Combinations Of Features For Predicting Soccer Results

Hung Tran
Department of Computer Science
University of Manitoba
Winnipeg, Canada
trannh@myumanitoba.ca

Duc Nghia Do
Department of Computer Science
University of Manitoba
Winnipeg, Canada
dodn@myumanitoba.ca

Fidelio Ciandy
Department of Computer Science
University of Manitoba
Winnipeg, Canada
ciandyf@myumanitoba.ca

Jay Nguyen
Department of Computer Science
University of Manitoba
Winnipeg, Canada
nguyen89@myumanitoba.ca

***Abstract*— In modern days, predicting sports outcomes has gained significant attention from researchers and spectators all around the world. Accurate game result prediction is important not only for competitive purposes but also for improving fan engagement and supporting decision-making for sports teams. Football, also known as soccer - one of the most popular sports in the world- contains a lot of data that is available to us. In this paper, we present a comparative analysis of different combinations of features that are essential for soccer match result prediction. Using data from English Premier League seasons, we will compare and analyze the results to find out which features individually, and which combinations of features have the most impact on the outcome of a game. The findings aim to advance the understanding of sports analytics and provide insights into practical applications in the football domain.**

Keywords—European football, English premier league, sport prediction, sport data mining, soccer data mining, sliding window algorithms, random forest, gradient boosting

I. INTRODUCTION

Data mining refers to the process of extracting potentially useful information and knowledge hidden in a large amount of incomplete, noisy, fuzzy, and random practical application data [1]. In the sports field, data mining applications span across various activities, such as performance assessment, player injury prediction, sports talent identification or game strategy evaluation [3].

These applications improve both fan experience and decision-making for teams, coaches, and analysts. However, one of the most interesting areas of sports data mining is the prediction of game outcomes. This is a challenging task as there are many factors that contribute to the result of a game.

Traditionally, sports predictions were made by human experts, such as commentators, former players, and coaches, who based their predictions on experience, intuition, and subjective reasoning [2]. While expert opinions can be valuable, their predictions often lack consistency and are mostly influenced by personal bias [2]. Because of this subjectivity, quantitative prediction models based on historical results have been created. These models can result in a more analytical approach with no bias to sports prediction by examining team performance over time and combining it with their achievement [2]. Unlike human predictions, data mining models can process a large amount of data to recognize patterns and trends that might not be clear for humans to find out, resulting in more consistent and accurate predictions.

In this paper, we focus on soccer, one of the most widely followed sports globally, and explore the impact of different key features on match result prediction. The features were selected based on their relevance to soccer analytics and their potential influence on game outcomes. Using data from past Premier League seasons, we evaluate the significance of each feature in determining match results, such as goal score, shot on target, number of red cards, etc. Our aim is to identify the most impactful factors as well as the best combinations of factors, providing insights into their contribution to soccer game prediction.

The remainder of this paper is organized as follows. The next section provides background information on soccer and an overview of related work in sports analytics. We then describe our methodology for analyzing the selected features and evaluating their impact individually on match predictions. After that, we find out the best combinations of features, present and discuss our findings, highlighting the relative importance of each feature. Finally, we conclude with the implications of our study and suggest directions for future research in sports analytics.

II. BACKGROUND

As this paper focuses on mining and predicting results for soccer games, we will provide some background knowledge about the sport, specifically in the professional scene. Soccer, also known as football outside of North America, is a team sport played between two teams of eleven players. There is a goalpost at both ends of a rectangle field, and the objective of each team is to score more goals than the opposing team. Soccer games have two halves of 45 minutes, with a short break in between and extra time if needed. The sport's popularity has resulted in many datasets of leagues, tournaments, player statistics, and match outcomes, making it an ideal field for data mining. In this paper, we will be focusing on the English Premier League (EPL) - one of the most famous leagues in the world. In this tournament, 20 teams are facing each other twice per season, once at home and once away, creating a dataset of 380 matches each season. This EPL dataset provides us with valuable information for predicting outcomes, as it includes detailed statistics on match results and team performance metrics.

In soccer, team roles and responsibilities are divided as follows:

1. Offensive Players: Primarily focused on moving the ball into the opponent's net, these players include forwards and attacking midfielders.

2. Midfielders: These players play in the midfield, support both offence and defense, distributing the ball, controlling possession, and connecting the defense and attack.

