National College of Ireland

A Machine Learning Framework to Scout Football Players

MSc Research Project
Data Analytics

Hashir Sayeed
Student ID: X21214611

School of Computing
National College of Ireland

Supervisor: Paul Stynes, Eugene McLaughlin, William Clifford

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Hashir Sayeed
Student ID:	X21214611
Programme:	Data Analytics
Year:	2023
Module:	MSc Research Project
Supervisor:	Paul Stynes, Eugene McLaughlin, William Clifford
Submission Due Date:	14/08/2023
Project Title:	A Machine Learning Framework to Scout Football Players
Word Count:	3848
Page Count:	16

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	25th August 2023

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

A Machine Learning Framework to Scout Football Players

Hashir Sayeed
X21214611

Abstract

Scouting football players involves selecting players that have potential to play at a national level for which they are being assessed and also the optimal position they can play based on their performance such as goals scored, accuracy of passes and so on. The current challenge is that the scouts may not identify all the performance factors during their assessment for example Command, Communication and so on. This research proposes a Machine Learning Framework to scout football players that have the potential to play at a national level. The proposed framework combines a classification model and a predictive model. The classification models identifies the optimal position of the player. The predictive models identifies best possible match in a national team for every position. The dataset is given by the FIFA organization, who are responsible for keeping updated statistics of the players.

1 Introduction

Scouting a suitable and talented players among the pool of professional players is a challenge. It takes a lot of time and effort to find a potential player and watch his performance in person and the performances in the past through video recordings and then decide whether the player is a good fit for the team or not. As mentioned in the article (Kidd; 2020), the whole process of scouting which includes 12 or 13 scouts which analyze at-least 130 games per players which are on recording and then go for live matches. This whole process is very time consuming. Also, scouts may not identify all the performance factors during their assessments and might neglect the factors such as Command, Communication and so on. Säfvenberg (2022) All teams strive to assemble their teams with the greatest players available at every position to improve their chances of success. The international world cup, the Euros which is an international tournament for European teams, and other major football events are among the most eagerly awaited occasions. Making a team with the highest calibre players is therefore really essential. Every country's selection committee must assess players who are currently competing in numerous leagues around the globe to determine which among them is the best fit to make the team. This makes the role of scouts very important and essential. With all the leagues and potential players, the task becomes more and more time consuming and difficult as they have to assess a lot of players which can or cannot be an asset to their respective teams

There are a lot of advancements in using the machine learning techniques with the data

generated by the sports to get crucial insights. Morciano et al. (2022) talks about the players and their performances according to their positions of play and roles and the impact they have on the overall outcome in a team sport like football or soccer. This supports the idea that creating a team with best possible players can enable a team to acquire more sophisticated tactical advancement which in turn boost the likelihood of winning and overall success of the team. Rajesh et al. (2020) indicated the process of finding positions of the players and building a team would be beneficial to the managements and help them reduce the costs by finding players of similar positions but with less wages.

The writers of the aforementioned publications have discussed the use of machine learning approaches to forecast game results, player injury risk, goal scoring trends, etc., all of which contribute to a football team's success. Supporting the idea of finding the right players for the club will significantly boost their chances of success, which will be profitable for management. Thus, making the task of scouts more and more relevant and important but at the same time making it more and more complicated, hefty and time consuming task.
The above research has influenced the following research question **How well Artificial Neural Networks can predict the position and value football players?**. The proposed solution will include a machine learning framework which consist of two stage. First stage will consist of a classification model which will predict the possible position of the player depending upon the statistics and skills. The second stage will consist of a regression model which will determine the value of the player which will lead to selection of players which are more valuable. To implement the framework, specified sets of research objectives were derived:

- Investigate the state of the art of predicting the position of the player given in the article Rajesh et al. (2020) and recreate the model defined.
- Design a framework which predicts the position of the players and then predicts the value of the players.
- Implement the framework
- Evaluate the classification model using confusion matrix and accuracy score and regression model using RMSE and MAE.

The further structure of the documents is as follows:

- Section 2 consists of related work which includes the in-depth critical analysis of papers which were previously published and are around the area of interest. The paper talks about the machine learning algorithm's implementation to predict the match results, players contributions, effects of attributes of players, risk of injury, effect of passing styles, value of players etc. Thus, relating to the idea of the research question and supporting it.
- Section 3 consists of the implementation of the framework. This includes the description of how the framework was created and how the pipeline of the framework was set and implemented and what results were acquired from each experiments that were performed during the implementation.

- Section 4 consists of evaluation of the models in the framework and how they performed and their significance. This also stated the different evaluation metrics used to evaluate the models and why they were used.
- Section 5 stated the conclusion of the who research and its significance in the respective domain and the potential future work.

2 Related Work

In general the sports domain have a complex nature due to numerous factors effecting the in-game direction and overall outcome. Thus, creating an alliance of the predictive analysis and the domain of sports. Danisik et al. (2018) stated in the article that the sports sector is constantly looking for ways to improve, so it incorporates insights from predictive analysis and machine learning to identify patterns and alter tactics to support growth. As a result, it draws people and businesses looking to win big by outsmarting the odds in gambling, adding to the reasons that make it a very attractive field.

