e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:08/August-2023 Impact Factor- 7.868 www.irjmets.com

COMPARISON OF DIFFERENT ML MODELS TO PREDICT FOOTBALL MATCH RESULTS

Arnav Jhamvar[*1]

[*1]HDFC School, Pune, Maharashtra, India.

DOI : https://www.doi.org/10.56726/IRJMETS44017

ABSTRACT

In this research paper, I will compare three ML models used to predict football match results. These models are from three different time frames, and a comprehensive review was done to decide which model was the best and gave the most accurate prediction.

Keywords: Models, Football, Research, Time Frames, Prediction.

I. INTRODUCTION

Football is the most popular sport in the world. Dating back centuries, this game has made significant evolution since. The emotion is at an all-time high, encompassing a rich history of some of the greatest clubs and their fans worldwide. So much so that it is much more than a game now; people's emotions and money ride on the results of football games. The betting industry has evolved with the games' increasing popularity. Predicting the result of a game is a complex task, taking into account various factors that could influence the result ranging from home advantage, player fitness and crowd.

1.1 MOTIVATION

The betting industry has been thriving since the sports increasing significance, and people want to test their luck in winning some extra cash by trying to predict match results. The unpredictable nature of Football makes it almost impossible to bet on these games. Thus came the role of machine learning to help analyse games and make calculated predictions instead of betting on luck. Other than that, fantasy leagues that do not necessarily deal in money also hold a significant share in predicting match results.

1.2 CURRENT STATE OF THE ART

Football analysis and studies of football patterns and strategies to improve a team's chance of winning has been carried out for over seven decades. Back in the 1950s, Charles Reep collected information by hand statistics suggesting the key to scoring goals was passing the ball around quickly from defence to offence leading to the "Long Ball Movement" in English Football.

We have come a long way since in collecting data and information to create a winning team. Technology has made it possible to collect large volumes of data, allowing for more extensive research analysis, and thus paved the way for various algorithms designed to extract information and learn from this data. The current technologies allow data collection of the events that take place during a match.

1.2.1 Soccer Logs

Describe events that occur during a match and are captured through proprietary tagging software.

1.2.2 Video tracking Data

Describes the movement of players during games and is collected from match recordings.

1.2.3 GPS Data

Describes the players' trajectories during games and training sessions. It also helps gather the physical aspects of a player's game, like top speed, heart rate, shot power etc. This is calculated with the help of tracking devices and motion sensors embedded in their equipment.

This data helps gather essential information that would be used in algorithms to make precise predictions for match results; algorithms like Support Vector Machines, XG boost, and Logistic Regression for data training are used to work with this data.

1.3 GOALS OF THE PROJECT

There are countless models to predict results in not only Football but other sports as well, but my objective is to focus primarily on football match predictions over the last few years and review the methods and improvements that increased the efficiency to predict results more accurately. I shall research a model from 2017, 2021, and 2022 to the present day. By delving deeper into the nuances of their models and processes, we can determine which model is the most effective in making the most accurate predictions, which is my project's ultimate goal.

II. RELATED WORK

Some uncountable processes and methods go into predicting match day results with varying levels of accuracy, ranging from algorithms like Logistic Regression, Neural Networks, and Random Forest with an accuracy of about 60%. Some algorithms also consider the teams' form in the last five games. For example, A Sure Bet by Stanford University used a sequential model (LSTM), which garnered only 47% accuracy.

III. METHODOLOGY

I would review three models over three years to determine the best and most accurate. However, the data can have different forms and nature. Some papers have looked at how to predict the final outcome of a match, given the events occurring during the match. Other works make their predictions given the previous matches played by both teams and taking into account the statistics of those matches, not only in terms of a high accuracy percentage but also the one which is most accurate, taking into consideration a lot of different factors that weigh in to sway a match result.

3.1 Model I, 2020

3.1.1 Data Description

This model contained a dataset across 63 countries or leagues for the 2016-17 season. The data included events during matches, like goals, fouls, and penalties. They had to transform the events to use this data as input for their model, so they all had the same size.

