ENHANCED KNOWLEDGE IN SCIENCES AND TECHNOLOGY
e-ISSN: 2773-6385

EKST

Vol. 5 No. 2 (2025) 492-504
https://publisher.uthm.edu.my/periodicals/index.php/ekst

Football Match Outcome Prediction Based on Team Rating System

Siti Munawarah Mohd Din[1], Logenthiran Machap[1*]

[1] *Department of Mathematics and Statistics, Faculty of Applied Sciences and Technology, UTHM Kampus Cawangan Pagoh, Hab Pendidikan Tinggi Pagoh, KM 1, Jalan Panchor, 84600, Pagoh, Muar, Johor, MALAYSIA.*

*Corresponding Author: logen@uthm.edu.my
DOI: https://doi.org/10.30880/ekst.2025.05.02.057

Article Info

Received: 30 December 2024
Accepted: 17 January 2025
Available online: 19 December 2025

Keywords

Football, Prediction, Elo Rating, Pi Rating, Machine Learning

Abstract

Football outcome prediction is one of critical part in sports analytics that gives coaches, analysts, and fans a means to make strategic planning easier and improve engagement. In this study, a novel approach of combining Elo and Pi rating systems with machine learning models to predict the outcome of Malaysia Super League (MSL) matches is presented. Team performance can be evaluated by Elo and Pi ratings which are well known methods, integration which of the static and dynamic rating systems gives a more comprehensive assessment. Football historical match dataset from 2015 to 2022 was used to analyse machine learning algorithms such as Logistic Regression, Naive Bayes, K-Nearest Neighbors (KNN) and XGBoost. The performance of these algorithms was evaluated by using critical metrics such as accuracy, Ranked Probability Score (RPS), precision, recall, and F1-score. The Pi Rating combined with Naive Bayes achieves the highest accuracy and the best F1 Score across all metrics, making it the most superior combination for predicting match outcomes. Future research is recommended to apply k-fold cross validation, expand the dataset to include more diverse attributes and player specific statistics. By enhancing prediction accuracy and reliability, these enhancements can significantly improve sports analytics by better strategic decisions and team performance evaluations.

1. Introduction

Football match outcome prediction is an important area of research in sports analytics, providing useful information to clubs, coaches, analysts, and spectators. Football's global prominence has increased the demand for precise prediction systems. These systems are critical for strategic decision-making, improving team performance, and increasing fan involvement. The combination of classic team rating systems, such as Elo and Pi ratings, with machine learning (ML) algorithms constitutes a big step forward in this field, providing a more complete and trustworthy method for predicting match results [1, 2].

Elo and Pi rating systems are commonly used to assess team performance using historical data. While Elo ratings measure club strength through dynamic score adjustments, Pi ratings provide a more complete analysis by taking into account other contextual aspects such as goal differentials and home versus away performances. However, when used independently, these systems frequently encounter constraints, such as difficulty in reacting to real-time changes or incorporating complicated linkages in data [3, 4]. Machine learning algorithms, on the other hand, excel in detecting patterns and relationships in datasets, with the potential to dramatically improve prediction accuracy when paired with traditional rating methods [5].

Despite progress, football match prediction remains a difficult undertaking due to the game's fluid and unpredictable nature. Existing models frequently rely primarily on past data and struggle to account for external events such as player injuries, tactical modifications, or squad makeup changes. Furthermore, issues such as dataset homogeneity, class imbalance (for example, underrepresentation of draws), and a lack of feature diversity restrict the creation of strong predictive systems. To address these limitations, a unique framework that combines standard rating systems with advanced ML techniques is required to reflect the multidimensional structure of football dynamics [6, 7].

This study tries to overcome these issues by employing ML classification algorithms that use both Elo and Pi rating systems to forecast Malaysia Super League (MSL) match results. The study compares the performance of Logistic Regression, Naive Bayes, K-Nearest Neighbors (KNN), and XGBoost utilizing crucial measures such as accuracy, recall, precision, F1-score, and Ranked Probability Score (RPS). Using historical match data from 2015 to 2022, this work aims to show how combining traditional rating systems with contemporary ML algorithms can overcome existing restrictions and enhance predicted accuracy [8]. The evaluation of Elo rating and Pi-rating based on accuracy, recall, precision, F1 score, and Rank Probability Score (RPS) aims to identify the best method for the Malaysia Super League dataset.

The importance of this research goes beyond theoretical contributions to actual applications in sports analytics. Accurate projections allow teams to create better game strategy, optimize player selections and modify tactics dynamically. Reliable forecasting allows analysts and bettors to gain deeper insights into team performance and make more informed decisions. Furthermore, improved prediction models promote audience engagement and entertainment, establishing a closer bond between viewers and the sport. By merging classic statistical models with modern computational methodologies, this study lays the groundwork for the future of football analytics, emphasizing the possibility for integrating data-driven strategies with current rating systems [9, 10].