2.1 Sports incorporating machine learning

In order to learn more, numerous studies involving various sports have been conducted. For example, Miguel et al. (2019) used machine learning to forecast the NBA draft using historical information about the players and drafts made. The author's study was based on the relationship between NBA selection order and career longevity. Author implemented the bayesian multilevel modeling to forecast the longevity of the player's career so that it can incorporate the prior effect of the variables and compute the probability of each one. In addition to it, the author also implemented the generalized additive model using Hamilton Monte Carlo. Consequently, stating that the pick affected the players' careers and that as the picks grow, the players' careers went shorter. Also, Tirtho et al. (2022) in their paper, made a forecast of the player's performances by factoring on their prior performances, and based on the outcomes, recommended the ideal squad. The players for each position and class was evaluated by the author, thus then created the future team with the highest level of performance. The application of machine learning in sports and how it can advance the existing state of the art are discussed in both of the aforementioned publications. Predicting the outcome of the game is the most popular interest. It might change depending on a variety of factors, and in order to arrive at findings, all of the aspects must be taken into consideration.

2.2 Match Result Prediction

The team's players are one of the most significant elements or factors influencing the outcome of the game. They are the ones whose performance can affect how the game turns out. Each football squad consists of 11 players. As a result, forecasting is extremely difficult because each player's engagement has a varied impact. Danisik et al. (2018) attempted to foretell the outcome of the game by examining each player's unique traits. To enhance the present mode, the authors decided to use a neural network using LSTM and back propagation. The regression model outperformed the classification model in the author's evaluation of the model as he performed experiments using both strategies. The model can accurately forecast the match results by 52.4 percent, according to the final

analysis. According to this, the players' actions have a greater than 50 percent impact on the final outcome. The model might become better with the introduction of other factors. The squad and players have a significant and extremely crucial impact on the outcome, which served as the primary driving force behind this article. Hence, it becomes evident why a team would focus primarily on strengthening their squad.

2.3 Improving the Squad

The management of the club must analyse all of the players who are currently on it and how they may improve by, if at all feasible, locating the best replacement on the market. Numerous considerations must be made in order to identify the best players for a certain role. To identify and assess these characteristics, many research have been conducted. The player's age is one of the things that affects them and their performance. Saavedra-García et al. (2019) emphasised the significance and varying effects of player age. The author employed an additive model and discovered an intriguing finding that showed a relationship between player height, performance, and age and year of competition. The first 90 days of the year, or a quarter, were selected by the author because they are frequently employed to discover or examine the relative age effect. The first quarter's performance was compared to the year as a whole. The study revealed that the RAE fluctuated throughout the year. Additionally, because younger participants had lower RAE and their RAE was approaching a stable section as they became older (approaching 35), there was no relevance found in their data. Säfvenberg (2022) added his findings to the research and stated that by taking into account several statistics like team composition, player position, goals, strategies used, tackles, throws, assists, free kicks, etc., it provided the player's peak performing age. The study backed up the notion that a player's peak performance is crucial, and that every player in every position has a different range of peak performing ages.Thus, managements can improve squads by taking into account the age factor. Research is going on to reveal the additional elements that can be used to strengthen the squad. For example, Rossi et al. (2022) emphasised the significance of top athletes' health under demanding and competitive conditions because it directly affects their performance and enhances comparative performance over time. Also, Robles-Palazón et al. (2023) brought out the significance of reducing injury risk among the players to assist them perform better by forecasting injury using the players' medical and fitness test data. In order to improve the unbalanced data, the paper included ensemble approaches to forecast the probability of injury among the players.All of these lend support to the squad's improvement idea. However, one of the most important elements in enhancing the squad might also be creating a higher-performing team with new and better players.

2.4 Building Team

Finding players with diverse attributes and figuring out the optimal composition are necessary when creating a team from scratch or adding new members to an existing one. This can change based on a long number of criteria. Al-Asadi and Tasdemír (2021) demonstrated the significance of improving a football squad. The author considered different results-affecting aspects to identify the best player for a given role before selecting the players for the composition. To lessen the disparity in the players in the minority position, such as the goalkeepers and defenders, several sampling techniques were used.

The research used data from the FIFA video game, which includes comparable player statistics from real life. The study identified the various characteristics that contributed to the various positions of players. Correspondingly, Cho et al. (2021) mentioned that when building a team, team chemistry is equally crucial. In order to create a team with a similar mentality and playing style, it is crucial to assess each player's passing style. This will boost the team's efficacy. The author used player and club statistics from many leagues, including the Barclay's Premier League, the Bundesliga, and the La Liga, among others. The passing points, proportion of long and short passes, source of the majority of the passes, and completion rate of the passes were used to produce the vectors in order to identify patterns in passing style. For this, a neural network was used as the modelling method. Finding the players who are most suited for the team, according to both articles, is crucial. As a result, the value of the players is crucial to team construction because it directly relates to each player's success.

