Data Mesh Approach to Predicting Soccer Match Outcomes with Dynamic In - Game Factors

Sajith Narayanan

Capital One Services, Richmond, Virginia, USA
Email: *sajith.narayanan[at]capitalone.com*

Abstract: ***The study looks at a model to predict the final number of goals in soccer matches by breaking the games down into minutes and looking at both fixed and changing factors. The study used data from the four biggest soccer leagues from the 2018/19 to 2021/22 seasons. A Machine Learning (ML) model called a Multilayer Perceptron (MLP) was trained and compared with other models like multiple linear regression and random forest to predict how many goals would be scored. The research also studied how a coach's decisions, like making substitutions or changing the team's lineup, affect the final score. The results showed that the MLP model worked better than the other models, with an improvement of 1.42% over linear regression and 0.41% over random forest. Adding in factors like substitutions and changes in strategy made the predictions even better. It found that increasing the number of substitutions reduced the total goals scored by both teams.***

Keywords: Machine Learning, Matches, Multilayer Perceptron, Soccer

1. Introduction

In recent years prediction models for sports during the game have become popular especially for predicting match results. These models uses both fixed factors and changing factors. Research shows that adding these dynamic factors during the game can help make predictions more accurate and give a better understanding of how a coach's decisions affect the match. Coaches have a significant impact on the conclusion of a game. One of the keyways coaches influence the game is by making substitutions or by swapping the team's strategy. For instance, changing an attacking player for a defensive one or the other way around can change how the team plays. However, the effect of these changes on the result in soccer has not been fully studied. Most research on soccer predictions has focused more on fixed factors and not enough on what happens during the game. Predicting the winner of a soccer match can be hard because many games end in ties. This makes it difficult to rely on win/loss predictions. For this reason, this study focuses on predicting the number of goals scored by both the home and away teams rather than just predicting the winner. Using goals as the prediction variable is better because it gives more detailed information about the strength of a team. For instance, a 2 - 1 win doesn't show as much dominance as a 10 - 1 win. Games with more goals are usually more exciting for fans. This can make the sport more popular. Therefore, the main goal of this study is to use a method called Artificial Neural Networks (ANN) [1], [2], [3], [4], [5], [6] to predict how many goals both the home and away teams will score. This method will consider both fixed factors and dynamic factors. It will help to create a better prediction model for soccer games. The study also focusses to explore how coach decisions change the game and affect the number of goals scored. These decisions provide valuable insights in how coaching impacts the match leads to new strategies for improving match outcomes and predicting goals more accurately.

This study emphasizes at some important questions about predicting soccer match results such as how does switching to a more defensive lineup affect the number of goals, and how does switching to a more attacking lineup change the outcome? Previous studies mostly looked at player changes to reduce fatigue, but they didn't focus on tactical changes like switching an attacker for a defender. Earlier research in soccer mostly predicted the winner of a match based on known factors. Also, what happened during the match like the studies by [7] and [8]. Some studies even broke the match into smaller time periods to make predictions, like [9] did. But this study tries to predict the total number of goals for both teams, not just who will win. sNevertheless, the impact of tactical substitutions has not been studied enough. This study aims to explore how these changes affect the final number of goals scored by the teams. The results of this research could be useful for different groups. Soccer teams and coaches could use real - time predictions from ANNs to make better decisions during the match. This will help them adjusting their player lineups based on what is shown to work. Organizations like FIFA could also use the findings to change the rules of the game to make matches more exciting. Lastly, the research might have a positive impact on viewers and society. More people may watch the games if matches become more exciting due to better coach decision.

The study is as follows; the related works will be shown in the next section. The materials and methods are described in Section III. The experimental analysis is carried out in Section IV. Additional experiments are carried out in Section V, and in Section VI, we wrap up the study with some conclusions and plans for future research.