3. Defensive Players: Including defenders and the goalkeeper, they focus on preventing the opposing team from scoring.

Furthermore, soccer data contains multiple statistics relevant to these roles, such as team points, shots, goals scored, goals conceded, fouls, and formations, as well as player-level data. Additionally, team dynamics, recent form, league-specific factors, and external elements (e.g., weather, injuries, and venue) also play an important role in determining match outcomes.

In this paper, we utilize EPL match data to evaluate the impact of key features, such as goals scored, goals conceded, head-to-head results, shots on target, red cards, and more, on match outcome predictions, specifically, wins or losses. By analyzing these features, we first aim to determine their individual contributions to match results and compare them to see which has the most significant influence on a match result. Then we will combine them and find out which combination of features will have the best prediction accuracy. Through this approach, we aim to identify the best features suited for analyzing and predicting results in soccer - an unpredictable sport with many variables..

III. Related Work

There have been many related works about using data mining and machine learning to predict outcomes for sports with different approaches and purposes [5], [6], [7].

Nicholas Utikal [5] used data from 2018 and 2019 for teams across various leagues, including those from England, Spain, Argentina, China, and others. The dataset included information such as the date, HomeTeam, AwayTeam, FTHG (Full-Time Home Team Goals), FTAG (Full-Time Away Team Goals), HST (Home Team Shots on Target), and AST (Away Team Shots on Target). First, he calculated AVGHTGDIFF (Average Home Team Goal Difference) and AVGFTHG (Average Full-Time Home Goals) by dividing the total by the last 10 games to get per-game values. Then, he evaluated the importance of each feature. He improved the accuracy using Random Forest optimization through the Random Search algorithm. The final accuracy achieved was 23% for predictions per goal and 51% for predicting a game's outcome (win, draw, or loss). For this paper, the number of attributes was 25 and the author did find each feature's importance. However, his attributes were not divided equally into different features and in a fixed pattern, for example, the HST feature has HST_1 to HST_10, while HTGDIFF has HTGDIFF_3 to HTGDIFF_8. Moreover, he did not consider how different combinations of features impact the final result, which is one of the main focuses of our paper.

Darwin Prasetio and Dra. Harlili [6] used the results of 5 EPL seasons from the 2010-2011 season to the 2014-2015 season as training data and the results of EPL seasons 2015-2016 as testing data. They used 4 attributes for their research: Home Offense, Home Defense, Away Offense, and Away Defense and obtained match records from football-data.co.uk and team strengths from sofifa.com. Then, they examined the impact of historical seasons on prediction accuracy using different combinations of training data and the Newton-Raphson algorithm to estimate regression coefficients to predict the win/loss for each match. Finally, they found an accuracy of 69.513% for the training record 2010-2015, 69.163% for the training record 2011-2015, 68.376% for the training record 2010-2016 and 68.005% for the training record 2011-2016. For this paper, the authors only considered a very small number of attributes. Moreover, they did not state a specific reason for choosing those 4 attributes and consider window size for their training data as well.

Bing Shen Choi, Lee Kien Foo and Sook-Ling Chua [7] used the data comprising the EPL matches spanning 10 seasons, starting from 2012-2013 to 2021-2022. Their goal was to train 4 models: random forest (RF), logistic regression (LR), Linear Support Vector Classifier, and XGBoost and compare their accuracy with and without stratification. To achieve this, they used 2 methods, which are correlation analysis to remove highly correlated features to avoid redundancy and the BORUTA Algorithm to identify features with significant contributions. They ended up with features: referee, shots, corners, points, UnbStreak, LastSeasonRank, Venme and PromotedMatchup. Then, they grouped draw and loss and "Non-Win" for simplicity and used stratified sampling to maintain class proportions and balanced sampling to ensure equal representation of each class. In the end, they came to the result of different accuracies for those models and types of sampling, with RF combined with balanced sampling and LR combined with balanced sampling having the highest accuracy of 0.666. For this paper, the authors had 8 features for their training data and did not consider different combinations of features as well. Moreover, they had a window size of 5, where they took the average for the match statistics. There was no statistical reason why they chose that window size and they did not consider the different number of window sizes, which resulted in the limitation that they encountered in their research that the constructed models were not able to make predictions for the first 5 matches of the season. In our approach, we took the number of window sizes into account and found what was the best window size statistically, so that we could avoid that limitation and improve the overall work further.

IV. Our Methology

In this section, we outline our sports data mining approach, focusing on two primary objectives. First, we aim to identify the attributes that contribute most significantly to predicting the outcome of a match. Second, we will determine the combination of features that yields the highest accuracy in predicting match outcomes. To achieve these goals, we followed structured methods that followed these major steps: ***(1) data collection***, ***(2) data preparation, (3) feature contribution analysis, and (4) feature combination analysis***. Figure 1 gives a more detailed step-by-step explanation in the form of a flow chart.