2.5 Selection of a Player

Multiple studies are being conducted to determine whether a player has the potential to succeed or not. Morciano et al. (2022) in his study, made an attempt to forecast the player's performance in the training session using the statistics from the preceding training sessions. The author concentrated on physiological parameters, and performance was anticipated based on those and improvements in those areas. The author employed 15 folds of cross-validated multivariate logistic regression. The outcomes demonstrate that the models worked effectively and provided accuracy between 92 and 95 percent.Consequently, determining if a player is a managerial asset or not. Likewise, Murugappan (2022) used the ensemble approach and attempted to foretell a player's suitability for a specific squad based on the player's position, skills, and historical fundamental data. To identify players who are comparable to the current player but less skilled, the author used Cosine and Euclidean distance. The results demonstrated that the Random Forest had the best outcome overall, providing accuracy of 60 percent. Rajesh et al. (2020) tried to develop the team based on the qualities and competencies of the players while determining the market value of the teams and players. To determine the players who were qualified for team selection, the author used techniques including Random Forest, Support Vector Machine, and Decision Tree. Following an evaluation of the models, the authors concluded that the Random Forest model, which had an accuracy of 83 percent, an F1 score of 92, and a Jaccard similarity of 83, produced the best results. On the other hand, Lindberg and Soderberg (2020) utilised time series data from the fantasy premier league, an official Barclays premier league programme where users create their own teams and attempt to assemble the greatest lineup possible for each game day based on the performance of the players on that game day. Because the data was categorised as a time series data, the author suggested using neural networks with LSTM. Using the data from the prior week, the model projected the highest performing player from each position, resulting in the top performing team for each week. Additionally, the results indicated that models can receive more than average points when both approaches, namely regression and classification, were taken into account. The models provided varying degrees of accuracy for various positions. With 70 and 83 percent respectively for the goalkeeper and midfielder, and 69.8 and 75.6 for the midfielder and forward, respectively, and the best performing model was MLP. However, on average, LSTM outperformed other methods by 1.5 percent. Consequently, it may be used to forecast which players

will meet a team's needs for each position.

2.6 Research Contribution

Taking inspiration from the paper reviews mentioned above and concluding that the team building is one of the most important aspects which determines the success of the team and thus, making the role of Scouts even more crucial. The contribution of the research will be helping scouts in making the decisions faster and more effectively by finding the suitable players which can be an asset to the team and reduce the workload of whole procedure by finding the best possible position of the player using the Artificial Neural Network and then predicting the value of the player giving information of which players have potential. This would be incorporated as a framework which will be comprised of two stages, first will be the classification stage for predicting the position and then the regression stage which will be used to predict the value.

3 Methodology

This research will incorporate the Knowledge Discovery in Database(KDD) methodology. The following figure 1 shows the structure of the KDD methodology. As the figure shows,

Figure 1: Knowledge Discovery in Database(KDD)

the process involving the KDD process. The KDD emphasises the "high-level" use of certain data mining techniques and refers to the broad process of discovering knowledge in data. The main goal of KDD is to extract or formulate the knowledge from the given data which can be of use for the research. The main extraction of knowledge is done by using the data mining and machine learning methods. To implement this, certain procedure have to followed to transform the data which can be used to fit in the methods for extraction of the knowledge. The following steps describe the incorporation of the KDD methodology in the research.

3.1 Data Pre-Processing

There are a lot non-essential and unwanted data in the dataset. Thus data pre-processing can help remove the unwanted data and improve the quality of the data. It will also create a better understanding of the data which in turn can open various patterns in the data. Also, there are a lot of missing data, thus cleaning of data is necessary.

3.2 Data Transformation

The current data is not suitable for implementation of any of the machine learning techniques as it has a lot of alpha numeric data also reduction of the dimension of the data is

essential for data mining and machine learning models to work effectively. Thus, transformation of data is essential step for the research.

3.3 Data Mining Methods and Evaluation

To achieve the final goal of the research, data mining and machine learning algorithms are to be implemented to predict and forecast the final factors to support the research and these results are to be evaluated using proper evaluation metrics to give a proper understanding of the result and how it will affect the current situation.
Thus, above explanation validates the use of KDD methodology in this research.

4 Design Specification

The following research used a machine learning frame work to help support the research and it's purpose. The frame work consists of two layers or stages as shown in the figure below. First stage includes a classification model using Artificial neural network.