2. Related Works

Prediction models in sports have made great progress and often did better than human experts at predicting outcomes. These models are used in various sports, like predicting how far a javelin will be thrown [10] or who will win a basketball game [11]. In soccer, [12] used a method called Bayesian networks to predict the outcome of a soccer match for one team. They found that their model, which was 59%

DOI: https://dx.doi.org/10.21275/SR25202194426

accurate, worked better than other models like decision trees and *K* - nearest neighbour models. Other researchers such as [13] have tried different ways to predict match outcomes used logistic regression and neural networks to predict soccer match winners and got a high accuracy of 95%. Later, [14] used Machine Learning (ML) models to predict soccer match outcomes from over 45, 000 matches in the top five European soccer leagues. They set up that random forest with an accuracy of 81% worked better than other models. Other models like gradient boosting, Support Vector Machine (SVM) and linear regression. [15] had similar results with random forest outperforming other models like logistic regression and SVM in predicting soccer match outcomes. In addition to traditional ML models, Deep Learning (DL) has become popular. [16] used a DL model called Multilayer Perceptron (MLP) to predict match outcomes. Their model had 73% accuracy, which was better than random forest, SVM, and Gaussian Naïve Bayes. This suggests that DL might work better than traditional ML models. Recently, there has been more interest in combining both known match features and dynamic in - game features for predictions. [17] created a model to predict the winner of one - day international cricket matches while the game was still going on. Using data from the first inning they were able to predict the final winner with 71% accuracy using multiple linear regression [18], [19], [20], [21], [22], [23], [24]. In conclusion, prediction models in soccer have advanced, but we still don't fully understand how coach decisions, like player substitutions and strategy changes, affect match outcomes and goals. More research in these areas could make soccer prediction models even more accurate and useful.

3. Materials and Methods

In sports, predicting the result of a match is often done by classifying the outcome as a home win, lose or a draw. This method is common, but it has some trouble. Major issue is that it can lose crucial information by making it difficult to find useful factors that affects the game. This is because classifying the outcome into just three categories reduces the accuracy of predictions. A better way, suggested by recent studies is to predict the exact number of goals scored by both the home and away teams instead of just the match result. This approach allows for more variation in the data by making predictions more accurate. Fixed features like the league, season and date of the match can help to report for differences between games. While the date is used to divide the data into training and testing sets it is not directly used to predict outcomes. A key factor used in the model is the Elo rating system. It measures how strong a team is and has been shown to help predict match results. Another important factor is the advantage as home teams tend to win more often than away teams. To report for this, the model makes separate predictions for the number of goals scored by the home and away teams. The model also considers factors that change during the game, as these can influence the result. The match is divided into 90 minutes, with each minute being treated as a separate data point. After every minute, the model is updated with key events, like red and yellow cards, which can affect the game's outcome. The current score is also included as it affects how teams play during the match. Two more factors related to coaching decisions are included in the model. First, player substitutions are considered as substituting players can reduce fatigue and change the outcome of the game. Second, changes in team strategy during the match are tracked. The team strategy score is updated based on the types of players substituted in attackers increase the score while defenders lower it. These factors help to capture how the team's playing style changes which can influence the number of goals scored. The proposed method's flowchart is displayed in Fig.1.

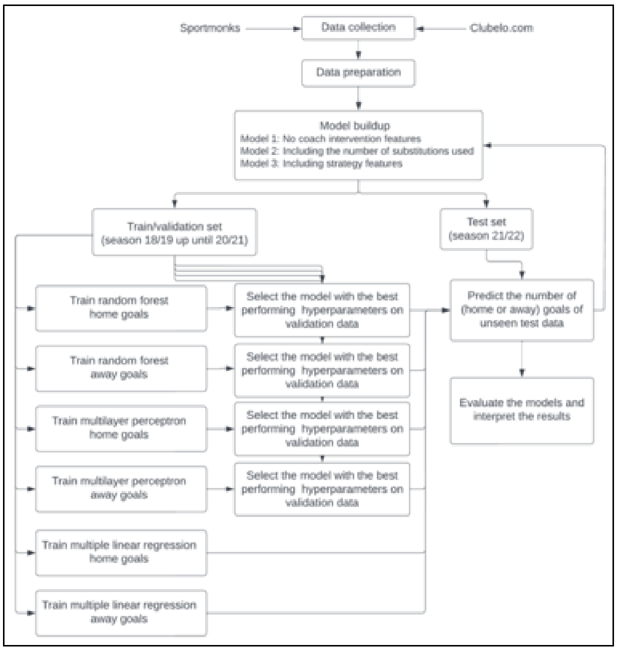
Sportmonks
Data collection
Clubelo.com
Data preparation
Model buildup
Model 1: No coach intervention features
Model 2: Including the number of substitutions used
Model 3: Including strategy features
Train/validation set (season 18/19 up until 20/21)
Test set (season 21/22)
Train random forest home goals
Train random forest away goals
Train multilayer perceptron home goals
Train multilayer perceptron away goals
Train multiple linear regression home goals
Train multiple linear regression away goals
Select the model with the best performing hyperparameters on validation data
Predict the number of (home or away) goals of unseen test data
Evaluate the models and interpret the results

