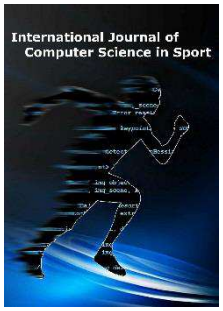
International Journal of Computer Science in Sport

Volume 24, Issue 1, 2025

Journal homepage: http://iacss.org/index.php?id=30

sciendo

DOI: 10.2478/ijcss-2025-0004



A Deep Learning Approach Based on Interpretable Feature Importance for Predicting Sports Results

Messaoud Bendiaf[1], Hakima Khelifi[1], Djamila Mohdeb[1], Mouhoub Belazzoug[1], Abdelhamid Saifi[1]

[1] Department of Computer Science,

University Mohamed El Bachir El Ibrahimi of Bordj Bou Arreridj, Bordj Bou Arreridj, Algeria

Abstract

Football match result prediction is a challenging task that has been the subject of much research. Traditionally, predictions have been made by team managers, fans, and analysts based on their knowledge and experience. However and recently there has been an increased interest in predicting match outcomes using statistical techniques and machine learning. These algorithms can learn from historical data to identify complex relationships between different variables, and then make predictions about the outcome of future matches. Accordingly, forecasting plays a pivotal role in assisting managers and clubs in making well-informed decisions geared toward securing victories in leagues and tournaments. In this paper, we presented an approach, which is generally applicable in all areas of sports, to forecast football match results based on three stages. The first stage involves identifying and collecting the occurred events during a football match. As a multi-class classification problem with three classes, each match can have three possible outcomes. Then, we applied multiple machine learning algorithms to compare the performance of those different models, and choose the one that performs the best. As a final step, this study goes through the critical aspect of model interpretability. We used the SHapley Additive exPlanations (SHAP) method to decipher the feature importance within our best model, focusing on the factors that influence match predictions. Experiment results indicate that the Multilayer Perceptron (MLP), a neural network algorithm, was effective when compared to various other models and produced competitive results with prior works. The MLP model has achieved 0.8342 for accuracy. The particular significance of this study lies in the use of the SHAP method to explain the predictions made by the MLP model. Specifically, by exploiting its graphical representation to illustrate the influence of each feature within our dataset in predicting the outcome of a football match.

KEYWORDS: FOOTBALL, MACHINE LEARNING, MULTILAYER PERCEPTRON, FEATURE IMPORTANCE, SHAPLEY ADDITIVE EXPLANATIONS (SHAP).

Introduction

Football, also known as soccer, is the most popular sport globally, with a massive fan base. As a result, vast amounts of data and statistics are aggregated for major sports tournaments. The growing availability of football-related data has become a focal point of study in disciplines such as statistics and applied economics, sparking significant research interest.

Football is highly stochastic, with unpredictable factors, such as a lucky hit or team dynamics, influencing match outcomes. This makes predicting football match results more complex. Strategies that work in one match may not succeed in future matches, even within the same league. Even strong teams can occasionally lose to weaker opponents. As a result, football managers and analysts increasingly use classification models to develop match-winning strategies. Accurate football match prediction is also invaluable to analysts, managers, and strategists.

A common approach to predicting match results is using historical data, including team and player statistics, past match results, and league standings. Statistical models analyze this data to identify patterns and predict future outcomes. Machine learning algorithms also play a key role in predicting match results. In this paper, we focus on using Multilayer Perceptron (MLP) models to predict football outcomes in the English Premier League (EPL), evaluating data from the 2005-2006 to 2022-2023 seasons. Given the variations in attributes across different leagues, we argue that a prediction model tailored for the EPL is not directly applicable to other football competitions.

The prediction of football match outcomes is an area of growing interest in sports analytics, driven by advancements in machine learning techniques. This study aims to develop and evaluate machine learning models for predicting football match outcomes, specifically focusing on EPL data. The research addresses the following questions: (1) Which machine learning model provides the highest predictive accuracy for football match outcomes? (2) How does the inclusion of match event data, such as Head-to-Head Home Win Rate and Overall Home Win Rate, impact model performance? (3) What is the comparative performance of traditional statistical methods versus modern machine learning algorithms in predicting match results? This research contributes to the growing literature on sports analytics, offering valuable insights for decision-makers and researchers.

The remainder of this article describes our approach to predict football match outcomes. We begin with a brief overview of the field and related work. Next, we present our methodology and the results we achieved. Finally, our paper is concluded by discussing our findings and future improvements.

Related work

Football holds the title of world's most beloved sport and comprises the fastest developing segment in the sports market (Dobson and Goddard, 2011). Many football teams today are worth billions of dollars, because of their success in leagues and tournaments. As a result, several researchers have developed various methods to forecast football results.

When studying match prediction in football, studies can be mainly separated into those that try to predict the match's outcome and those that attempt to predict the final scores (Ren and Susnjak, 2022). Among those that attempt to predict the outcome of a match, we can distinguish between those that predict three classes (win, loss, draw) and those that choose to discard tied matches to simplify the prediction process. However, anticipating results poses a formidable challenge due to the multitude of variables that must be considered, many of which resist precise

quantification or formal modeling (Hucaljuk and Rakipovic, 2011).

In (Bilek and Ulas, 2019), a comprehensive research inquiry was conducted to explore the influential factors that significantly impact the outcome of matches (i.e., win, loss, or draw) in relation to the quality of the opposing teams. This investigation utilized data encompassing situational variables and performance metrics from matches played during the 2017-2018 season of the English Premier League. Various methods, including statistical techniques such as One-way ANOVA and Tukey's Honestly Significant Difference (HSD) test, as well as machine learning approaches like k-means clustering and decision tree algorithms, were employed to conduct the analytical assessments.

There are further studies available in the literature that cover predict matches results in a championship and of other team sports. One such investigation, referenced as (Ulmer and Fernandez, 2013), explored a range of predictive techniques, including baseline methods, Gaussian naïve Bayes, hidden Markov models, multimodal naïve Bayes, Support Vector Machine (SVM), random forest (RF), and One vs All SGD. This study used historical goal-scoring data from ten seasons, in the context of the English Championship. Additionally, (Hucaljuk and Rakipovic, 2011) conducted a comprehensive examination, through scored goals, of result prediction for the UEFA Champions League using different machine learning algorithms. Moreover, (Igiri, 2015) utilized SVM in an investigation centered on scoring patterns within the English Championship.