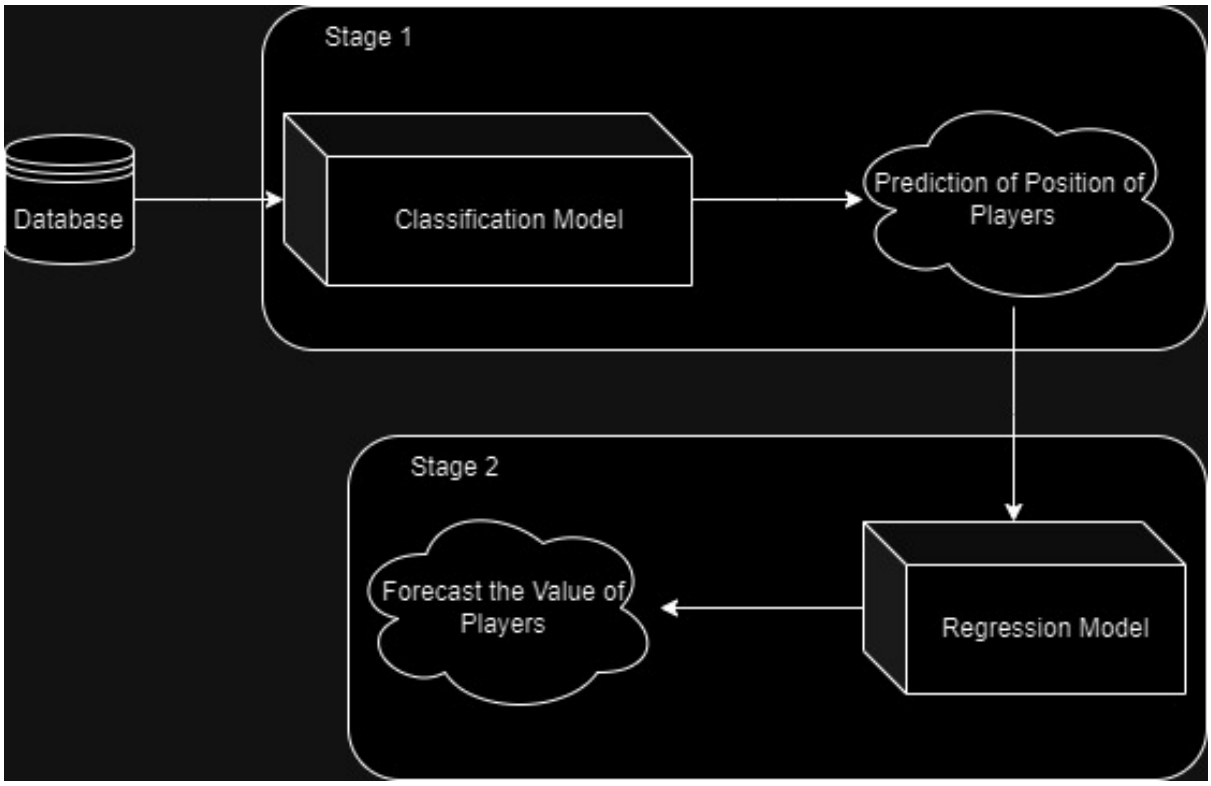

Figure 2: Framework

This classification model predicts the position of the player depending upon the statistics that were provided of the player in the following database. The number of classes were confined to 9 major positions that was"Striker" (ST), "Midfielder" (MF), "Center Attacking Midfielder" (CAM), "Center Defensive Midfielder" (CDM), "Center Midfielder" (CM),"Winger" (WN), "Center Back" (CB),"Defense" (DF) and "Goalkeeper" (GK). The model is supposed to predict one of these positions. The Artificial Model used is made using the python library named Tensorflow. The model have an input sequential

layer followed by the two hidden layers and finally an output layer. The first three layer have an activation function called "Relu". The final output layer has an activation function of "Softmax". All of the layers are being followed by a dropout layer up until the final output player with 0.1 percent as the dropout rate. The proper structure of the model can be seen in the figure shown below. For compilation of the model, "Categorical

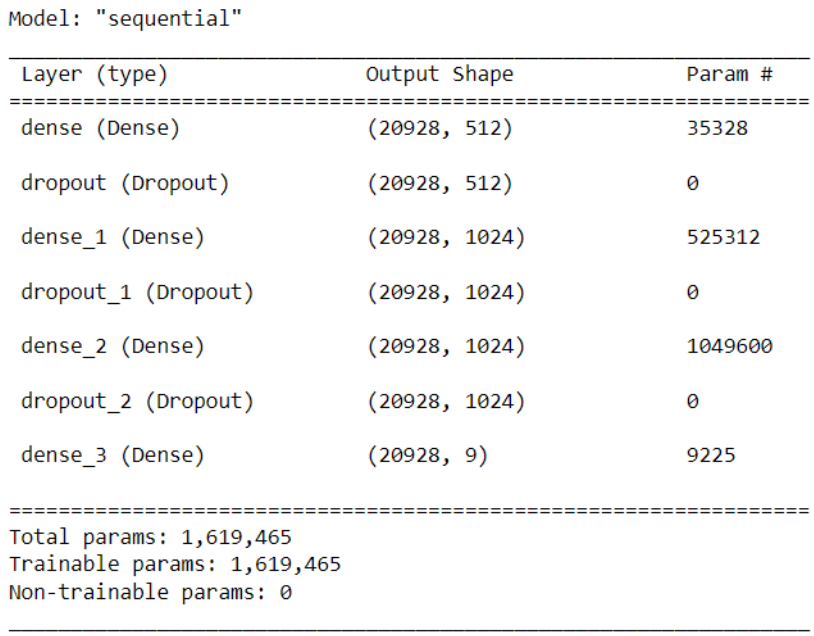

Figure 3: Model Structure

crossentropy" loss function was used as the model is a classification model with multiple classes. The optimizer used was "SGD" optimizer and the test and train ratio used for the model was 80 percent train and 20 percent test data. The model was train for 500 epochs with a batch size of 20 per epoch.