Figure 1: Flowchart showing the procedures from gathering data to assessing the model

a) Dataset Analysis

This study looks at data from the last four seasons of the four biggest football leagues in Europe. It uses information from Sportmonks[1] and clubelo. com[2] to understand how things like goals, yellow/red cards, and substitutions affect the game and how teams change their strategies during matches. This helps measure how a team's strategy changes when players are swapped. The study also gets the Elo rating (a system for ranking teams) for both the home and away teams from clubelo. com. The data is then combined into one big set, where each match is broken down into 90 "observations", one for each minute of the match. For every minute the match features stay the same but events like yellow cards, red cards, goals and substitutions are updated as they happen. For instance, if a yellow card happens at the 6th minute the number of yellow cards for the home team is updated from that point onward. The study also looks at how substitutions change the team's strategy by comparing the positions of the player being substituted and the one coming in. The difference in positions is calculated and added to the data to show how much the strategy changes. The study shows some basic statistics about the data, such as the average number of goals scored by both home and away teams, and how often substitutions are made in Figs.2 to 4. On average, home teams score 1.382 goals per match, which is a little more than away teams, who score 1.14 goals per match. This shows that home teams have a slight advantage.

[1] https://www.sportmonks.com/
[2] http://clubelo.com/

The study also finds that coaches make fewer substitutions in the first half, but after the 60th minute, substitutions become more common. After halftime, both teams' strategy scores tend to go up, but then they decrease as the match comes to an end.

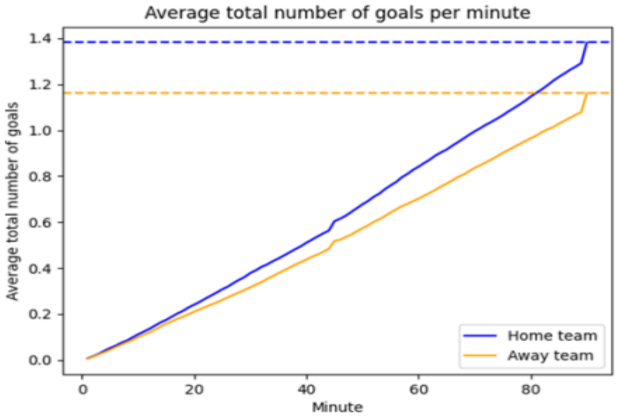

Figure 2: Total goals scored by the home and away teams per minute

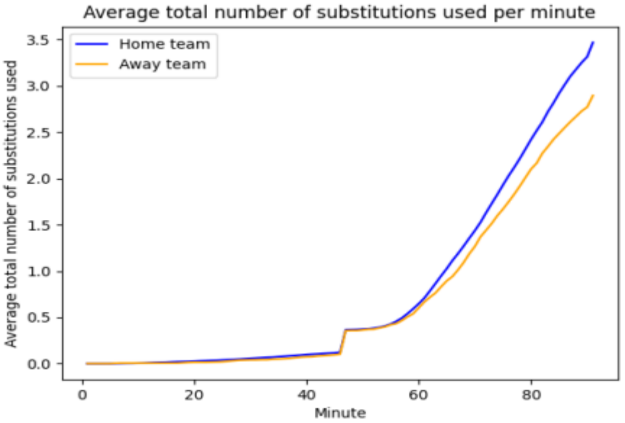

Figure 3: The total number of substitutions made by the home and away teams per minute