In the same context, an approach focuses on the analysis of match data to predict the results. This investigation encompassed various combinations of pair-wise match results, including win-draw, win-defeat, and draw-defeat scenarios within multiple championship settings. The suggested system analyzes and defines match results based on a polynomial algorithm, which yielded notable enhancements in accuracy compared to the other machine learning algorithms (Martins et al., 2017).

In a recent study, (Razali et al., 2017) utilized Bayesian methodologies to develop a predictive model based on Bayesian Networks (BNs). Their objective was to forecast the outcomes of football matches within the English Premier League (EPL) throughout three seasons: 2010-2011, 2011-2012, and 2012-2013. The study rigorously assessed the accuracy of their prediction model using K-fold cross-validation. Impressively, the Bayesian Networks demonstrated an average predictive accuracy of 75.09% through these three seasons, marking a substantial achievement in the realm of football match outcome prediction.

Work in (Constantinou, 2019) introduces Dolores, a cutting-edge model engineered for forecasting football match results in a specific country based on checking matches of football in several countries. This work utilizes a mixture of two different techniques that are distinct from each other: (a) dynamic ratings and (b) Hybrid Bayesian Networks. Its creation stems from active participation in the international special issue competition known as Machine Learning for Soccer. Dolores stands apart from conventional academic literature, which typically concentrates on individual leagues or tournaments. In a remarkable departure, Dolores is trained using a unified dataset, incorporating match outcomes, and addressing missing data as part of the competition's challenge. This dataset spans an impressive 52 football leagues from various corners of the globe. The challenge posed to the developers was the daunting task of employing a single model to predict the outcomes of 206 upcoming matches, spanning 26 diverse leagues, all scheduled between March 31 and April 9 in the year 2017.

In another study, the players' attributes were used to predict football match results. Modern data processing techniques, coupled with the formidable computational prowess of modern computers, empower the anticipation of forthcoming match outcomes via a cutting-edge

machine learning algorithms to meticulously gathered data. This study introduces a pioneering approach to preprocess input data, enabling the prediction of match results from input attributes associated with the participating players of the match with limited information about match history itself. This novel model holds the potential to facilitate the selection of ideal players, based on their individual attributes, for a given match or team (Danisik et al., 2018).

Basic neural networks have made significant contributions to various domains, as exemplified by a notable study (Aslan and Inceoglu, 2007). This research prominently highlighted the predictive capabilities of neural networks, particularly their proficiency in employing black-box modeling to forecast football match outcomes. Similarly, our approach also leverages neural networks for football match prediction; however, we enhance the interpretability of the model using Explainable Artificial Intelligence (XAI) techniques such as SHAP, which was not a focus in their study. While Aslan and Inceoglu's work compared neural network performance with conventional statistical methods, our study extends this comparison by incorporating modern machine learning techniques, such as Multilayer Perceptron (MLP) and Logistic Regression, and assessing their performance specifically on English Premier League data. Additionally, our approach introduces the use of recent match event data, including shots and fouls, to further refine predictive accuracy, which distinguishes our work from their more general approach. The novelty of our approach lies in combining state-of-the-art machine learning models with interpretable AI techniques to not only predict outcomes but also explain the factors driving those predictions, thereby offering practical insights for football analysts and coaches.

Explainable AI (XAI) has become increasingly important in sports analytics, particularly for enhancing the transparency and interpretability of machine learning models. In the context of team sports, like football, XAI techniques help in making AI models more understandable, which is critical for ensuring that decisions are trusted by stakeholders, such as team managers and analysts.

Recent research (Procopiou and Piki, 2023) highlights the potential of XAI in the football industry, emphasizing how explainable models can assist both players and clubs. The study discusses how XAI can make AI models more transparent and interpretable, thereby aiding in tactical analysis and other football-related applications. This research underscores the importance of XAI in managing the complexity of football as a multi-layered system, where understanding AI decisions is crucial for effective deployment.

Another study applied explainable machine learning techniques to investigate factors influencing the physical development and performance of young football players. By utilizing SHAP method, the study identified critical anthropometric and hematological parameters that affect player performance. This research demonstrates the effectiveness of XAI methods in predicting athlete performance, and its findings can be incorporated into fitness assessments and regional policy development (Kelum et al., 2024).

In the realm of another team sport, basketball, XAI has also been employed, in (Yuanchen et al., 2022), to analyze the match style and gameplay of the National Basketball Association (NBA). A study utilized Local Interpretable Model-agnostic Explanations (LIME) to interpret the results of supervised-learning AI models, providing insights into the reasoning behind predictions. This application of XAI in basketball showcases its potential to enhance the understanding of AI-driven predictions in sports and can be adapted for use in football as well.

In (Moustakidis et al., 2023), another key study focused on football matches applied XAI techniques to identify performance indicators that contribute to match outcomes. By using SHAP, the researchers were able to pinpoint key team-level variables, such as ball possession and passing behaviors, that significantly impact scoring performance. This study highlights how

XAI can provide valuable insights into team performance, enabling targeted interventions by coaches and analysts to improve outcomes.

Methodology

The following section outlines the methodology employed in our experimental research. It contains an exposition of the dataset employed, along with a comprehensive account of the essential data analysis and preprocessing steps undertaken before model application. After the proper partitioning of our data, we used many proposed models along with a baseline model for comparison. Next, we applied proper metrics to our model outputs to provide results for comparison and analysis.

Data collection

This study focuses on the English Premier League (EPL), the top tier of English professional football, with 20 teams competing for the championship. The Premier League season consists of 38 rounds, in which each team plays each other team once at home and once away. The team with the most points at the end of the season wins the Premier League title. Three points are awarded for a win, one point for a draw, and no points for a loss. If multiple teams finish with the same number of points, their position in the league table is decided by goal difference, then goals scored, then head-to-head record, and finally away goals scored in head-to-head matches. The bottom three teams in the league table at the end of the season are relegated to the Championship, the second tier of English professional football, and are substituted with the three teams promoted from the Championship; the teams that finish in first and second place and the third via the end-of-season playoffs (Premier League).

In our study, we used the Premier League matches data from the Football-Data website (Football-Data). In the pre-processing stage of our dataset, we removed data from the 1993-1994 to 2004-2005 seasons due to a lack of necessary features, ensuring that only 18 seasons from 2005-2006 to 2022-2023 were retained. This selection ensures consistency across the dataset by focusing on seasons with common features required for analysis. Additionally, attributes such as "League Division", "Match Date", and "Match Referee" were excluded, as they were deemed irrelevant for our predictive models. Notably, the dataset is complete, with no missing values, allowing for a robust and clean analysis. We used only 15 essential features that are common in all 18 seasons which are shown in Table 1. We notice that the first three features are about the key to results data, the rest are concerning the match statistics.