3.1.2 Data Processing

They trained the LSTM algorithm for this model, which comes under Recurrent Neural Network (RNN) classification. They proposed Deep Embedding inspired by word2vec, which worked well for NLP. Second place was the One-Hot Vector with the attribute of all events meaning that it had an empty column for all attributes. The final option was Concatenated Embedding Vectors for all attributes. Along the same lines as the one-hot vector as each event is embedded. Ultimately Concatenated Embedding Vectors was chosen, and each sample entered was an event in the match. The input consists of a vector of ten variables. These variables are then transformed to a one-hot encoding for each attribute and then concatenated or to an embedding lookup for each feature and concatenated to form a more significant vector. During training, they proposed seven case studies, of which six used LSTM models with different architectures but using the "Many-to-Many" error. The last one, however, had the "Many-to-One" model. They experimented with these models at 15-minute intervals during the match; the best model was the "Many-to-One" model, with two hidden layers and 256 LSTM units with an accuracy of 98% by the time the game ended.

3.2 Model II, 2021

3.2.1 Data Description

This model contains a dataset of combined features from the 2019-20 and 2020-21 Premier League seasons. It used the sequential data format (sample time step, sequence). Each sample contained the Home team features(H), away tea features(A), and odds from their last five matches.

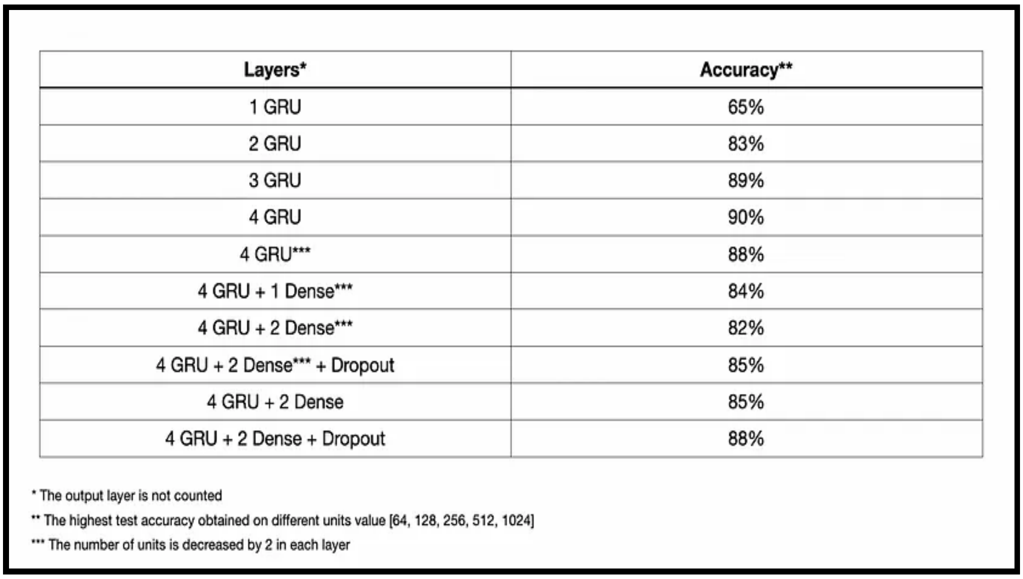
3.2.2 Data Processing

The Gated Recurrent Unit(GRU) is the model used in this algorithm, again an RNN classification. This is a neural network designed to process sequential data. It uses gates to select valuable data. An update gate determines how much data should be passed into the next state.

Even though LSTMs are the most popular recurrent neural network, GRU has proved to work better with less training data and converged faster with less complexity. The GRU testing outperformed LSTM by 20%, with an accuracy of 92%.

Different architectures, with the help of Dense Layers and descending units, and compared based on the achieved accuracy to find the optimal performance with the most remarkable accuracy. The model was tested with 1GRU, which attained an accuracy of 65%. 2GRUs obtained an accuracy of 83%. 3GRUs obtained an accuracy of 89%. 4GRUs obtained an accuracy of 90%. The accuracy lowered When further tested with dense layers and descending units. However, after a grid search, the training parameters listed below gave a stunning 92% accuracy on testing data. Parameters used during testing were: Dropout = 0.3, Learning rate = 0.0002, Batch = 30, Epochs = 75.

