A Machine Learning Model for Analysis and Prediction of Football Match Outcomes in the English Premier League

Emmanuel Bamidele Ajulo[1*], Qayyum Adekunle Tiamiyu[2]

Department of Computer Science, Federal University of Technology, Akure, Ondo State, Nigeria.

***Corresponding Author**

DOI: https://doi.org/10.51584/IJRIAS.2026.11010020

Received: 27 December 2025; Accepted: 03 January 2026; Published: 24 January 2026

ABSTRACT

Football stands as the world's most popular sport, captivating billions globally. The English Premier League, in particular, is widely regarded as the pinnacle of professional football, boasting immense global viewership and attracting widespread interest. Its dynamic and unpredictable nature fuels a massive industry built around match analysis, reflecting the deep desire to anticipate match outcomes. Early attempts at football match prediction often relied on static historical data, assumed independence among events, failed to adapt quickly to football's rapid evolution, and lacked the capacity to capture complex nonlinear interactions among multiple features. This study develops a machine learning model for football match analysis in the English Premier League to predict match outcomes, addressing gaps in previous models by using ensemble machine learning algorithms to provide timely, accurate, and real-time analysis. The study utilised Random Forest (RF), XGBoost, and LightGBM. Performance evaluation using standard classification metrics, including Accuracy, Precision, Recall, F1-Score, and ROC-AUC, showed that Random Forest achieved the best overall performance, with an accuracy of 87.14% and an ROC-AUC of 99.00%. The ensemble model further enhanced prediction consistency by combining the strengths of the three machine learning models. This study demonstrates the effectiveness of machine learning for match predictions and, from an industry perspective, offers practical recommendations for football to enhance retention, efficiency, and competitiveness.

Keywords: Football, Machine learning, Dataset, Random Forest, XGBoost , LightGBM

INTRODUCTION

Football is the world's most popular sport, captivating billions worldwide (Almarri et al., 2022). The English Premier League, in particular, holds a premier position as the pinnacle of professional football, with immense global viewership and widespread commercial interest. Its dynamic and often unpredictable nature fuels a massive industry built around match analysis, fan engagement, and sports betting. Consequently, a vast array of predictions, both casual and professional, circulate before matches, reflecting the deep desire to anticipate outcomes in this highly competitive arena. Commentators, pundits, and dedicated sports channels like ESPN regularly engage in pre- and post-match analysis, underscoring the perennial challenge of forecasting results in this high-stakes environment (Razali et al., 2017).

Historically, football prediction has relied heavily on human intuition, expert opinions, and rudimentary statistical methods. While these approaches offer some insights, they often struggle to capture the complex interplay of numerous variables that influence match outcomes, such as team dynamics, individual player performance fluctuations, tactical shifts, and contextual factors (Atitallah et al., 2022).

The advent of Artificial Intelligence and Machine Learning has revolutionised predictive analytics across various domains, offering unprecedented capabilities for highly accurate forecasting (Eryarsoy and Delen, 2019). This technological leap provides a transformative opportunity to elevate the precision and depth of football match analysis, moving beyond subjective assessments to data-driven insights. Machine learning models, with their ability to process vast datasets and learn from historical trends, are increasingly being applied to sports to enhance understanding and predict future events.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue I January 2026

Despite the widespread interest and extensive efforts to predict football match outcomes, a significant challenge persists: the lack of a robust, dynamic, and data-driven approach that fully leverages modern AI and machine learning techniques to achieve high predictive accuracy. Existing prediction models often suffer from several limitations. Many are based on static historical data, failing to adapt quickly to the rapid evolution of team form, player transfers, injury impacts, coaching strategies, and in-game tactical adjustments. This reliance on outdated or incomplete information significantly compromises their predictive power in a league as volatile as the English Premier League (Baboota and Kaur, 2018).

Furthermore, traditional models often overlook the complex variables that collectively determine match outcomes, including player statistics, specific tactical setups, environmental factors, and even the subtle influence of referee decisions (Atitallah et al., 2022). The sheer volume and complexity of these variables make it challenging for conventional methods to account for their combined effect. Consequently, there is an urgent and critical need for a comprehensive predictive model. Such a model must not only utilise extensive historical data but also integrate the most up-to-date, quasi-real-time statistics and advanced machine learning algorithms. This integration is crucial to provide timely, accurate, and actionable insights into match outcomes, thereby addressing the limitations of current approaches and offering a more sophisticated understanding of football's inherent unpredictability.