Table 1: Description of features.

Feature Name	*Description*
HomeTeam	Home Team
AwayTeam	Away Team
FTR	Full Time Result (H=Home Win, D=Draw, A=Away Win)
HS	Home Team Shots
AS	Away Team Shots
HST	Home Team Shots on Target
AST	Away Team Shots on Target
HC	Home Team Corners
AC	Away Team Corners
HF	Home Team Fouls Committed
AF	Away Team Fouls Committed
HY	Home Team Yellow Cards
AY	Away Team Yellow Cards
HR	Home Team Red Cards
AR	Away Team Red Cards

From 2005 to 2023, a cumulative 42 football teams have taken part in the EPL. These teams have been designated with unique identifier IDs ranging from 1 to 42 (so we can represent 'HomeTeam' and 'AwayTeam' feature values with their ID numbers instead of their names), which correspond to their involvement across consecutive seasons. For teams commencing their participation in the inaugural season, the ID starts at 1, and subsequent teams are assigned sequential IDs. Remarkably, only seven teams have demonstrated unwavering consistency by participating in every single EPL season during this period. The ID values increment to reflect the order in which teams joined or reentered the league, ensuring a systematic and distinct reference for each team's appearance data.

Our dataset contains 6840 matches in which there are 3,144 occurrences of a home team winning, 1,662 draws, and 2,034 occurrences of the away team winning.
Creating relevant and informative features from the existing features (mentioned in Table 1) could help us to increase a model's efficiency and lead to more accurate results in our next experimentations. We are adding two relevant features to our dataset: *Head-to-Head Home Win Rate (HHWR)* and *Overall Home Win Rate (OHWR)*.

- *Head-to-Head Home Win Rate (HHWR)*: Calculate the percentage of home wins for the home team against the specific away team up to the current match day. It is derived by dividing the number of home wins by the total number of matches played between the two teams leading up to the present match day.
- *Overall Home Win Rate (OHWR)*: This feature calculates the percentage of home wins by the home team against all opponents at their home venue up to the current match day. It is calculated by dividing the total number of home wins achieved by the home team in all matches played at home by the total number of home matches the team has participated in up to the current match day.

Teams have distinct playing styles and rivalries. Even if a team is superior in all aspects, it can lose due to a rivalry (e.g., a derby match), incompatible playing styles or clubs tactics. Additionally, teams have psychological and physiological advantages when playing at home, with the support of local fans, while away teams face disadvantages such as switching time zones or climates, and the difficulties encountered during travel. A recent statistical study (Smith, 2017) revealed a steady decline in the home advantage in English football since the Football League's inception, for reasons that remain unclear. Possible explanations include reduced crowd hostility and teams becoming more accustomed to travel.

As a final step of data preprocessing, before building our models, we need to encode the target of our dataset (Full Time Result - FTR) to numeric values for each class: **1** = *Home Win*, **0** = *Draw*, and **-1** = *Away Win*. We finalize the preparation process, and our dataset contains all features as numerical values.

Modeling process

In this section, we will present the general design of our methodology and the choices we have made. The detailed steps of our technical work are presented in Figure 1. We propose to implement and make a comparison between six different classification models:

LR: Logistic Regression,
NB: Naïve Bayes,
SVM: Support Vector Machine,
MLP: Multilayer Perceptron,
RF: Random Forest,
XGB: XGBoost.

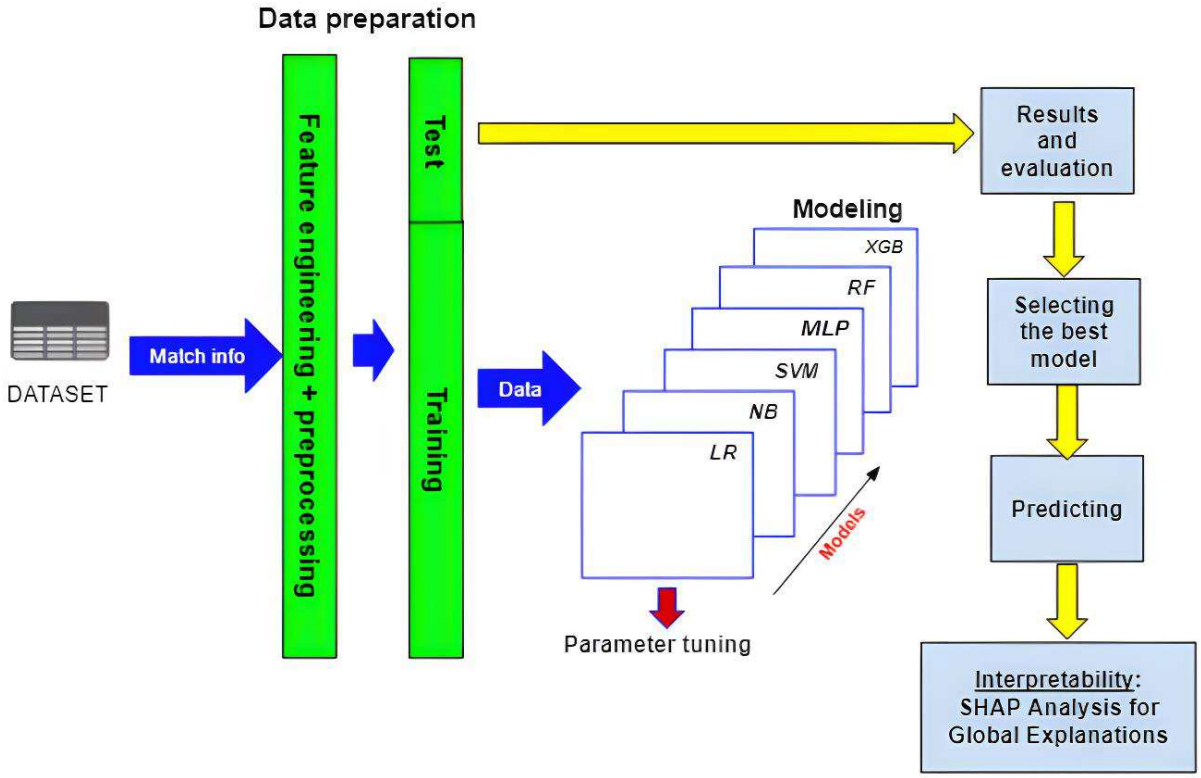


Figure 1. The proposed methodology.

