AI-driven sports analysis

Adam Gábor*

FIIT STU, Ilkovičova 2, Bratislava, Slovakia
xgabora@stuba.sk

Abstract. This paper explores the integration of advanced machine learning techniques for predicting football match outcomes through a novel multi-layered framework. We develop a hierarchical ensemble architecture that combines static pre-match features with dynamic in-play data to generate real-time predictions. The system incorporates deep neural networks and gradient boosting models (LightGBM and XGBoost). The meta-learning approach dynamically adjusts the prediction weights based on match conditions, scenario similarities, and model confidence metrics. Using our custom-built dataset of 230,000 matches in 42 leagues and 27 countries spanning 2000 to 2025, our framework achieves 67.4% accuracy for three-way match outcomes, outperforming traditional statistical methods by 9.3 percentage points, and improving baseline machine learning approaches by 2-3 percentage points. For exact score prediction, our ensemble approach reaches 26.7% accuracy with an R^2 of 0.510, representing a 40.5% improvement over conventional Poisson distribution models. Evaluation in multiple metrics (accuracy, F1 score, MSE, MAE, Brier score, ROC-AUC) demonstrates consistent performance improvements across different prediction tasks and match scenarios, advancing sports analytics by providing actionable predictions for stakeholders in the football industry. The custom built dataset with comprehensive documentation can be found at: https://github.com/xgabora/Club-Football-Match-Data-2000-2025.

Keywords: football prediction · machine learning · ensemble methods · match clustering · real-time data analysis · sports data mining.

1 Introduction

Sports match analysis has evolved significantly with increasing data availability, placing football, the world's most popular sport, at the forefront [1]. Predicting the outcomes of football matches is a key challenge in sports analytics, impacting team strategy, and prediction markets. Accurate predictions offer competitive advantages across the multibillion dollar global football industry, making this a compelling area for applying advanced machine learning. The dynamic nature

* Bachelor study programme in field: Informatics. Supervisors: doc. Dr. Ing. Michal Ries, Ing. William Brach, Institute of Informatics, Information Systems and Software Engineering, Faculty of Informatics and Information Technologies STU in Bratislava

of football provides a testbed for techniques that can be transferred to other complex prediction domains.

Traditional approaches to football prediction face challenges. The sport's unpredictability limits simple statistical models. Most existing methods treat pre-match and in-play predictions separately, failing to capture evolving match conditions. Previous solutions are either statistical (focused on pre-match factors) or reactive (analyzing in-play data), missing an opportunity to create a unified framework that can seamlessly transition between pre-match and in-play scenarios while accounting for the evolving nature of matches.

Furthermore, most existing models fail to recognize that different match types (e.g., high-scoring offensive games versus tactical defensive battles) require specialized prediction approaches. A key limitation is the scope of existing studies, with many focusing on a small group of matches, such as a single league or a particular season. This prevents a generalized solution and limits the adaptability of the model to real-world scenarios.

This paper introduces a multi-layered machine learning framework to address these limitations. Our approach combines specialized models with meta-learning for dynamic optimization and uses match clustering for situation-specific refinement. Leveraging gradient boosting, neural networks, and ensemble methods, we improve predictive accuracy, though computational complexity and further optimization of the framework remains a challenge.

2 Related Works

The challenge of predicting sports outcomes has attracted significant research attention in multiple disciplines. Early work focused primarily on statistical approaches, with Dixon and Coles [2] demonstrating that football goal distributions closely follow Poisson patterns.

The adaptation of the Elo rating system from chess to football by Hvattum and Arntzen [3] marked an advance in team strength assessment. Although effective for long-term performance evaluation, the system's limitations in capturing short-term variations led researchers to explore more sophisticated approaches.

Recent years have seen a shift toward machine learning methods. Igiri [4] achieved improvements using logistic regression models that incorporated multiple features beyond the historical results. Their work demonstrated the importance of considering team composition and external factors in the prediction.

Deep learning approaches have shown particular promise in capturing complex patterns. Muszaidi et al. [6] implemented recurrent neural networks for match prediction, achieving accuracy rates of up to 78% in specific prediction tasks. Their work highlighted the potential of temporal pattern recognition in football analysis, although computational requirements remain significant.

