actual and predicted values)

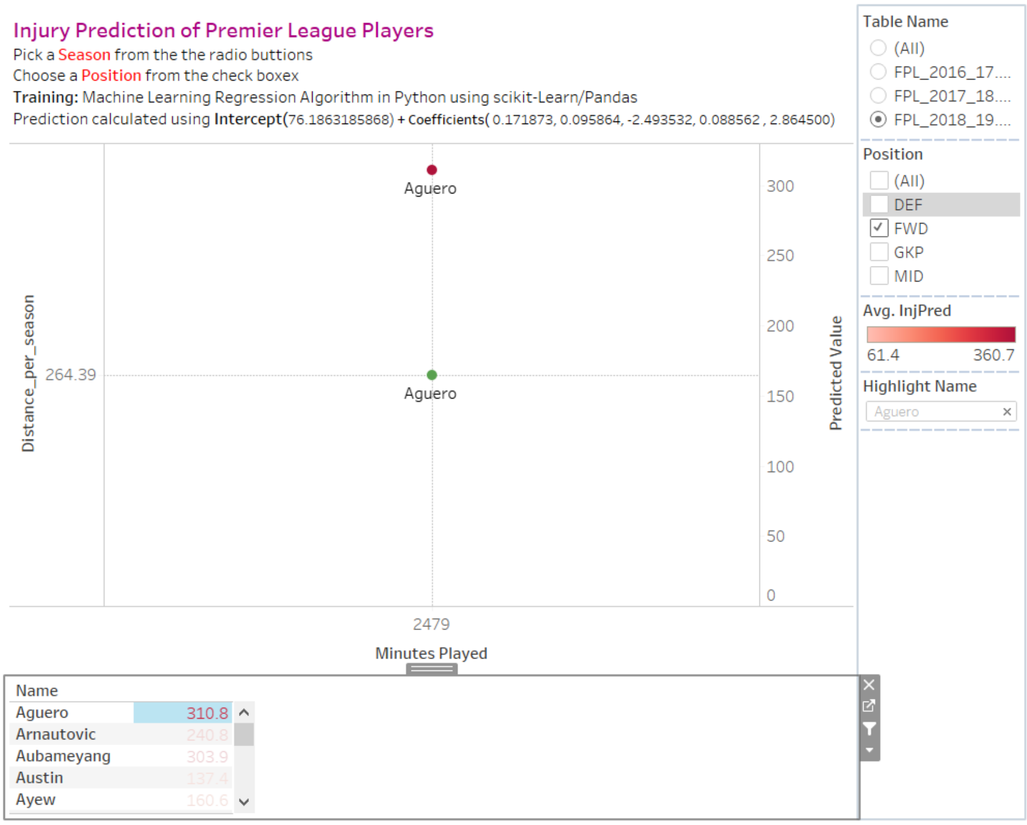

Fig. 9 Player named "Aguero" selected from the list

Additionally, each player will be placed into the low, medium or high risk profile categories. The range for each of the categories is determined by:

maximum_predicted_value – minimum_predicted_value / 3

Therefore, from Fig. 8, the range for each category is 360.7 – 61.4 / 3 = 99.76, where the low risk profile range is from 61.4 – 161.16 (indicated by light-shade red-dots), the medium risk profile is from 161.17 – 260.93 (indicated by medium-shade red-dots) and high risk profile is from 260.94 – 360.7 (indicated by dark-shaded red-dots).

An individual player can be selected either by finding the name from the list of player's menu, as seen in Fig. 8 or by clicking on the <Highlight Name> textbox in the right panel and searching for a player by name. When the <Highlight Name> textbox is activated, a list of players will be displayed from which a player can be selected. A player can be selected from the list presented or by manually entering the player's name as seen in Fig. 9.

From Fig. 9, the predicted value for the high risk profile player named Aguero is 310.8, while the actual value is 264.39. This means that the actual distance travelled by this player is 264.39 km during the season. The computed predicted value of 310.8 km is indicating that whenever this player reaches (or surpasses) this number of distance travelled during the season, the player has a 76% chance of picking up an injury while playing. Since a player generally travels 10.5 km during a standard 90-minute game [25] in the premier league, then this player has a 76% likelihood of being injured within the next five games played based on our model.

IV. Conclusions

The use of this application can be very beneficial for soccer team coaches and training staff as they prepare their teams to compete in the EPL. The injury prediction indicator can give a birds-eye view of each player on a team as well as allowing the coach to determine when to implement intervention strategies for a particular player. For example, the coach may decide to implement playing-time restriction per game for a particular player; substitute a player in the later stage of a match; rest a player for certain number of games; change a player position so they travel less distance during a game; or recommend a particular training regimen for a player who has a higher risk injury profile. This would work best for various sets of players playing too many games, and players covering a large amount of distance during games. The training staff can apply this software to give players the right type of training and rest schedule to reduce the risk of the player getting an injury. This application can also be used by team scouts who are sent around the world to find new players for their teams. Scouts can analyze a player of interest and determine the players risk profile to decide if the player will add value to their team and to understand the chances of such a player playing consistently throughout a season of play.

V. Future Work

Currently, the data sources used by the application are static data files. The next step in the evolution of this application will be to make the data sets dynamic by retrieving them in real time. This would provide the end-user with the most recent iteration of player information and thereby making the prediction more current. Another improvement is to add more soccer leagues from across the world as currently only players in the EPL are activated. Additionally, it would be nice to include data sets that could examine players in the youth academy of a team to track how injury prone they are from an early age and to implement injury preventative actions. Furthermore, we could incorporate options for other kinds of professional sports such as basketball, baseball and athletics.