To evaluate these models, we employed a systematic data partitioning approach, ensuring a clear distinction between training, validation, and test sets to maintain temporal consistency and optimize model performance (Montesinos et al., 2022). We split the dataset into training (seasons 1–16), validation (season 17), and test (season 18). Hyperparameter tuning was performed using the validation set, and the model was retrained on the combined training and validation data (seasons 1–17) before being evaluated on the test set. Grid search or similar optimization was applied to all models to ensure fair comparison, and this methodology was consistently followed for each model. The results indicated that the retraining approach, with hyperparameter tuning using a dedicated validation set, resulted in more reliable performance metrics, as shown in Table 2.

The performance observed with this approach can be attributed to the effective utilization of validation data for hyperparameter tuning, which enhanced model generalization. Additionally, ensuring fairness by tuning hyperparameters for all models reduced potential biases in model comparisons.

It is crucial to preserve the chronological order of football match data when forecasting upcoming matches during the season. Standard cross-validation and shuffling techniques are not suitable for this task, as they would disrupt the temporal relationships between matches (Pedregosa et al., 2011). For this reason, we decided to attribute the first 16 seasons of our dataset to the training set, the second-to-last season to the validation set, and the last season to the test set. A fine-tuning process was performed using the grid search method to compute the best parameters for each model, ensuring a fair and systematic evaluation.

Different testing metrics have been used for evaluating the performance of each model: Accuracy, Precision, Recall, F1-score, and Area Under the Curve (AUC). In addition to these metrics, Receiver Operating Characteristic (ROC) graphs, as described in reference (Fawcett, 2006), were employed for classifiers' organization and the visualization of their performance. For the ROC curve analysis, we applied the One-vs-Rest (OvR) strategy across all models.

This approach allowed us to calculate ROC curves for each class by converting the multiclass problem into multiple binary classification tasks, which is a standard technique in multiclass ROC analysis. The use of the OvR strategy does not affect the training process but is employed for effective evaluation of the model's performance. ROC graphs can provide a richer measure of classification performance. This is a measure of the algorithm's ability to distinguish between the three classes (Home Win, Draw, and Away Win).
ROC curves are conventionally characterized by positioning the True Positive Rate (TPR) along the vertical axis, while the False Positive Rate (FPR) is represented on the horizontal axis. Consequently, the upper-left extremity of the plot signifies the theoretical optimum, characterized by an FPR of zero and a TPR of one. While this ideal scenario is rarely attainable in practice, a greater Area Under the Curve (AUC) typically signifies superior model performance. The "steepness" of ROC curves is also essential, since it is optimal to maximize the TPR (Sensitivity) and minimize the FPR (1 - Specificity). In our case, multiclass classification, a notion of TPR or FPR is attained only after binarizing the output (Pedregosa *et al.*, 2011). Multiclass ROC curves can be computed using two different strategies: One-vs-Rest (OvR) and One-vs-One (OvO).

- OvR: For each class, the classifier is trained to discern that class from the collective set of other classes. This results in a separate ROC curve for each class.
- OvO: For every pair of classes, the classifier is trained to distinguish between those two classes. This results in a separate ROC curve for each pair of classes.

We adopt the OvR strategy in this work, which is more computationally efficient than OvO and has been shown to be effective in practice.

After that, and based on the evaluation metrics, we can select the best (most accurate and reliable) model by comparing the performance of our different models. The final step is to interpret global feature importance of the best algorithm. This shows which features are most important for the algorithm to make accurate predictions. To do so, we conducted a SHapley Additive exPlanations (SHAP) analysis for model interpretation.

SHAP (Lundberg and Lee, 2017), a game-theoretic approach to machine learning model interpretation based on the Shapley value (Winter, 2002), connects optimal credit allocation with local explanations. SHAP has been integrated as a tool for visualization that scores each feature regarding its effect on the final outcome.

Results and discussion

After training, the performances of our models were tested using a predefined test dataset. The last season 2022-2023 of the EPL in our dataset was used for testing, which means predicting 380 matches played in this season. The Table 2 below shows models performance. The Accuracy specified in the table is the overall accuracy for Home Win, Away Win and Draw classifications, as well as for other performance indicators.

Table 2: Performance comparison of the six models in our study.

Classifiers	Accuracy	Precision	Recall	F1-score	AUC
Logistic Regression	0.64	0.62	0.64	0.63	0.81
Naïve Bayes	0.60	0.60	0.60	0.60	0.75
Support Vector Machine	0.64	0.63	0.64	0.64	0.82
Multilayer Perceptron	**0.67**	0.66	0.67	0.67	**0.83**
Random Forest	0.64	0.62	0.64	0.62	0.81
XGBoost	0.66	0.63	0.66	0.64	0.82

The performance comparison results in Table 2 demonstrate that the Multilayer Perceptron (MLP) model outperforms the other classifiers in terms of accuracy, achieving an accuracy of 0.67, followed closely by XGBoost at 0.66. MLP also shows the highest precision (0.66), recall (0.67), F1-score (0.67), and AUC (0.83), highlighting its ability to handle class imbalance effectively. This indicates that the MLP model is particularly proficient in correctly identifying all class categories, balancing precision and recall more efficiently than the other models.

With an accuracy of 0.64, the Logistic Regression, Support Vector Machine (SVM), and Random Forest models exhibit similar performance, with precision, recall, and F1-scores slightly lower than MLP. Logistic Regression achieves precision and recall values of 0.62 and 0.64, respectively, while SVM and Random Forest both have precision and recall values around 0.62 to 0.64. While these models perform well, their lower precision and recall compared to the MLP suggest they are less effective in addressing class imbalance.

Naïve Bayes, with an accuracy of 0.60, shows the lowest performance across all metrics. Its precision, recall, F1-score, and AUC values are the smallest among all models, indicating its relative difficulty in effectively classifying the data.
The AUC values for all classifiers are above 0.80, indicating that all models exhibit a good level of discriminative ability. However, the MLP stands out with the highest AUC (0.83), suggesting a superior ability to distinguish between the different match outcomes (Home Win, Draw, Away Win) across all classes.

These results highlight the importance of considering precision, recall, F1-score, and AUC when evaluating model performance, particularly for imbalanced datasets. The MLP's overall superior performance reflects its ability to manage the trade-offs between these metrics and handle class imbalance more effectively than the other models.