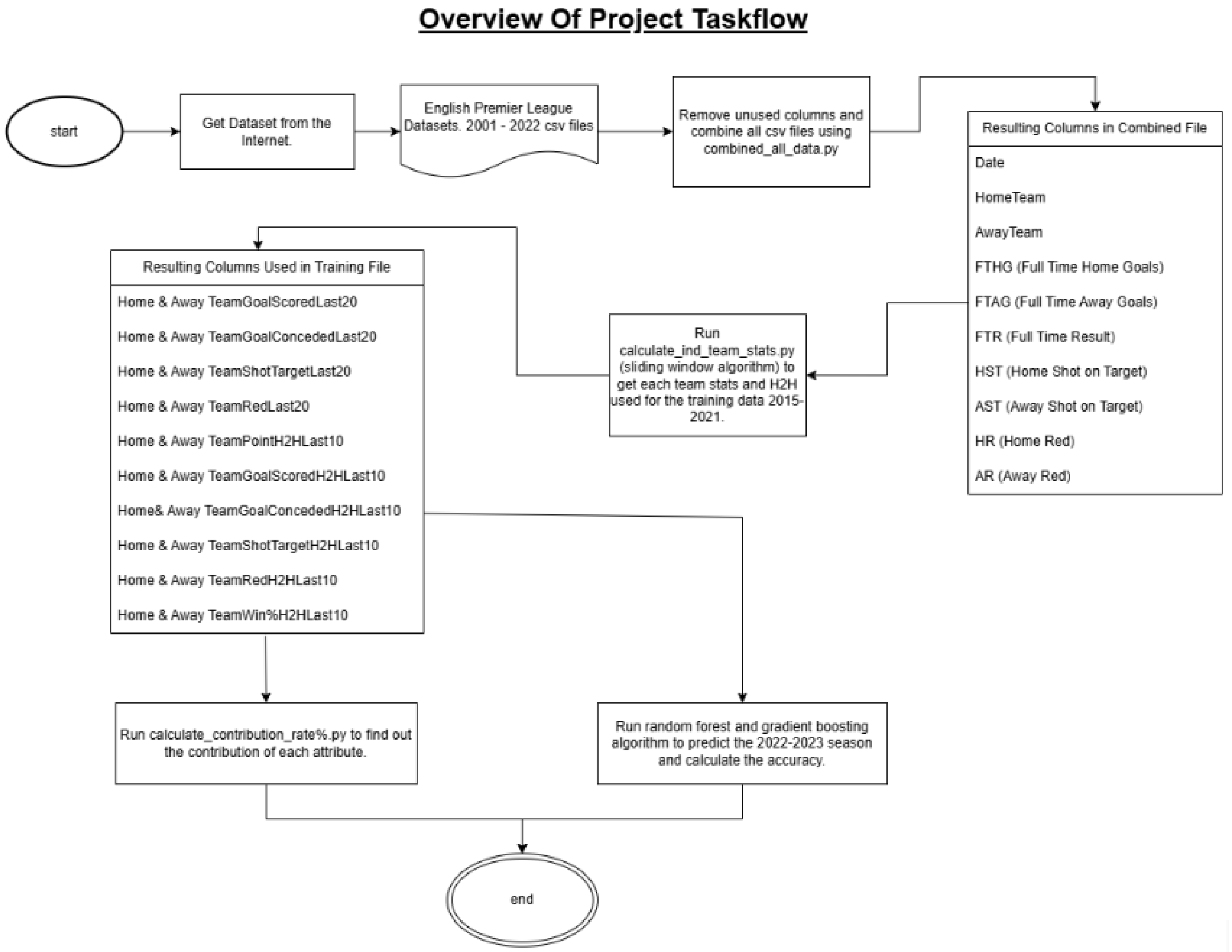

Fig. 1. Overview of Project Task Flow

4.1 Data Collection

We utilized the English Premier League dataset spanning the 2015/2016 to 2020/2021 seasons [https://www.kaggle.com/datasets/saife245/english-premier-league]. The English Premier League is chosen particularly for this study due to its status as one of the most popular and widely watched football leagues in the world. Additionally, the dataset is comprehensive and well-suited for our project, containing thousands of recorded games from 2000 to 2022.

This dataset not only includes a large volume of matches, but it also provides detailed statistics for each game, such as full-time scores, half-time scores, fouls, and other metrics. Thus, these attributes make it ideal for conducting extensive analysis and achieving our goals for this study.