2. Materials and Methods

2.1 Data Description

The dataset used for this study was obtained from https://www. rsssf.org/tablesm/malay2015.html#super used in the study by [1] on prediction of Malaysian Super League (MSL). The dataset includes 990 samples gathered from eight seasons of the league (2015 – 2022) and contains additional descriptors including date of match, names of teams, goals scored and match outcome as shown in Table 1.

Table 1 *The description features for the dataset*

Feature	Description
Round	The league round in which the match was played
Date	Date of the match
Home Team	Name of the home team
Away Team	Name of the away team
Full-Time Home Goals (FTHG)	Number of goals scored by the home team
Full-Time Away Goals (FTAG)	Number of goals scored by the away team
Match Outcome (FTR)	Result of the match: Win, Draw, or Loss

2.2 Data Pre-Processing

The data was also divided for predictive modelling, 2015–2021: 858 records for training which accounts for 86.67% and 2022: 132 records for testing the outcome which accounts for 13.33% of the total dataset. Some of the data preparation procedures applied included data cleansing and cleaning, replacing the missing values and normalizing of data shown in Fig. 1.

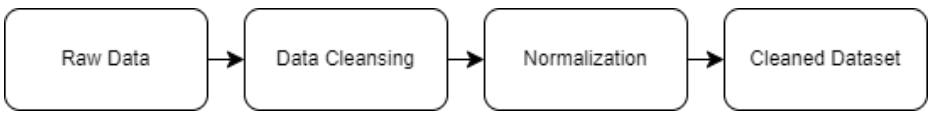

Fig. 1 The flowchart for pre-processing data

2.3 Team Ratings

To quantify team performance and competitiveness, a combined statistical method of the Elo rating and Pi-rating systems were used. Elo provides a robust baseline for team strength, while Pi offers contextual enhancements, making their integration valuable for capturing both static and dynamic performance aspects. These ratings were used as input features to train two types of predictive models: those based on the machine learning algorithms.

2.3.1 Elo Ratings

The Elo rating system, first created for chess and personalized to football to determine team strength, as well as chance of match outcome. In the case of this system, it is assumed that the rating difference between two teams would indicate the expected result of their match. If there are two teams playing, the ratings of two teams will be updated based on an outcome (win, lose or draw) and the points are transferred among these two teams. The point exchange depends on the expected outcome and the goal difference. The numerical encoding of match results, where outcomes are represented as 0 (away win), 1 (draw), or 2 (home win). The ratings of both teams are changed after the match as in equations (1) and (2) [11].

$$R_H^{t+1} = R_H^t + k(1+\delta)(S_H - E_H) \quad (1)$$

$$R_A^{t+1} = R_A^t + k(1+\delta)(S_H - E_H) \quad (2)$$

where:

- R_H^t and R_A^t are the current ratings of the home and away teams
- R_H^{t+1} and R_A^{t+1} are the updated ratings, k represents a learning rate
- δ is the absolute goal difference
- γ is a meta parameter that scales the effect of the goal difference on the change in ranking

2.3.2 Pi Ratings

The pi-rating system is a detailed approach of assessing football team performances, with the specific aim of comparing expected results with actual results and considering factors such as the relative strengths of the two teams, the goal difference and whether a team is at home or away. Assuming a match between home team α and away team β, the home and away ratings are respectively updated cumulatively as in equations (3) to (6) [12].

$$\hat{R}_{\alpha H} = R_{\alpha H} + \phi_H(e) \times \lambda \quad (3)$$

$$\hat{R}_{A\alpha} = R_{A\alpha} + (R_{H\alpha} - R_{A\alpha}) \times \gamma \quad (4)$$

$$\hat{R}_{A\beta} = R_{A\beta} + \phi_A(e) \times \lambda \quad (5)$$

$$\hat{R}_{H\beta} = R_{H\beta} + (\hat{R}_{A\beta} - R_{H\beta}) \times \gamma \quad (6)$$

where:

- $R_{\alpha H}$ and $R_{A\alpha}$ are the current home and away ratings for team α,
- $R_{H\beta}$ and $R_{A\beta}$ are the current home and away ratings of team β,
- $\widehat{R_{\alpha H}}$, $\widehat{R_{A\alpha}}$, $\widehat{R_{H\beta}}$, and $\widehat{R_{A\beta}}$ are the respective revised ratings,
- e is the error between predicted and observed goal difference,
- $\phi_H(e)$and $\phi_A(e)$ are functions of e,
- λ and γ are the learning rates.

2.4 Machine Learning Algorithm

2.4.1 Logistic Regression

Logistic Regression was chosen for efficiency with binary classification problems in predicting probabilities of match outcomes like home win, away win or draw. We applied a sigmoid function to a linear combination of input features to get a model that output probabilities and mapped them to class labels. Also, Logistic Regression's

simplicity and interpretability were adequate for such datasets where linear relationships between features and the outcome were identified [13]. The logit function is defined in equation (7).

$$logit(Y) = \ln\left(\frac{p}{1-p}\right) = \alpha + \beta X \tag{7}$$

The probability of the occurrence of the result can be calculated by taking the antilog of equation (7) and can be written as follows in equation (8)

$$p = \frac{e^{\alpha+\beta X}}{1+e^{\alpha+\beta X}} \tag{8}$$

where *p* is the probability of an outcome (Y = outcome of interest |X = x, a specific value of X), α = Y intercept, β = regression coefficient and e = base of the natural logarithm.