Ensemble methods have become particularly effective. Stübinger [7] demonstrated that the combination of multiple models can significantly improve prediction accuracy, achieving rates above 80% in certain two-way prediction (home

or away win) scenarios. This success has inspired further research into hybrid approaches that take advantage of the strengths of different methodological frameworks for different microtasks.

The most recent developments focus on real-time prediction updates. However, as noted by Yao et al. [8], the challenge of integrating live match data with pre-match predictions remains largely unresolved. This gap in current approaches, particularly in handling the dynamic nature of football matches, motivates our present research.

Clustering techniques have emerged as valuable tools in football analysis, offering insights beyond traditional prediction methods. Yi et al. [9] applied k-means clustering to classify teams in the 2018 FIFA World Cup, distinguishing possession-based from direct-play styles, with possession-oriented teams exhibiting higher goal-scoring efficiency.

Moura et al. [10] utilized principal component and cluster analysis on 2006 World Cup data, successfully grouping teams based on match results, with 70.3% of winners classified within the same cluster. Expanding on this, Diquigiovanni and Scarpa [11] developed a clustering approach to analyze playing styles and demonstrated its efficacy in match prediction, effectively linking tactical patterns to goal-scoring success. These studies confirm that clustering can reveal tactical patterns and team characteristics that traditional models may overlook, providing context for more accurate outcome prediction.

3 Data

3.1 Data Sources and Collection

This study utilizes our custom-built dataset of approximately 230,000 professional football matches from 42 leagues across 27 countries, spanning July 2000 to March 2025, making it one of the largest open-source football datasets. Match results and statistics were sourced from Football-Data.co.uk [14], while team strength metrics were obtained from the ClubElo API [15].

A custom Python pipeline ensured data consistency across competitions. Match data was systematically extracted from Football-Data.co.uk, while Elo ratings for 800 clubs were sampled twice a month from ClubElo. The final dataset including clear structured documentation and processing pipeline is publicly available on GitHub [16], Kaggle [17] and HuggingFace [18] under the MIT license to support further football analytics research.

3.2 Data Structure and Features

The dataset consists of two primary components: match data and Elo ratings. The match data table includes 76 features per match, covering key categories such as match identifiers (division, date, time), team details (home and away teams), results (full-time and half-time scores), performance metrics (shots, fouls, corners, cards), team strength indicators (Elo ratings, recent form), and market data (prediction market odds for outcomes and goal expectations). Missing

values, primarily in match statistics and Elo ratings, were excluded rather than imputed to preserve data integrity.

3.3 Match Clustering

A key innovation in our dataset is match clustering, categorizing games based on statistical characteristics using the K-means algorithm. Feature engineering transformed raw match data into meaningful dimensions such as shot efficiency, possession dominance, game physicality, tempo, and more.

The analysis identified six distinct match types by examining key characteristics that significantly deviate from average match statistics, revealing underlying game patterns and tactical approaches. After identifying these distinctive patterns, we developed three-letter codes for efficient reference and classification.

The clusters were named based on their most prominent characteristics: Low Tempo Home-oriented (LTH) for matches with reduced pace favoring home teams, High Tempo Balanced (HTB) for fast-paced action-rich evenly matched contests, Low Tempo Away-oriented (LTA) for slower matches with away team accuracy advantage, Visibly Away-Dominated (VAD) and Visibly Home-Dominated (VHD) for games with clear team superiority, and Physical Balanced (PHB) for matches characterized by above-average physical intensity across teams. Cluster probabilities were derived using a softmax transformation, allowing nuanced classification that captures matches exhibiting characteristics of multiple clusters simultaneously.