For the next stage that was stage 2, a regression artificial model was created. For this stage the model has to predict the value of the players which is related to success of the player and thus, providing valuable information about the player's importance. For this model, the first layer having "Relu" as the activation layer followed by two hidden layers having "Relu" as their activation function as well. For the final output layer, a "Linear" activation function was used. All the layers were followed by the dropout layers up-to the final output layer and have a 0.1 percent of dropout rate. This model also has the same structure as the one for the classification model which is shown in the figure 3 above. For the compilation of the model, "Mean Squared Error" loss was taken as the model is a regression model. As for the optimizer, "Adam" optimizer was used. Model was trained for 500 epochs using 20 as the batch size per epoch. Thus, by combining these two stages, the final framework was created.

5 Implementation

5.1 Data Description

The data was taken from an open source website named as Kaggle. The data has 74 columns and consists of statistics of the football players given by the organization called FIFA who are responsible for keeping the updated statistics of all the players who are registered under their jurisdictions. FIFA is the biggest organization which controls and organizes the events related to football all around the world. They are also partnered with several organization who developed the game named FIFA which represents the actual players in a video game with real life statistics. This is another motivation for the organization to keep their statistics up to date.

5.2 Data Pre-Processing

The data have 74 columns which indicates certain aspect and information about a player. Some of the columns are Name of the player, Age, Clubs they play in, their value, their current wage and all the statistics related to their performance like acceleration, agility composure etc. These columns are essential as they provide critical information about a player, its play-style and his performance in the past. Some of the columns are non-essential for example, link of the photo of the player, link to their club's logo and so on which are non-essential towards the research. Thus, for the initial step, all the non-essential and unwanted columns were removed from the data. Some of the players have more than one preferred positions as players tend to play different positions under different managers as they like to play with different compositions and have different play-styles. For example, Cristiano Ronaldo, a very famous Portuguese player have two different positions set as his preferred position as he tends to play on "Left Wing" in his club named Real Madrid and played as "Striker" in his National team. This is mainly due to two different factors, first was his ability to play in multiple positions and second was the manager's preference. Thus, to incorporate both sides, the table was flattened and made as all the possible positions were made into different rows. After this, all the possible positions of the players were narrowed down to just 9 major positions. This was done to reduce the complexity of the model and classes for the predictions. These 9 classes are "Striker" (ST), "Midfielder" (MF), "Center Attacking Midfielder" (CAM), "Center Defensive Mid-fielder" (CDM), "Center Midfielder" (CM), "Winger" (WN), "Center Back" (CB), "Defense" (DF) and "Goalkeeper" (GK). Labels were created for all the positions to use them as classes. Unwanted columns like, "photo", "flag", "club logo" and so on which are non-essential and ineffective towards the final results were removed. The "Value" and "Wage" columns had some characters to represents the values in euros. The columns were stripped off of any such character and type was changed to a integer format for the models. The column named the "Nationality" which was a character type column, was labeled and changed into factors. Same process was followed for the "Club Names" column. All the NaN and empty values were filled in accordance with the factors for that particular column.

Some of the columns in statistics had values like "70+9" which is representation of combination of old and updated statistics of that player. It indicates how much the player have improved or worsen from his past performance. The first numeric values represent the previous performance and the next numeric value represents the updated values. For the model, all the values have to be added or subtracted accordingly. Thus, a function

was created to add all the values which had "+" in the value and similarly to subtract if "-" was found in the value. For final transformation of data for the artificial neural network, the final prediction values which is the "Position" of the players were changed to a category which is a representation of a list containing 0's and 1's. 1's corresponds to the positions of the player and rest are represented as 0's.

5.3 Model Implementation

Several model's were implemented in the following research. These are two models used in a framework in addition to this, base paper's model was also implemented which was taken from the author Rajesh et al. (2020). Thus the implantation was divided into following experiments that are listed below.

5.3.1 Experiment 1

The first experiment was totally dedicated to recreate the base model paper by the author. For this, the data was divided into two part after the pre-processing Rajesh et al. (2020). The two parts were the depended data which is used to train the model and second was the independent data, which is used to train the model for predictions. After dividing the data, Random Forest Classifier was implemented with 1000 estimators. The following model was the best performing model according to the author to predict the positions of the players among several other classification models. Predictions were made after training the model and results were evaluated.

5.3.2 Experiment 2

For the next experiment, the artificial neural network was used in-place of the Random Forest Classifier. As discussed in the "related work" section, the author Inana and Cavas (2021) implemented the ANN with LSTM to predict the value of players which gave great result and low error as compared with the results obtained by the author Al-Asadi and Tasdemır (2022) in which he used different regression models. Thus, supporting the idea of using the ANN model. For this experiment, the structure of the neural network that is defined in the design specification section was used. The model was trained using "Adam" optimizer for 500 epochs. Then the trained model was used to predict the results and the results were evaluated and compared with the base model paper.

