forecasting

MDPI

Article

Is Football Unpredictable? Predicting Matches Using Neural Networks

Luiz E. Luiz [1], Gabriel Fialho [1,2] and João P. Teixeira [1,*]

[1] Research Centre in Digitalization and Intelligent Robotics (CeDRI), Laboratório Associado para a Sustentabilidade e Tecnologia em Regiões de Montanha (SusTEC), Instituto Politécnico de Bragança, 5300-253 Braganza, Portugal; luiz.luiz@ipb.pt (L.E.L.); a41446@alunos.ipb.pt (G.F.)

[2] Celso Suckow da Fonseca Federal Centre for Technological Education (CEFET-RJ), Rio de Janeiro 20271-110, Brazil

* Correspondence: joaopt@ipb.pt

Abstract: The growing sports betting market works on the premise that sports are unpredictable, making it more likely to be wrong than right, as the user has to choose between win, draw, or lose. So could football, the world's most popular sport, be predictable? This article studies this question using deep neural networks to predict the outcome of football matches using publicly available data. Data from 24,760 matches from 13 leagues over 2 to 10 years were used as input for the neural network and to generate a state-of-the-art validated feature, the pi-rating, and the parameters proposed in this work, such as relative attack, defence, and mid power. The data were pre-processed to improve the network's interpretation and deal with missing or inconsistent data. With the validated pi-rating, data organisation methods were evaluated to find the most fitting option for this prediction system. The final network has four layers with 100, 80, 5, and 3 neurons, respectively, applying the dropout technique to reduce overfitting errors. The results showed that the most influential features are the proposed relative defending, playmaking, and midfield power, and the home team goal expectancy features, surpassing the pi-rating. Finally, the proposed model obtained an accuracy of 52.8% in 2589 matches, reaching 80.3% in specific situations. These results prove that football can be predictable and that some leagues are more predictable than others.

Keywords: football forecasting; soccer prediction; deep neural network; sports betting; pi-rating; feature importance

Citation: Luiz, L.E.; Fialho, G.; Teixeira, J.P. Is Football Unpredictable? Predicting Matches Using Neural Networks. *Forecasting* **2024**, *6*, 1152–1168. https://doi.org/10.3390/forecast6040057

Academic Editor: Kauko Leiviskä

Received: 13 November 2024
Revised: 9 December 2024
Accepted: 10 December 2024
Published: 12 December 2024

1. Introduction

Football is the world's most popular sport, with around 5 billion people engaged in the 2022 Qatar World Cup (https://www.fifa.com/fifa-world-cup-qatar-2022-commercial, accessed on 9 October 2024). A reason for this is that the sport is very dynamic and anything can happen in a match. Some people believe that football is unpredictable, which has made it thrive [1]. A particularly notable example of this occurred at the 2014 World Cup in Brazil when the Brazilian team with five World Cups lost 1–7 to Germany despite having won only three World Cups. Before this match, Brazil and Germany had met 21 times, with twelve wins for the Brazilians, four for the Germans, and five draws. It was the biggest defeat suffered by Brazil in a World Cup [2], being considered one of the most remarkable results in football history.

Nonetheless, people fight against this intrinsic characteristic and try to predict matches using different methods. The interest in trying to know the future of football matches is overwhelming. There is a specific market where people earn money when they correctly predict an outcome, and it is called the football betting market. Sports betting is a way of testing sports knowledge, with the possibility of making money on successful guessing. This market has been growing year after year, and it was estimated that the volume of online bets placed in Brazil alone in 2023 was approximately USD 11 billion (www1.folha.uol.com.br/internacional/en/business/2024/01/brazilians-spent-over-11-billion-on-online-betting-in-

2023.shtml, accessed on 9 October 2024), leading governments (e.g., Portugal (https://diariodarepublica.pt/dr/detalhe/decreto-lei/66-2015-67098359, accessed on 9 October 2024) and Brazil (www.planalto.gov.br/ccivil_03/_ato2023-2026/2023/lei/l14790.htm, accessed on 9 October 2024)) to regulate this market. This leaves us with the following question: is football predictable or not?

Creating a model to predict football matches is a highly difficult mission, as this sport has a large set of characteristics that are directly or indirectly related to the result, resulting in difficulty for humans to consider all the features and predict a football match with high accuracy. On the other hand, as the technology has expanded exponentially, it is possible to use Artificial Intelligence in sports prediction, as it can process a large volume of data systemically to statistically calculate the most likely result. This technology had an aggressive expansion of its accuracy, and artificial neural networks can now outperform humans in many areas [3], with forecasting being one of them.

Furthermore, analysing the data used to feed forecasting models is a crucial factor in their final accuracy, as numerous variables can be considered. Given the objective of creating a model to predict results in an unbiased way, it must avoid common mistakes compared to when it is done by humans, where, for example, studies showed that the decision could be biased even by the colour a team is wearing [4]. Therefore, an analysis of the importance of each parameter in the model's decision-making process is essential.

This study aims to use artificial neural networks to predict football match results using previous statistical data. The results provide information beyond the sole forecasting, namely, methods of data organisation, dealing with missing data, feature creation based on raw data, and how to measure the importance of these created features in the model training. Therefore, beyond the football theme, the results can be used in multidisciplinary developments.

This paper begins with a review of some important works on this subject, followed by an explanation of the database: extraction, analysis, and processing. Then, the known and proposed features are explained, followed by measurements of their importance in the system training. Subsequently, the neural network model is presented and, lastly, the results and conclusion of the article are described.

2. Related Work

Several researchers have carried out studies with football forecasts. The work [5] presents an Artificial Intelligence system for predicting the results of football matches that beat the bookmakers' odds. He used a database containing data from 15 years of the Dutch football competition with the following features: goals scored, goals conceded, and average number of points gained. He tested the model's performance and concluded that an average of the 20 past matches was the optimal number of matches to be considered, and the performance of the chosen classifier was 55%. Nevertheless, it was shown that the right betting strategy can lead to a profit for the bookmakers in the long term.

In Ref. [6], a novel and simple approach is proposed, which the authors called pi-rating, for dynamically rating Association Football teams solely based on the relative discrepancies in scores through relevant match instances. They affirm that the pi-rating system applies to any other sport where the score is considered a good indicator for prediction purposes, as well as determining the relative performances between adversaries. In addition, they concluded that the pi-rating system generates performance values of higher accuracy when compared to the popular and widely accepted ELO rating system while keeping the rating complexity and computational power required at roughly the same levels.