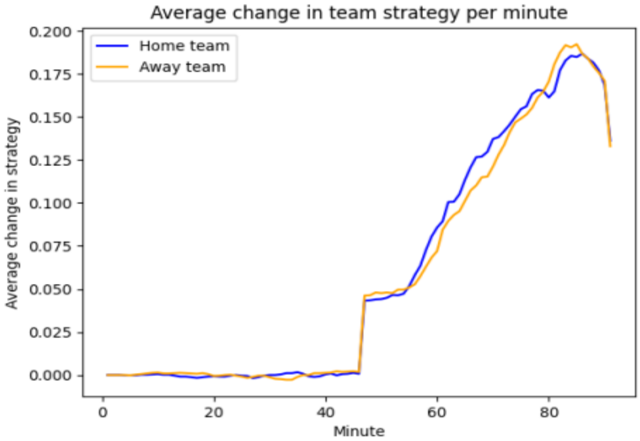

Figure 4: The home and teams' average strategy score change per minute

b) Model Analysis

In sports analytics predicting the outcomes of games such as the number of goals in soccer is a growing area of interest. This study aims to compare different models to see which one is best at predicting how many goals will be scored in soccer matches. The study focuses on three prediction models which are multiple linear regression, random forest and MLP. Multiple linear regression is a common statistical method. It helps to understand the relationship between a dependent variable and multiple independent variables. Linear regression is easy to use and helps to explain the data clearly. It assumes that the relationship between the variables is linear and the errors are evenly spread out. The variables are not too closely related to each other. If these assumptions are not met the results may not be reliable. Random forest is a method that builds many decision trees based on different parts of the data and selects the best trees for making predictions. Unlike linear regression, random forests can find more complex relationships between the variables. They work well for predicting soccer match results and avoid problems like overfitting, which happens when a model is too closely fitted to the training data. However, the number of decision trees in the random forest needs to be carefully adjusted. Adding more trees doesn't always improve performance so the settings (hyperparameters) must be fine - tuned using methods like random grid search. A MLP is a type of ANNs. It has an input layer one or more hidden layers and an output layer. The input layer takes in the data and the hidden layers process the data to make a prediction. The MLP can model complex relationships that are not just straight lines. It has been shown to be accurate in predicting sports outcomes including soccer. The MLP uses different functions called activation functions. It processes the data in the hidden layers. Rectified Linear Unit (ReLU) is often used in the hidden layers because it works well for avoiding certain problems in training. The output layer uses an identity function to make predictions about the number of goals which is a continuous number. To test how well the models works the data is divided into two parts. One for training the models and the other for testing them. The training set includes data from three soccer seasons (2018 - 2021) while the test set includes data from the most recent season (2021 - 2022). Because soccer data is ordered by time, traditional cross - validation methods can't be used because they would allow the model to "cheat" by using future information during training. Instead, a special method called time series cross - validation is used, which simulates real - world conditions where the model only knows past data when making predictions. The performance of the models is measured using different metrics such as Mean Squared Error (MSE), R - squared and adjusted R - squared. The model with the lowest MSE is considered the best. The study also looks at whether adding features related to coach changes (like a new manager) can improve the predictions. For multiple linear regression, a test called the likelihood ratio test[3] is used. For random forest and MLP a Diebold - Mariano test[4] checks if adding coach features improves accuracy. The three models are then tested to see how well they predict the number of goals scored by both home and away teams. The predicted number of goals is rounded to the nearest whole number to determine if the match ends in a win, lose or draw. The accuracy of the models is compared to previous studies to understand how well they perform. The goal is to find the most effective model for predicting

[3] The likelihood ratio test determines whether statistical model is a better fit by comparing the likelihoods of the two models.

[4] By comparing the forecast errors of two models and taking statistical significance for performance differences into account, the Diebold-Mariano test assesses predictive accuracy.

soccer match outcomes with a focus on how well the model can make predictions for new data.

4. Experimental Analysis

For the multiple linear regression model to work correctly certain assumptions need to be met. One of these is that the features (independent variables) should not be too closely related to each other. In this study, the "minute" variable is too strongly related to other features. So, it is removed from the model. The model also assumes that the errors (residuals) should follow a normal distribution. However, the test showed that this assumption is not met. Even after changing the data the problem did not go away. The data also showed varying spread of errors which means the errors were not consistent. Because of these issues, the results of this model should be interpreted carefully. The study tested three models with increasing complexity. The first model did not include coach interventions, the second one added coach strategy changes, and the third one added both strategy changes and substitutions. Table I shows how close the predictions are to the actual results. The lower the MSE, the better the model. For the multiple linear regression model, without coach interventions, the MSE for home goals was 0.834 and for away goals was 0.695. When substitutions were added, the MSE improved to 0.816 for home goals and 0.690 for away goals. Adding strategy changes improved the model even more. In the random forest model, the MSE was 0.792 for home goals and 0.671 for away goals without coach interventions. Adding substitutions lowered the MSE to 0.776 for home goals and 0.656 for away goals, and adding strategy changes improved the accuracy further. For the MLP model, the initial MSE was 0.771 for home goals and 0.666 for away goals. Adding substitutions brought the MSE down to 0.762 for home goals and 0.652 for away goals, and adding strategy changes improved the predictions even more. In all three models, adding coach interventions (substitutions and strategy changes) led to better predictions.