Finally, the importance of tracking players using motion sensor technology could greatly optimize and increase the strength of the analysis and prediction that can be performed on a player in real time. This would involve a player being attached with a tracking device close to their chest or on their arm to capture player information that can then be used by the injury predictor application in real time.

References

[1] B. Alamar, Sports analytics: a guide for coaches, managers, and other decision makers. New York: Columbia University Press, 2013.
[2] E. Alpaydın, Introduction to machine learning second edition. Cambridge, MA: The MIT Press, 2014.
[3] R. J. Bar and J. F. X. DeSouza, "Tracking Plasticity: Effects of Long-Term Rehearsal in Expert Dancers Encoding Music to Movement," PlosOne, 29-Jan-2016.
[4] F. Wunderlich and D. Memmert, "The Betting Odds Rating System: Using soccer forecasts to forecast soccer," PlosOne, 05-June-2018.
[5] K. Pelechrinis and E. Papalexakis, "The Anatomy of American Football: Evidence from 7 Years of NFL Game Data," PlosOne, 22-Dec-2016.
[6] M. Bush, C. Barnes, D. T. Archer, B. Hogg, and P. S. Bradley, "Evolution of match performance parameters for various playing positions in the English Premier League," ScienceDirect, Feb-2015.
[7] B. Pang, L. Lee, and S. Vithyanathan, "Thumbs up? Sentiment Classification using Machine Learning" Association for Computational Linguistics, July-2002.
[8] K. Ozcan, A. Mahabalagiri, and S. Velipasalar, "Autonomous tracking and counting of footsteps by mobile phone cameras," An introduction to biometric recognition - IEEE Journals & Magazine, Nov-2015.
[9] J. Castellano, D. Casamichana, and C. Lago, "The Use of Match Statistics that Discriminate Between Successful and Unsuccessful Soccer Teams," Economics and Culture, 23-Oct-2018.
[10] H. Chen, P. B. Rinde, L. She, S. Sutjahjo, C. Sommer, and D. Neely, "Expert Prediction, symbolic learning, and neural networks," IEE-Expert Volume 9 Issue 6, Dec-1994.
[11] V. Di Salvo, W. Gregson, G. Atkinson, P. Tordoff, and B. Drust, "Analysis of high intensity activity in Premier League soccer," Current neurology and neuroscience reports., Mar-2009.
[12] Roiger R. (2016). Basic Data Mining Techniques. In Cohen R. (Ed.), Data Mining a Tutorial-Based Primer (pp. 64-95). Boca Raton, Florida: CRC Press.
[13] A. Heuer and O. Rubner, "Optimizing the Prediction Process: From Statistical Concepts to the Case Study of Soccer," PlosOne, 08-Sept-2014.
[14] T. Nishimoto, K. Mukaigawa, S. Tominaga, N. Lubbe, T. Kiuchi, T. Motomura, and H. Matsumoto, "Serious injury prediction algorithm based on large-scale data and under-triage control," ScienceDirect, Jan 2017.
[15] Nationwide Children's Hospital. "Injuries to High School Baseball Players Becoming More Serious," ScienceDaily, 03-Jun-2008.
[16] J. Kang and D.-S. E. Joonbeom Lee, "Smartphone-Based Traveled Distance Estimation Using Individual Walking Patterns for Indoor Localization," American Journal of Drug and Alcohol Abuse, 18-Sep-

2018.
[17] O.-S. Kwon and D. C. Knill, “The brain uses adaptive internal models of scene statistics for sensorimotor estimation and planning,” PNAS, 12-Mar-2013.
[18] C. Lago-Penas, J. Lago-Ballesteros, A. Dellal, and M. Gomez, “Game-Related Statistics that Discriminated Winning, Drawing and Losing Teams from the Spanish Soccer League,” Journal of Sports Science and Medicine, 01-June-2010.
[19] G. D. Myer, K. R. Ford, J. Khoury, P. Succop, and T. E. Hewett, “Biomechanics laboratory-based prediction algorithm to identify female athletes with high knee loads that increase risk of ACL injury,” British Journal of Sports Medicine, 17-June-2010.
[20] D. Memmert and R. Rein, “Match Analysis, Big Data and Tactics: Current Trends in Elite Soccer,” Research Gate, March-2018.
[21] D. Memmert and J. Perl, “Game creativity analysis using neural networks,” Journal for Sports Sciences, 19-Jan-2009.
[22] AFS Enterprises. (2020). English Premier League Statistics of Player’s Age filtered by club (Data File). Retrieved from https://www.11v11.com/teams/arsenal/tab/stats/option/oldest/
[23] Transfermarkt GmbH & Co. KG. (2020). English Premier League Statistics of Player’s Injury History (Data File). Retrieved from https://www.transfermarkt.com/kieran-trippier/verletzungen/spieler/95810
[24] Premier League. (2020). English Premier League Statistics of Player Attributes (Data File). Retrieved from https://www.premierleague.com/stats/player-comparison
[25] Fenn, A. (2017, September 22). Premier League v amateur fitness. FourFourTwo. https://www.fourfourtwo.com/us/performance/training/premier-league-v-amateur-fitness