2.4.2 Naïve Bayes

Since Naïve Bayes is known to work well even when high dimensionality data is present and it is computationally efficient, it was applied to a treatment. Bayes' theorem, feature independence, with conditional probabilities being calculated, led the model to rely on feature independence. Even with its simplifying assumptions, Naïve Bayes seemed to work well for datasets that were large and had few highly correlated features in their relationship with the features of the model. In this study, the model was tested using its application, which showed its robustness in predicting outcomes of given football matches [14].

If S is a set of the training dataset and |S| is the total number of conditioning factors, the factors can be grouped into n classes Si (i = 1, 2, . . ., n). |Si| is the number of conditioning factors belonging to the class Si [15]. The classification of S can be calculated based on the expected entropy as follows in equation (9).

$$\text{Entropy}(S) = -\sum_{i=1}^{n}\left(\frac{|S_i|}{|S|}\log_2\left(\frac{|S_i|}{|S|}\right)\right) \tag{9}$$

Consider attribute A, for example aspect, in set S . The expected entropy can be expressed as in equation (10).

$$\text{Entropy}_A(S) = -\sum_{i=1}^{n}\frac{|S_i|}{|S|}\times \text{Info}\left(S_i\right) \tag{10}$$

The difference between Entropy (S) and Entropy_A (S) is represented as the information gain (InfoGain) in equation (11).

$$\text{InfoGain}(A) = \text{Entropy}(S) - \text{Entropy}_A(S) \tag{11}$$

The information gain ratio (IGR) is calculated according to the following equation (12).

$$\text{GainRatio}(A) = \frac{\text{InfoGain}(A)}{\text{SplitInfo}(A)} = \text{Entropy}(S) - \text{Entropy}_A(S) - \sum_{i=1}^{n}\left(\frac{|S_i|}{|S|}\log_2\left(\frac{|S_i|}{|S|}\right)\right) \tag{12}$$

2.4.3 K-Nearest Neighbors (KNN)

For example, as a non-parametric algorithm, KNN was used to classify match outcomes on the basis of the proximity of a data point to its neighbors in the feature space. The model which scores labels for the neighboring training points and assigns the label of the majority of these points to the test point gives highest scores. Very much because of the simplicity, the adaptability of KNN, yet could be sensitive to noisy data as well as imbalanced

class distributions [16]. The optimal number of neighbors, k, is a critical parameter and can be estimated using the formula shown in equation (13):

$$k = \sqrt{n} \tag{13}$$

where:

- k: Number of nearest neighbors to consider
- n: Total number of samples in the dataset

After determining k, the algorithm calculates the distance between the query point and all other data points in the dataset to identify the k-nearest neighbors. The most frequently used metric, Euclidean distance, is mathematically represented in equation (14):

$$d = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{14}$$

where:

- d: Distance between two points
- x_i: Coordinate of the first point along the i-th dimension
- y_i: Coordinate of the second point along the i-th dimension
- n: Number of dimensions in the feature space

2.4.4 XGBoost

For its efficiency and the ability to work with structured data, we included XGBoost, or Extreme Gradient Boosting. It is an ensemble learning method that sequentially corrects the weak models' residual errors (decision trees) to increase overall prediction accuracy. XGBoost is known to be very robust and flexible, but we needed to spend a lot of time tuning its parameters to get the best out of it on the dataset. The application showed the feasibility of advanced machine learning technology in sports analytics [17].

2.5 Evaluation Metrics

To comprehensively evaluate the performance of the machine learning models used in this study, five key metrics were employed: Accuracy, Precision, Recall, F1-Score, and Ranked Probability Score (RPS). These metrics provide a multidimensional understanding of the models' ability to predict football match outcomes, considering both overall correctness and the reliability of probabilistic predictions. Each metric serves a distinct purpose and contributes to a thorough assessment of the models' strengths and limitations. Table 2 summarizes the purpose and application of each metric in the context of this research.

Table 2 *The purposes for metric*

Metric	Purpose and Application
Accuracy	Measures the proportion of correctly predicted outcomes among all predictions.
Precision	Focuses on the accuracy of positive predictions.
Recall	Evaluates the ability to identify true positives.
F1-Score	Provides a balanced metric combining precision and recall.
Ranked Probability Score (RPS)	Assesses the quality of probabilistic predictions for ordered categories, critical in sports analytics.