4.2 Data Preparation

4.2.1 Data Cleaning & Data Generation

The initial dataset contains numerous columns, including full-time result statistics, half-time result statistics, betting odds, and more. However, since our primary objective is to analyze and predict the full-time result of a match before it begins, we focused solely on the columns related to full-time statistics. Additionally, the initial dataset is divided into separate CSV files for each season, ranging from 2000 to 2022. To simplify our analysis, we decided to combine all these files into a single dataset. After cleaning the columns and merging the files, we finalized a set of columns, as shown in the picture below.

Fig. 2. Combined File Columns

Furthermore, since one of the primary objectives of this study is to predict match outcomes before the season starts, using match-specific statistics is not feasible, as these matches have not yet occurred. Instead, we modify the dataset to include historical statistics up to the point of prediction, focusing on the recent performance of teams. Therefore, we decided to use the data ranging from the season 2015/2016 to 2020/2021 as the older data might not give us relevant performance of teams since their strength might change over a long span of time.

To prepare the data, we apply a sliding window algorithm to aggregate features from prior matches using the columns from Figure 2. Specifically, for each team, we compute features based on their last 20 matches, such as Points, Goals Scored, Goals Conceded, Red Cards, and Shots on Target. Additionally, for head-to-head (H2H) statistics, which capture the historical performance between the two competing teams, we use a smaller window of 10 recent matchups. These features include Points H2H, Goals H2H, Goals Conceded H2H, Red Cards H2H, and Shots on Target H2H.

The columns generated, as shown in Figure 3, will be used in our analysis to identify the attributes with the highest contribution to match outcomes. Additionally, these columns will serve as inputs for training our model, allowing us to determine the combinations of features that yield the highest accuracy in predicting match results.

Resulting Columns Used in Training File
Home & Away TeamGoalScoredLast20
Home & Away TeamGoalConcededLast20
Home & Away TeamShotTargetLast20
Home & Away TeamRedLast20
Home & Away TeamPointH2HLast10
Home & Away TeamGoalScoredH2HLast10
Home& Away TeamGoalConcededH2HLast10
Home & Away TeamShotTargetH2HLast10
Home & Away TeamRedH2HLast10
Home & Away TeamWin%H2HLast10

Fig. 3. Training Data Columns

4.2.2 Features Explanation

The features are categorized into two types:

Performance-Based Features

These features summarize the recent overall performance of a team across their last 20 matches:

- Points: Total points earned, based on wins (+3), draws (+1), and losses (+0).
- Goals Scored: The number of goals the team has scored.
- Goals Conceded: The number of goals the team has allowed opponents to score.
- Shots on Target: The total number of accurate shots directed towards the opponent's goal.
- Red Cards: The number of red cards issued to the team. When a player gets a red card, they are suspended for the rest of the game.

Head-to-Head (H2H) Features

These features capture the recent historical performance of both teams against each other in their last 10 direct matchups:

- Points H2H: Total points earned in the last 10 encounters with the opposing team, based on wins (+3), draws (+1), and losses (+0).
- Goals Scored H2H: Goals scored against the opposing team in their direct matchups.
- Goals Conceded H2H: Goals conceded to the opposing team in their direct matchups.
- Shots on Target H2H: Accurate shots directed at the opposing team's goal in head-to-head matches.
- Red Cards H2H: The number of red cards received in matchups with the opposing team. When a player gets a red card, they are suspended for the rest of the game

4.2.3 Window Size Selection

The selection of window sizes was decided by using both related works and experimental results. Specifically, we referenced insights from Nicholas Utikal [5] and conducted experiments using Random Forest and Gradient Boosting models to determine the optimal configuration. These experiments were conducted using data from the 2015/2016 to 2020/2021 seasons as training data, while the 2022/2023 season was used for testing. This allowed us to evaluate the models' ability to predict match outcomes, including draws, using the features described earlier.

We tested various combinations of window sizes for normal features and head-to-head (H2H) features. The results, shown in the table below, summarize the accuracy achieved for each combination:

Normal Window size	H2H Window Size	Random Forest	Gradient Boosting
5	5	48.95%	47.89%
10	5	50.00%	49.21%
15	5	52.37%	54.74%
15	10	52.63%	51.84%
20	10	54.47%	52.37%
30	10	52.37%	51.84%

Table. 1. Result of Different Window Sizes

The combination of a 20-match window for general features and a 10-match window for H2H features yields the best accuracy, particularly for Random Forest (54.47%) and Gradient Boosting (52.37%).

