Prediction System for National Football Team Match Results

Gustavo C. Bento Laurindo[1], **Ivan Fernando R. Villalba**[1], **Murilo C. Naldi**[1]

[1]Departamento de Computação – Universidade Federal de São Carlos (UFSCar)
P.O. Box 676 – 13.565-905 – São Carlos – SP – Brazil

{gustavocbl,ivanrizzi}@estudante.ufscar.br, naldi@ufscar.br

Abstract. *This work presents a supervised learning system for predicting the outcomes of football matches between national teams, classifying the results as home win, draw, or away win. Three classifiers were evaluated: XGBoost, Logistic Regression, and an MLP Neural Network. Records from 1872 to 2026 were used to compute Elo ratings and performance indicators, with learning applied to matches from 2015 onward, enriched with overall FIFA video game player ratings aggregated by national team. Twelve configurations per model were tested using Rolling and Expanding Window validation. The results indicate that Elo-based features are the most informative predictors. XGBoost achieved the best accuracy and probabilistic calibration, whereas the balanced MLP achieved the highest macro F1-score and the best performance in drawing identification.*

1. Introduction

Predicting the outcomes of football matches is a classical supervised classification problem that has been extensively studied in the machine learning literature. Due to the inherently probabilistic nature of the sport, modeling match outcomes involves significant uncertainty, making it a relevant challenge from both theoretical and practical perspectives. In this work, three classification techniques — XGBoost, Logistic Regression, and Multi-Layer Perceptron (MLP) Neural Networks — are evaluated with the objective of predicting the results of matches between national teams, categorizing them as home win, draw, or away win. The objective of this study is to compare the performance of these approaches in a real-world scenario of sports prediction.

The main contribution of this work lies in the data engineering process employed for football match prediction, which incorporates FIFA video game ratings in combination with attributes such as Elo ratings. Furthermore, a comprehensive analysis of national teams historical match records was conducted from a broad perspective, encompassing multiple international competitions rather than focusing on a single tournament.

It should be emphasized that the gain obtained by the proposed models over the Elo-based baseline lies mainly in probabilistic calibration and in the identification of draws, rather than in raw accuracy, which improves only modestly. It is also worth noting that the FIFA `overall` ratings used in this work reflect individual player quality and are therefore subject to an inherent distortion when used to represent a team, since they do not capture the tactical cohesion and on-field understanding among players, a limitation that is particularly relevant for national teams, whose squads train together far less often than club sides.

The methodological details, as well as the complete description of the experiments and the analysis of the results, are presented in the following sections. Additionally, the Jupyter Notebook used in this work, including references to the original datasets, is available in the repository [Laurindo et al. 2026].

2. Related Work

The prediction of football match outcomes using different classification techniques has been extensively explored in the literature. The studies most closely related to this work can be organized around two main topics: the use of rating systems to represent team strength and the application of machine learning techniques combined with feature engineering.

[Hvattum and Arntzen 2010] used the Elo rating system to predict association football match outcomes, constructing covariates from score differences and feeding them into ordered logistic regression models, and compared this approach with naïve benchmarks and betting-market implied probabilities. [Baboota and Kaur 2019] built a predictive system for the English Premier League using feature engineering and machine learning (Gradient Boosting, Random Forest, SVM), notably incorporating FIFA-based player attributes (attack, midfield, defense, overall scores) as predictors. Also for the Premier League, [Rodrigues and Ângelo Pinto 2022] combined historical match statistics with FIFA-derived player attributes to develop classification models, evaluating eight algorithms (including Naïve Bayes, KNN, SVM, Random Forest, XGBoost, Neural Networks) and applying a two-stage feature selection strategy. More recently, [Przybyła-Kasperek and Wesecki 2025] examined Premier League outcome prediction with ensemble methods and explicit feature selection, comparing Random Forest, XGBoost, and LightGBM with embedded feature selection against an SVM classifier paired with LASSO, PCA, RFE, and ANOVA-based feature selection.