2.5.1 Accuracy

Accuracy measured the proportion of correctly predicted match outcomes among all predictions, providing a straightforward metric for model performance. This metric is particularly meaningful when the class was evenly distributed [18]. Equation (15) is the formula for accuracy.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{15}$$

where:

- TP: True Positive
- TN: True Negative
- FP: False Positive
- FN: False Negative

2.5.2 Precision

Precision focused on the accuracy of positive prediction, reflecting model's ability to minimize false positives. This metric highlights the model's ability to avoid false positives, making it critical in scenarios like medical diagnostics or fraud detection [19]. The formula for precision is shown in equation (16).

$$\text{Precision} = \frac{TP}{TP + FP} \tag{16}$$

2.5.3 Recall

Recall evaluated the proportion of true positive predictions among all actual positive instances, highlighting the model's sensitivity to relevant cases. This metric is particularly important in applications were missing a positive case (false negative) is costly, such as in identifying diseases or fraud [20]. The formula for precision is shown in equation (17).

$$\text{Recall} = \frac{TP}{TP + FN} \tag{17}$$

2.5.4 F1-Score

The F1-Score, a harmonic mean of precision and recall, provided a balanced metric that accounted for both false positives and false negatives. The F1-Score is especially useful for datasets with imbalanced class distributions, where a balance between precision and recall is critical [19,20]. The formula for precision is shown in equation (18).

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{18}$$

2.5.5 Rank Probability Score (RPS)

RPS evaluates the quality if probabilistic predictions for ordered categories. It is particularly suited scenarios like sports or weather forecasting, where outcomes can be ranked. Lower RPS values indicate better probabilistic predictions. This metric is essential for assessing models' reliability in providing accurate probabilities [1]. RPS evaluates the accuracy of predicted probabilities in scenarios with ordinal outcomes, such as match results (Win, Draw, Loss). A lower RPS value indicates better model reliability. The formula for RPS is in equation (19).

$$RPS = \frac{1}{K-1} \sum_{k=1}^{K-1} \left(\sum_{j=1}^{k} F_j - O_j \right)^2 \tag{19}$$

where K is the total number of categories, Fj is the forecast probability for category j, Oj is the observed probability for category j, which is 1 if the event occurred in category j and 0 otherwise and the inner sum-k (F j – Oj) calculates the difference between the cumulative forecast and cumulative observed probabilities up to category k.

Enhanced Knowledge in Sciences and Technology Vol. 5 No. 2(2025) p. 492-504

3. Results and Discussions

3.1 Exploratory Data Analysis

An exploratory data analysis was done before actually going deeper into the analysis of the data to understand the general structure of the data and detect any outliers. This preliminary step was crucial for getting the reliable and accurate analyses as well as to fit the data collected to the aims of the current research.

Table 3 *The descriptive Analysis of dataset*

Metric	FTHG	FTAG
Mean	1.694	1.309
Median	1.000	1.000
Variance (Var)	1.756	1.431
Standard Deviation (Std)	1.325	1.196
Min	0.000	0.000
Max	7.000	7.000
IQR	1.000	2.000

The descriptive analysis for full-time home goals (FTHG) and full-time away goals (FTAG) on table 1 provide a valuable information about the dataset. The mean values suggest that in general, home teams were able to score 1.694 goals, and the away teams was able to score 1.309 goals, suggesting a consistent home advantage. Factors such as crowd support, familiarity with the pitch, and reduced travel fatigue likely contribute to this phenomenon. The median for both FTHG and FTAG is 1, which means that majority of the games that were played saw each team scoring 1 goal. It also can be seen the variability of the observed scores through variance, and here we have home goals (1.756) with slightly more variability than away goals (1.431). Also, the standard deviation, which calculates the average deviation of the data set from the mean, is also higher for FTHG (1.325) than for FTAG (1.196), indicating that home goals are less reliable. The least number of goals that can be scored in a match is zero and the most that has been scored in both halves is seven. Matches with exceptionally high scores (maximum of 7 goals) highlight the unpredictable nature of football. These outliers underline the importance of incorporating features that can explain anomalous results, such as tactical shifts or player performance spikes. Lastly, the IQR shows that 50% of home goals are within 1 goal difference, but for away goals, the range is 2 showing that away team have a wider distribution of goals. This could reflect differences in adaptability and resilience when playing under adverse conditions.

The primary features of interest in this study are Full Time Home Goals (FTHG) and Full Time Away Goals (FTAG), because they are directly correlated with match outcomes. These features are readily available and consistently recorded and are therefore reliable indicators of team performance. FTHG and FTAG are included to have a simple and interpretable model that is consistent with the objectives of this research. Since the current dataset did not contain such detailed data, there were other potential features for consideration, such as players' statistics like individual performance metrics, injuries, or substitutions, as well as tactical adjustments, for example the formations or play style. Additional granularity and increased predictive accuracy could be provided by these attributes by capturing the dynamic aspects of match outcomes. Future studies should therefore focus on integrating these features to obtain a more nuanced analysis and to improve model robustness.