4.3 Each Feature Contribution Calculation

After aggregating the data for the 10 key features, we calculate the contribution percentage of each feature to a team's likelihood of winning. This is done using the following formula:

$$\%\text{Win Feature} = \left(\frac{Win\ when\ Feature(teamA)\ better\ than\ Feature(teamB)}{Total\ games}\right) \times 100\%$$

Example 1 (Feature: Points):

We evaluate the number of matches where Team A's Points are greater than Team B's Points in their last 20 performances and resulted in a win for Team A. Using the formula:

$$\%\text{ of winning when points is better} = \left(\frac{Wins\ \ when\ Points(teamA) > Points(teamB)}{Total\ games}\right) \times 100\%$$

Example 2 (Feature: Goals Concede):

We evaluate the number of matches where Team A's Goals Conceded are less than Team B's Goals Conceded in their last 20 performances and resulted in a win for Team A. Using the formula:

$$\text{\% of winning when concede less} = \left(\frac{Wins\ when\ GoalsConcede(teamA) < GoalsConcede(teamB)}{Total\ games}\right) \times 100\%$$

This process is repeated for all 10 features, allowing us to determine which features are the most significant contributors to a team's success.

After calculating the contribution percentages for all features, we ranked them based on their predictive accuracy. Figure 4 shows the rankings taken from the calculated percentages.

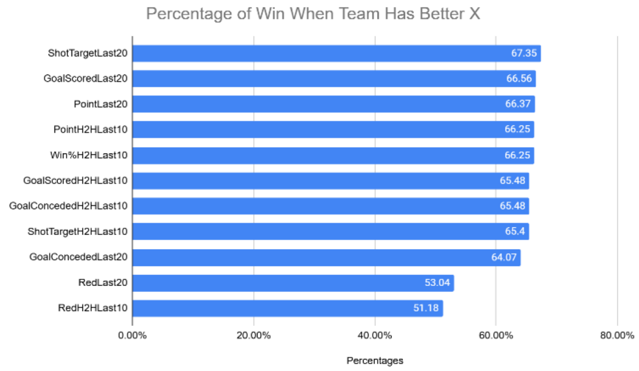

Fig. 4. Individual Feature Contribution Result

4.4 Applying The Algorithms

Using the rankings in Figure 4, we will experiment with different combinations of features to predict the 2022/2023 Premier League season outcomes.

Each combination will be tested with Random Forest and Gradient Boosting models. By analyzing the results, we aim to identify which combination of features and algorithms will have the highest accuracy.

4.4.1 Random Forest

Random Forest is an ensemble learning method that combines multiple decision trees to make predictions that are more accurate and robust than individual trees. It uses a process called bagging (Bootstrap Aggregating), where each decision tree is trained on a different random subset of the data. The final prediction is made by aggregating the results—averaging for regression tasks or majority voting for classification tasks. [4]

4.4.2 Gradient Boosting

Gradient Boosting is another ensemble technique, but it differs from Random Forest in that its decision trees are built sequentially. Each new tree corrects the errors of the previous ones by optimizing a loss function. At every iteration, a new tree is added to minimize the residual errors between the predictions and the actual outcomes. [4]

4.4.3 Why We Use Random Forest/ Gradient Boosting

Both algorithms are well-suited for soccer match outcome predictions due to their ability to handle complex and non-linear relationships, robust performance on diverse datasets, and they are easy to implement. Furthermore, these two algorithms have been proven to perform well in other works such as "Predicting and analyzing football match outcomes with match and players statistics using several machine learning algorithms" [9] where Jasper Remmen compare Random Forest and Gradient Boosting with Decision Trees, and both algorithms outperform Decision Trees. They are also suggested by Xing et al [10] when it comes to Football Match Prediction

4.5 Finding The Best Features/ Combinations

Having determined the individual contributions of each feature and the selected algorithms, the next step is to identify the optimal combination of features and algorithms that yields the best prediction results.

Step 1: Generating Feature Combinations

With a total of 10 features, the number of possible non-empty combinations is 210 -1 or 1023 combinations. These combinations represent all subsets of the feature set, excluding the empty set. Each subset will be evaluated to determine its contribution to the overall prediction accuracy.

Step 2: Testing with Algorithms

Each feature combination will be evaluated using the Random Forest and Gradient Boosting algorithms. For consistency and fair evaluation, we will:

1) Use data from the 2015/2016 to 2020/2021 seasons as the training set.
2) Reserve the 2022/2023 season as the testing set to evaluate the performance of the models with the combination of the features.