Table I: Predictive Performance on the Test Data Across the Models Measured Via MSE

Time	Multiple Linear Regression	Random Forest	Multilayer Perceptron
	Model 1	Model 2	All Features
	Home	Away	Home
Overall	0.8341	0.6953	0.8156
90min	0.2443	0.1505	0.1435
75min	0.4407	0.3248	0.3965
50min	0.7210	0.6018	0.7166
25min	1.1177	0.9487	1.1074
0	1.5351	1.3520	1.5512

The study also checked how well the models could predict the winner of the match as shown in Table II. When coach interventions were added, the accuracy of all models improved. For multiple linear regression, the accuracy went up from 61.38% to 62.56%. The accuracy for the random forest model increased from 63.01% to 63.57%, and the MLP model's accuracy increased from 62.79% to 63.98%. This shows that adding features related to coach decisions improves the predictions of match outcomes. The study also looked at how coach decisions affected the number of goals scored. Substitutions had a small negative effect on goals. For instance, after each substitution, the home team's goals decreased by 0.0889, and the away team's goals decreased by 0.0072. Both random forest and MLP models found similar results. Changes in strategy had a bigger impact. Changing the away team's strategy did not affect the goals, but changing the home team's strategy did. A more defensive strategy by the home team led to fewer goals for both teams. On the other hand, a more attacking strategy increased home goals and reduced away goals. These results were also seen in the other models.

Table II: Predictive performance on the test data across the models measured via accuracy

Time	Multiple Linear Regression	Random Forest	Multilayer Perceptron
	Model 1	Model 2	All Features
Overall	0.6138	0.6196	0.6256
90min	0.9029	0.9076	0.9164
75min	0.7443	0.7495	0.7571
50min	0.6236	0.6320	0.6307
25min	0.525	0.5303	0.5464
0	0.4129	0.4184	0.4264

5. Additional Experiments

The analysis of the models starts by checking for errors and comparing how well the models predict on both the training data and testing data as shown in Fig.5. This venture helps to see if the models work well on new data and not just the data they were trained on. The models are trained using time series cross - validation. This ensures that they do not underfit or overfit. The performance of the top three models is further checked by looking at the mean squared residuals per minute as shown in Fig.6. This is a measure of how accurate the predictions are. For the first 20 to 60 minutes all three models predict similarly. However, after 60 minutes the non - linear models do a better job of predicting the final goal count for both the home and away teams. This suggests that the non - linear models are better at understanding the match as it goes on. Next, the results from the first 140 matches are compared with the results from the last 140 matches. This comparison shows that all the models perform better on the last 140 matches. This improvement is partly because the Elo rating system gets more accurate towards the end of the season as most player changes happen at the start of the season. Even though the models improve, their overall performance stays similar, which makes it fair to compare them.