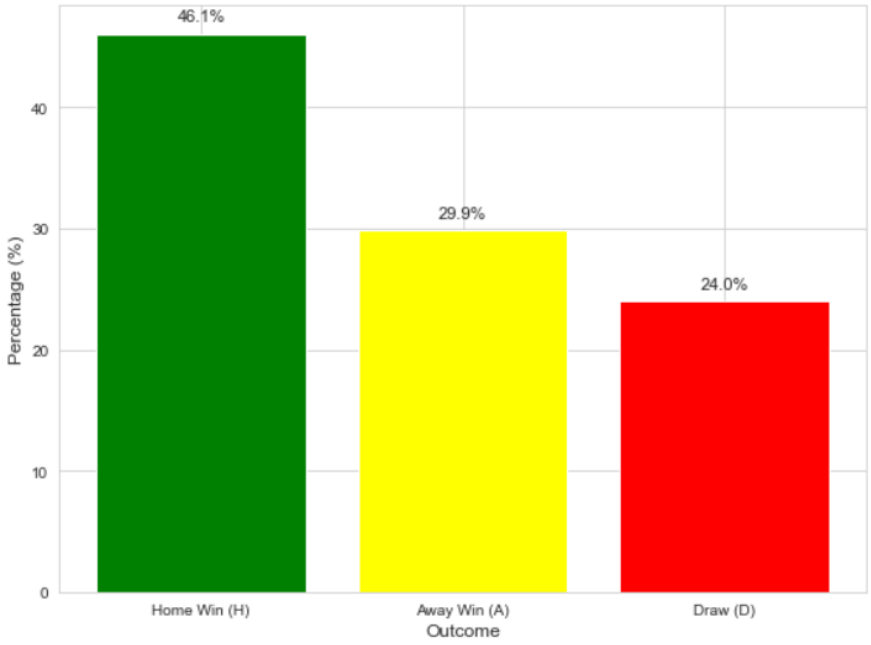

Fig. 2 The distribution of match outcome (Home Win, Away Win, Draw)

Fig. 2 illustrates the distribution of football match outcomes categorized into three possible results: H, A and D for home wins, away wins and draws respectively. Home field advantage had a big impact as home wins were the majority of matches, with 46.1%. Often attributed to factors such as local crowd support, familiarity with arena, and home team travel fatigue, this phenomenon has been proven to be caused by so many factors. Teams playing in less familiar conditions and often under psychological pressure were faced with 29.9% of matches being away wins. Only 24.0% of the games saw the draw, implying that quite often one side is the better side and will win the game. The distribution of this data emphasizes the necessity of including home field advantage in predictive models and the necessity to consider draws in machine learning algorithms. The ability to exactly capture these trends is important for developing robust prediction systems.

3.2 Elo Rating

Elo ratings for the match outcomes and team performance analyses were generated based on the objectives proposed in this study. These ratings were intended to measure and forecast team performance using statistical processing embedded in a dynamic system as much data about previous matches was integrated into this concept.

Table 4 *The excerpt dataset with Elo Rating*

Round	Date	Home Team	Away Team	FTHG	FTAG	FTR	Home ELO Rating	Away ELO Rating	Elo Diff
1	31/1/2015	JDT	Sri Pahang FC	2	0	2	1500.0	1500.0	0.0
1	7/2/2015	Young Lions	PDRM FA	5	3	2	1500.0	1500.0	0.0
1	7/2/2015	Perak FC	Sime Darby FC	2	0	2	1500.0	1500.0	0.0
1	7/2/2015	Terengganu FC	Selangor FC	2	0	2	1500.0	1500.0	0.0
1	7/2/2015	Felda United FC	Sarawak FA	3	3	1	1500.0	1500.0	0.0
1	7/2/2015	ATM FA	Kelantan FA	0	2	0	1500.0	1500.0	0.0

2	14/2/2015	Felda United FC	Sime Darby FC	3	0	2	1500.0	1492.262	7.738
2	14/2/2015	Sri Pahang FC	Selangor FC	1	1	1	1492.262	1492.262	0.0
2	14/2/2015	PDRM FA	JDT	1	0	2	1492.262	1507.738	-15.476
2	14/2/2015	Sarawak FA	Terengganu FC	3	1	2	1500.0	1507.738	-7.738

Table 4 shows the excerpt dataset with Elo Rating for the 2015 Malaysia Super League (MSL) season, Elo ratings for all teams were set to a baseline of 1500, since there was no previous matches to influence prediction. JDT, Young Lions, Perak FC and Terengganu FC won in the first round, while Felda United FC and Sarawak FA drew 3-3. By Round 2, Elo ratings began to track performance differences between teams as they adapt dynamically based on prior results. For example, Felda United FC's 3-0 win over Sime Darby FC had an Elo ratings predicted Elo advantage of +7.738, signifying that Elo ratings express expected performance in terms of historical results. On the other hand, PDRM FA managed to buck expectations and outdo a rating deficit of -15.476 to beat the higher rated JDT 1-0. Sarawak FA also did it, although they had a -7.738 rating handicap, beating Terengganu FC 3-1, showing how Elo predictions can be wrong in actual match results.