Unlike prior studies, this work focuses on national teams, which introduces challenges such as fewer matches per team, greater competition heterogeneity, and a lack of reliable in-game performance metrics. It also uniquely combines two information sources usually studied separately: Elo ratings and FIFA video game data.

3. Methodology

3.1. Dataset and Exploratory Analysis

The dataset integrates three data sources. The main source is a historical record of international football matches since 1872 [Jürisoo 2023], including date, teams, scoreline, competition type, and location. Structured FIFA World Cup data [Dahatonde 2023] were added to identify the match stage (group stage, round of 16, quarter-finals, semi-finals, final). Player ratings from the FIFA video game series [Capital 2024, Nyagami 2024, Rovnez 2025] were used to estimate national team strength by field area. The target variable is *result*: home win (0), draw (1), or away win (2).

Based on these sources, a set of pre-match features was constructed without future information leakage, using only data available before each match. These features include home and away Elo ratings and their difference (`elo_diff`); average goals scored and conceded over the previous five matches; win rate over the previous ten matches

(three points for a win, one for a draw); average goal difference; and indicators of recent momentum, such as cumulative Elo variation over 20- and 80-match windows. Contextual variables include home-field advantage, competition-type indicators (FIFA World Cup and continental competitions), and the `stage` variable, which classifies matches as friendlies, continental tournaments, Nations League, or qualification games. FIFA `overall` ratings were also aggregated at the national-team level using three criteria: the top 23 eligible players, the top 11 highest-rated players, and the top 5 players in each area (attack, midfield, defense).

The Elo rating system follows the methodology available at `https://www.footballdatabase.com/methodology`, updated after each match according to:

$$R' = R + K \cdot G \cdot (W - W_e), \quad W_e = \frac{1}{10^{-\Delta R/400} + 1} \tag{1}$$

where K ranges from 10 (friendly matches) to 50 (FIFA World Cup matches), G is the goal-difference factor, W is the observed result, and W_e is the expected result given ΔR, the Elo rating difference between the two teams. This feature is particularly informative because it implicitly reflects the strength of the team through match outcomes, goal margins, and the importance of competition.

Four analyses were performed on the constructed *dataset*. The first, shown in Figure 1, is a correlation heatmap of the numerical attributes. For visualization, variables were abbreviated: H and A for home and away teams, and the suffixes ELO, WR, GB, and M for Elo rating, win rate, goal balance, and recent momentum. Most attribute correlations are low to moderate, with some redundancy among variables derived from the same measures. The feature `elo_diff` has the highest absolute correlation with `result`, making it the most informative attribute for the classification task.

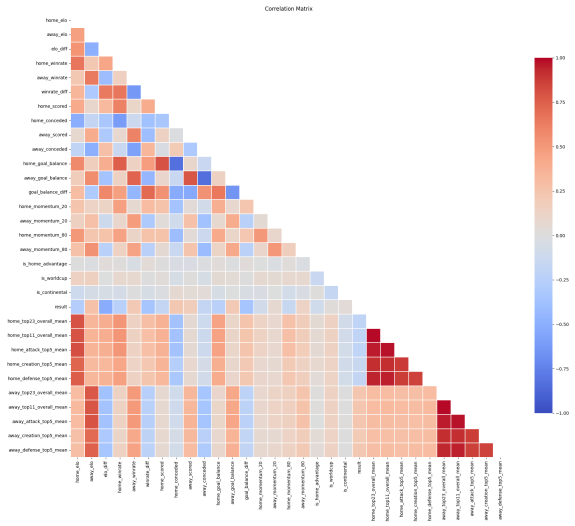


Figure 1. Correlation heatmap of the numerical attributes in the dataset.

The distribution of match outcomes (Figure 2) confirms class imbalance in the dataset: home wins are most frequent, followed by away wins, with draws being rarest. Because this mirrors real-world football results, no resampling was applied and the original class distribution was preserved.