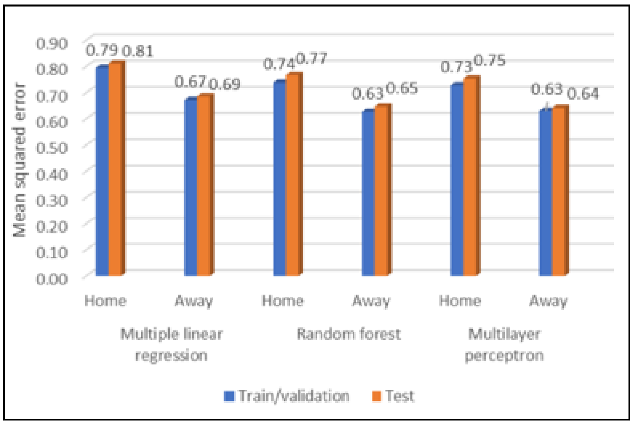

Figure 5: Each model's average MSE on the test and train/validation data

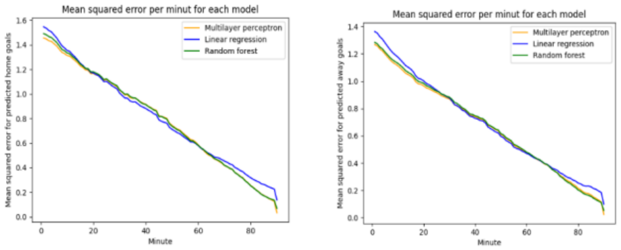

Figure 6: Each model's MSE per minute for the anticipated home and away goals

6. Conclusion and Future Works

This study investigates the prediction of final goals score in a soccer match for both home and away teams. It uses both fixed match factors and changing features during the game. MLP made errors of 0.658 goals for home teams and 0.599 goals for away teams with an overall prediction accuracy of 63.98%. MLP model performed better than other ML models like linear regression and random forest. However, substituting players had a negative impact on goals possibly because it helps to reduce player fatigue. Changing to a more defensive lineup increase goals for both teams. While switching to a more attacking lineup resulted in fewer goals. It gives coaches new ways to use data to make better decisions during games. However, there are some limitations such as missing in - game data and not testing across different seasons or leagues which should be looked at in future research.

Declarations

a) ***Funding:*** No funds, grants, or other support was received.
b) ***Conflict of Interest:*** The authors declare that they have no known competing for financial interests or personal relationships that could have appeared to influence the work reported in this paper.
c) ***Data Availability:*** Data will be made on reasonable request.
d) ***Code Availability:*** *Code will be made on reasonable request.*

References

[1] M. Kanojia, P. Kamani, G. S. Kashyap, S. Naz, S. Wazir, and A. Chauhan, "Alternative Agriculture Land - Use Transformation Pathways by Partial - Equilibrium Agricultural Sector Model: A Mathematical Approach, " Aug.2023, Accessed: Sep.16, 2023. [Online]. Available: https: //arxiv. org/abs/2308.11632v1

[2] H. Habib, G. S. Kashyap, N. Tabassum, and T. Nafis, "Stock Price Prediction Using Artificial Intelligence Based on LSTM– Deep Learning Model, " in *Artificial Intelligence & Blockchain in Cyber Physical Systems: Technologies & Applications*, CRC Press, 2023, pp.93–99. doi: 10.1201/9781003190301 - 6.

[3] G. S. Kashyap, A. Siddiqui, R. Siddiqui, K. Malik, S. Wazir, and A. E. I. Brownlee, "Prediction of Suicidal Risk Using Machine Learning Models, " Dec.25, 2021. Accessed: Feb.04, 2024. [Online]. Available: https: //papers. ssrn. com/abstract=4709789

[4] N. Marwah, V. K. Singh, G. S. Kashyap, and S. Wazir, "An analysis of the robustness of UAV agriculture field coverage using multi - agent reinforcement learning, " *Int. J. Inf. Technol.*, vol.15, no.4, pp.2317–2327, May 2023, doi: 10.1007/s41870 - 023 - 01264 - 0.

[5] F. Alharbi and G. S. Kashyap, "Empowering Network Security through Advanced Analysis of Malware Samples: Leveraging System Metrics and Network Log Data for Informed Decision - Making, " *Int. J. Networked Distrib. Comput.*, pp.1–15, Jun.2024, doi: 10.1007/s44227 - 024 - 00032 - 1.

[6] G. S. Kashyap, K. Malik, S. Wazir, and R. Khan, "Using Machine Learning to Quantify the Multimedia Risk Due to Fuzzing, " *Multimed. Tools Appl.*, vol.81, no.25, pp.36685–36698, Oct.2022, doi: 10.1007/s11042 - 021 - 11558 - 9.

[7] M. Spann and B. Skiera, "Sports forecasting: A comparison of the forecast accuracy of prediction markets, betting odds and tipsters, " *J. Forecast.*, vol.28, no.1, pp.55–72, Jan.2009, doi: 10.1002/for.1091.