Step 3: Using Subsets of Features

For each combination of features:

1) We will use the existing dataset but consider only the features in the selected combination. This does not require creating new datasets; instead, the same dataset is filtered to include only the relevant columns corresponding to the features in the current combination.
2) This filtered subset of the dataset will be used as input to the Random Forest and Gradient Boosting algorithms.

Step 4: Evaluating Results

For each feature combination and algorithm:

1) Model Training: Train the model using the selected features on the training data.
2) Model Testing: Test the trained model on the 2022/2023 season data, recording the accuracy for each combination.
3) Comparison and Ranking: Rank all feature combinations based on their predictive accuracy for each algorithm.

V. RESULT

After running every possible combination of features for the two algorithms: Random Forest and Gradient Boosting, we find out the best 5 features combination of each algorithm.

Random Forest:

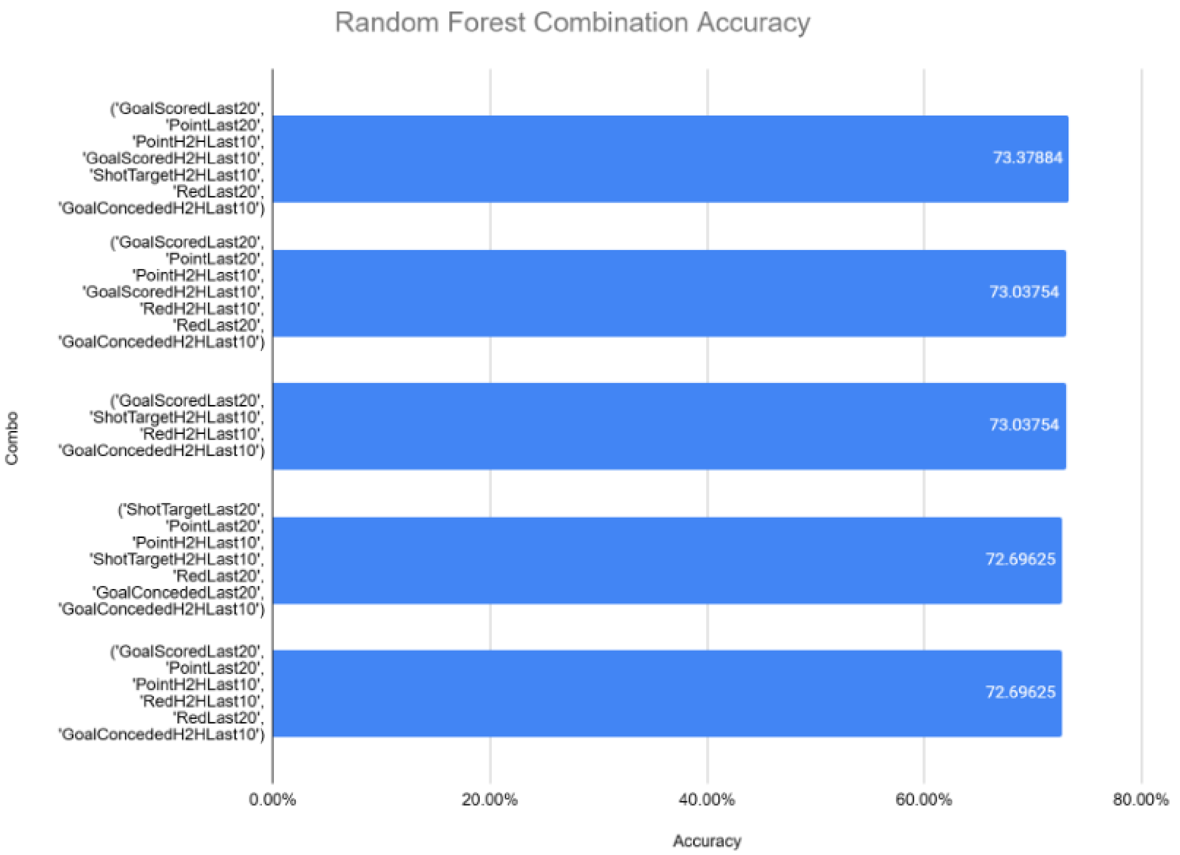

Fig. 5. Random Forest Combination Accuracy Result

For the Random Forest algorithm, as seen in Figure 5, we can see that the best combination that can be used to predict a result of a match with the highest percentage (73.38%) consists of GoalScoredLast20, PointLast20, PointH2HLast10, GoalScoredH2HLast10, ShotTargetH2HLast10, RedLast20, GoalConcededH2HLast10. This indicates that, if these attribute statistics between two teams are available, the Random Forest can be utilized to predict the result of a match with an accuracy of approximately 73.38%.