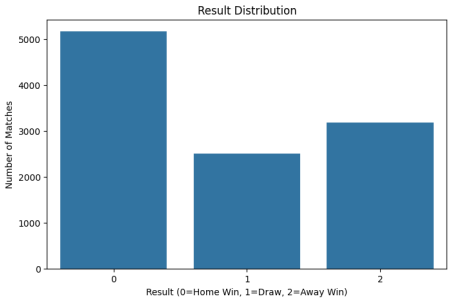


Figure 2. Distribution of outcome classes.

The boxplots of `elo_diff` and `winrate_diff` (Figure 3) show medians near zero, roughly symmetric distributions, and outliers in both directions, as expected given football's probabilistic nature.



Figure 3. Boxplots of `elo_diff` and `winrate_diff`.

Finally, the scatter plot (Figure 4) relating `elo_diff` and `winrate_diff` indicates a certain degree of separation between home-win and away-win outcomes, whereas draws are concentrated in the intermediate region and exhibit substantial overlap with the other classes.

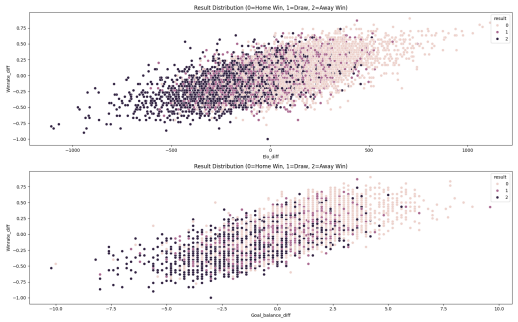


Figure 4. Scatter plot of `elo_diff` and `winrate_diff` by outcome class.

This pattern is expected, as matches between teams of similar strength tend to be more balanced. Therefore, distinguishing draws from the remaining classes constitutes one of the main challenges of the prediction problem.

3.2. Supervised Learning

The experiments were conducted in a Jupyter Notebook using Python with Scikit-learn, XGBoost, Pandas, NumPy, and Matplotlib. Because the data are time-dependent,

train–test splits were chronological to prevent information leakage. Two validation strategies were used: *Rolling Window* and *Expanding Window*. In the *Rolling Window* approach, the model is trained on a fixed 4-year window and tested on the following 2 years, shifting the window by 1 year at each fold to emphasize recent team performance. In the *Expanding Window* approach, training starts with 4 years of data and expands every 2 years to include all available history. In both strategies, features are standardized (`StandardScaler`) within each fold, fitting the scaler only on the training data.

Three classification models were evaluated. *XGBoost* is an ensemble of decision trees trained via gradient boosting with `n_estimators=150`, `max_depth=2`, `learning_rate=0.05`, `subsample=0.6`, and `colsample_bytree=0.8`. The multiclass *Logistic Regression* model uses softmax with parameters estimated by log-loss minimization (`C = 1.0`, `solver = lbfgs`, `max_iter = 1000`). The *MLP Neural Network* is a Multi-Layer Perceptron with three hidden layers (128, 64, and 32 ReLU units), optimized with Adam, L2 regularization (`alpha = 1e-4`), a learning rate of 0.001, and early stopping (patience 20, up to 500 epochs). All used a fixed random seed of 42.

For each classifier, 12 configurations were tested to assess the contribution of different features to predictive performance. Three feature sets were used: (i) all features, including One-Hot Encoding of `stage`; (ii) only `elo_diff`, identified as the variable most correlated with `result`; and (iii) all features excluding Elo-based variables, to isolate the impact of recent-performance attributes and FIFA ratings. Two temporal validation strategies were combined with two balance conditions of the class (no adjustment or weight of 1.4 for the draws by `sample_weight`). The metrics included precision, precision, recall, per-class F1, macro F1, weighted F1, and log-loss, calculated per fold and averaged.

4. Results

Before presenting the results, note that RW and EW refer to the *Rolling Window* and *Expanding Window* temporal validation strategies, respectively, and that "+ bal" denotes configurations where class balancing was applied to the draw class.