Elo ratings captured the dynamic change of a team's rating over the course of the season, allowing us to explore performance trends with both expected outcomes and surprising upsets. These results underscore the broader implication of using Elo ratings in predictive analytics: while they do an excellent job of capturing team strength, and give you probabilistic forecasts, football is a dynamic sport: unexpected things can happen. Elo ratings are so flexible that stakeholders like coaches and analysts can see how a team's competitive status is changing throughout a season. Additionally, Elo predictions are also deviated from, which emphasizes the necessity of considering contextual factors, including tactical decisions, player availability, and situational dynamics, when refining predictive models. Elo ratings are ultimately a great way to assess team performance and to use for strategic planning and to understand the competitive dynamics of football more.

3.3 Pi Rating

The analysis of match outcomes and team performance also incorporated ELO ratings, developed according to the objectives of this study. ELO ratings were utilized to dynamically evaluate team strength by updating scores based on prior match results.

Table 5 *The excerpt dataset with Pi Rating*

Round	Date	Home Team	Away Team	FTHG	FTAG	FTR	Home Home Rating	Home Away Rating	Away Home Rating	Away Away Rating	Exp Pi Diff	Pi Diff
1	31/1/2015	JDT	Sri Pahang FC	2	0	2	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1	7/2/2015	Young Lions	PDRM FA	5	3	2	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1	7/2/2015	Perak FC	Sime Darby FC	2	0	2	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1	7/2/2015	Terengganu FC	Selangor FC	2	0	2	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1	7/2/2015	Felda United FC	Sarawak FA	3	3	1	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1	7/2/2015	ATM FA	Kelantan FA	0	2	0	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
2	14/2/2015	Felda United FC	Sime Darby FC	3	0	2	0.0000	0.0000	-0.0011	-0.0802	0.0166	0.0802
2	14/2/2015	Sri Pahang FC	Selangor FC	1	1	1	-0.0011	-0.0802	-0.0011	-0.0802	0.0164	0.0790
2	14/2/2015	PDRM FA	JDT	1	0	2	-0.0011	-0.0802	0.0802	0.0011	-0.0004	-0.0023
2	14/2/2015	Sarawak FA	Terengganu FC	3	1	2	0.0000	0.0000	0.0802	0.0011	-0.0002	-0.0011

Table 5 shows that Pi ratings changed dynamically in response to match outcomes during Rounds 1 and 2 of the 2015 Malaysia Super League (MSL) and were able to predict interesting patterns about who will win any future

match. In Round 1, all teams started with baseline ratings of 0, so all teams had neutral expected and actual Pi rating differences. By Round 2, past results started affecting predictions, for instance, Felda United FC's 3–0 win over Sime Darby FC where the actual rating difference (0.0802) was larger than the expected one (0.0166), which indicates a better performance than expected. The 1-1 draw between Sri Pahang FC and Selangor FC also had small expected (0.0164) and actual (0.0790) differences, consistent with the balance of the match. Pi ratings were sensitive to underperformance, not only because PDRM FA defeated JDT 1-0, but also because the actual difference (-0.0023) between the two teams' pi ratings was greater than the difference in their initial ratings. Conversely, Sarawak FA's 3-1 win over Terengganu FC exhibited a consistent difference between expected (-0.0002) and actual (-0.0011), which while unexpected, is explainable. The implications of Pi ratings extend to their dynamic and context sensitive evaluation of team performance. They are helpful for discovering unexpected outcomes, underperformances and strengths, providing actionable insights for coaches, analysts and stakeholders. Pi ratings incorporate factors such as home/away performance, and goal margins and provide for real time adjustments to strategies and a better preparation of future matches. Some deviations are rare, but highlight the inherent unpredictability of football, and the possibility of improving predictive models by accounting for additional such features as player form or tactical changes. Overall, Pi rates are an important instrument for performance evaluation and strategic decision making, which accommodates the gap between statistical analysis and football's dynamic activities.

3.4 Elo Versus Pi rating

Team performance can be evaluated in different ways, two being Elo and Pi Ratings with their own strengths and weaknesses. On the other hand, Elo ratings pay less attention to the match specific dynamics and prize more on the all-round strength of a team. As they can only incrementally change, they're stable and reliable solution for long term performance tracking. However, they fail to be able to capture contextual nuances such as goal margin and home/away, preventing their utility for analysing short-term fluctuations and unpredictable outcomes. In contrast however, Pi ratings are highly sensitive to the match specific factors, eg, particulars associated with the match as well as home/away performance details like goal differences. This dynamic responsiveness of Pi ratings enables them to adjust to the particularities of each match, so more sophisticated insights into the impact of recent performance changes are available.

A big advantage is that Elo ratings are very stable, because their incremental adjustments prevent individual matches from influencing the Elo Ratings too much. That makes them perfect for spotting consistent team performance trends over multiple seasons. However, Pi ratings' sensitivity to individual matches adds volatility, which makes them well suited for short term performance shifts analysis, but less reliable for long term trend analysis. Elo ratings are simpler to compute and use very little resources, making them well suited for finding out the best teams across an entire season. However, pi ratings are computationally more complex as they require the addition of goal margins and home / away adjustments. It's a limitation, but this complexity creates richer, more detailed insights for decision making through Pi ratings.