Table 1: **Key Characteristics of Football Match Clusters**

Characteristic	**LTH**	**LTA**	**VHD**	**VAD**	**HTB**	**PHB**
Dataset %	20	18	13	14	19	16
Shot Efficiency	+23	-24	+6	-4	0	+1
Home Possession %	51	39	70	38	55	54
Total Fouls	24.90	24.93	24.81	25.30	25.27	30.40
Shot Count	22.46	22.53	22.76	22.15	24.32	23.49
Elo Diff (H-A)	+7	0	+200	-200	0	-4
Percentage Differences from Dataset Average						
Home Shots	-4.1	-2.9	+3.2	-3.1	+5.2	+2.9
Away Shots	-3.1	-3.5	-1.3	+3.0	+5.6	+1.3
Home Fouls	-4.1	-3.9	-3.4	-0.8	-1.6	+18.4
Away Fouls	-4.3	-3.5	-3.3	-1.4	-2.1	+18.0
Home Yellow Cards	-3.9	-4.1	-2.2	+4.3	-0.4	+13.8
Away Yellow Cards	-3.5	-2.8	-0.5	+2.5	+0.2	+12.2
Home Red Cards	-1.0	-5.1	-2.2	+5.1	-2.4	+17.5
Away Red Cards	-3.0	-1.4	-4.8	-5.5	-3.1	+25.8

4 Methods

4.1 Framework Architecture

The framework consists of four main layers - input layer, processing layer, prediction layer and output layer, as shown in Figure 1 below.

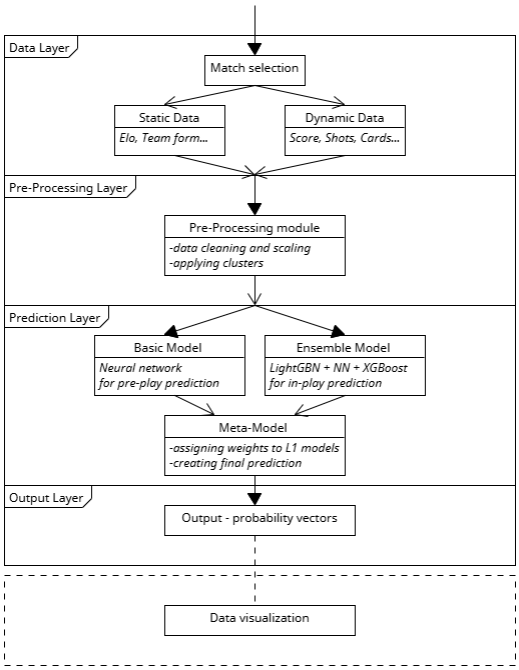

Fig. 1. Proposed framework architecture.

The proposed prediction layer implements a hierarchical ensemble approach with a two-level architecture designed to capture multiple aspects of match dynamics while facilitating integration between pre-match and in-play predictions.

Level 1 comprises specialized models that address distinct aspects of the prediction task. This level contains two primary components: First, a "classic" deep neural network processes static pre-match features and employs multiple dense layers with batch normalization and dropout regularization. The architecture uses ReLU activation functions in hidden layers and a softmax output layer for probabilistic outcome prediction. The second component consists of an ensemble of gradient boosting models XGBoost [12] and LightGBM [13] alongside a specialized neural network for processing dynamic match features, with LightGBM and XGBoost selected and trained specifically as "expert" models for predicting challenging scenarios such as low scoring matches and draws.

Level 2 implements a meta-learning approach that dynamically combines predictions from Level 1 models. This meta-model employs a weighted ensemble mechanism where weights are determined through multiple factors, such as historical performance in similar scenarios based on validation set performance, current match conditions, cluster membership, match progression, and individual model confidence metrics. The weighting algorithm incorporates both linear and non-linear components, and a hyperbolic tangent transformation for cluster membership strength.

4.2 Training methodology and real-time optimization

Model training followed a structured approach to prevent overfitting and ensure generalization. The dataset was split temporally, reserving the most recent 20% matches for testing to mirror real-world prediction scenarios. Within training, a sliding window validation method maintained chronological order, preventing data leakage and preserving the evolving nature of team performance.

Hyperparameter optimization was performed through an extensive grid search process, balancing accuracy and computational efficiency. For gradient boosting models, we tuned key parameters, including learning rates (optimal range 0.005-0.01), tree depths (optimal values 5-7), regularization terms (alpha = 0.1, lambda = 1) and sampling strategies (subsample = 0.85, colsample_bytree = 0.85). These optimizations were crucial for controlling the complexity of the model without sacrificing predictive power. The neural network architecture was refined through iterative experimentation, resulting in a configuration with four hidden layers (512, 256, 128, 64 nodes) using tanh activation functions. Regularization was implemented through dropout (rate = 0.2) and L2 penalty (alpha=0.0005). Early stopping (patience=10 epochs) and the Adam optimizer (learning rate=0.001) with batch size optimization (64 samples) ensured stable convergence while preventing overfitting.