[8] J. Ashar *et al.*, "RoboCup Standard Platform League - rUNSWift 2010", Accessed: Apr.30, 2024. [Online]. Available: http: //www.atlassian. com/

[9] S. Ramamoorthy *et al.*, "Team Edinferno Description Paper for RoboCup 2011 SPL", Accessed: Apr.30, 2024. [Online]. Available: https: //www.researchgate. net/publication/266043176

[10] R. H. Koning and R. Zijm, "Betting market efficiency and prediction in binary choice models, " *Ann. Oper. Res.*, vol.325, no.1, pp.135–148, Jun.2023, doi: 10.1007/s10479 - 022 - 04722 - 3.

[11] P. Stone and M. Veloso, "Multiagent systems: a survey from a machine learning perspective, " *Auton. Robots*, vol.8, no.3, pp.345–383, Jun.2000, doi: 10.1023/A: 1008942012299.

[12] R. Bunker and T. Susnjak, "The Application of Machine Learning Techniques for Predicting Match Results in Team Sport: A Review, " Apr.14, 2022, *AI Access Foundation*. doi: 10.1613/jair.1.13509.

[13] F. Archontakis and E. Osborne, "Playing It Safe? A Fibonacci Strategy for Soccer Betting, " *J. Sports Econom.*, vol.8, no.3, pp.295–308, Jun.2007, doi: 10.1177/1527002506286775.

[14] H. Rue and Ø. Salvesen, "Prediction and retrospective analysis of soccer matches in a league, " *J. R. Stat. Soc. Ser. D Stat.*, vol.49, no.3, pp.399–418, Sep.2000, doi: 10.1111/1467 - 9884.00243.

[15] F. Palomino, L. Renneboog, and C. Zhang, "Information salience, investor sentiment, and stock returns: The case of British soccer betting, " *J. Corp. Financ.*, vol.15, no.3, pp.368–387, Jun.2009, doi: 10.1016/j. jcorpfin.2008.12.001.

[16] R. Stefani, "Improved Least Squares Football, Basketball, and Soccer Predictions, " *IEEE Trans. Syst. Man. Cybern.*, vol.10, no.2, pp.116–123, Jul.2008, doi: 10.1109/tsmc.1980.4308442.

[17] K. Croxson and J. J. Reade, "Information and efficiency: Goal arrival in soccer betting, " *Econ. J.*, vol.124, no.575, pp.62–91, Mar.2014, doi: 10.1111/ecoj.12033.

[18] P. Kaur, G. S. Kashyap, A. Kumar, M. T. Nafis, S. Kumar, and V. Shokeen, "From Text to

Transformation: A Comprehensive Review of Large Language Models' Versatility, " Feb.2024, Accessed: Mar.21, 2024. [Online]. Available: https: //arxiv. org/abs/2402.16142v1

[19] S. Naz and G. S. Kashyap, "Enhancing the predictive capability of a mathematical model for pseudomonas aeruginosa through artificial neural networks, " *Int. J. Inf. Technol.2024*, pp.1–10, Feb.2024, doi: 10.1007/S41870 - 023 - 01721 - W.

[20] G. S. Kashyap *et al.*, "Detection of a facemask in real - time using deep learning methods: Prevention of Covid 19, " Jan.2024, Accessed: Feb.04, 2024. [Online]. Available: https: //arxiv. org/abs/2401.15675v1

[21] F. Alharbi, G. S. Kashyap, and B. A. Allehyani, "Automated Ruleset Generation for 'HTTPS Everywhere': Challenges, Implementation, and Insights, " *Int. J. Inf. Secur. Priv.*, vol.18, no.1, pp.1–14, Jan.2024, doi: 10.4018/IJISP.347330.

[22] S. Wazir, G. S. Kashyap, and P. Saxena, "MLOps: A Review, " Aug.2023, Accessed: Sep.16, 2023. [Online]. Available: https: //arxiv. org/abs/2308.10908v1

[23] G. S. Kashyap *et al.*, "Revolutionizing Agriculture: A Comprehensive Review of Artificial Intelligence Techniques in Farming, " Feb.2024, doi: 10.21203/RS.3. RS - 3984385/V1.

[24] S. Wazir, G. S. Kashyap, K. Malik, and A. E. I. Brownlee, "Predicting the Infection Level of COVID - 19 Virus Using Normal Distribution - Based Approximation Model and PSO, " Springer, Cham, 2023, pp.75–91. doi: 10.1007/978 - 3 - 031 - 33183 - 1_5.