4.1. Simple Classifier Based on *elo_diff*

As an initial benchmark, a simple classifier based solely on the Elo rating difference between home and away teams (*elo_diff*) was evaluated on 10,835 matches from 2015 to 2026. It uses only the sign of the rating difference: if the home team's Elo rating is higher, it predicts a home win; otherwise, an away win. Table 1 reports the overall performance metrics for this period.

This rule-based classifier requires no training, so it was not evaluated with the Rolling or Expanding Window schemes of Section 3.2 used for learned models. Its predictions were computed over the full 2015–2026 period and broken down by calendar year to assess temporal variability. The classifier achieved an overall accuracy of 58.41%, well above random guessing (33.3%). However, the classification report in Table 2 reveals strong asymmetry in its predictions, indicating substantial room for improvement.

The model cannot predict draws (class 1), yielding zero precision and recall for this class. This stems from the decision rule: since the classifier uses only the sign of

Table 1. Global performance metrics of the baseline classifier (2015–2026)

Metric	Value
Accuracy	0,5841
Macro F1-score	0,4343
Weighted F1-score	0,5110
Log-Loss	14,9896

Table 2. Baseline classifier performance by class

Class	Precision	Recall	F1-score	Support
Home Win (0)	0,66	0,76	0,71	5.162
Draw (1)	0,00	0,00	0,00	2.501
Away Win (2)	0,49	0,75	0,60	3.172

elo_diff, it never assigns the draw class. Thus, all 2,501 drawn matches (23.1% of the sample) are labeled as home or away wins. This bias explains the high recall for home wins (76%, as most home victories occur when the home team has a higher Elo rating) and for away wins. The high Log-Loss (14.99) results from the model's deterministic probabilities (0 or 1), which severely penalize wrong predictions made with full confidence, as expected from a non-probabilistic baseline. A temporal analysis (Table 3) shows substantial year-to-year variation: the best accuracy was 62.69% in 2021 and the worst 53.18% in 2018. The 2021 accuracy approaches the informational upper bound of the `elo_diff` signal, indicating that the relation between rating difference and match outcome was particularly linear that year.

Table 3. Year-by-year accuracy of the baseline classifier (2015–2026)

Year	Acc.	Year	Acc.	Year	Acc.
2015	0,6054	2019	0,6094	2023	0,6063
2016	0,5750	2020	0,5245	2024	0,5858
2017	0,5422	2021	0,6269	2025	0,6058
2018	0,5318	2022	0,5604	2026	0,5455

The global confusion matrix (Table 4) shows that 1,224 home wins were incorrectly classified as away wins, while 781 away wins were incorrectly classified as home wins. No draws were correctly identified, which is fully consistent with the binary decision rule adopted by the classifier.

4.2. XGBoost

According to the results presented in Figure 5, the best accuracy performance was achieved by the Complete + Expanding configuration (60.93%), a value slightly higher than that of the Elo-based baseline (58.41%). However, the most informative metric for this imbalanced classification problem is the macro F1-score.

Table 4. Confusion matrix of the baseline classifier

Actual \Predicted	Home Win (0)	Draw (1)	Away Win (2)
Home Win (0)	3938	0	1224
Draw (1)	1280	0	1221
Away Win (2)	781	0	2391

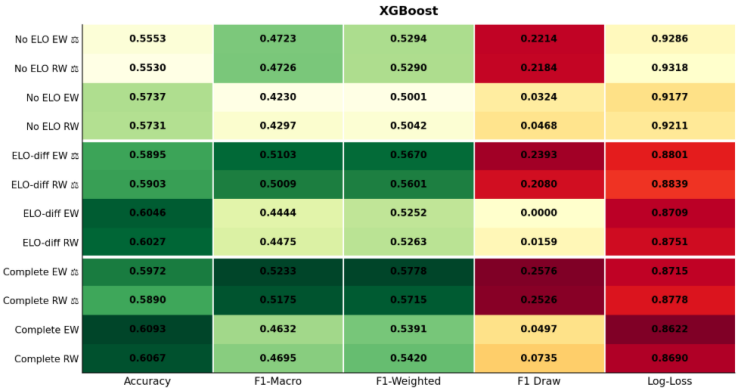


Figure 5. Heatmap of the overall performance of the XGBoost configurations.