The authors of [7] used machine learning techniques to predict football matches. He prepared the dataset at the player level, match level, and team level. First, he started with 308 performance attributes and observed that by selecting only 34, the model had the best performance. He obtained the best prediction at an accuracy of 53.4% when using the weighted average of ratings from the past seven matches.

Furthermore, the authors of [8] won the Soccer Prediction Challenge of 2017 with an Artificial Intelligence model composed of gradient-boosted trees, reaching an accuracy of 53.88%. The challenge is organised in conjunction with the MLJ's special issue on machine learning for Soccer [9], and its goal was to predict the outcomes of future matches within a selected time frame from different leagues worldwide. They used a dataset of over 200,000 match outcomes, and the features contained the historical strength of the team, current form, pi-rating, and league, among others [10].

Shifting away from models predicting which team will win, the authors of [11] instead researched the probability of both teams scoring at least one goal, which is also used in sports betting systems, obtaining encouraging results and, in some cases, outperforming bookmakers with machine learning classifiers.

The authors of [12] presented a method based on players' abilities on each team to predict the results of football matches, allowing a dynamic rating framework based on publicly available data from each player. Furthermore, they introduce a model to consider the interaction between opposing players concerning the different team's formations. The resulting model requires a large and renewable database but can achieve results similar to bookmakers, with an accuracy of 52.89% on a dataset of 1350 matches.

In Ref. [13], the performance of prediction models that used only team parameters, only player parameters, and both combined were studied. The results instigate the use of player and team ratings combined, as it results in better forecasting accuracy. Meanwhile, the use of only one parameter performed similarly without regard to the source, whether it was player- or team-based.

The state-of-the-art shows that predicting the outcome of a sports match with relatively high accuracy requires a specific model for the problem, a large set of historical data collected, various methods of data analysis, and various pre-processing techniques.

The objective of the work presented here is to converge state-of-the-art techniques and methodologies with features proposed in this work, combining team and player ratings, finding the most promising configuration of data, data combination, and model architecture, improving the accuracy of football prediction systems, and defining the most important features in the decision-making process.

The method of averaging the data from the last 20 matches, proposed by Ref. [5], is compared with averaging the last seven matches, proposed by Ref. [7], and with the exponential weighted moving average, a widely used method for finance, where recent values have a greater weight than old values. Furthermore, the pi-rating feature, created by Ref. [6], is generated to evaluate its potential against other methodologies and the proposed features. The difference between using normalisation and standardisation techniques during data pre-processing is also analysed to find the most accurate. Furthermore, successful state-of-the-art methods for data organisation and outlier mitigation are used, seeking a resulting methodology that embraces the most successful approaches to deal with the different prediction challenges. Finally, the grid-search method is used to test multiple model hyperparameters to find the most accurate system with the most promising selected data features.

Regarding data type, the proposed work also pursues covering a number of data levels that are normally assessed in different state-of-the-art works: team-related data [11], statistical generated features in pre-processing [10], and individual player data [12]. To find the most important features and to increase the system's accuracy, the three categories of data will be used together.

Finding the most accurate system would demonstrate the true potential of deep neural networks for predicting complex systems while verifying the possibility of predicting a supposedly unpredictable sport.

Moreover, since different methodologies for data analysis, manipulation, and organisation are compared, the results can be applied to data from other complex prediction systems.

Furthermore, the relevance of objective-driven generated features being compared with original raw data, embracing the processing of raw data to generate complex features that

help the algorithm understand how multiple parameters affect each other, is an important methodology approach that can be applied to other complex event prediction developments.

3. Dataset Manipulation

Considering the stated requirements for predicting the outcome of a sports match, the first stage is to build the dataset. Therefore, the data must first be extracted. From these data, combinations are made to create more meaningful features, followed by an organisation process of this whole package of raw data and driven-generated features, so it only uses data from before a specific match to predict its result, allowing correct interpretation by the neural network [14].

3.1. Data Extraction

The creation of a dataset is undoubtedly the main step in developing a learning model; thus, all the information contained must be studied and proportionally adapted to the complexity of the problem. In the previous section, it became clear that a major difficulty for the authors is maintaining a balance between a relatively large number of samples and a significant amount of statistical information. This happens because it is not simple to find statistics from football games and collect them properly.

The data were extracted from the WhoScored website (www.whoscored.com, accessed on 9 October 2024), which specialises in the in-depth analysis of detailed football data. The information was not in a format that machine learning methods could directly analyse, so computer science techniques were used to bring all the available data together in more manageable formats, such as CSV or XLS files. Table 1 shows all the features collected.

The data shown in Table 1 were taken from 24,760 matches from 13 different leagues, as can be seen in Table 2.

Leagues with the same point-ranking system were selected to avoid competitions where the match is decided by playoffs, such as quarterfinals, semi-finals, and finals.

Furthermore, the website has no uniformity as to the year in which it started recording match statistics, so there is a difference between the first year collected from the leagues. Some have data since 2009, and others only since 2017, as shown in Table 2.

Table 1. Features collected from the website.

Feature	Feature Description
Playing Ground	Which team was the home team and which was the away team
Goals	Number of goals during the whole match
Players Ratings	Individual Evaluation of each player that played during the match
Shots	Shots taken by the team
Shots on Target	Shots on goal taken by the team
Successful Passes	Successful passes made by each team
Aerial Duel Success	Total number of direct aerial wins over an opponent of each team
Dribbles Success	Total number of successful dribbles (passing the opponent with the ball) made by each team
Recovery Attempts	Total number of attempts to take the ball away from the opponent by each team
Possession	Total ball possession as a percentage
Team Rating	Overall quality of the team
Dribbles Success	Total percentage of successful dribbles (passing the opponent with the ball)
Successful Recoveries	Total percentage of ball recoveries successful
Interceptions	Number of opponent's pass interception before reaching the destination
Dispossessed	Losing possession without trying to pass to the opponent
Corners	Total number of team corners
League	League in which the match took place

Table 2. Leagues and number of matches collected.