Table 1. Model II results



Layers*	Accuracy**
1 GRU	65%
2 GRU	83%
3 GRU	89%
4 GRU	90%
4 GRU***	88%
4 GRU + 1 Dense***	84%
4 GRU + 2 Dense***	82%
4 GRU + 2 Dense*** + Dropout	85%
4 GRU + 2 Dense	85%
4 GRU + 2 Dense + Dropout	88%

* The output layer is not counted
** The highest test accuracy obtained on different units value [64, 128, 256, 512, 1024]
*** The number of units is decreased by 2 in each layer

3.3 Model III, 2022

3.3.1 Data Description

This dataset contained 1900 football matches spanning over five seasons, from 2013-14 to 2018-19 in the English Premier League. Besides statistical data and performance information on both teams, additional features, including individual player features and skills, were also included. This dataset separated data into training and testing sets as it considered how the performance would vary throughout the season—considering that a team would have a dip in their performance by the end of the season as fatigue can kick in or the players are rested or not utilising their full potential upon reaching their desired goals.

3.3.2 Data Processing

Several models with different characteristics were tested to find the best classification model on the data set to check for the best fit. These, along with precision, calculated the profit made in euros. Before testing the different methods, data was normalised to eliminate the effects of significant variations. This "data cleaning" was done to have the models only use the most relevant variables and guarantee better, more precise results. The Support Vector Machines(SVM)-SVM method from the e1070 package was the best algorithm to get the most accurate results in this data set with an accuracy of 61.32% with a 95.06€ profit.

Table 2. Model III results

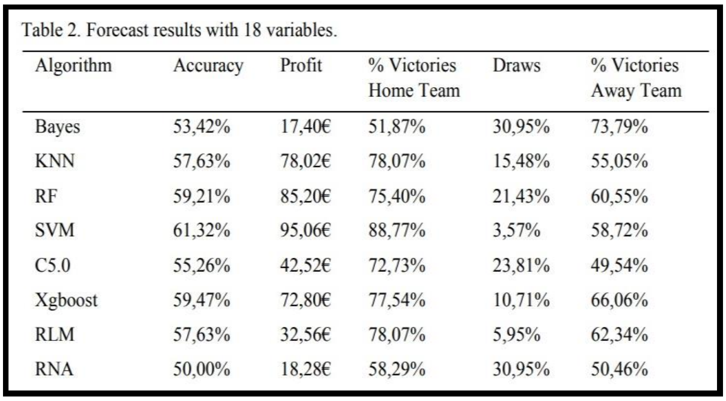


Table 2. Forecast results with 18 variables.

Algorithm	Accuracy	Profit	% Victories Home Team	Draws	% Victories Away Team
Bayes	53,42%	17,40€	51,87%	30,95%	73,79%
KNN	57,63%	78,02€	78,07%	15,48%	55,05%
RF	59,21%	85,20€	75,40%	21,43%	60,55%
SVM	61,32%	95,06€	88,77%	3,57%	58,72%
C5.0	55,26%	42,52€	72,73%	23,81%	49,54%
Xgboost	59,47%	72,80€	77,54%	10,71%	66,06%
RLM	57,63%	32,56€	78,07%	5,95%	62,34%
RNA	50,00%	18,28€	58,29%	30,95%	50,46%

IV. THREATS TO VALIDITY

Among the three different models, a number of different factors were considered to make predictions, but only some of those considered factors occurred during the match. For example, injuries sustained by any player during the course of the match. That could completely turn the tides of a team winning or losing, especially when a marquee or star player comes down with an injury. The odds and bets would thus be compromised if any such event would occur. The other unforeseeable circumstance would be extreme weather conditions, like snow and heavy rains. These can alter the flow of events and, thereby, the match results. Not considering these factors can get us an inaccurate prediction.

V. CONCLUSION

We have reviewed three different models from 2020, 2021, 2022. Model 1 from 2020, using the LSTM model, attained an accuracy of 98%. Model 2 from 2021, using the GRU model, acquired an accuracy of 92%. Model 3 from 2022, using the SVM method, observed an accuracy of 61.32%.

VI. SUMMARY OF RESULTS

As we can see, the accuracy with every passing year decreases. One of the most significant reasons is the consideration of more factors over time. Even though the algorithms' precision decreases, the actual result's accuracy increases.

The first model was used across 63 leagues and countries, focusing more on the number of leagues than determining accurate results taking multiple factors into consideration. It used a many-to-many function, calculated the accuracy at 15-minute intervals, and kept increasing.