LITERATURE REVIEW

Historically, football analysis was predominantly qualitative, relying on subjective expert opinions and post-match video breakdowns. The introduction of statistical recording in the mid-20th century marked the first quantitative shift, allowing for rudimentary comparisons based on goals, shots, and possession. The digital age, however, has ushered in an era of big data in sports. Sophisticated data collection systems, from opti-tracking cameras to GPS vests, now capture granular event data (e.g., passes, tackles, runs, touches) and physical metrics, providing unprecedented detail. This explosion of data has been the primary catalyst for the adoption of more advanced analytical tools, transforming football analysis from subjective interpretation to an increasingly data-driven science (Atitallah et al., 2022).

The accurate prediction of football match outcomes, particularly in highly competitive leagues such as the English Premier League (EPL), has been a persistent area of research. Driven by the increasing availability of detailed match data and advancements in Artificial Intelligence and Machine Learning, numerous studies have explored various methodologies to enhance predictive accuracy. This comprehensive review analyses recent contributions within the last five years (June 2020 – June 2025), summarising their approaches, processes, and key findings in the context of football match prediction.

Researchers have utilised a variety of datasets, predominantly scraped from public football statistics websites (e.g., Football-Data.co.uk, historical league tables, stats aggregators), to train and test their predictive models (Atitallah et al., 2022). Standard features engineered from this data include team form (recent wins/losses/draws), goal differences, home/away advantages, head-to-head records, league positions, and disciplinary statistics. A significant portion of recent research has focused on the application of supervised learning algorithms and ensemble methods, given their robustness and ability to capture complex nonlinear relationships in football data.

Ensemble Learning Approaches

Narayanan et al. (2024) examined the efficacy of XGBoost and LightGBM in predicting EPL match outcomes, using features such as team form, player market values, and historical results. Their study highlighted the superior performance of boosting algorithms on structured football data, achieving higher accuracy than traditional models. However, it could not outperform bookmakers, and its limited feature selection limited its predictive power. Štemberk et al. (2023) conducted a comparative analysis of Random Forest, Gradient Boosting, and Logistic Regression for predicting football matches across European leagues, including the EPL. They emphasised the importance of feature engineering, particularly dynamically updated team strength metrics, and found that ensemble methods generally outperformed single classifiers. As good as it was, it did not compare its performance to real-world betting strategies.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue I January 2026

Rane et al. (2022) developed an ensemble model combining Random Forest and Support Vector Machine (SVM) classifiers for predicting EPL match results. Their work underscored that feature selection based on statistical significance improved predictive performance, demonstrating the benefits of combining diverse models. Shen et al. (2025) developed a real-time football match prediction platform using Random Forest, SVM, neural networks, and a stacking method. They created a platform that updates predictions dynamically using real-time data. It was limited to sofa score data, which was not sufficient to improve accuracy. Mittal et al. (2020) investigated the use of Gradient Boosting models for EPL outcome prediction, incorporating a wide array of team statistics and pre-match odds as features. Their findings suggested that sophisticated boosting techniques could produce results competitive with bookmaker predictions, albeit without consistently guaranteeing profit.

Traditional Machine Learning and Deep Learning Applications

Khan et al. (2024) utilised Logistic Regression and Artificial Neural Networks (ANNs) for predicting EPL match results, focusing on readily available team-level statistics. Their study contributed to understanding the baseline performance of simpler models against more complex neural network architectures in this domain. Almalki et al. (2023) explored the application of Deep Neural Networks (DNNs) for predicting football match outcomes by processing historical match data. Their research demonstrated that while DNNs could learn complex patterns, proper feature engineering remained crucial for optimal performance, sometimes even more so than model complexity.

Bhattacharya et al. (2022) applied various ML algorithms, including Naive Bayes, SVM, and Random Forest, for EPL match prediction, highlighting the importance of data preprocessing and feature selection in achieving reliable results. Their work provided a comparative study of different algorithms on a consistent dataset. In the same vein, Das and Das (2021) focused on feature importance for EPL prediction using Random Forest and XGBoost, identifying key statistical features such as goal difference, home form, and recent win streaks as the most influential. This contributes to the interpretability of predictive models.