League	Matches	Interval Collected
Argentina First Division	1064	2016/2019
Brazilian Championship	2142	2013/2018
China Super League	407	2016/2018
England Championship	933	2017/2019
England Premier League	3507	2009/2019
France League	3282	2009/2019
Germany Bundesliga	3638	2009/2019
Italy Serie A	3266	2009/2019
Eredivise	259	2017/2019
Portugal NOS League	810	2016/2019
Russia Premier League	1063	2013/2019
Spain League	3186	2009/2018
Süper Lig Turkey	1203	2014/2018
Total	24,760	

3.2. Feature Generation

To help the model interpret the information more realistically, ten new types of data were created, correlating information that normally influences each other, which provides a path to better and more meaningful weighting.

3.2.1. Goal Difference

The first feature is the difference between goals made and goals conceded in the previous games. This variable is important because it expresses attack and defence ability in only one value.

3.2.2. Pi-Rating

The second parameter is the teams' pi-rating. This parameter proposes dynamically evaluating a team based only on the discrepancies between the predicted and actual goals, as shown above in the related works. Each team has a pi-rating for home and away matches, and their value is updated for each match according to the result. This statistic, created by Ref. [6], proved to be extremely effective in predicting football match results and surpassing other ways of assessing teams, such as the ELO system, which was originally invented as an improved chess rating system [15]. Still, it is also used as a rating system for multiplayer competition in video games, football, basketball, baseball, and table tennis [16].

Therefore, the authors of [6] followed three rules for the creation of this rating system: the advantage when a team is playing at home, the fact that more recent results are more important than old ones in estimating a team's ability, and that winning is more important to a team than increasing their goal difference. To predict the match result, the pi-ratings from both teams are compared, resulting in a prediction of goal difference.

To adapt to the first rule, both teams have two different pi-ratings, one for home scores and another for away scores. As for the second rule, a constant value (λ) weights the match resulting goal difference before summing with the previous pi-rating, defining how much influence the new data have over the old data. The authors performed tests to find the most accurate constant, resulting in 0.35% to adjust the pi-rating used in the match and 70% to adjust the team's missing pi-rating. Finally, to deal with the third rule, the goal difference has an exponentially decreasing weight in the pi-rating, following the equation

$$\Psi(e) = c \times log_{10}(1 + e) \quad (1)$$

where 'c' is the maximum weight, set to 3; e is the absolute error between the predicted and real goal difference; and Ψ is the weight in the final pi-rating update equation. Since e is an absolute value, the $log_{10}(1 + e)$ value will always be a real positive number. Therefore, the Ψ weight will always increment the team's pi-ratings, even if its predicted goal difference

is lower than the real goal difference. For that reason, the Ψ value will be used as a positive weight when the team performs better than expected and as a negative weight when it performs worse than predicted. This translates to

$$\Psi_H(e) = \begin{cases} \Psi(e), & \text{if predicted goal < real goal difference} \\ -\Psi(e), & \text{otherwise} \end{cases} \quad (2)$$

for the home team and

$$\Psi_A(e) = \begin{cases} \Psi(e), & \text{if predicted goal > real goal difference} \\ -\Psi(e), & \text{otherwise} \end{cases} \quad (3)$$

for the away team.

To achieve the actual goal difference prediction, the teams' respective home and away pi-ratings must be translated into each team's expected goal difference. Since the home team has the advantage from rule 1, its value is positive, indicating a goal difference in favour of the home team, and the away team's pi-rating is negative, indicating an unfavourable result. Then, the home team's value is subtracted by the away team's value, resulting in

$$G_D = (10^{\frac{|R_H|}{c}} - 1) - (-10^{\frac{|R_A|}{c}} - 1) \quad (4)$$

where 'c' is again set to 3, R_H is the home team's pi-rating, and R_A is the away team's pi-rating. After the match, this value is subtracted by the actual goal difference value (G_R) to determine the absolute goal error (e), resulting in

$$e = |G_D - G_R| \quad (5)$$

The pi-rating adjusted after every match is then calculated with

$$R_{new} = R_{old} + \Psi(e) \times \lambda \quad (6)$$

where 'R' is the team's away pi-rating if it played away or the team's home pi-rating if it played at home, and λ is 0.35%. Afterwards, the missing pi-rating (P) is adjusted regarding this already-adjusted value

$$P_{new} = P_{old} + (R_{new} - R_{old}) \times \lambda \quad (7)$$

considering that λ changes to 70%.

Applying this logic to one situation as an example, considering a match between the home team ρ and the away team α; before the match, the home team's pi-ratings are $R_{\rho H} = 1.2$ and $R_{\rho A} = 0.5$ for home games and away games, respectively, and the away team's pi-ratings are $R_{\alpha H} = 0.1$ and $R_{\alpha A} = -1$, again, for home and away games. From these values, considering that team ρ is playing at home and team α is playing as an away team, the ratings that will be used to predict this match's result are $R_{\rho H} = 1.2$ and $R_{\alpha A} = -1$.

With these data, using Equation (4), the expected goal difference is estimated, resulting in

$$\begin{aligned} G_D &= (10^{\frac{|1.2|}{3}} - 1) - (-10^{\frac{|-1|}{3}} - 1) \\ &= (1.51) - (-1.15) \\ &= 2.66 \end{aligned}$$

Supposing that the match result was 4-1 for the home team, it is translated to a real goal difference (G_R) of +3. The error (e) is then calculated using Equation (5):

$$e = |2.66 - 3|$$
$$= 0.33$$

With the prediction goal difference error, the Ψ weight is obtained using Equation (1):

$$\Psi(e) = 3 \times log_{10}(1 + 0.33)$$
$$= 0.37$$

Based on Equation (2), since the home team won by more than expected, its Ψ value is positive; further, based on Equation (3), since the away team lost by more than expected, its Ψ value is negative.

Finally, to update ρ's pi-ratings, Equation (6) is used for its home pi-rating, since it played at home

$$R_{\rho H} = 1.2 + 0.37 \times 0.035$$
$$= 1.21$$

and Equation (7) for its away pi-rating,

$$R_{\rho A} = 0.5 + (1.21 - 1.2) \times 0.7$$
$$= 0.51$$

The same way, to update α's pi-ratings, Equations (6) and (7) are used:

$$R_{\alpha A} = -1 - 0.37 \times 0.035$$
$$= -1.01$$
$$R_{\alpha H} = 0.1 + (-1.01 - (-1)) \times 0.7$$
$$= 0.09$$

After the updates, the resulting pi-ratings are $R_{\rho H} = 1.21$ and $R_{\rho A} = 0.51$ for the ρ team and $R_{\alpha H} = 0.09$ and $R_{\alpha A} = -1.01$ for the α team. This shows that since the ρ team performed better than expected in goal difference, its pi-rating increased. Similarly, since the α team performed worse than expected, its pi-ratings decreased.