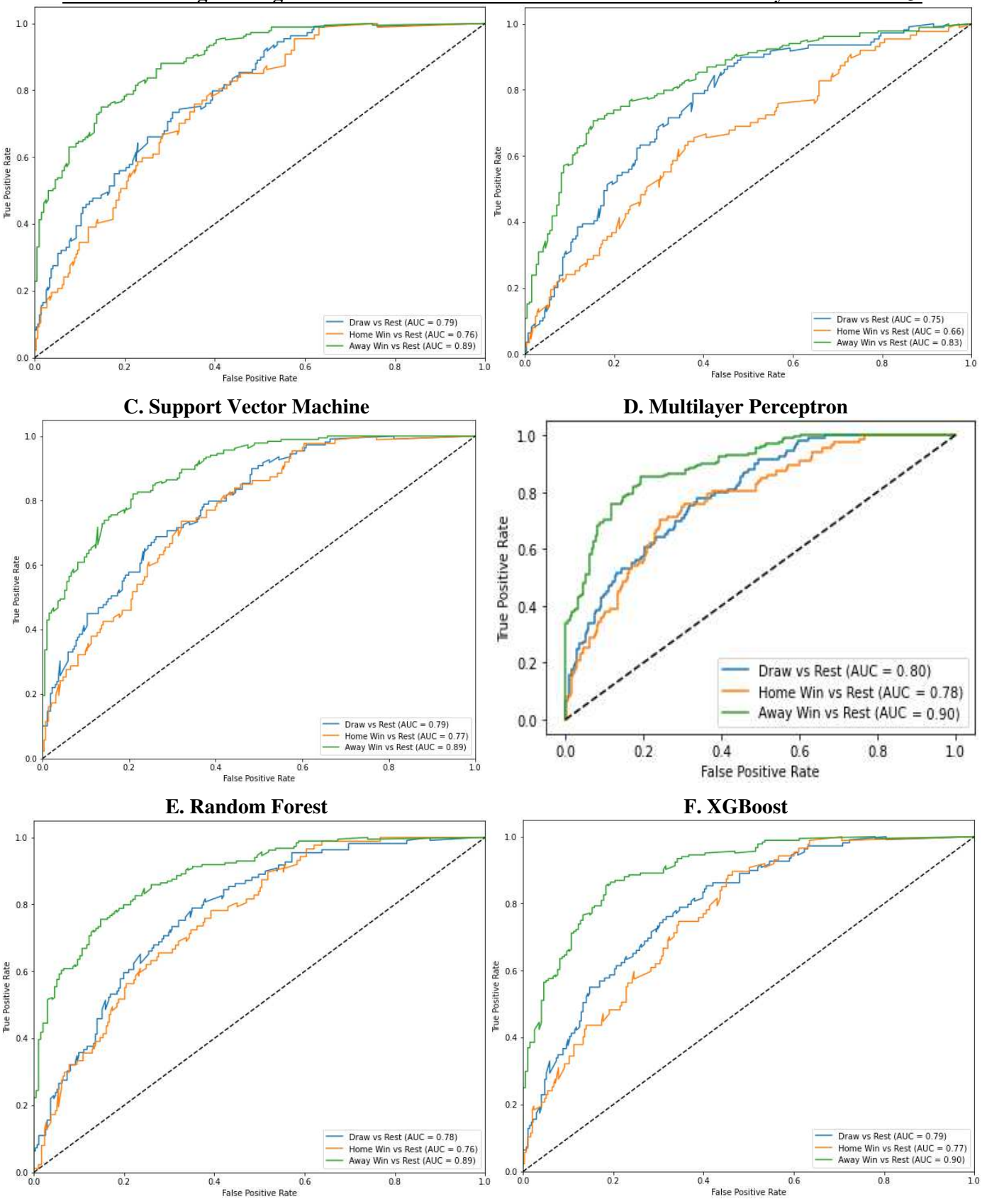


Figure 2. Receiver Operating Characteristic (ROC) curves of the six models.

ROC curves can help us understand the performance characteristics for each class individually. For each class in our multi-class prediction problem, we can create a ROC curve. Each curve represents the performance of the classifier for a specific class. As shown in Figure 2, which presents Receiver Operating Characteristic (ROC) curves for the six machine learning models, each evaluating their performance on three key outcomes: *Draw vs Rest, Home Win vs Rest,* and *Away Win vs Rest*. The Area Under the Curve (AUC) is a metric that reflects a model's ability to distinguish between classes, with a higher AUC (closer to 1) indicating better performance. In the sub-figures, Logistic Regression shows solid performance, particularly for "Away Win vs Rest" with an AUC of 0.89. Naïve Bayes performs poorly on "Home Win vs Rest" with an AUC of 0.66, reflecting its difficulty with this outcome. Support Vector Machine (SVM) mirrors Logistic Regression, excelling in "Away Win vs Rest" with a similar AUC of 0.89. Multilayer Perceptron (MLP) outperforms all models, with an AUC of 0.90 for "Away Win vs Rest" and strong scores across all outcomes. Random Forest performs well, particularly for "Away Win vs Rest", with a consistent AUC of 0.89. XGBoost also shows strong performance, matching MLP with an AUC of 0.90 for "Away Win vs Rest." Overall, MLP and XGBoost demonstrate the best performance, especially for "Away Win vs Rest." Across all models, predicting "Away Win vs Rest" is easier than predicting "Draw vs Rest" or "Home Win vs Rest", indicating the challenge of class imbalance. "Home Win vs Rest" consistently exhibits the lowest AUC, suggesting it is the most difficult outcome to predict accurately.

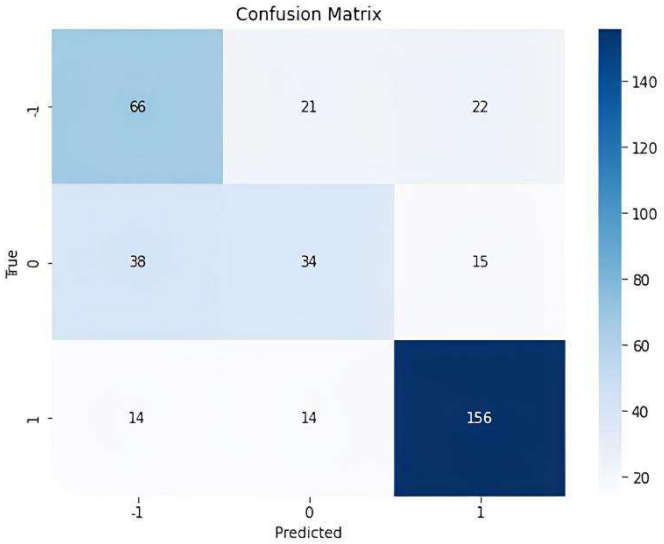


Figure 3. Confusion matrix of the MLP model with values -1 (Away Win), 0 (Draw), and 1 (Home Win).