The real-time prediction system processes match data in stages, initially relying on static features (team strength, form, and odds) and integrating in-play statistics through time-scaling adjustments. For example, shot counts at 30 minutes are scaled to full-time equivalents, with nonlinear adjustments for stats that typically accelerate later in matches. Feature engineering during live prediction generates momentum indicators, efficiency ratios, game dominance indices, and draw likelihood estimates along the probabilities of the match cluster.

During implementation, we found that employing separate models for home and away goals, rather than a single multi-output predictor, improved accuracy by allowing the system to capture the asymmetric nature of scoring patterns in football.

4.3 Evaluation Metrics

System performance was evaluated using multiple metrics: accuracy and F1-score for categorical predictions (match outcomes), and MSE and MAE alongside R^2 for numerical predictions (goal totals). The Brier score assessed the accuracy of

predicted probabilities, while ROC-AUC measured the model's ability to distinguish between outcomes.

We chose a comprehensive evaluation approach to address the multifaceted nature of football prediction. Accuracy provides an intuitive success rate, while F1-score balances precision and recall, crucial given the class imbalance (fewer draws). For numerical predictions, MSE penalizes large deviations, and MAE offers an interpretable error scale. R^2 quantifies the proportion of outcome variance explained.

The Brier score evaluates probabilistic calibration-essential for applications where accurate probability estimates are more valuable than binary classifications. ROC-AUC assesses discrimination ability across decision thresholds, showing the model's capacity to separate outcomes regardless of the chosen threshold. This multimetric approach ensures we capture both practical predictive utility and theoretical model performance, as well as its strengths and weaknesses.

5 Results

5.1 Match Result Prediction Performance

To evaluate the effectiveness of our framework, we benchmarked various prediction approaches against our proposed model. Table 2 presents performance metrics for both baseline methods and machine learning models in prediction of three-way match outcome (home win, draw, away win).

Table 2. Match Result Prediction Performance

Method	Accuracy	F1	Home Acc.	Draw Acc.	Away Acc.
Baseline Methods					
Random Prediction	0.334	0.328	0.333	0.334	0.335
Always Predict Home Win	0.444	0.205	1.000	0.000	0.000
Odds-Weighted Prediction	0.480	0.446	0.562	0.312	0.421
Elo-Weighted Prediction	0.473	0.447	0.553	0.319	0.413
Form-Weighted Prediction	0.410	0.369	0.512	0.231	0.392
Half-time Result	0.581	0.569	0.622	0.547	0.526
Machine Learning Models					
Logistic Regression	0.646	0.567	0.743	0.419	0.538
LightGBM	0.648	0.610	0.739	0.443	0.570
XGBoost	0.638	0.587	0.736	**0.458**	0.574
Neural Network	0.653	0.614	0.748	0.442	0.573
Meta model	**0.674**	**0.635**	**0.763**	0.443	**0.612**

The results demonstrate a clear performance hierarchy, with machine learning approaches substantially outperforming baseline methods. Among the baseline

approaches, the use of half-time results achieved the highest accuracy (58.1%), highlighting the significant predictive value of the in-play data.

In the machine learning category, our cluster-supported meta-model dominated, achieving 67.4% overall accuracy and 63.5% F1 score. This represents a remarkable 9.3 percentage point improvement over the best baseline method. The cluster-enhanced neural network also demonstrated superior performance in predicting home wins (76.3% accuracy) and away wins (61.2% accuracy), while XGBoost showed particular strength in identifying draws (45.8% accuracy) - traditionally the most challenging outcome to predict in football.

5.2 Exact Score Prediction Performance

Predicting exact match scores represents a significantly more challenging task than three-way outcome prediction. We evaluated various approaches to this task, with the results presented in Table 3.