3.2.3. Player Statistics Features

Based on player-wise evaluation, the next three features are calculated—the relative attack (R_A), defence (R_D), and midfield (R_M) power—between both competing teams. Regarding the home team, this was estimated using

$$R_A = \frac{P_{HA1} + P_{HA2} + P_{HA3}}{P_{AD1} + P_{AD2} + P_{AD3}} \quad (8)$$

$$R_D = \frac{P_{HD1} + P_{HD2} + P_{HD3}}{P_{AA1} + P_{AA2} + P_{AA3}} \quad (9)$$

$$R_M = \frac{P_{HM1} + P_{HM2} + P_{HM3} + P_{HM4}}{P_{AM1} + P_{AM2} + P_{AM3} + P_{AM4}} \quad (10)$$

where the 3 most advanced players are considered strikers, where the x team's n striker is represented by P_xAn. Similarly, the 3 most recessed players (not considering the goalkeeper) are considered the defenders, where the x team's n defender is represented by P_xDn. Finally, the midfielders are the 4 in-between players, where the x team's n midfielder is represented by P_xMn.

The sixth feature uses the goal-keeper evaluation to estimate team A's goal expectancy (G_{EC}) by comparing the evaluation from the goal-keeper of team B against team A's shots on goal using

$$G_{EA} = \frac{S_A}{P_{GB}} \tag{11}$$

and the other way around, for the B team's goal expectancy, considering its shots on goal and the other team's goalkeeper.

3.2.4. Team Statistics Features

The seventh generated feature is the goal shot accuracy (G_{SA}), measuring how many shots on goal a team has for every goal it makes. It is estimated using

$$G_{SA} = \frac{G_M}{S_M} \tag{12}$$

where G_M is the average number of goals and S_M is the average number of shots on goal.

The eighth feature is the relative corner power (R_{CP}) for the match. This parameter is also calculated regarding the home team using

$$R_{CP} = \frac{C_H \times AD_H}{C_A \times AD_A} \tag{13}$$

where C_x is the x team's number of corners and AD_x is the X team's number of aerial duels.

The ninth feature is the relative ball loss (R_BL) regarding the home team. It is estimated using

$$R_{BL} = \frac{BL_H \times BS_A}{BL_A \times BS_H} \tag{14}$$

where BL_x is the x team's number ball losses and BS_x is the X team's number of ball steals.

The tenth and final feature is the relative playmaking power (R_{PP}) regarding the home team. It is estimated using

$$R_{PP} = \frac{\frac{S_{PA}}{I_{PB}}}{\frac{S_{PB}}{I_{PA}}} \tag{15}$$

where S_{Px} is the x team's percentage of successful passes and I_{Px} is the X team's number of pass interceptions.

Ending the features generation, Table 3 presents all the features and their respective index.

Table 3. Generated features presented.

Feature Index [1]	Feature
1	Goal difference
2.1	Pi-rating Home Team
2.2	Pi-rating Away Team
3	Relative Attacking Power
4	Relative Defending Power
5	Relative Midfield Power
6.1	Goal Expectancy Home Team
6.2	Goal Expectancy Away Team
7.1	Shot Accuracy Home Team
7.2	Shot Accuracy Away Team
8	Relative Corner Power

Table 3. *Cont.*

Feature Index [1]	Feature
9	Relative Ball Loss
10	Relative Playmaking Power

[1] Feature index refers to the order of appearance of the generated feature in Section 3.2.

3.3. Data Organisation

With the raw dataset created and the features generated from it, it now has to be organised and adjusted to be fed into the algorithm for training and further accuracy evaluation.

Apart from the pi-rating, which is already a complete system, being updated after each match, the rest of the dataset and generated features still need to be turned into a single value that summarises the historical behaviour.

To predict a future value, it is necessary to group data from only previous games in a way that tries to show a trend to the future. There are different methods for performing this calculation, such as the simple average, weighted average, or exponential average of previous matches.

Ref. [5] proved that the simple average from the last twenty matches, or less if the team had not completed twenty matches yet, would be ideal for predicting the outcome of a football match with his dataset.

Ref. [7] bet on the weighted average, where he gave more weight to recent matches using

$$x_t = \frac{nx_{t-1} + (n-1)x_{t-2} + (n-2)x_{t-3} + ... + 1x_{t-n}}{n + (n-1) + (n-2) + ... + 1} \tag{16}$$

for a given statistical datum 'x' in a match 't' for a window of 'n' matches.

After tests, Ref. [7] concluded that when performing the weighted average with a window of seven matches, the best result is obtained.

Another method widely used in finance is the exponential weighted moving average. This technique gives more weight to recent values by a variable called 'the smoothing factor'; thus, it becomes more sensitive to new information, trends, and patterns. This is one of the main differences with other methods as it responds more quickly to changes in values. The exponential weighted moving average is calculated through

$$S_t = \alpha x_{t-1} + (1-a)S_{t-1} \tag{17}$$

where it is stated that the value of the moving average 'S_t' at time 't' is a mixture between the raw value 'x' at time '$t-1$' and the previous value of the moving average itself, and 'α' is the weighting factor between the previous average and the 'x-value' at time '$t-1$'.

Therefore, the mentioned methods will be tested and compared to achieve the best performance for the model.

3.3.1. Home and Away Fields

Due to the home advantage, the data are calculated separately depending on whether the team plays at home or away. Therefore, if a team is going to play on the home field, the statistical data are calculated using the last matches played also at home. When the team plays away from home, the data are calculated using the last matches also played away from home.

3.3.2. Outliers

Outliers are data with values much smaller or larger than most other values in a dataset. These values must be handled because they can affect how data are perceived [17].

Historical data are used to capture trends. For example, a team that has scored five goals in the last five matches tends to score one goal per match. Therefore, it is necessary to

treat outliers before averaging, as they are not values that are contained in the trend; rather, they are abnormal values that affect the behaviour of a team.