However, when analyzing the confusion matrix of the MLP model in Figure 3, it reveals that the model performs well overall, as indicated by the True Positives (diagonal elements). For instance, it correctly predicts 66 Away Wins as Away Wins, 34 Draws as Draws, and 156 Home Wins as Home Wins. However, there are some misclassifications. Away Wins are sometimes incorrectly predicted as Draws (21 times) or Home Wins (22 times). Draws are often misclassified as Away Wins (38 times) or Home Wins (15 times), and Home Wins are mistakenly predicted as Away Wins (14 times) or Draws (14 times). Despite an overall accuracy of 67%, the model struggles to differentiate between Away Wins, Draws, and Home Wins, with particular difficulty in distinguishing between Draws and Away Wins.

In the confusion matrix above, the model correctly predicts the Home Win outcomes most frequently, likely due to the strong influence of the features added to the dataset, such as Head-to-Head Home Win Rate (HHWR) and Overall Home Win Rate (OHWR). These features help

the model effectively capture patterns that favor home teams, which is reflected in the model's ability to predict Home Wins accurately.

The Away Win class is predicted reasonably well, with 66 correct predictions out of 109, reflecting a relatively high accuracy for this class. This strong performance can be attributed to the influence of features such as Head-to-Head Home Win Rate (HHWR) and Overall Home Win Rate (OHWR), which, while primarily capturing patterns favoring the home team, also indirectly highlight situations where the away team overcomes the home advantage.

The Draw class is the worst-performing class (only 34 out of 87 actual Draws are correctly predicted), it is often confused with Away Wins, as shown in the confusion matrix, where Draws are misclassified as Away Wins 38 times. This misclassification suggests that the model struggles more with identifying Draws, possibly because a Draw represents a more complex outcome that doesn't favor either team strongly. Overall, this study highlights the difficulty of predicting Draws, as they sit in between Home Wins and Away Wins, reflecting the unpredictable and intermediate nature of such outcomes in football.

Model interpretability

SHapley Additive exPlanations

In this section, we selected the MLP model with the highest mean AUC score. To enhance the interpretability of our best classification algorithm, we performed a SHAP analysis and visualized the feature importance results. SHAP is a technique that calculates the average contribution of each feature value to each match's classification. This provides us with insights into the magnitude and direction of each feature's contribution to the complete probability of a Home Win, Draw, or Away Win.

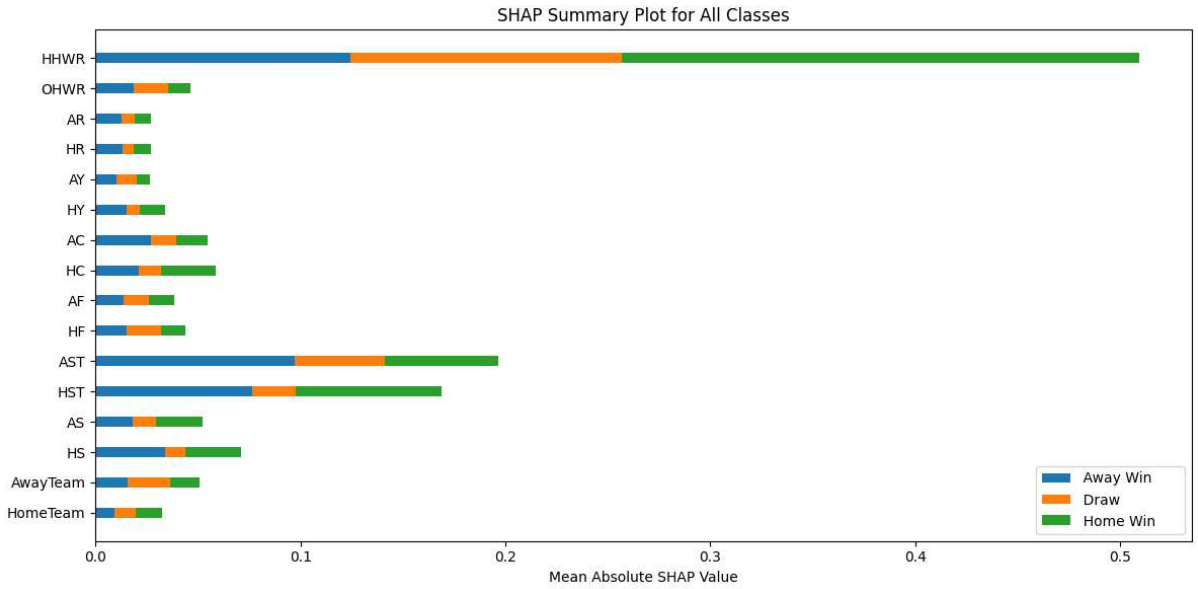


Figure 4. SHAP summary plot.

SHAP plot in Figure 4 illustrates the feature importance in our model. A bar plot that shows the mean absolute SHAP value for each feature. The bar heights represent the importance of each feature in predicting the output of the model. Each bar is divided, at most, into three parts with different colors; represent the influence of the feature for each output class: Home Win (green), Draw (orange), and Away Win (blue). The most influential feature is HHWR (Home Team Historical Win Rate), which strongly drives predictions for Home Wins, moderately affects

Draws, and has minimal impact on Away Wins. Performance metrics like AST (Away Team Shots on Target) and HST (Home Team Shots on Target) are also critical, primarily predicting Away Wins and Home Wins, respectively, with little influence on Draws. Features such as HS (Home Team Shots) and AS (Away Team Shots) further contribute to predictions, aiding in distinguishing Home Wins and Away Wins. Moderate contributors include HC (Home Team Corners) and AC (Away Team Corners), which impact Home Wins and Away Wins, respectively. OHWR (Overall Home Win Rate) is another key feature, primarily influencing Away Wins, with smaller impacts on Home Wins and minimal effect on Draws. Other features, such as disciplinary metrics like AY (Away Team Yellow Cards) and HY (Home Team Yellow Cards), have lower but distributed influence across all three classes. This analysis highlights the dominant role of HHWR, alongside performance metrics like AST and HST, complemented by additional features, in enhancing the model's predictive accuracy.

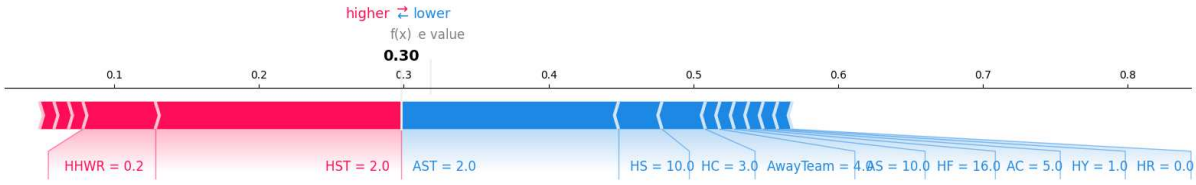


Figure 5. Force plot for the first prediction's explanation.