The difference between the highest-performing combination and fifth-best combinations (73.38% vs. 72.70%) is minimal, suggesting that the model is relatively robust across different feature sets. This consistency indicates that the inclusion or exclusion of certain features has only a minor impact on overall prediction accuracy.

Gradient Boosting:

For the Gradient Boosting algorithm, as seen in Figure 6, the best combination that can be used to predict the outcome of a match consists of PointLast20, PointH2HLast10, RedH2HLast10, and RedLast20 with the highest percentage (72.35%). Notably, the Gradient Boosting algorithm requires few features, utilizing only four attributes in its top-performing combination compared to seven attributes in the top-performing combination for the Random Forest algorithm.

The difference in accuracy between the best-performing combinations of the two algorithms is minimal, with Random Forest achieving 73.38% and Gradient Boosting achieving 72.35%, representing a margin of approximately 1%. This suggests that, in situations where all seven attributes necessary for Random Forest are unavailable, Gradient Boosting can serve as an alternative, which also offers relatively comparable performance with fewer features required.

Comparing Results:

In the article “Predicting football match results with logistic regression” by Darwin Prasetio and Dra. Harlili built their model using variations of training data from the 2010/2011 season until 2015/2016. Similarly, our research utilizes data over the span of six years from 2015/2016 to 2020/2021. Both studies exclude the prediction of draws. Therefore, this is a fair comparison.

Algorithm	Accuracy
Random Forest	73.37%
Gradient Boosting	72.35%
Choi et al approach	69.51%

Table. 2. Accuracy Comparison Table

Prasetio and Harlili’s model uses only 4 attributes - Home Offense, Home Defense, Away Offense, and Away Defense - and achieves a maximum prediction accuracy of 69.51% []. In contrast, our study demonstrated better performance, with the