To solve this problem, the data that would be used for the average (e.g., previous 7 or 20 matches data) were distributed to find the normal values. From the selected period, the lowest and highest values were taken and subtracted, and the samples were considered normal if they were between 25% and 75% of the difference. Values lower than 25% or higher than 75% were considered outliers and were replaced by limit values within the normal distribution of each data.

3.3.3. Missing Values

As shown, the data are organised using previous matches and fields; however, then, starting from the first match played at home and away, a team will not have the necessary data to calculate the average. To fix this for the second match and onward, the existing matches will be used. For example, in the fourth match of a given team, the average of the last three matches will be used, while in the second match, only the value from the first match will be used.

As for the first match, they should be removed from the dataset so that there are no misinterpretations of the learning model.

On the other hand, the players' level is more complex. In a football league, the players' rotation is large, meaning that players will be playing their first match for a team on several occasions. Consequently, as mentioned above, there will be no past statistics for the team; the only difference is that the number of matches with no data from at least one player is much higher. If we use the same technique of removing matches, much information would be lost, so another method is necessary to deal with these situations.

Therefore, in those cases where past player information is unavailable, the missing values will be replaced by the overall average of all players, so no matches will be removed again.

This is based on the rationale that it is not possible to affirm that a new player in a new team would perform better than the team's normal performance. This results in a player having no impact on the team's players' average performance. Furthermore, the theoretical support for mean substitution also refers to the fact that the mean is a reasonable estimate for a randomly selected observation from a normal distribution.

3.3.4. Normalisation or Standardisation

Normalisation or standardisation is required so that all inputs are in a comparable range.

For example, consider a neural network with two inputs, 'x_1' and 'x_2'. The first input ranges from 0 to 0.5 and the second ranges from 0 to 1000. If 'x_1' changes its value from 0 to 0.5, it is a change of 100%, but does not have a big effect on the output since it only changed by 0.5. But, if 'x_2' changes from 0 to 5, it is only a change of 0.5% of its value, but the final result would have a much bigger change.

Thus, this process is necessary for the correct functioning of the model since the different types of data have different ranges of variation, and it is not expected to make assumptions or allow one data to have a greater impact than any other on the final prediction.

Normalisation changes the scale on which a value ranges from its original scale to a limit of 0 to 1. During this process, the values were normalised through

$$\bar{x} = \frac{x - min(x)}{max(x) - min(x)} \tag{18}$$

where $\bar{x}$ are the normalised data and $min(x)$ and $max(x)$ are, respectively, the minimum and maximum possible values for this parameter or these data.

The standardisation adapts the data within a standard normal distribution with a median value (μ) of 0 and a standard deviation (σ) of 1. The values are standardised through

$$\tilde{x} = \frac{x - \mu}{\sigma} \tag{19}$$

where $\tilde{x}$ are the standardised data, μ is the period median value, and σ is the period standard deviation.

Normalisation and standardisation techniques are applied to compare them and find the solution that best suits them.

3.3.5. Categorical Data

Categorical variables describe a certain attribute of an object, system, or entity. The dataset contains categorical data corresponding to the league in which the football match occurs since it can have a high impact on the prediction [13].

Since artificial neural networks do not support categorical data in their input, this needs to be adapted to numerical values. Still, since there is no specific necessary order or relation between the leagues, it is not possible to enumerate the leagues.

Therefore, the One-Hot Encode method will be used to encode the league into numeric values. This is made by transforming the data into binary number vectors, where each column represents a specific league. For example, if the game is from the England Premier League, which corresponds to the second column of the binary vector, then the value of this column will be '1' and the rest will be '0'. Using this technique, the league is defined, but it does not tend to use an algorithm to interpret the leagues as an obligatory order.

4. Model

The predictability of a phenomenon is based on finding a pattern in the variables that encompass it. Football has several variables involved in the outcome, and the relationship between them and the result is not intuitive. In addition, it is a sport made up of humans, where players' behaviour affects the result. Therefore, the problem is highly complex and the neural network should be structured to support this.

A Deep Feed-Forward neural network (multiple hidden layers) will be used. In fact, due to the extra layers, the network can identify complex nonlinear relationships between the variables, as the input is expressed as a composition in primitive layers in which each network layer models an element. The network consists of 100 neurons in the first layer, 80 in the second, five in the third, and three in the output layer. In addition, a regularisation technique called dropout will be used to reduce the effects of overfitting [18]. The network's response consists of a vector of three binary values representing a result, as shown in Table 4.

Table 4. Correlation between network response and result.

Vector Position	**1**	**2**	**3**
Home team win	1	0	0
Draw	0	1	0
Away team win	0	0	1

In addition, the total outcomes are unbalanced, meaning there are more home wins than draws and losses. Furthermore, the network will tend to predict much more home team wins and less for the other cases. To correct this, weights to each class (home team win, draw, and away team win) will be applied proportionally to the number of samples for each result. Thus, the error function becomes weighted, where the weight of each sample is specified. With this technique, the model outputs will be more distributed, and consequently, there will be better learning. This method was compared with undersampling (i.e., forcing all classes to have the same number of samples as the smaller one), which

resulted in a higher result bias. Meanwhile, keeping the unbalanced dataset and penalising the accuracy of the most common result maintains the expected behaviour of a match, with higher chances of a home win.

The training set consists of 20,712 matches, and the validation and testing set has 2558 matches from all the leagues mentioned. The datasets are organised so that the league's concentration is the same; for example, if the training set consists of 25% matches from the England Premier League, the validation and test set will also have 25%. In addition, the dates are increasingly organised, from the training, validation, and test sets, in that order. With this, the network will predict future outcomes using past data.

4.1. First Results

The model was tested differently to determine the best data organisation performance set. Accuracy for the test sets is shown in Tables 5 and 6. The accuracy (A_c) is calculated by the sum of correct predicted matches (p_c) divided by the total number of predicted matches (p_t):

$$A_c = \frac{p_c}{p_t} \tag{20}$$

Table 5. Results of different types of organisation using standardisation as a pre-processing method.

Organisation Technique	Alpha	Accuracy
Weighted Mean of last 7 matches	-	49.30%
Simple mean of last 20 matches	-	51.30%
Exponential Weighted Moving Average	0.2	49.40%
Exponential Weighted Moving Average	0.4	50.90%
Exponential Weighted Moving Average	0.6	49.70%
Exponential Weighted Moving Average	0.8	49.30%

Table 6. Results of different types of organisation using normalisation as a pre-processing method.