Liang et al. (2024) explored a dynamic prediction model that incorporated "real-time" aspects by updating team strength ratings after every match. While not truly "live in-game," their approach of continuous model adaptation showcased an advancement in leveraging evolving data. Consequently, Silva and Pinto (2022) investigated the impact of various tactical and player-level statistics on EPL match outcomes using ML models. Their research highlighted the potential for more granular data to enhance predictive power, moving beyond aggregated team statistics.

DATA AND METHODS

The study establishes a transparent framework for building a data-driven prediction system, from data acquisition and preprocessing to feature engineering, model training, and evaluation. The system architecture is shown in Figure 1.

Figure 1: The System Architecture

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue I January 2026

Data Collection and Acquisition

A structured tabular dataset containing historical English Premier League (EPL) match statistics from 2021 to 2025 was obtained. The model predicts three outcomes: Home Win, Draw, or Away Win as target variables. The data was scraped and downloaded in CSV format from reputable sites, including Football-Data.co.uk, Kaggle, and various public sports APIs.

Data Preprocessing

Raw statistics were transformed into a format suitable for machine learning. Data cleaning was carried out to address missing values through imputation and to remove irrelevant features, such as match IDs, to reduce noise. Foundational metrics, such as Goal Difference and Total Goals, were created. Categorical encoding was applied to team names and referees, converting them into numerical values using One-Hot Encoding. Feature scaling was achieved by applying Z-score normalisation to numerical features, ensuring all variables have a mean of 0 and a standard deviation of 1.

Feature Engineering

This phase focused on capturing "real-time" team form and tactical nuances. To calculate rolling averages, performance indicators such as goals, shots, and possession were calculated over the last 3, 5, and 10 matches. Strength Metrics rating of dynamic skill levels, such as attack/defence ratios, and head-to-head historical statistics, were considered. Contextual factors, such as home-ground advantage and "days since last match" to capture fatigue, were also accounted for

Dataset Splitting

The data was divided using a temporal split to prevent data leakage and to simulate real-world forecasting: Training (70%): Earliest seasons used for model learning; Validation (15%): Middle period used for hyperparameter tuning; Test (15%): Most recent season(s) used for final, unbiased evaluation; and Stratification: Ensured the distribution of wins, draws, and losses remained consistent across all three sets.

Model Selection

Ensemble learning was chosen for its ability to handle complex, non-linear relationships in tabular sports data. Two specific algorithms were implemented: XGBoost (Extreme Gradient Boosting): A sequential boosting method where each new tree corrects the errors of previous ones. It includes built-in regularisation to prevent overfitting and Random Forest, a bagging method that aggregates predictions from multiple independent decision trees to reduce variance and improve generalisation.

Model Training and Optimisation

To ensure peak performance, the models underwent a rigorous training process: Hyperparameter Tuning: Techniques like Grid Search and Random Search with k-fold Cross-Validation were used to find optimal settings (e.g., tree depth, learning rate, and number of estimators); Probabilistic Output: Models were configured to output probabilities for each outcome (Home Win, Draw, Away Win) rather than simple labels; and Early Stopping: Specifically for XGBoost, training was halted when performance on validation data stopped improving, further protecting against overfitting.

Evaluation Metrics

The models are assessed using several industry-standard metrics to ensure accuracy and reliability. This is shown in Table 1.

Table 1: Evaluation Metrics

Metric	**Purpose**	**Formular**
Accuracy	Measures the overall percentage of correct predictions.	Accuracy = $\frac{Current\ Predictions}{Total\ Predictions}$

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue I January 2026

Precision Recall	Evaluates the "exactness" and "completeness" of predictions for each specific outcome.	Precision = $\frac{TP}{TP+FP}$ Recall = $\frac{TP}{TP+FN}$
F1-Score	Provides a balanced harmonic mean of precision and recall, useful for uneven class distributions.	F1-Score = 2 x $\frac{Precision\ xRecall}{Precision+Recall}$
ROC-AUC	Measures the model's ability to distinguish between classes across various thresholds using a "one-vs-rest" approach.	
Confusion Matrix	Provides a visual breakdown of True Positives vs. False Positives to identify specific prediction trends or errors.	