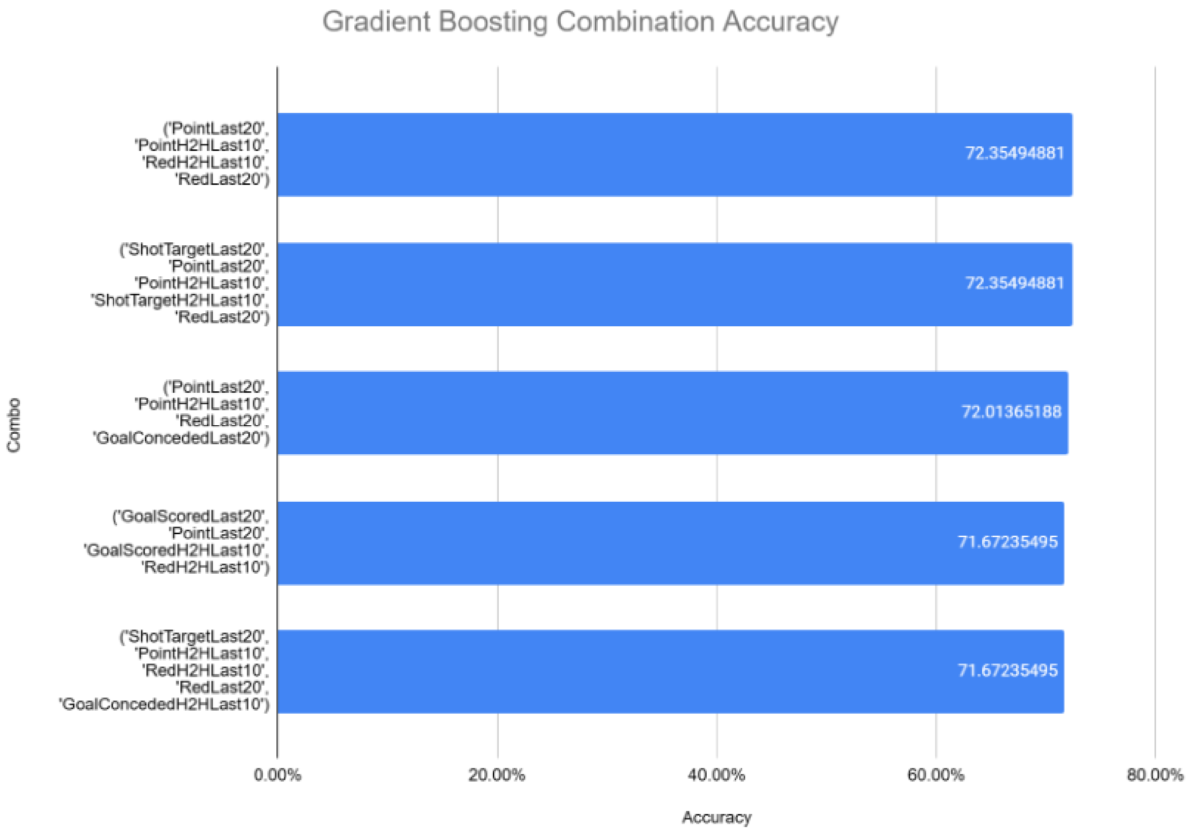

Fig. 6. Gradient Boosting Combination Accuracy Result

Random Forest algorithm achieving an accuracy of 73.38% and the Gradient Boosting algorithm achieving 72.35%.

This comparison indicates that the feature combinations utilized in our research yield more accurate predictions, underscoring the effectiveness of our approach

VI. Conclusions

In this paper, we conducted sports data mining, in particular, we focused on football. Specifically, we designed and implemented a sports data mining algorithm to find the best feature combinations for predicting soccer game wins and losses. The algorithm takes into account features like shot on target, goal scored, point, goal conceded, and red cards of both historical matches in general and historical matches against a specific opponent (i.e. head-to-head matches). We evaluate our algorithm by using historical data from the Premier League (2015/2016 to 2020/2021) to train the model then use the model to predict the results for the Premier League dataset for season 2022/2023. Comparisons of the predictions against the actual football match results demonstrate the practicality of our sports data mining algorithm.

As ongoing and future work, we would incorporate predicting draw matches, and weight corresponding to recent matches. Moreover, we would also come up with our own algorithm to see if the prediction will be better than Random Forest and Gradient Boosting.

References

[1] W. T. Wu, Y. J. Li, A. Z. Feng, et al., "Data mining in clinical big data: the frequently used databases, steps, and methodological models," Military Medicine Research, vol. 8, no. 44, 2021. [Online]. Available: https://doi.org/10.1186/s40779-021-00338-z

[2] C. K. Leung and K. W. Joseph, "Sports Data Mining: Predicting Results for the College Football Games," Procedia Computer Science, vol. 35, pp. 710–719, 2014. [Online]. Available: https://doi.org/10.1016/j.procs.2014.08.153

[3] M. Haghighat and H. Rastegari, "A Review of Data Mining Techniques for Result Prediction in Sports," Advances in Computer Science: An International Journal, vol. 2, 2013.

[4] H. Idrees, "Gradient Boosting vs. Random Forest: Which Ensemble Method Should You Use?," Medium. [Online]. Available: https://medium.com/@hassaanidrees7/gradient-boosting-vs-random-forest-which-ensemble-method-should-you-use-9f2ee294d9c6

[5] N. Utikal, "Predict football results with random forest," Medium. [Online]. Available: https://medium.com/@nicholasutikal/predict-football-results-with-random-forest-c3e6f6e2ee58

[6] D. Prasetio and D. Harlili, "Predicting football match results with logistic regression," in 2016 International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA), Penang, Malaysia, 2016, pp. 1–5. doi: 10.1109/ICAICTA.2016.7803111

[7] B. Choi, L.-K. Foo, and S.-L. Chua, "Predicting Football Match Outcomes with Machine Learning Approaches," MENDEL, vol. 29, pp. 229–236, 2023. doi: 10.13164/mendel.2023.2.229

[8] S. Alam, "English Premier League Dataset," Kaggle, [Online]. Available: https://www.kaggle.com/datasets/saife245/english-premier-league.

[9] J. Remmen, "Predicting and Analyzing Football Outcomes with Match and Player Statistics Using Several Machine Learning Algorithms," M.S. thesis, Dept. of Cognitive Science & Artificial Intelligence, Tilburg Univ., Tilburg, Netherlands, Jun. 2022. [Online]. Available: https://arno.uvt.nl/show.cgi?fid=161491

[10] A. C. J. Xing, E. A. H. Shaker, C. Qiyang, Z. A. bin A. Salam, and Y. M. A. Al-Fakih, "Football Match Prediction Using Random Forest Classifier," Journal of Applied Technology and Innovation, vol. 8, no. 1, pp. 25–29, Jan. 2024. [Online]. Available: https://jati.sites.apiit.edu.my/files/2024/01/Volume8_Issue1_Paper05_2024.pdf