Organisation Technique	Alpha	Accuracy
Weighted Mean of last 7 matches	-	49.90%
Simple mean of last 20 matches	-	51.80%
Exponential Weighted Moving Average	0.2	50.60%
Exponential Weighted Moving Average	0.4	51.30%
Exponential Weighted Moving Average	0.6	51.20%
Exponential Weighted Moving Average	0.8	50.50%

According to Tables 5 and 6, the simple average of the data from the last 20 matches presents better performance than the others. Also, the normalisation between minimum and maximum proved to be better than standardisation. This observation is important for choosing the temporal organisation and processing technique. With the highest performance set, the statistical data to improve are investigated further.

4.2. Feature Importance Analysis

Knowing the organisational technique that best suits the given problem, decision trees are used to obtain an overview of the importance of raw data compared to the pi-rating feature. One benefit of this technique is that once trees are built, it is relatively easy to calculate the importance of each attribute, allowing a better understanding of the difference between using raw data and adjustable features. Decision trees are also useful in this topic as they allow a direct comparison with other literature works that present their respective importance trees for the decision-making process.

In this first analysis, the pi-rating was selected since it is proved, by previously state-of-the-art studies, to have an accuracy superior to raw data prediction.

The importance of each data input is calculated separately for each of the three prediction situations: home team win, draw, and away team win. This provides a score that indicates how useful or relevant each data were in building decision trees within the model.

The more specific data used to make tree-branch decisions, the greater its relative importance. This importance is calculated and quantified for each attribute in the dataset, allowing them to be classified and compared to each other. Thus, it is possible to identify which variables are the most important for predicting each of the three cases. The three resulting weight graphs are shown in Figures 1–3.

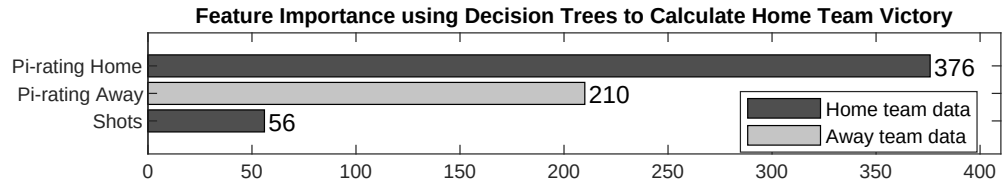

Figure 1. Feature importance using decision trees to calculate home team victory.

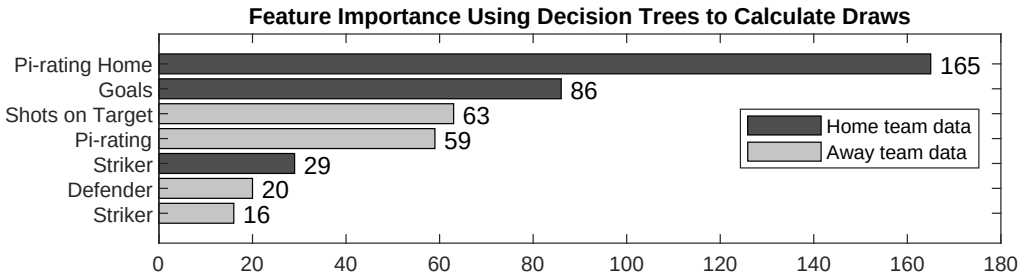

Figure 2. Feature importance using decision trees to calculate draws.

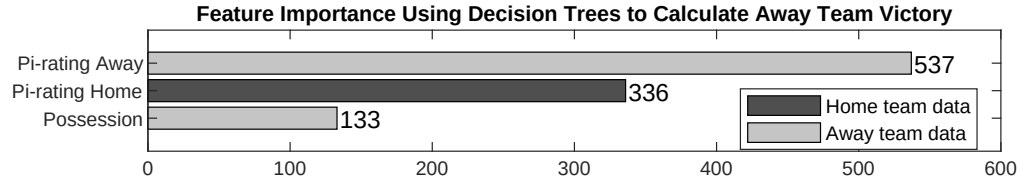

Figure 3. Feature importance using decision trees to calculate away team victory.

From the graphs, only three features were relevant in the home team and away team wins, represented by Figures 1 and 3. On the other hand, a bigger number of variables are relevant for draw prediction, represented in Figure 2, showing a bigger uncertainty in detecting draws.

It can also be noted that pi-ratings are much more important than raw parameters, such as corners or kicks. Therefore, with only this initial analysis, it can be observed that the network is probably not able to interpret the information contained in the raw data. This proves that data manipulation has an advantage in improving the interpretation and learning of the neural network. Data can contain valuable information that is only visible in its transformed state [19].

After the first analysis, which confirms the advantage of the pi-rating method, a deeper analysis is made to estimate the impact that each datum and created feature has on the neural network.

The importance of each parameter can be calculated by observing the model's changes in accuracy before and after the parameter is randomly shuffled, as the relationship between this variable and the result no longer exists [20,21]. In other words, a feature (or raw data) is important if, after randomly shuffling its values, the model's error increases. That is because, in this case, the model depends on the variable to correctly predict. However, a feature is unimportant if the model's error does not change after randomly exchanging variable values.

With all the parameters correctly used, the model's normal performance ($\in_c$) was calculated. After that, a feature was shuffled, and the new model's performance ($\in_n$) was obtained. The feature importance is calculated by subtracting the second performance from the first ($\in_c - \in_n$). These steps were repeated for all the features, and the data that were unimportant to the model were excluded. The resulting importance of each parameter is shown in Table 7.

The most interesting result is that the features proposed in this work—namely, the relative defending, playmaking, midfield power, and goal expectancy of the home team—are more important to the system's final decision than the pi-ratings from both teams. This proves that these four proposed features are a more reliable data source for predicting a match result than existing state-of-the-art parameters.

Furthermore, all the proposed features can be validated in this comparison to the state-of-the-art features, showing if the features can have a positive, neutral, or negative impact on the system accuracy. Based on Table 7, considering the pi-rating importance of 0.73, features 4, 10, 5, and 6.1 were successfully validated positively as their importance was higher than 0.73. Features 6.2, 7.2, 7.1, and 9 were validated as neutral to negative impact on the accuracy, as their importance was lower but close to 0.73. Finally, the remaining features were considered insufficiently useful for decision-making.