Another way to visualize the same explanation is to use a force plot. Figure 5 depicts the first prediction's explanation with a force plot. We will examine SHAP values for a single row of the dataset (we selected arbitrarily the first match in the last season). Blue features push the prediction lower (to the left) while red features push the prediction higher (to the right). The features within each group are ranked by their importance, and the most important features are labeled. The size of the visualization represents the magnitude of each feature's impact. Through this representation, starting from a baseline score of 0.30, which represents the average prediction across all instances, the final score remains at 0.30, indicating a balance of opposing feature contributions. Positive influences (red), such as HHWR = 0.2 (Head-to-Head Home Win Rate), which highlights a slight historical advantage for the home team, and HST = 2.0 (Home Shots on Target), which contributes moderately, push the prediction upward. Conversely, negative influences (blue), like AST = 2.0 (Away Shots on Target), signaling strong offensive performance by the away team, and HS = 10.0 (Home Shots), HC = 3.0 (Home Corners), and AwayTeam = 4.0, pull the prediction downward. Additional factors, such as HF = 16.0 (Home Fouls), AC = 5.0 (Away Corners), and HY = 1.0 (Home Yellow Cards), also exert minor negative impacts. The final prediction reflects a nuanced interplay between these opposing forces, with positive contributions from home-related features balanced by negative contributions from away-related performance metrics, resulting in a neutral prediction score of 0.30.

Correlation matrix

The correlation matrix is a square matrix that shows the strength and orientation of the linear relationship between each pair of variables in a dataset. The values in the matrix range from -1 to 1, with -1 indicating a perfect negative correlation, 1 indicating a perfect positive correlation, and 0 indicating no correlation. It should be highlighted that the existence of a correlation between two variables does not imply that one causes the other.

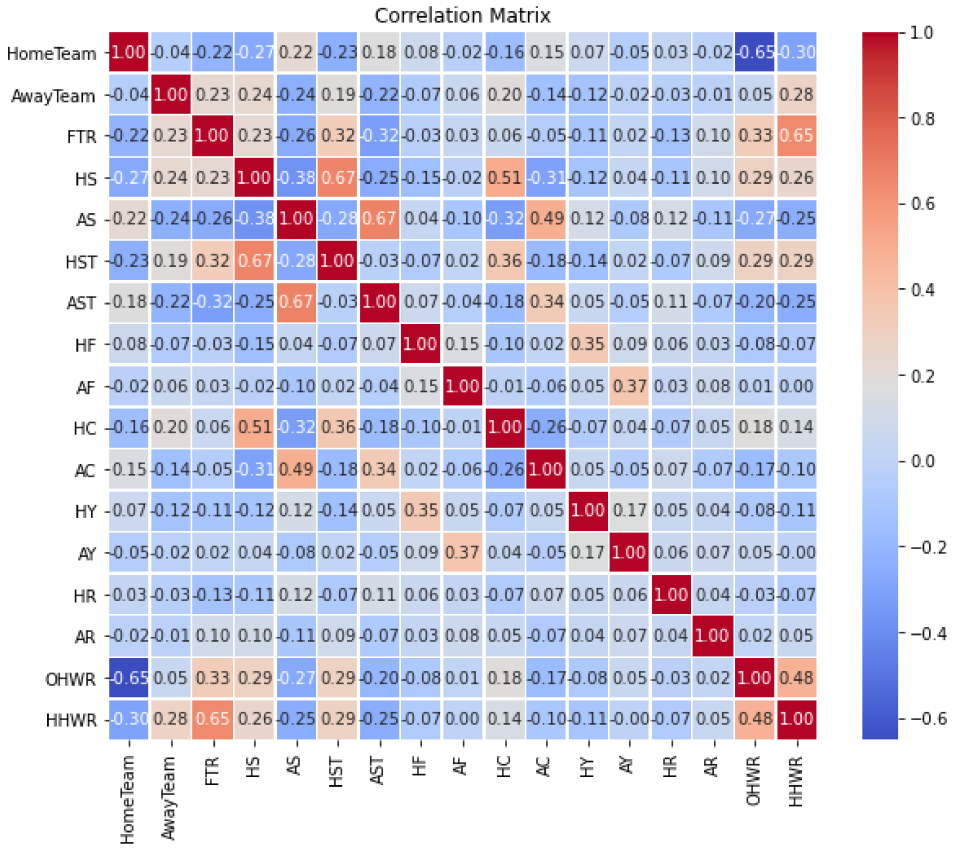


Figure 6. Correlation matrix between features.

Now, we want to understand which input features are most strongly related to the target variable. Including the target feature in the correlation matrix can provide insights into how it correlates with the input features. However, remember that the correlation between any variable and itself will always be 1. In Figure 6, we can see that The correlation between OHWR (Overall Home Win Rate) and HHWR (Head-to-Head Home Win Rate) is moderate (0.48), suggesting that a team's past home performances against specific opponents have some influence on their overall home success. However, it is interesting to note that OHWR has a stronger correlation with HomeTeam (0.65) than with HHWR, indicating that a team's overall home winning record might be more strongly linked to its general home advantage than its specific head-to-head records. Also, FTR (Full-Time Result) shows a moderate positive correlation with both OHWR (0.33) and HHWR (0.65). This indicates that the overall and head-to-head home win rates are somewhat predictive of match outcomes, with a slightly stronger association with HHWR. Key match statistics like HST (Home Shots on Target) and HS (Home Shots) show moderate positive correlations with FTR (0.32 and 0.23, respectively), indicating that a stronger attacking performance from the home team increases the likelihood of winning. Conversely, AST (Away Shots on Target) and AS (Away Shots) show negative correlations with FTR (-0.32 and -0.26, respectively), as better away performance reduces the home team's chances of winning. Features such as HC (Home Corners), HY (Home Yellow Cards), and HR (Home Red Cards) show weak correlations with FTR, suggesting that while they influence match dynamics, they have a lesser impact on the final result.