System Specifications

The development environment was configured to ensure efficient data processing and model deployment. The model was developed on a macOS system with 16GB RAM and 50GB SSD storage. Utilising Python 3.8+ within Jupyter Notebooks. The developed models were integrated into interactive interfaces using Streamlit and Gradio.

Core Libraries and Frameworks

A specialised Python ecosystem was used for different stages of the study. This is shown in Table 2.

Table 2: Core Libraries and Frameworks

Library	**Primary Role**
Pandas	Data manipulation, loading CSVs, and handling missing values.
NumPy	Numerical computing and mathematical array operations.
Matplotlib / Seaborn	Data visualization (histograms, distribution plots, and correlation heatmaps).
Scikit-learn	Data splitting, feature extraction, and calculating evaluation metrics.
Transformers	Integration with Hugging Face for tokenization and classification tasks.

Data Description and Attributes

The dataset consists of five EPL seasons (2021–2025) sourced from football-data.co.uk. It contains detailed match statistics used to train the model, including: Temporal & Identity: Season, Match Date, Home Team, and Away Team; Scoring Data: Full-time and Half-time goals/results (Home Win, Away Win, Draw); Match Performance: Total shots and shots on target for both sides; and Set Pieces & Discipline: Corner counts, fouls committed, and yellow/red cards issued for each team.

RESULTS AND DISCUSSION

The obtained results from the different machine learning models, andom Forest (RF), XGBoost (XGB), and LightGBoost (LGB) are shown in Table 3.

Table 3: Comparative Performance Analysis of Model

Model	**Accuracy (%)**	**F1-Score (%)**	**Precision (%)**	**Recall (%)**	**ROC (%)**
R F	87.14	84.92	86.70	84.70	99.00
XGB	86.64	85.09	85.85	84.62	98.40
LGB	85.67	83.89	85.09	83.26	97.96

Two primary visual tools were used to interpret model results: the confusion matrix and the ROC curve. The confusion matrix provided insights into how well each model distinguished between true and predicted

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue I January 2026

outcomes, highlighting true positives, false positives, true negatives, and false negatives. The ROC curve, on the other hand, illustrated the models' ability to discriminate among the three classes across different outcomes, providing a visual understanding of sensitivity and specificity. These tools collectively supported both quantitative and qualitative analysis of model performance. Figures 2–7 clearly visualise the confusion matrices and ROC curves for each model.