Elo ratings are easier and more stable to use in real applications to evaluate long term performance and league standings. For example, they do a good job of capturing how consistent certain team such as JDT is over a season. On the other hand, Pi ratings perform well in the context sensitive evaluations, for example, by analysing recent performance or predicting a particular match outcome. For example, Pi ratings showed how PDRM FA's surprising win over JDT was dynamically adjusted for match specific factors, like home advantage and performance gaps. Elo ratings offer a stable, long-term perspective and Pi ratings give you granular, context driven insights.

To put things in a nutshell, an Elo rating is best for simplicity and stability, and for long term assessments; whilst a Pi rating is better for capturing short term dynamics and match specific nuances. They complement each other and together they provide a complete view of how team performance can better predict modelling and decision making in football analytics.

3.5 Comparative Analysis

To achieve the third objective of this study, machine learning algorithms were implemented to evaluate the performance of ELO Rating and Pi Rating systems in predicting match outcomes. This approach allowed for a systematic comparison of key metrics across different models, providing insights into the predictive accuracy and reliability of each rating system.

Enhanced Knowledge in Sciences and Technology Vol. 5 No. 2(2025) p. 492-504

Table 6 *Comparative result*

	Elo Rating				Pi Rating			
Metric	LR	NB	XGB	KNN	LR	NB	XGB	KNN
Accuracy	**0.5152**	0.5076	0.4394	0.4318	0.5152	**0.5227**	0.4394	0.4545
Precision	0.4034	0.3959	0.4286	0.4257	0.4058	0.4118	0.4377	0.4568
Recall	0.5152	0.5076	0.4394	0.4318	0.5152	0.5227	0.4394	0.4545
F1 Score	**0.4434**	0.4350	0.4329	0.4285	0.4518	**0.4560**	0.4365	0.4531
RPS	0.2141	0.2150	0.2523	0.2444	0.2158	0.2157	0.2666	0.2571

Table 6 presented the comparison result for Elo Rating and Pi Rating to determine which system was better at predicting match outcomes and Pi Rating has a clear advantage in key metrics. Naive Bayes (NB) demonstrated unparalleled performance with Pi Rating, achieving the highest accuracy of 0.5227 and the best F1 score of 0.4560, surpassing all other models in predictive effectiveness. In contrast, the top-performing model for Elo Rating, Logistic Regression (LR), exhibited comparatively lower metrics, including an accuracy of 0.5152 and an F1 score of 0.4434. These findings underscore the exceptional consistency and precision of Pi Rating in predicting match outcomes. While Logistic Regression for Elo Rating delivered a slightly better Ranked Probability Score (RPS) of 0.2141 compared to 0.2157 for Naive Bayes with Pi Rating, the marginal difference does not diminish the superior overall performance of Pi Rating. The higher F1 score and accuracy of Pi Rating illustrate its exceptional ability to capture the intricacies of match outcomes more effectively and to meaningfully reflect actual performance disparities among teams. Pi Rating's superior performance is attributed to its advanced adaptability in accounting for match-specific factors, enabling a more accurate alignment between predictions and actual outcomes. This establishes Pi Rating with Naive Bayes as the most powerful, effective, and reliable system for achieving the objectives of this analysis, delivering superior accuracy and predictive modelling of competitive and outcome probabilities.

4. Conclusion

The objectives of this study were met using Elo and Pi registered rating systems coupled with machine learning models to forecast football match outcomes, demonstrating the potential of data analysis in the domain of sports for generating meaningful insights. With machine learning models, this study evaluated the effectiveness of Elo and Pi ratings in predicting football match outcomes. Results consistently showed that Pi ratings outperformed Elo ratings in terms of accuracy and F1 score, particularly when combined with Naive Bayes, which achieved the highest accuracy of 0.5227. Pi ratings were able to capture contextual nuances, such as goal margins and home/away performance, making them more effective in reflecting match-specific dynamics. In contrast, Elo ratings proved to be less adaptable, lacking the ability to incorporate situational factors effectively, which limited their accuracy and responsiveness. These findings highlight the limitations of Elo ratings for short-term predictions and suggest that combining both systems could improve the robustness of predictive models for sports analytics.

The models were moderately accurate in predicting team performance and provide useful information about the strengths and weaknesses of teams which can be used by coaches, analysts, and stakeholders to make tactical decisions and improve strategic planning. Our research is limited, however, and leaves room for future research in how restrictions, like excluding player specific statistics and tactical changes, impact our perspective of fatality. Additional features such as player performance metrics, injuries, and real time tactical adjustments can eventually improve model precision and applicability even further by expanding the dataset. Additionally, incorporation of sophisticated approaches, including ensemble learning or deep learning, could improve predictive performance.

This research has broader implications for real world applications in sports analytics. Match predictions can be accurate to help teams make better decisions, engage fans better, and even help betting systems and media content creation. Continuously refining predictive systems in this field will lead to increasingly more reliable insights into team dynamics and competitive outcomes. Future work will embed contextual factors beyond teams' ratings, as well as testing hybrid models combining the strengths of multiple rating system and machine learning technology. These improvements would not only make analytic prediction more accurate but also cement the place of analytics in the reigning meteoric rise of modern football.