From this perspective, the configurations with class balancing achieved substantially better performance: the Complete + Expanding + Balanced configuration obtained a macro F1-score of 0.5233, compared with only 0.4632 for the unbalanced version. The improvement is even more pronounced in draw detection (class 1), whose draw F1-score increased from 0.0497 (without balancing) to 0.2576 (with balancing).

Class balancing lowers overall accuracy by about one percentage point across configurations, as seen for Complete + Expanding (60.93% to 59.72%). This trade-off is acceptable, since improving draw prediction (the minority class) was the main challenge. Removing all Elo-related variables clearly harms performance: the best No Elo setup reaches a macro F1-score of 0.4726, versus 0.5233 for the complete version. Thus, Elo ratings provide predictive information not captured by the other variables. Table 5 reports the class-wise performance of these configurations.

Table 5. Class-wise performance for the best configurations

Configuration	F1-C0	F1-C1	F1-C2	Macro F1
Complete + Expanding	0,7260	0,0512	0,6129	0,4634
Complete + Expanding + Bal	0,7238	0,2603	0,5819	0,5220
No Elo + Rolling + Bal	0,6931	0,2208	0,5116	0,4752

To complement the XGBoost analysis, Figure 6 presents the plot of the importance of the feature according to the average gain criterion, calculated from the models trained in the experimental folds (without class balancing). The gain criterion represents the average contribution of each feature to reducing the impurity of the decision-tree leaves and is generally considered more informative than simply counting how often a feature is used. The plot shows that `elo_diff` is the most important feature of the model.

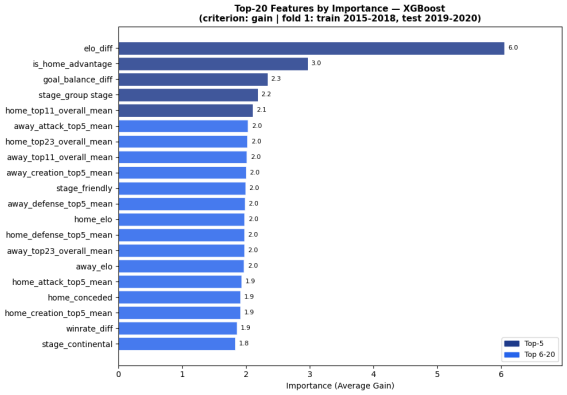


Figure 6. Feature importance according to average gain.

4.3. Logistic Regression

From Figure 7, the Complete Features + Rolling Window + No Balancing configuration achieved the highest accuracy (60.47%). However, as in the previous case, the macro F1-score is the most appropriate metric due to class imbalance.

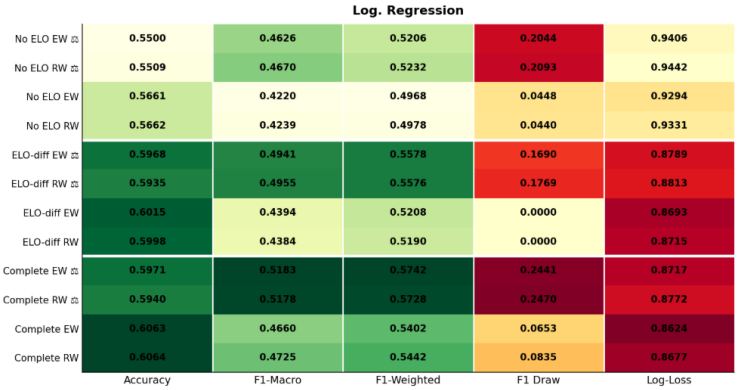


Figure 7. Heatmap of the overall performance of the Logistic Regression configurations.

Furthermore, class balancing substantially increases the macro F1-score, from 0.4725 to 0.5178, representing a relative improvement of 9.6%. This gain comes at the cost of a small reduction in accuracy (from 60.6% to 59.4%). In addition, Table 6 shows that the F1-score for draws (class 1) increased from 0.0835 to 0.2470 when balancing was applied, representing a significant improvement.