5.3.3 Experiment 3

In order to improve the current Artificial Neural Network, another model was made by adding another hidden layer of 2048 neurons having "Relu" as an activation function. The new model was trained for 500 epochs using "SGD" optimizer. Then the model was used to predict the positions and the results were compared with the previous model and base paper's model.

5.3.4 Experiment 4

As for the next experiment, a neural network of similar structure was used that was used in the experiment 2. This model was an regression model to predict the value of the player, which was the next stage of the framework. The model was trained for 500

epochs using the “Mean Squared Error” loss function and “Adam” optimizer. Then the model was used to predict the value of the players and the final results were evaluated.

6 Evaluation

The framework was evaluated separately and by using different evaluation metrics as it has two levels or two stages with different type of models. For the first stage/level, the model used is a classification model.
Thus, for the classification model there are numerous evaluation metrics. Among the many, the most important one is the confusion matrix. The confusion matrix is the evaluation tool which evaluates the performance of the classification model. It give out the evaluating number like true positive, true negative, false positive and false negative. The “True Positive” represents the values which were predicted by the model as True and the actual value was also True. The “False Positive” represents the value which was predicted by the model as False but actually it was True. The “True Negative” represents the value which was predicted by the model as False and the actual value was False. The “False Negative” represents the value which was predicted by the model as False and the actually it was True.
Another evaluation matrix that was used was the classification report which gives the “Precision”. “Recall” and “F1 Score” of the model and also the accuracy of the model. The “Precision” depicts the quality of the predictions that were made by the model which were positive. The “Recall” give the information about the ratio of the predictions that were correctly predicted by the model. The “F1 Score” gives a sense about the accuracy of the model in predicting a correct class throughout the entire dataset by combining the “Precision” and “Recall” scores of the model.
As for the next stage/level, the model was a regression model. Thus, requiring different evaluation metrics as compared to classification models. Thus, the metrics used were “Mean Squared Error”, “Mean Absolute Percentage Error” and “Mean Absolute Error”. The “Mean Squared Error(MSE)” dictates the average error between the squared values of the predicted and actual values. The “Mean Absolute Percentage Error(MAPE)” calculates the magnitude of the error made or generated by the model the model or in other words how far were the prediction from the actual values. The “Mean Absolute Error” also calculates the difference between the predicted and actual values thus, stating the error made by the model in the predictions.

6.1 Experiment 1

The first experiment was to recreate the base paper model. Thus the random forest model was used to predict the positions of the players as described in the paper by the author Rajesh et al. (2020). For evaluation of this model, a confusion matrix was created which is shown below in the fig. 4. It shows the matrix with respect to the ”9 major positions” that the model was trained to predict. These are in a sequence from 0 to 8 which represents “CAM”, “CB”, “CDM”, “CF”, “CM”, “DF”, “MF”, “ST” and “WN” respectively. The colour shows the ability of the model to predict the particular class. The darker the colour the worse are the predictions. For example, the third class which represents ”CF” have black colour throughout the range of predictions thus stating that the model had difficulty in predicting this position. This could be because of data imbalance.

A classification report was also made to show the accuracy of the model which was given as 88 percentage. All together with "Precision", "Recall" and "F1 scores" of each class.