Acknowledgement

The authors would like to thank the Faculty of Applied Sciences and Technology, Universiti Tun Hussein Onn Malaysia for its support.

Conflict of Interest

Authors declare that there is no conflict of interests regarding the publication of the paper.

Author Contribution

The authors confirm their contribution to the paper as follows: ***study conception and design, data collection, analysis and interpretation of results, and draft manuscript preparation****: Siti Munawarah Mohd Din, Logenthiran Machap. All authors reviewed the results and approved the final version of the manuscript.*

References

[1] Razali, N., Mustapha, A., Aziz, A. Q. A. A., & Mostafa, S. A. (2023). Machine learning approach for Malaysia Super League football match outcomes prediction based on Elo rating system. In *Innovation and Technology in Sports: Proceedings of the International Conference on Innovation and Technology in Sports (ICITS) 2022, Malaysia* (pp. 169–176). Springer.

[2] Wheatcroft, E. (2021). Forecasting football matches by predicting match statistics. *Journal of Sports Analytics, 7*(2), 77–97.

[3] Van Eetvelde, H., Mendonça, L. D., Ley, C., Seil, R., & Tischer, T. (2021). Machine learning methods in sport injury prediction and prevention: A systematic review. *Journal of Experimental Orthopaedics, 8*(1), 1–15.

[4] Harish, S., Abishek Kevin, A., Harsha Vardhan, U., & Sharon Femi, P. (2023). Expected goals prediction in football using XGBoost. *PLOS One, 18*(4), e0282295.

[5] Arntzen, H., & Hvattum, L. M. (2021). Predicting match outcomes in association football using team ratings and player ratings. *Statistical Modelling, 21*(5), 449–470.

[6] Li, Y., & Hong, Y. (2022). Prediction of football match results based on edge computing and machine learning technology. *International Journal of Mobile Computing and Multimedia Communications (IJMCMC, 13*(2), 1–10.

[7] Reade, J., Singleton, C., & Brown, A. (2021). Evaluating strange forecasts: The curious case of football match scorelines. *Scottish Journal of Political Economy, 68*(2), 261–285.

[8] Saribekyan, G., & Yarovoy, N. (2024). Football prediction model based on the teams' Elo ratings and scoring indicators. *Journal of Sports Analytics, 12*(3), 243–256.

[9] Razali, N., Mustapha, A., Mostafa, S. A., & Gunasekaran, S. S. (2022). Football matches outcomes prediction based on gradient boosting algorithms and football rating system. In *Human Factors in Software and Systems Engineering* (pp. 57–68).

[10] Beal, R., Middleton, S. E., Norman, T. J., & Ramchurn, S. D. (2021). Combining machine learning and human experts to predict match outcomes in football: A baseline model. In *Proceedings of the AAAI Conference on Artificial Intelligence, 35*, 15447–15451.

[11] Szczecinski, L., & Djebbi, A. (2019). Understanding and pushing the limits of the Elo rating algorithm. In *arXiv: Statistics Theory*.

[12] Constantinou, A., & Fenton, N. (2013). Determining the level of ability of football teams by dynamic ratings based on the relative discrepancies in scores between adversaries. In *Journal of Quantitative Analysis in Sports, 9*(1), 37–50.

[13] Mishra, V. N., Kumar, V., Prasad, R., & Punia, M. (2021). Geographically weighted method integrated with logistic regression for analyzing spatially varying accuracy measures of remote sensing image classification. Journal of the Indian Society of Remote Sensing, 49, 1189–1199.

[14] Basani, Y., Sibuea, H. V., Sianipar, S. I. P., & Samosir, J. P. (2019). Application of sentiment analysis on product review e-commerce. In Journal of Physics: Conference Series, 1175. IOP Publishing. p. 012103.

[15] Nhu, V.-H., Shirzadi, A., Shahabi, H., Singh, S. K., Al-Ansari, N., Clague, J. J., Jaafari, A., Chen, W., Miraki, S., Dou, J., et al. (2020). Shallow landslide susceptibility mapping: A comparison between logistic model tree, logistic regression, naïve Bayes tree, artificial neural network, and support vector machine algorithms. In *International Journal of Environmental Research and Public Health, 17*(8), 2749.

[16] Uddin, S., Haque, I., Lu, H., Moni, M. A., & Gide, E. (2022). Comparative performance analysis of k-nearest neighbour (KNN) algorithm and its different variants for disease prediction. Scientific Reports, 12(1), 6256.

[17] Asselman, A., Khaldi, M., & Aammou, S. (2023). Enhancing the prediction of student performance based on the machine learning XGBoost algorithm. Interactive Learning Environments, 31(6), 3360–3379.

[18] Abercrombie, G. (2021). Topic-centric sentiment analysis of UK parliamentary debates. The University of Manchester (United Kingdom).

[19] Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PloS one, 10(3), e0118432.
[20] Powers, D. M. W. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. Journal of Machine Learning Technologies, 2(1), 37–63.