Table 6. Class-wise performance for the best configurations

Configuration	F1-C0	F1-C1	F1-C2	Macro F1
Complete + Rolling	0,7262	0,0835	0,6078	0,4725
Complete + Rolling + Bal	0,7224	0,2470	0,5839	0,5178
Elo Only + Rolling + Bal	0,7226	0,1769	0,5869	0,4955

4.4. Neural Networks

Figure 8 shows that the neural network without class balancing achieved a maximum accuracy of 60.36%, a value similar to those obtained by the linear models. However, when balancing was applied, the network's macro F1-score (0.5271) slightly outperformed both Logistic Regression (0.5183) and XGBoost (0.5233), indicating that the greater modeling capacity of neural networks becomes advantageous when class imbalance is addressed. Table 7 shows that the F1-score for draws (class 1) increased from 0.1234 to 0.2840 with balancing, representing the best performance among all models evaluated for this class.

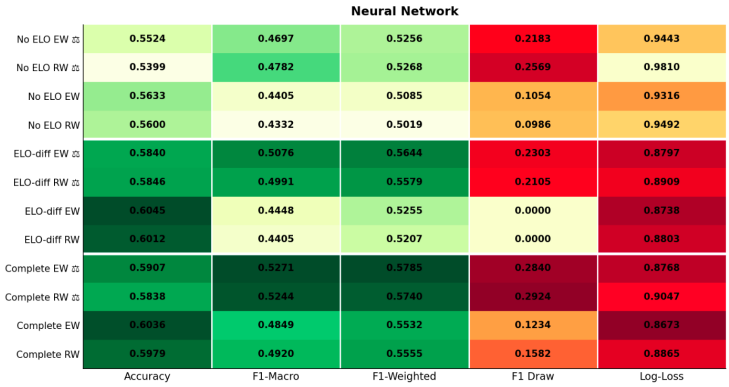


Figure 8. Heatmap of the overall performance of the Neural Network model configurations.

Table 7. Class-wise performance for the best configurations

Configuration	F1-C0	F1-C1	F1-C2	Macro F1
Complete + Expanding	0,7263	0,1234	0,6050	0,4849
Complete + Expanding + Bal	0,7192	0,2840	0,5780	0,5271
Complete + Rolling + Bal	0,7138	0,2924	0,5670	0,5244

4.5. Model Comparison

The Neural Network achieved the highest macro F1-score (0.5271) and the best draw prediction (F1 = 0.2840) among all models. Compared to logistic regression, this is a 1.7% gain in macro F1-score and 16.4% in draw F1-score. All supervised models clearly outperformed the Elo sign baseline. The best model (Balanced Neural Network) improved the macro F1-score by 15.6% over baseline (0.5271 vs. 0.4559). Unlike the baseline, which could not predict draws, the neural network achieved a draw F1-score of 0.2840.

Figure 9 shows that XGBoost and Logistic Regression are the most robust general-purpose models, with the best accuracy and probability calibration, but they struggle to detect the Draw class. In contrast, the Neural Network better balances all classes and identifies specific outcomes, achieving the highest macro F1-score and clearly outperforming the other models on draw prediction.

Figure 10 illustrates the relationship between the macro F1-score and the F1 draw score. The results confirm that models trained with class balancing achieve better per-

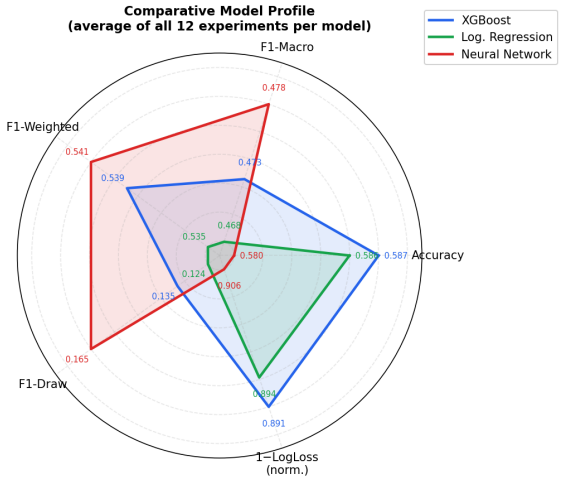


Figure 9. Radar chart comparing the evaluated models.