Table 3. Score Prediction Performance

Model	Exact Score Acc.	R^2	MAE	Brier
Poisson Distribution	0.190	0.426	0.642	0.190
Linear Regression	0.238	0.492	0.575	0.174
LightGBM	0.237	0.493	0.575	0.177
Neural Network	0.244	0.489	0.573	0.174
Ensemble (NN+XGB+LGB)	**0.267**	**0.510**	**0.550**	**0.167**

The ensemble approach combining neural networks with gradient-boosted trees demonstrated superior performance across all metrics, achieving 26.7% accuracy for exact score prediction. This represents a 7.7 percentage point improvement over the widely used Poisson distribution model (19.0%), which is the traditional standard in football score modeling. The ensemble's R^2 value of 0.510 indicates that it explains more than half of the variance in goal scoring, a significant achievement given football's inherent unpredictability.

The mean absolute error (MAE) of 0.550 achieved by the ensemble approach indicates that on average the predictions were within approximately half a goal of the actual score per team. The lower Brier score 0.167 further confirms that the ensemble model produced well-calibrated probability estimates across different score lines.

The performance gap between the Poisson distribution and machine learning models underscores the limitations of traditional statistical approaches that assume goal independence and rely primarily on team strength indicators. Our ensemble framework addresses these limitations by incorporating match dynamics, tactical considerations, and specialized model components for different scoring patterns.

6 Discussion

The results demonstrate our multi-layered framework's effectiveness, significantly improving football match prediction accuracy compared to traditional approaches and remaining competitive with state-of-the-art methods. This stems from effectively integrating pre-match and in-play data through specialized Level 1 models and our dynamic meta-learning approach. Our ensemble approach for score prediction substantially outperformed traditional methods, achieving 26.7% exact score accuracy versus the commonly used Poisson distribution model's 19.0%, representing a 40.5% relative improvement in predictive power.

Despite these strengths, our approach has several important limitations. The multi-layered architecture introduces significant computational complexity, particularly for real-time applications. The performance of the system is highly dependent on the quality of the data, which varies considerably between leagues and seasons. The modest overall accuracy improvement of the meta-model (+0.3% to +2.5%) also suggests potential diminishing returns from architectural complexity and indicates opportunities for more efficient model design. The system furthermore demonstrates inconsistent performance across different match phases, with predictions in the first 15 minutes of matches showing substantially higher error rates than those made at half-time or later.

The performance of our framework compares favorably with that of the existing literature. Traditional statistical methods typically achieve an accuracy of 45-55% for three-way predictions [2] [3], a benchmark that we surpass. Compared to machine learning approaches, our results are competitive, with our logistic regression matching Igiri's [4] 64.6% and our gradient boosting implementation (64.8%) aligning with Yao et al.'s [8] reported levels. Muszaidi et al. [6] reported 67-78% accuracy for goal difference prediction, slightly higher than our 67.4% for three-way outcomes, though direct comparison is limited by differing tasks and datasets.

Our future work will focus on enhancing performance through Bayesian hyperparameter optimization instead of grid search. Preliminary tests suggest that the automated architecture search could increase accuracy by 1-3%, particularly for cluster-specific models. Computational efficiency could be improved through model distillation and quantization, enabling deployment on edge devices.

We plan to explore sophisticated ensemble strategies, including hierarchical stacking and attention mechanisms that dynamically weight models based on match conditions. Implementing online learning would allow continuous improvement during deployment. Reinforcement learning could optimize prediction timing and confidence thresholds. The match clustering methodology could benefit from deep embedding approaches and self-supervised learning to identify more nuanced tactical patterns. Improving the framework's explainability would make predictions more actionable for stakeholders.

Although our framework demonstrates significant improvements over baseline approaches, substantial opportunities remain to enhance its accuracy, efficiency, and applicability in diverse football contexts.

7 Conclusion

This paper presented a multi-layered framework for football match prediction that successfully integrates pre-match data with in-play statistics through a hierarchical ensemble architecture. Our approach improved accuracy, with the neural network model achieving more than 67% for match outcomes and the ensemble reaching 26.7% for exact score prediction. The novel match clustering methodology proved valuable, identifying six distinct game patterns and enabling specialized prediction strategies that improved performance. Although the system shows promising results, computational complexity remains a challenge for real-time deployment in resource-constrained environments. Future work will focus on model compression techniques, online learning capabilities, and enhanced feature engineering to further improve both prediction accuracy and operational efficiency. This framework advances the field of sports analytics by demonstrating how specialized model ensembles can effectively capture the complex and dynamic nature of football matches.