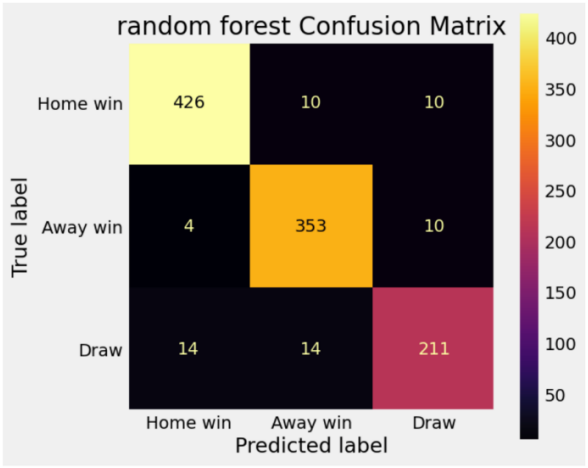

Figure 2: Random Forest Confusion Matrix

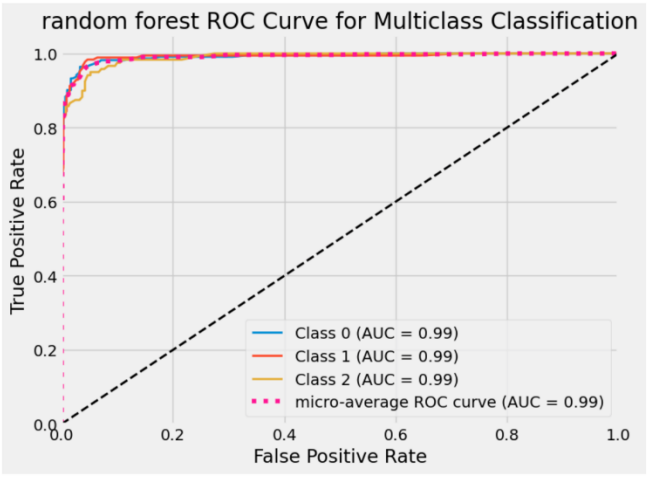

Figure 3: Random Forest (RF) ROC Curve

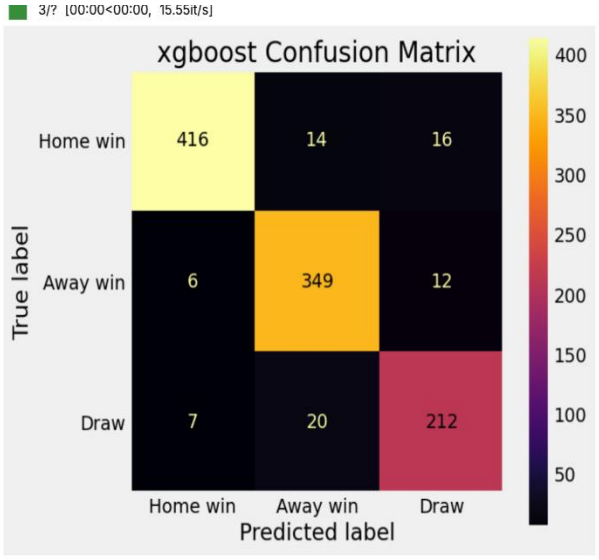

Figure 4: XGBoost Confusion Matrix

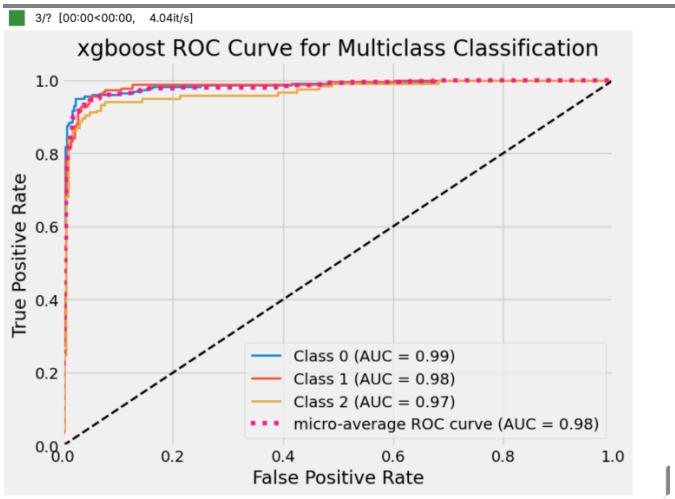

Figure 5: XGBoost ROC Curve

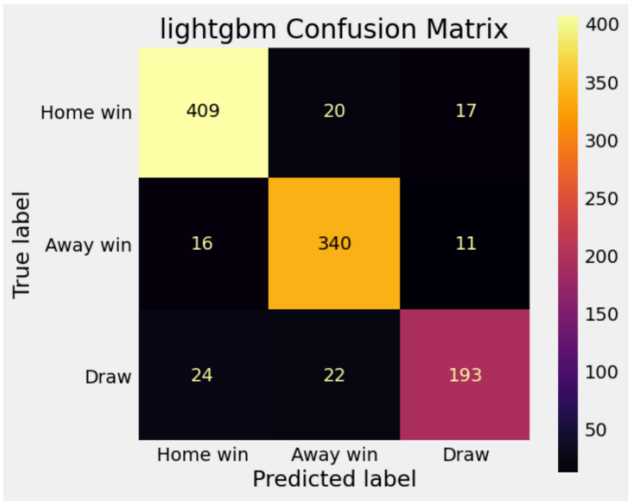

Figure 6: LightGB Confusion Matrix

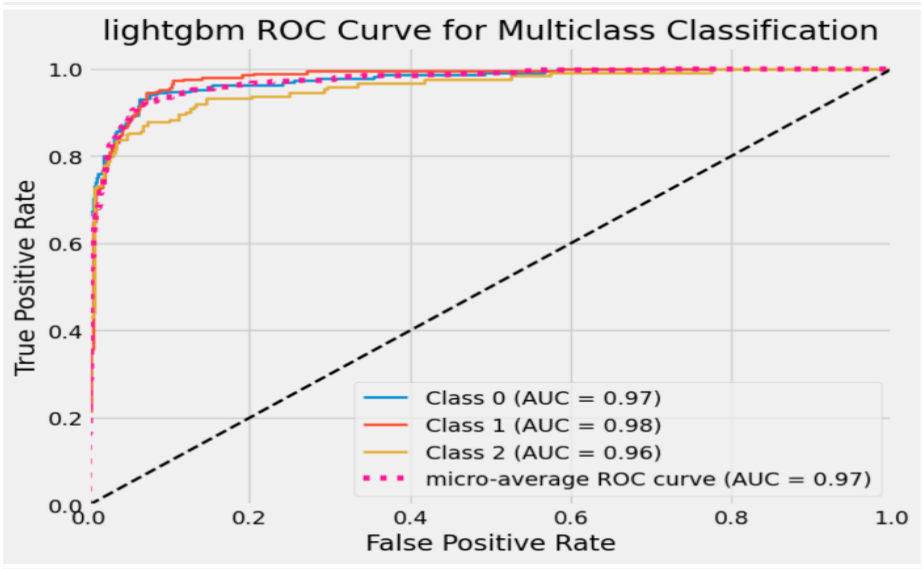

Figure 7: LightGB ROC Curve

Results and Performance Analysis

Random Forest served as a baseline model. It achieved an overall accuracy of 87.14%, with a precision of 86.70% and a recall of 84.70%. The F1-score was 84.92%, and the ROC-AUC was 99.00%, indicating a

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue I January 2026

moderate ability to discriminate between actual outcome and predicted outcome. The model performed exceptionally well on the three classes, but struggled with false positives. Despite its limitations, the model offered a strong benchmark for subsequent evaluations.

The XGBoost model significantly outperformed the Random Forest benchmark, achieving 86.64% accuracy, 85.85% precision, 84.62% recall, and 85.09% F1-score. The ROC-AUC score reached 98.40%, indicating robust discriminative performance. Unlike the Random Forest model, the XGBoost demonstrated a better balance between precision and recall, improving its ability to identify important features while minimising false positives.

The LightGB model achieved an accuracy of 85.67%, precision of 85.09%, recall of 83.26%, and an F1-score of 83.89%. The ROC-AUC was 97.96%, slightly lower than XGBoost but still demonstrating strong classification performance. LightGB was particularly effective at maintaining high memory speed, which is critical for match prediction, ensuring that real-time data is correctly identified. Its ability to model sequential dependencies also contributed to its robust performance.

Comparative Analysis