formance in detecting draws. In addition, a positive correlation can be observed between macro F1-score and draw F1-score, indicating that improvements in overall class-balanced performance tend to enhance draw identification as well. The Neural Network achieved the best trade-off between overall predictive performance and draw detection.

It should be noted that the gain-based feature importance shown in Figure 6 is a native property of tree ensembles and is therefore not directly available for the Logistic Regression and Neural Network models. Nevertheless, the comparison between the three sets of characteristics used throughout this study (complete features, `elo_diff` only, and non-Elo features) provides an interpretability analysis that applies to all three models and points to their most influential attributes. This analysis confirms that Elo ratings remain the dominant predictor regardless of which model is regarded as best: removing all Elo-derived variables reduces the balanced Neural Network's macro F1-score from 0.5271 to 0.4782, a drop comparable in magnitude to the one observed for XGBoost (from 0.5233 to 0.4726). This consistency across model families reinforces the conclusion, drawn from the gain-based analysis, that the Elo rating difference is the single most influential attribute in this prediction task.

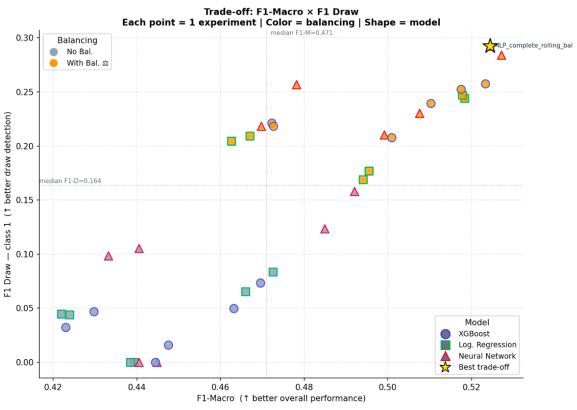


Figure 10. Macro F1-score versus Draw F1-score.

5. Conclusion

This study compared three machine learning methods (XGBoost, Logistic Regression, and an MLP Neural Network) in 12 experimental setups to predict football match outcomes, using an Elo-based naive classifier as the baseline. All supervised models slightly improved global accuracy (60.5–60.8% vs. 58.4%), but the main gain was in probabilistic calibration: Log-Loss decreased from 15 to 0.87, indicating that the models effectively learned to distribute probabilities between the three classes.

The configuration analysis yielded three main insights. First, Elo-based information is the dominant predictor. As Figures 5, 7 and 8 show, using only `elo_diff` already yields accuracy and macro F1 close to the full feature set, with other attributes adding only 0.013–0.024 macro F1, mainly for draws. In contrast, removing all Elo-derived variables reduces accuracy by 3–4 percentage points and macro F1 by 0.03–0.06, confirming that Elo-based information is the main source of predictive power. Second, class balancing is crucial for the three-class task: without it, draw recall stays below 11% in all models. Third, the three algorithms perform very similarly, indicating an approximately linear decision boundary; XGBoost's advantage over Logistic Regression comes mainly from capturing residual non-linear interactions.

The models show complementary strengths. XGBoost and Logistic Regression maximize overall accuracy and temporal stability, making them suitable when the goal is to predict the most likely outcome. The Neural Network achieves the highest draw recall and best macro F1, so it is preferable when detecting the minority class is more important. Comparison with the literature is difficult, as prior work focuses on club leagues with many yearly matches, while this study covers national teams, which play fewer games. This lower match frequency affects models that are sensitive to class distribution.