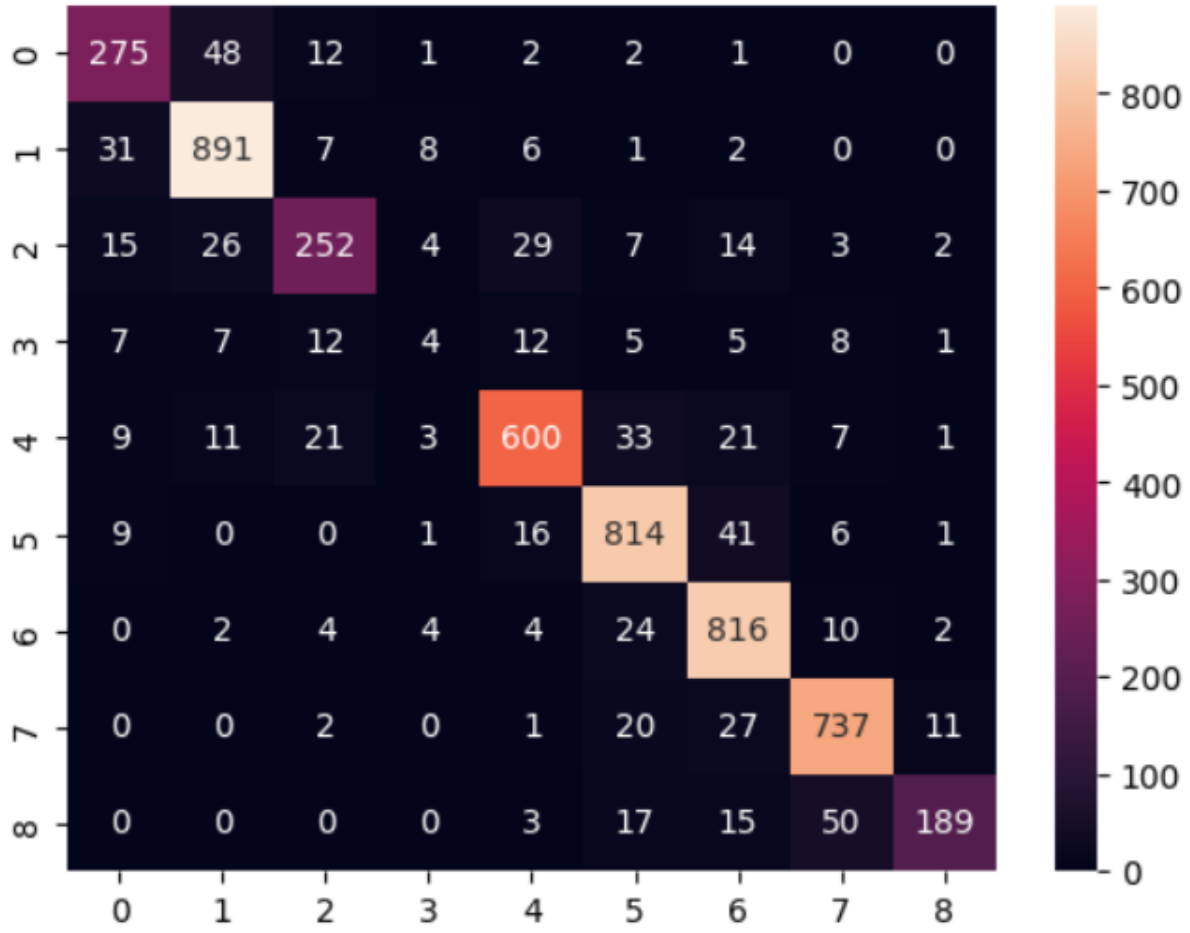

Figure 4: Confusion Matrix of Random Forest

6.2 Experiment 2

For the next experiment, the model was changed from the Random forest to Artificial Neural Network with three hidden layers and "Softmax" and the final layer's activation function. The model was compiled using "Categorical Cross Entropy" and "ADAM" optimizer. For evaluation of the model, accuracy scores were taken into account. For this model, on train data of 80 percent the final accuracy score after 500 epochs, achieved was 65 percent.

6.3 Experiment 3

For the next experiment, the Neural network was changed a little by adding another hidden layer and the optimizer was changed from "ADAM" to "SGD". The final accuracy score achieved for this model on train data of 80 percent after 500 epochs was 72 percent. Also, the confusion matrix was created to compare the results with the base model paper. The matrix is shown in the figure below. Each array represents each class.

```
array([[[4461,  414],
        [ 302,   55]],

       [[3742,  499],
        [ 428,  563]],

       [[4557,  333],
        [ 285,   57]],

       [[5139,   44],
        [  47,    2]],

       [[4200,  325],
        [ 591,  116]],

       [[3832,  556],
        [ 434,  410]],

       [[3694,  668],
        [ 578,  292]],

       [[4103,  324],
        [ 437,  368]],

       [[4782,  183],
        [ 244,   23]]], dtype=int64)
```

Figure 5: Confusion Matrix of ANN

6.4 Experiment 4

For the next experiment, the next level/stage of the framework was implemented. Artificial neural network was created with three hidden layer's having "relu" as an activation function with final layer with "linear" activation function. The model was compiled using "mean squared error" and "ADAM" optimizer as this is an regression model for predicting the value of the players. After 500 epochs, the accuracy achieved was 66 percent and loss was 0.9. The "Mean Squared Error" for the model was calculated which came out to be 0.63 and "Mean Absolute Error" came out to be 0.3 which was very less indicating the predictions of the value of the players were very close.

6.5 Discussion

The base paper model was difficult to recreate due to number of reasons. First major problem with the article was no information about the dataset. The information about the dataset was incomplete, for example the year of the data used and from where it was acquired. The second major issue with the research which created a hindrance and variation in the final result was the lack of description of data pre-processing. The research paper was lacking the in-depth description of the data pre-processing process which added to the change in results as the final accuracy was different with respect to the results depicted in the research paper. In the paper the accuracy acquired was 85 percent while in the research, after recreating all the steps provided and taking some necessary steps, the final accuracy acquired was 88 percent which is 3 percent more than the base paper. The accuracy of the artificial neural network achieved in the classification model was 72 percent which is less than the base paper. This can be due to imbalance in the dataset as the positions are not equally distributed throughout the data and also

Models	Accuracy(%)
Random Forest(Base Paper Model)	88
Artificial Neural Network(3 hidden layers and "Adam" optimizer with 500 epochs)	65
Artificial Neural Network(4 hidden layers and "SGD" optimizer with 500 epochs)	71

Figure 6: Accuracy of models

the classes that were created to reduce the number of positions to only nine can be a factor. The final results are represented in the table shown in Fig. 5. As for the final regression model, the error achieved in the model was very low indicating that the value of the players were predicted accurately thus, supporting the idea of finding the probable players which could be valuable to the team in particular positions thus, reducing the time of the scouts in narrowing the choices they have to evaluate.