The second model considered considerably more stuff, like the home team crowd and its effect, xG for the home and away teams, attacking styles, defensive styles and passes. This model sees a decrease in accuracy as many more factors are considered to weigh in the games' results. It also showed a linear increasing graph when tested with a different number of GRU units. When tested with the dense methods, it showed a linear decreasing graph. Over all it made a more accurate prediction of an actual game result.

In Model 3, besides free statistical data on football matches, information about the performance of both teams was also collected, including goals scored, shots made, number of corners, number of faults committed, number of yellow and red cards, odds for each match, final result of match and match referee. In addition to these data, data extracted from the website sofifa.com, containing descriptive statistics of individual features and skills of all football players (e.g., pass accuracy, agility, reaction, aggression) and statistical information on football teams' quality, were also used. All these variables made predicting the match result even more challenging but more accurate. The greater the number of variables, the more combinations to be tested to make a more accurate prediction.

The accuracy percentage has decreased due to the additional factors considered. Model 1, even though with the highest accuracy percentage, is most unlikely to be the correct prediction, as the others consider significantly more.

Table 3. Comparison of all three models

YEAR	MODEL	ACCURACY
2020	**LSTM (RNN)**	**98%**
2021	**GRU (RNN)**	**92%**
2022	**SVM**	**61.32%**

VII. FUTURE WORK

For future work, we can investigate more variables that could be useful to make predictions with greater accuracy, like injuries and considering weather conditions. We can also take individual player stats using FIFA and PRO Evolution. More data on individual team playing styles should be given more focus. And lastly, physiological match day variables like distance/time and acceleration. Taking all these into consideration would help to make future predictions more accurate.

VIII. ETHICAL IMPLICATIONS AND RECOMMENDATIONS

In this modern day and age, using all these data scraping methods to collect information and make predictions can be a boon or bane. If used in favour of a team, it can go a long way in improving a team's performance. While if a team cannot afford to use modern-day technology to improve their game, they stand at a distant disadvantage in terms of gameplay. Thus, to remain morally fair, each team should be given a fair share of funds to either allow data scraping to improve their performance or not allow any team to do so.

ACKNOWLEDGEMENT

I would like to express my gratitude to Professor Janyl Jumadinova, Chair of Computer Science and Information Science, Allegheny College. She helped me understand the basics of Machine Learning and helped me complete my project with the utmost efficiency, and offered help and inputs to improve my project.

IX. REFERENCES

[1] Rodrigues, Fátima, and Ângelo Pinto. "Prediction of football match results with Machine Learning." Procedia Computer Science 204 (2022): 463-470.

[2] Jain, Vivek, et al. "An Exploratory Study of ML Techniques in Football Match's Result Prediction." Journal of Advanced Research in Embedded System 9.3&4 (2023): 1-4.

[3] Pruñonosa Soler, Guillem. A study on the performance of machine learning algorithms for predicting professional football match outcomes. Diss. Universitat Politècnica de València, 2023.

[4] Głowania, Szymon, Jan Kozak, and Przemysław Juszczuk. "Knowledge Discovery in Databases for a Football Match Result." Electronics 12.12 (2023): 2712.

[5] Nestoruk, Roman, and Grzegorz Slowinski. "Prediction of Football Games Results." CS&P. 2021.

[6] Carloni, Luca, et al. "A machine learning approach to football match result prediction." HCI International 2021-Posters: 23rd HCI International Conference, HCII 2021, Virtual Event, July 24–29, 2021, Proceedings, Part II 23. Springer International Publishing, 2021.

[7] Bunker, Rory P., and Fadi Thabtah. "A machine learning framework for sport result prediction." Applied computing and informatics 15.1 (2019): 27-33.

[8] Matheus Kempa. "Machine Learning Algorithms for Football Predictions." Towards Data Science. 2020.

[9] Maciej Balawejder. "Premier League Predictions Using Artificial Intelligence." Nerd for Tech. 2021.

[10] Herold, Mat, et al. "Machine learning in men's professional football: Current applications and future directions for improving attacking play." International Journal of Sports Science & Coaching 14.6 (2019): 798-817.

[11] Baboota, Rahul, and Harleen Kaur. "Predictive analysis and modelling football results using machine learning approach for English Premier League." International Journal of Forecasting 35.2 (2019): 741-755.

[12] chat.openai.com