Limitations

This study provides valuable insights into predicting football match outcomes using machine learning models, particularly with the use of the MLP algorithm and SHAP analysis for

interpretability. However, there are some limitations to consider. First, the dataset used is limited to the English Premier League, meaning the model's performance and generalizability to other leagues or competitions remain untested. Additionally, the inclusion of only structured match event data omits potentially influential contextual factors such as player injuries, weather conditions, or psychological aspects like team morale, which can significantly affect match outcomes. The fixed train-validation-test split method, while maintaining temporal consistency, may restrict the model's ability to leverage modern cross-validation techniques. Lastly, the reliance on historical match statistics could introduce biases when predicting highly dynamic or atypical matches. Addressing these limitations in future work, such as incorporating real-time data or testing on diverse datasets, could further enhance the model's robustness and applicability.

Conclusions and future work

This research has addressed the problem of predicting the outcomes of football matches using the English Premier League matches data from the 2005-2006 to 2022-2023 seasons. The proposed approach explored multiple machine learning algorithms for predicting the outcomes of the matches. The last season of our dataset was used for testing the performance of those algorithms. Our experiments have shown that the MLP algorithm outperformed all other algorithms. On the other hand, the interpretability of the feature importance, of the best model, using SHAP method has been done.

While our results demonstrate a commendable predictive accuracy, we acknowledge the existence of latent variables that elude our current model. Factors such as fan support, financial backing, player recruitment strategies, and the inherent unpredictability of the game remain unaccounted for in our analysis. Future enhancements to our model could include these variables to augment predictive performance. For instance, the model could be enhanced by adding more relevant match event data, such as possession percentage, substitutions, and key passes. Additionally, player-centric elements and expected goals models could be incorporated to better understand the anticipated number of goals a team is predicted to score in a game. Finally, team data such as possession, number of crosses, tackles, and headers could be used to classify teams into different playing style categories (team profiles).

References

Aslan, B. G., and Inceoglu, M. M. (2007), "A comparative study on neural network based soccer result prediction", *Seventh International Conference on Intelligent Systems Design and Applications (ISDA 2007),* pp. 545-550, IEEE.

Bilek, G., and Ulas, E. (2019), "Predicting match outcome according to the quality of opponent in the English Premier League using situational variables and team performance indicators", *International Journal of Performance Analysis in Sport*, Vol. 24(7), pp. 1-12.

Constantinou, A. C. (2019), "Dolores: a model that predicts football match outcomes from all over the world", *Machine Learning*, Vol. 108, pp. 49-75.

Danisik, N., Lacko, P., and Farkas, M. (2018), "Football match prediction using players attributes", *2018 World Symposium on Digital Intelligence for Systems and Machines (DISA)*, pp. 201-206.

Dobson, S., and Goddard, J. (2011), "The economics of football", *Cambridge University Press*.

Fawcett, T. (2006), "An introduction to ROC analysis", *Pattern recognition letters*, Volume 27, Issue 8, pp. 861-874, ISSN 0167-8655.

Football-Data, available at: https://www.football-data.co.uk/englandm.php/ (accessed August 2023).

Hucaljuk, J., and Rakipovic, A. (2011), "Predicting football scores using machine learning techniques", *In MIPRO, 2011 proceedings of the 34th international convention*, pp. 1623–1627, IEEE.

Igiri, C. P. (2015), "Support vector machine-based prediction system for a football match result", *IOSR Journal of Computer Engineering (IOSR-JCE)*, 17(3), pp. 21–26.

Kelum, S., Shanika, A., Valery, O., Kenjabek, U. , and Upaka, R. (2024), "Explainable artificial intelligence for fitness prediction of young athletes living in unfavorable environmental conditions", *Results in Engineering*, Volume 23, 102592, ISSN 2590-1230, https://doi.org/10.1016/j.rineng.2024.102592.

Lundberg, S. M., and Lee, S.-I. (2017), "A unified approach to interpreting model predictions", *Advances in Neural Information Processing Systems 30 (NIPS 2017).*

Martins, R. G., Martins, A. S., Neves, L. A., Lima, L. V., Flores, E. L., and do Nascimento, M. Z. (2017), "Exploring polynomial classifier to predict match results in football championships", *Expert Systems with Applications*, 83, 79-93.

Montesinos López, O.A., Montesinos López, A., Crossa, J. (2022). "Overfitting, Model Tuning, and Evaluation of Prediction Performance". In: *Multivariate Statistical Machine Learning Methods for Genomic Prediction*. Springer, Cham. https://doi.org/10.1007/978-3-030-89010-0_4

Moustakidis, S., Plakias, S., Kokkotis, C., Tsatalas, T., and Tsaopoulos, D. (2023), "Predicting Football Team Performance with Explainable AI: Leveraging SHAP to Identify Key Team-Level Performance Metrics", *Future Internet*, 15, no. 5: 174, https://doi.org/10.3390/fi15050174.

Pedregosa, F., Varoquaux, G., Gramfort, A., *et al.* (2011), "Scikit-learn: Machine learning in Python", *Journal of machine learning research*, Vol. 12, pp. 2825-2830.

Premier League, available at: https://www.premierleague.com/ (accessed August 2023).

Procopiou, A. and Piki, A. (2023), "The 12th Player: Explainable Artificial Intelligence (XAI) in Football: Conceptualisation, Applications, Challenges and Future Directions", *Proceedings of the 11th International Conference on Sport Sciences Research and Technology Support* - icSPORTS; ISBN 978-989-758-673-6, ISSN 2184-3201, SciTePress, pages 213-220, DOI: 10.5220/0012233800003587.

Razali, N., Mustapha, A., Yatim, F., and Aziz, R. (2017), "Predicting football matches results using Bayesian networks for English Premier League (EPL)", *IOP Conference Series: Materials Science and Engineering*, 226, DOI: 10.1088/1757-899X/226/1/012099.

Ren, Y., and Susnjak, T. (2022), "Predicting football match outcomes with explainable machine learning and the Kelly Index", *arXiv preprint arXiv:2211.15734*.

Smith, A. (2017), "Sky Sports bust common football myths: Home advantage?", available at: https://www.skysports.com/football/news/11096/10955089/sky-sports-bust-common-football-myths-home-advantage/ (accessed August 2023).

Ulmer, B., and Fernandez, M. (2013), "Predicting soccer match results in the English Premier League", *Stanford,* Ph.D. thesis.

Winter, E. (2002), "The Shapley value", *Handbook of game theory with economic applications*, Vol. 3, pp. 2025-2054, Elsevier.

Yuanchen, W., Weibo, L., and Xiaohui, L. (2022), "Explainable AI techniques with application to NBA gameplay prediction", *Neurocomputing*, Volume 483, Pages 59-71, ISSN 0925-2312, https://doi.org/10.1016/j.neucom.2022.01.098.