References

1. Lolli, L., Bauer, P., Irving, C., Bonanno, D., Höner, O., Gregson, W., Di Salvo, V.: Data analytics in the football industry: A survey investigating operational frameworks and practices in professional clubs and national federations from around the world. Science and Medicine in Football, **2024**, 1–10 (2024). https://doi.org/10.1080/24733938.2024.2341837
2. Dixon, M.J., Coles, S.G.: Modelling Association Football Scores and Inefficiencies in the Football Betting Market. Journal of the Royal Statistical Society Series C: Applied Statistics **46**(2), 265–280 (2002). https://doi.org/10.1111/1467-9876.00065
3. Hvattum, L.M., Arntzen, H.: Using ELO ratings for match result prediction in association football. International Journal of Forecasting **26**(3), 460–470 (2010).
4. Igiri, C.: An Improved Prediction System for Football Match Result. IOSR Journal of Engineering **04**, 12–020 (2014). https://doi.org/10.9790/3021-04124012020
5. Meng, X.: Soccer match outcome prediction with random forest and gradient boosting models. Applied and Computational Engineering **40**(1), 99–107 (2024). https://doi.org/10.54254/2755-2721/40/20230634
6. Muszaidi, M., Mustapha, A., Ismail, S., Razali, N.: Deep Learning Approach for Football Match Classification of English Premier League (EPL) Based on Full-Time Results. In: Applications of Science and Mathematics 2021, pp. 339–350. Springer, Singapore (2022). https://doi.org/10.1007/978-981-16-8903-1-30
7. Stübinger, J., Mangold, B., Knoll, J.: Machine Learning in Football Betting: Prediction of Match Results Based on Player Characteristics. Applied Sciences **10**(1), 46 (2020). https://doi.org/10.3390/app10010046
8. Yao, W., Wang, Y., Zhu, M., Cao, Y., Zeng, D.: Goal or Miss? A Bernoulli Distribution for In-Game Outcome Prediction in Soccer. Entropy **24**(7), 971 (2022). https://doi.org/10.3390/e24070971
9. Yi, Q., Gómez, M.A., Wang, L., Huang, G., Zhang, H., Liu, H.: Technical and physical match performance of teams in the 2018 FIFA World Cup: Effects of two different playing styles. Journal of Sports Sciences **37**(22), 2569–2577 (2019). https://doi.org/10.1080/02640414.2019.1648120

10. Moura, F.A., Martins, L.E.B., Cunha, S.A.: Analysis of football game-related statistics using multivariate techniques. Journal of Sports Sciences **32**(20), 1881–1887 (2014). https://doi.org/10.1080/02640414.2013.853130
11. Diquigiovanni, J., Scarpa, B.: Analysis of association football playing styles: An innovative method to cluster networks. Statistical Modelling **19**(1), 28–54 (2018). https://doi.org/10.1177/1471082x18808628
12. Chen, T., Guestrin, C.: XGBoost: A Scalable Tree Boosting System. arXiv preprint arXiv:1603.02754 (2016). https://doi.org/10.48550/ARXIV.1603.02754
13. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.: LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In: Advances in Neural Information Processing Systems, vol. 30. Curran Associates, Inc. (2017)
14. Football Data Download, https://www.football-data.co.uk/downloadm.php. Last accessed 1 Mar 2025
15. Club Elo, http://clubelo.com/. Last accessed 1 Mar 2025
16. Club Football Match Data 2000-2025, https://github.com/xgabora/Club-Football-Match-Data-2000-2025. Last accessed 1 Mar 2025
17. Club Football Match Data 2000-2025, https://www.kaggle.com/datasets/adamgbor/club-football-match-data-2000-2025. Last accessed 1 Mar 2025
18. Club Football Match Data 2000-2025, https://huggingface.co/datasets/xgabora/Club-Football-Match-Data-2000-2025. Last accessed 1 Mar 2025