The performance differences between models are relatively narrow (1.47% range in accuracy), suggesting that all three algorithms demonstrate competitive predictive capabilities for football match prediction. However, several key distinctions emerge. Random Forest achieved the highest performance across all metrics, demonstrating superior generalisation capabilities with an accuracy of 87.14% and an F1-score of 84.92%. XGBoost closely followed with 86.64% accuracy and 84.62% recall, while maintaining competitive precision (85.85%). LightGBM achieved 85.67% accuracy while offering the fastest training and prediction times, making it ideal for real-time applications. The comparative analysis of the developed models based on the results obtained is shown in Table 4.

Table 4: Comparative Analysis of the Developed Models

Model	Strengths	Weaknesses
Random Forest	Highest Accuracy: Achieved 87.14%. Robust Selection: Reduces overfitting via feature randomness. Stability: Less sensitive to hyperparameter tuning. Parallelism: Efficient parallel tree construction.	Memory: Large ensembles require significant storage. Temporal Limits: Struggles with sequential match dependencies. Feature Bias: May underperform on continuous numerical data.
XGBoost	Regularisation: Built-in L1/L2 prevents overfitting. Interpretability: Strong metrics and SHAP value integration. Optimisation: Iterative error correction improves accuracy. Flexibility: Extensive tuning options and built-in cross-validation.	Complexity: Slower training than RF or LightGBM. Sensitivity: Requires very precise hyperparameter tuning. Resource Intensive: High memory usage during training. Sequential: Limited parallelisation.
LightGBM	Speed: Fastest training and prediction times. Efficiency: Low memory usage via histogram algorithms. Real-time: Ideal for live match prediction updates.	Overfitting: High risk on smaller datasets. Stability: Performance varies across different random seeds. Preprocessing: Often requires extra handling for categorical features.