7 Conclusion and Future Work

This research was done to help and support the scouts to find the players which can be valuable to the national team and improve the national team. This main focus was to reduce the overall time taken in the whole process of scouting. To tackle the problem, the framework was created. The first level of the framework gave 72 percent accuracy which is less than the base model paper which at the same time important as it can be of help as the research was focused on helping the scouts and not recreating was is already stated. In other words, less accuracy indicates the variation in the positions of the players with respect to actual position they currently have. The variation is in-accordance with all the statistics that the player currently have thus, creating a new point of view for the scouts to look at and new possibilities with the players. The accuracy is not low at the same time not too accurate to state that the results can be looked at or to be considered for future reference. The second stage of the framework predicted the value of the players with very less error and with MSE as 0.6 which is comparable and even better than most of the other research paper that were discussed above thus, indicating the ability of the model to find the probable players which can be of asset to the scouts. Thus together, the framework would provide essential information to the scouts which could be used to make important decisions and help the scouts to ease their work and complexity by

narrowing down the field of probable selection area of the players. As for the future work, the current framework can be improved in certain ways. For example, increasing the number of epochs or adding new layers to find the effect on the final results. New models can be used to replace the artificial neural network in regression model which is in stage/level 2 of the framework like additive model and compare the performance. New statistics of the players can be used in the first stage. For example, adding heat-map of the players and combining the results with current model. Also adding the statistics for the training performance of the players in the dataset to provide more attributes.

References

Al-Asadi, M. A. and Tasdemír, S. (2021). Empirical comparisons for combining balancing and feature selection strategies for characterizing football players using fifa video game system, *IEEE Access* **9**(1): 149266–149286.

Al-Asadi, M. A. and Tasdemır, S. (2022). Predict the value of football players using fifa video game data and machine learning techniques, *IEEE Access* **10**(1): 149266–149286.

Cho, H., Ryu, H. and Song, M. (2021). Analyzing soccer players' passing style using deep learning, *Sage Journals* **17**(2): 355–365.

Danisik, N., Lacko, P. and Farkas, M. (2018). Football match prediction using players attributes, *IEEE Conferences* **1**(1): 201–206.

Inana, T. and Cavas, L. (2021). Estimation of market values of football players through artificial neural network: A model study from the turkish super league, *APPLIED ARTIFICIAL INTELLIGENCE 2021* **35**(13): 1022–1042.

Kidd, R. (2020). How soccer scouting has changed, and why it's never going back. **URL:** *https://www.forbes.com/sites/robertkidd/2020/05/15/how-soccer-scouting-has-changed-and-why-its-never-going-back/?sh=5524fa7c1a1d*

Lindberg, A. and Soderberg, D. (2020). Comparison of machine learning approaches applied to predicting football players performance, *Master's thesis* **1**(1): 1–100.

Miguel, C. G., Milan, F. J., Soares, A. L. A., Quinaud, R. T., Kós, L. D., Palheta, C. E., Mendes, F. G. and Carvalho, H. M. (2019). Modelling the relationship between nba draft and the career longevity of players using generalized additive models, *Journal of Sport Psychology* **28**(1): 65–70.

Morciano, G., Zingoni, A., Morachioli, A. and Calabrò, G. (2022). Machine learning prediction of the expected performance of football player during training, *IEEE Conferences* (1): 574–579.

Murugappan, M. (2022). Football player selection based on positions and skills using ensemble machine learning and similarity measure techniques, *Master's thesis* **1**(1): 1–22.

Rajesh, P., Bharadwaj, Alam, M. and Tahernezhadi, M. (2020). A data science approach to football team player selection, *IEEE Conference* **9**(1): 175–183.

Robles-Palazón, F. J., Puerta-Callejón, J. M., Gámez, J. A., Croix, M. D. S., Cejudo, A., Santonja, F., de Baranda, P. S. and Ayala, F. (2023). Predicting injury risk using machine learning in male youth soccer players, *Chaos, Solitons Fractals* **167**(1): 1–11.

Rossi, A., Perri, E., Pappalardo, L., Cintia, P., Alberti, G., Norman, D. and Iaia, F. M. (2022). Wellness forecasting by external and internal workloads in elite soccer players: A machine learning approach, *Front. Physiol., Sec. Exercise Physiology* **13**(1): 1–11.

Saavedra-García, M., Matabuena, M., Montero-Seoane, A. and Fernández-Romero, J. J. (2019). A new approach to study the relative age effect with the use of additive logistic regression models: A case of study of fifa football tournaments (1908-2012), *PLOS ONE* (1): 1–12.

Säfvenberg, R. (2022). Age of peak performance among swedish football players, *Master's thesis, Linköping University* **1**(1): 1–72.

Tirtho, D., Rahman, S. and Mahbub, M. S. (2022). Cricketer's tournament-wise performance prediction and squad selection using machine learning and multi-objective optimization, *Applied Soft Computing* **129**(109526): 1–14.