Despite these constraints, the results align with the literature. [Baboota and Kaur 2019] reported a maximum F1-score of 0.31 for the draw class using Gradient Boosting, while in this work the MLP neural network achieved 0.2924 for the same class. This suggests that combining Elo ratings with aggregated FIFA features partly offsets the scarcity of national-team data. However, as noted in the Introduction, this compensation is limited by the tactical-cohesion distortion in FIFA `overall` ratings, which reflect individual quality but not players' on-field understanding.

The value of combining historical and player-related information extends beyond football. In basketball, [Papageorgiou et al. 2024] found that integrating historical and player-level data improves predictive performance, and combining Elo ratings with aggregated FIFA attributes yields competitive results for national teams. Similar patterns appear in tennis, where [Gao and Kowalczyk 2021] identified serve strength as the key predictor of match outcome. Likewise, Elo difference was the dominant predictor in this study, indicating that sports forecasts often rely on a small set of highly informative variables.

This study's results align with recent work on football prediction. Rodrigues et al. [Rodrigues and Ângelo Pinto 2022] reported 61.32% accuracy before feature selection and 65.26% after a Boruta-RFE pipeline. Although datasets and experimental setups differ, the models here achieved 60.5–60.8% accuracy without explicit feature selection. Given the lower match frequency and higher variability in national-

team competitions, these results are consistent with the literature and show that combining Elo ratings, aggregated FIFA attributes, and contextual variables provides a strong representation for predicting match outcomes. Future work may examine feature-selection methods, which can enhance performance and interpretability in sports prediction [Przybyła-Kasperek and Wesecki 2025], particularly for heterogeneous numerical and categorical data [Warkiani and Moattar 2025]. The proposed framework is not restricted to national teams, aligns with club-football findings, and appears adaptable to other sports prediction tasks using historical and player-level data.

References

Baboota, R. and Kaur, H. (2019). Predictive analysis and modelling football results using machine learning approach for english premier league. *International Journal of Forecasting*, 35(2):741–755.

Capital, J. B. (2024). Fifa players – fifa 15 to fifa 24. Kaggle dataset. Accessed: 2026.

Dahatonde, A. (2023). Fifa world cup all dataset. Kaggle dataset. Accessed: 2026.

Gao, Z. and Kowalczyk, A. (2021). Random forest model identifies serve strength as a key predictor of tennis match outcome. *Journal of Sports Analytics*, 7(4):255–262.

Hvattum, L. M. and Arntzen, H. (2010). Using elo ratings for match result prediction in association football. *International Journal of Forecasting*, 26(3):460–470. Sports Forecasting.

Jürisoo, M. (2023). International football results from 1872 to 2017. Kaggle dataset. Accessed: 2026.

Laurindo, G. C. B., Villalba, I. F. R., and Naldi, M. C. (2026). Prediction System for National Football Team Match Results: Jupyter notebook and final dataset. `https://github.com/gustavocblaurindo/National-Teams-Football-Prediction.git`. Accessed: 2026.

Nyagami (2024). Ea sports fc 25 database: Ratings and stats. Available on Kaggle. Accessed: 2026.

Papageorgiou, G., Sarlis, V., and Tjortjis, C. (2024). Evaluating the effectiveness of machine learning models for performance forecasting in basketball: a comparative study. *Knowledge and Information Systems*, 66(7):4333–4375.

Przybyła-Kasperek, M. and Wesecki, M. (2025). A comparative study of ensemble methods and feature selection techniques for predicting english premier league match outcome. *Procedia Computer Science*, 270:4822–4829. 29th International Conference on Knowledge-Based and Intelligent Information Engineering Systems (KES 2025).

Rodrigues, F. and Ângelo Pinto (2022). Prediction of football match results with machine learning. *Procedia Computer Science*, 204:463–470. International Conference on Industry Sciences and Computer Science Innovation.

Rovnez (2025). Fc 26 / fifa 26 player data. Available on Kaggle. Accessed: 2026.

Warkiani, M. E. and Moattar, M. H. (2025). A comprehensive survey on recent feature selection methods for mixed data: Challenges, solutions and future directions. *Neurocomputing*, 623:129372.