Table 7. Result of importance analysis by variable shuffling.

Feature Index [1]	Parameter	Importance $\in_c - \in_n$
4	Relative Defending Power	0.85
10	Relative Playmaking Power	0.81
5	Relative Midfield Power	0.77
6.1	Goal Expectancy Home Team	0.77
2.1	Pi-rating Home Team	0.73
2.2	Pi-rating Away Team	0.73
6.2	Goal Expectancy Away Team	0.70
7.2	Shot Accuracy Away Team	0.64
7.1	Shot Accuracy Home Team	0.62
9	Relative Ball Loss	0.62
1	Goal difference	0.46
3	Relative Attacking Power	0.46
8	Relative Corner Power	0.35

[1] Feature index refers to the order of appearance of the generate feature in Section 3.2.

Finally, to obtain the best neural network architecture, a technique called grid-search is used. Grid-search creates a model for every combination of neural network hyperparameters, storing the performance for each one [22].

These hyperparameters are, namely, the number of neurons in each layer, the number of hidden layers between input and output, the activation function of each layer, the learning rate, the maximum number of learning iterations, the training algorithm, and the regularisation method to mitigate overfitting. In the end, the best set of hyperparameters for the model is determined.

5. Results and Discussion

After the improvements, the model achieved 60.9% accuracy on the 82 matches of the 2019 Portuguese NOS League, 58.1% on the 122 matches of the 2018/2019 Turkish Super League, and 57.2% on the 360 matches of the Premier League of England. The total results of all leagues are shown in Table 8.

The model's overall performance achieved 52.8% accuracy on the 2589 matches from the test set. Furthermore, when the neural network had a probability greater than 50% that the home team would be the winner, the model obtained an accuracy of 78.8%. When the probability was greater than 60%, the accuracy increased up to 80.3%. Moreover, for the

away team, the accuracy was 76.6% when the probability of winning was greater than 50%. On the other hand, the model did not perform well in predicting draws, hitting only 32%.

As seen from Table 8, the model showed a significant difference between each league's accuracy. This difference may have two possible explanations: there is a difference in the match results' concentration in the test set or some leagues are more predictable than others due to external factors.

In the first hypothesis, if the model showed a tendency to predict more results as a home team win, then with data mostly composed by home team wins, the model accuracy would be high; however, when the concentration of draws or losses are bigger, the model's accuracy would be worse. For example, suppose a model that only predicts the result as a home team wins, regardless of its input. In a dataset of 100 matches, where 90 were home team wins, the model would perform 90%, but if the dataset had 10% home team wins, the model accuracy would be 10%. This is an example of a "learning model" whose results do not match its actual performance.

To evaluate this first hypothesis, the concentration of match results from the test set was analysed. The result can be seen in Table 8.

Table 8. Concentration of match outcomes of different leagues from the test set.

		Test Set Matches per Result			
League	**Test Matches**	**Home Wins**	**Draws**	**Away Wins**	**Accuracy**
Portuguese Nos League	82	45.1%	23.1%	31.8%	60.9%
Turkey Super Lig	122	49.1%	22.9%	28.0%	58.1%
England Premier League	360	49.1%	18.3%	32.6%	57.2%
Italy Serie A	357	44.2%	26.3%	29.5%	56.5%
England Championship	94	50.0%	19.1%	30.9%	56.3%
La Liga Spain	321	47.6%	23.6%	28.8%	55.4%
Brazilian Championship	215	53.0%	28.8%	18.2%	52.5%
Super League China	62	54.8%	20.9%	24.3%	50.0%
Eredivise	28	46.4%	28.5%	25.1%	46.4%
Premier League Russia	114	39.4%	28.9%	31.7%	45.6%
French League	356	43.2%	29.2%	27.6%	45.2%
Argentine First Division	108	41.6%	32.4%	26.0%	44.4%
Germany Bundesliga	370	41.0%	29.7%	29.3%	41.3%
Total	2589	46.5%	25.52%	27.98%	52.8%

From Table 8, it is clear that the results concentration is not directly linked to the model's accuracy. For example, the Portuguese Nos League has a concentration of home team wins of 45.1% and an accuracy of 60%, while the Brazilian Championship presents a 53% concentration and only 52% accuracy. The same is true for the away team wins concentration.

It can be observed that there is a small tendency pattern with the draw concentration; the lower the draw concentration, the better the accuracy. However, this relationship is not very expressive and is not maintained for all leagues. This can be explained by the fact that the model does not obtain a good performance in predicting draws, as mentioned earlier.

Therefore, the first hypothesis was discarded, and the difference in the accuracy between leagues is justified by the fact that some leagues are more predictable than others; this can be due to several factors, such as the difference between league gameplay, playing style, league visibility, or other external factors. This effect was also noted by the work developed in [23]. This is positive for the model as it shows that its performance is unrelated to the concentration of the test set data and the model is actually predicting.

6. Conclusions

This study aimed to verify if it is possible to predict football match results using artificial neural networks. Several steps and techniques were performed to create the final prediction model. The model comprises a Deep Neural Network with three hidden layers, and its hyperparameters were automatically chosen for its best performance. In addition,

the dataset was pre-processed to optimise network training, and several new datasets were calculated to improve performance.

The data organisation method proved to be the best, in terms of accuracy, when compared to others presented in the state-of-the-art. Furthermore, four proposed features resulted in higher importance when compared to state-of-the-art features, such as the pi-rating, and another four features, even with lower importance, were close to the pi-rating. This is especially useful when considering the processor demand from the pi-rating.

The proposed model obtained 52.8% accuracy in the set of 2589 matches, reaching up to 80.3% for specific cases. After analysing the network's performance for the different leagues, it was concluded that the model's accuracy is unrelated to the concentration of the test set outcomes. That proves the successful learning of the neural network to predict football match results.

Football has three possible outcomes: home team win, draw, or away team win. If this sport was unpredictable, any prediction method would have a maximum of 33.3% accuracy, as there are three possible outcomes for one choice only (1/3 = 0.33 = 33%). The proposed model obtained 52.8% accuracy in the set of 2589 matches from 2018 and 2019 with training data before these dates. In other words, the neural network was able, with a positive margin of 19.5% (52.8% predicted by the model minus 33% if the sport was unpredictable), to identify the main parameters that affect the result and thus create a mathematical equation to predict matches with past statistical data. Looking further at the results, as shown, this margin can increase up to 47.5% for specific cases (when the neural network predicted that the home team would win with a 60% chance).