CONCLUSION

This study implemented and evaluated a comprehensive machine learning framework for predicting English Premier League (EPL) football matches outcome using three distinct algorithmic approaches: Random Forest, Extreme Gradient Boost (XGBoost), and Light Gradient Boost Machine (LightGBM). The study demonstrates the practical application of ensemble learning methods in sports analytics, explicitly addressing the complex challenge of predicting football match outcomes. The implementation process encompassed data acquisition

and preprocessing, as well as model development, training, evaluation, and deployment. Using five seasons of EPL data (2021-2025), the study processed comprehensive match statistics, including goals, shots, corners, fouls, and disciplinary actions, to engineer meaningful predictive features.

The results showed that XGBoost achieved better performance, followed by lightGB and Random Forest. The study confirms that machine learning approaches can effectively capture the complex patterns inherent in football match dynamics, providing valuable insights for sports analytics, betting markets, and strategic team management. The achieved accuracy levels of 85-87% represent significant performance in the inherently unpredictable domain of football match prediction.

Ethical Considerations: Ethics declaration not applicable.

Conflict of Interest: There is no conflict of interest.

Data Availability: Data obtained from football-data.co.uk and Kaggle.com

REFERENCES

1. Almarri, M. M., Alotaibi, S. A., & Al-Thani, A. (2022). Ensemble-based machine learning for classification and prediction of diabetic patients' status using a Saudi Arabian dataset: pre-diabetes, Tidm, and T2DM. Computers in Biology and Medicine, 147, 105757. https://doi.org/10.1016/j.compbiomed.2022.105757
2. Atitallah, S. B., Driss, M., & Almomani, I. (2022). A novel detection and multi-classification approach for IoT-malware using random forest voting of fine-tuning convolutional neural networks. Sensors, 22(11), 4302. https://doi.org/10.3390/s22114302
3. Baboota, R., & Kaur, H. (2018). Predictive analysis and modelling football results using machine learning approach for English Premier League. International Journal of Forecasting. Advance online publication. https://doi.org/10.1016/j.ijforecast.2018.01.003
4. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32. https://doi.org/10.1023/A:1010933404324
5. Eryarsoy, E., & Delen, D. (2019). Predicting the outcome of a football game: A comparative analysis of single and ensemble analytics methods. In Proceedings of the 52nd Hawaii International Conference on System Sciences (p. 1107). https://hdl.handle.net/10125/59550
6. FiveThirtyEight. (n.d.). FiveThirtyEight football predictions. Retrieved from https://projects.fivethirtyeight.com/soccer-predictions/
7. Forebet. (2018). Mathematical football predictions, Tips, Statistics, Previews. Retrieved from https://www.forebet.com
8. Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. The Annals of Statistics, 29(5), 1189-1232. https://projecteuclid.org/journals/annals-of-statistics/volume-29/issue-5/Greedy-function-approximation-A-gradient-boosting-machine/10.1214/aos/1013203451.full
9. Hubáček, O., Šourek, G., & Železný, F. (2019). Learning to predict soccer results from relational data with gradient boosted trees. Machine Learning, 108(1), 29-47. https://doi.org/10.1007/s10994-018-5704-6
10. Kaggle. (n.d.). Kaggle European Soccer Database. Retrieved from https://www.kaggle.com/hugomathien/soccer12
11. Opta Sport. (n.d.). Opta Sport data provider. Retrieved from http://www.optasports.com/
12. Razali, N., Mustapha, A., Yatim, F. A., & Ab Aziz, R. (2017). Predicting football matches results using Bayesian networks for English Premier League (EPL). IOP Conference Series: Materials Science and Engineering, 226(1), 012099. https://doi.org/10.1088/1757-899X/226/1/012099.
13. Ulmer, B., Fernandez, M., & Peterson, M. (2013). Predicting soccer match results in the English Premier League. Stanford University CS229 Final Project.
14. Wunderlich, F., & Memmert, D. (2016). Analysis of the predictive qualities of betting odds and FIFA rankings in forecasting the results of football matches. PLOS ONE, 11(2), e0148982. https://doi.org/10.1371/journal.pone.0148982