The accuracy is not a linear parameter that allows direct comparison between the frameworks due to the different validation datasets. Still, it is worth noting that the result is very close to the Soccer Prediction Challenge champion model and at the same level or higher than the state of the art presented, especially regarding the most recently used one [12], in which the model described obtained similar accuracy but considering almost double the number of matches.

The proposed future works involve bringing new objectives and aspects to the forecasting of football match results with the positively validated features of this work. Namely, it is proposed to set an objective of finding the best accuracy using these features and different models' architecture (e.g., RNN and LSTM), and the objective of a deeper analysis on why some categorical features have such an important impact on the accuracy results, especially the difference in the competing leagues.

From the results, it can be stated that football is not unpredictable, and it is possible to predict football match results with good accuracy using artificial neural networks and past match statistics.

Author Contributions: L.E.L. formal analysis, original draft—review and editing, and visualization. G.F.: methodology, software, investigation, and data curation. J.P.T.: conceptualization, validation, resources, supervision, project administration, and funding acquisition. All authors have read and agreed to the published version of the manuscript.

Funding: The authors are grateful to the Foundation for Science and Technology (FCT, Portugal) for financial support through national funds FCT/MCTES (PIDDAC) to CeDRI, UIDB/05757/2-020 (DOI: 10.54499/UIDB/05757/2020) and UIDP/05757/2020 (DOI: 10.54499/UIDB/05757/2020), and SusTEC, LA/P/0007/2020 (DOI: 10.54499/LA/P/0007/2020). The authors are also grateful for the funding granted by the project NanoStim—Nanomaterials for wearable-based integrated biostimulation (POCI-01-0247-FEDER-045908).

Data Availability Statement: The original data presented in the study are openly available at the WhoScored website (www.whoscored.com).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Buraimo, B.; Simmons, R. Uncertainty of Outcome or Star Quality? Television Audience Demand for English Premier League Football. *Int. J. Econ. Bus.* **2015**, *22*, 449–469. [CrossRef]
2. Chari, T. Discursive constructions of the Germany–Brazil semi-final match during the FIFA 2014 World Cup: The limits of football as a soft power resource. *Communicatio* **2015**, *41*, 405–422. [CrossRef]
3. Jain, A.K.; Mao, J.; Mohiuddin, K.M. Artificial neural networks: A tutorial. *Computer* **1996**, *29*, 31–44. [CrossRef]
4. Rahman, M.A. A deep learning framework for football match prediction. *SN Appl. Sci.* **2020**, *2*, 165. [CrossRef]
5. Buursma, D. Predicting sports events from past results. In Proceedings of the 14th Twente Student Conference on IT, Enschede, The Netherlands, 21 January 2011; Volume 21, pp. 1–8.
6. Constantinou, A.C.; Fenton, N.E. Determining the level of ability of football teams by dynamic ratings based on the relative discrepancies in scores between adversaries. *J. Quant. Anal. Sport.* **2013**, *9*, 37–50. [CrossRef]
7. Kumar, G. Machine Learning for Soccer Analytics. Master's Thesis, Katholieke Universiteit Leuven, Leuven, Belgium, 2013. [CrossRef]
8. Tsokos, A.; Narayanan, S.; Kosmidis, I.; Baio, G.; Cucuringu, M.; Whitaker, G.; Király, F. Modeling outcomes of soccer matches. *Mach. Learn.* **2019**, *108*, 77–95. [CrossRef]
9. Dubitzky, W.; Lopes, P.; Davis, J.; Berrar, D. The Open International Soccer Database for machine learning. *Mach. Learn.* **2019**, *108*, 9–28. [CrossRef]
10. Hubáček, O.; Šourek, G.; Železný, F. Learning to predict soccer results from relational data with gradient boosted trees. *Mach. Learn.* **2019**, *108*, 29–47. [CrossRef]
11. da Costa, I.B.; Marinho, L.B.; Pires, C.E.S. Forecasting football results and exploiting betting markets: The case of "both teams to score". *Int. J. Forecast.* **2022**, *38*, 895–909. [CrossRef]
12. Holmes, B.; McHale, I.G. Forecasting football match results using a player rating based model. *Int. J. Forecast.* **2024**, *40*, 302–312. [CrossRef]
13. Arntzen, H.; Hvattum, L.M. Predicting match outcomes in association football using team ratings and player ratings. *Stat. Model.* **2021**, *21*, 449–470. [CrossRef]
14. Fialho, G.; Manhães, A.; Teixeira, J.P. Predicting Sports Results with Artificial Intelligence—A Proposal Framework for Soccer Games. *Procedia Comput. Sci.* **2019**, *164*, 131–136. [CrossRef]
15. Elo, A.E. *The Rating of Chessplayers, Past and Present*; Arco Pub: Houston, TX, USA, 1978.
16. Pelánek, R. Applications of the Elo rating system in adaptive educational systems. *Comput. Educ.* **2016**, *98*, 169–179. [CrossRef]
17. Hawkins, D.M. Outliers from the linear model. In *Identification of Outliers*; Springer: Dordrecht, The Netherlands, 1980; pp. 85–103. [CrossRef]
18. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958. [CrossRef]
19. Beck, M.B.; Lin, Z. Transforming data into information. *Water Sci. Technol.* **2003**, *47*, 43–51. [CrossRef] [PubMed]
20. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
21. Fisher, A.; Rudin, C.; Dominici, F. All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. *J. Mach. Learn. Res.* **2019**, *20*, 1–81. [CrossRef]
22. Alibrahim, H.; Ludwig, S.A. Hyperparameter Optimization: Comparing Genetic Algorithm against Grid Search and Bayesian Optimization. In Proceedings of the 2021 IEEE Congress on Evolutionary Computation (CEC), Kraków, Poland, 28 June–1 July 2021; pp. 1551–1559. [CrossRef]
23. Štrumbelj, E.; Šikonja, M.R. Online bookmakers' odds as forecasts: The case of European soccer leagues. *Int. J. Forecast.* **2010**, *26*, 482–488. [CrossRef]