Cognitive Computation (2025) 17:88
https://doi.org/10.1007/s12559-025-10443-z

RESEARCH

Comparative Analysis of Hybrid and Ensemble Machine Learning Approaches in Predicting Football Player Transfer Values

Wenjing Zhang[1] · **Dan Cao**[2]

Received: 18 November 2024 / Accepted: 11 March 2025

Abstract
In football economics, a player's transfer market value extends beyond performance metrics, with popularity playing a crucial role in clubs' decisions. Reputation indexes, reflecting a player's standing in the industry, are derived from various sources. Traditional metrics include goals, assists, and defensive prowess, while social media activity (likes on Facebook and Instagram), press citations, and Wikipedia page views add a new dimension. This study utilized Fédération Internationale de Football Association 19 data and a real-world statistical dataset, encompassing 54 features for 491 players across various leagues. After adding valuable data and removing ineffective features and outliers, two filtering-based feature selection methods identified the 20 most critical features for predicting market value. The study applied Extreme Gradient Boosting and Adaptive Boosting regression models, along with their hybrid forms optimized by metaheuristic algorithms. The Extreme Gradient Boosting optimized with the Ali Baba and Forty Thieves algorithm model showed the best performance, with a 99% match to actual values and a misestimation of around €1.9 million. Ensemble models, averaging predictions from all hybrid models, provided reliable market value estimates. These insights help managers make informed decisions to improve team performance and secure financial benefits for the club.

Keywords Market valuation · International reputation index · Football superstars · Filtering feature selection · Boosting Tree Regression · Metaheuristic optimization algorithms

Introduction

The sports industry is a hub where diverse interests, spanning from political to public, intersect. The growing global appeal of football is influenced significantly by both traditional and social media [1, 2]. An illustration of football's widespread viewership is the UEFA Euros final in 2021, attracting an average live audience of 328 million people [3]. Success on the field is crucial for football clubs, leading to increased financial gains when they progress to the knockout stage. Deloitte [4] reports that the top 20 clubs in turnover, belonging to the "Big Five" European football leagues, collectively generated revenues surpassing €9.200 million in the 2021–2022 season, slightly below the pre-COVID revenues of €9.283 million in 2018–2019 season [5].

Football clubs aim to improve their performance by recruiting new players, with options including loans, free agent signings, or outright purchases. The latter involves one club paying another a transfer fee for the player's services, necessitating negotiation between the involved clubs [6]. Striking a balance between the actual transfer fee and the perceived value of a player is crucial to minimize potential losses if the player underperforms. Beyond club considerations, the transfer value is of interest to fans and analysts, who evaluate whether the paid fee aligns with the player's abilities when a new player is acquired [7, 8].

In this context, researchers from different areas of knowledge have begun to specialize in evaluating players and studying the factors that affect the market value to predict transfer fees [9, 10]. The player's performance, position (forward, midfielder, defender, or goalkeeper), club, and physical characteristics (e.g., height and age) are the variables most often used in such studies [11–13]. Moreover, Barbuscak [14] investigated football player market values using

✉ Dan Cao
syyxytyb1@163.com

[1] Institute of Physical Education, Liaoning Finance and Trade College, Huludao, Liaoning 125105, China

[2] Physical Education Department, Shenyang Medical College, Shenyang, Liaoning 110032, China

Published online: 31 March 2025

data from transfermarkt.com. Employing a linear regression analysis, the study found that factors such as remaining years on contracts significantly influenced market values. This aligns with previous research by Carmichael et al. [15] and Frick [16].

Furthermore, the noteworthy influence of popularity on market value, with implications for predicting transfer fees, has been recognized. Academic theory on superstardom, as outlined by Rosen [17] and Adler [18], proposed that the emergence of superstars is only partially determined by actual talent in impacting sports competition outcomes; additional factors such as popularity are also pertinent. Player popularity is also an indicator for football clubs and extends its influence to jersey and ticket sales [19]. In their study of the Spanish football league, Garcia del Barrio and Pujol [20] identified both performance and popularity, measured through Google search hits, as determinants of football player market value. Kiefer [21] utilized Facebook likes to gauge player popularity and investigate its correlation with performance in the Euro 2012 tournament. Mueller et al. [11] adopted various popularity metrics, including Reddit mentions and YouTube appearances for market value evaluation, and others considered data from Google Trends [22] and Wikipedia views [23]. These studies consistently reveal the statistically significant impact of popularity in estimating market value and predicting football player transfer fees. Consequently, it is reasonable to consider popularity as a crucial factor affecting the transfer value of players, particularly due to its accessibility.

Artificial intelligence and machine learning techniques have become transformative tools in the fields of long-term player development, performance evaluation, and injury prevention [24, 25]. These technologies enable more accurate analysis and prediction of player performance, as well as the identification of potential injury risks, significantly enhancing decision-making processes in sports management. For instance, Teixeira et al. [26] tested a machine learning model for predicting high-intensity actions and body impacts in youth football training. Mandadapu et al. [27] applied machine learning algorithms to predict Premier League match outcomes by analyzing historical data and identifying key features that influence results. The research also aims to assist in setting bookmaker odds, providing insights into the role of various variables in shaping match outcomes and opening new opportunities for decision-making in sports forecasting and betting. Yang et al. [28] advanced previous studies on transfer fees in European football by applying machine learning methods, such as random forest and quantile additive models, to capture non-linear effects. Analyzing data from transfermarkt.de, they trained models on pre-COVID-19 transfers and compared prediction accuracy before and during the pandemic. Their findings revealed that models trained pre-COVID-19 significantly underestimated transfer fees during COVID-19, especially for high- and medium-priced players, questioning the existence of a cooling-off effect in the transfer market. Additionally, optimization algorithms have gained prominence for their efficiency in solving complex high-dimensional optimization problems [29, 30]. Recently, a metaheuristic optimization algorithm named football optimization algorithms has been developed based on tactical gameplay elements like short passes, long passes, and positional adjustments to balance exploration and exploitation within the solution space [31]. Also, many researchers have utilized optimization algorithms for hybrid prediction model development in football-related fields. For instance, Morciano et al. [32] predicted the above-team-average performance of football players using supervised machine learning algorithms. The algorithms were trained and tested on four biometric parameters as inputs and seven performance indicators as labels, optimized using grid-search and two versions of the whale optimization algorithm, one standard and another proposed by the authors incorporating Euclidean distance. The analysis accounted for player roles (strikers, midfielders, and defenders) to address the varying skill requirements.

Incorporating explainable machine learning techniques into sports analytics offers a transformative approach to understanding the multifaceted factors influencing football player valuation. Unlike traditional predictive methods, explainable machine learning techniques emphasize transparency and provide stakeholders with a clearer understanding of how model predictions are generated [33, 34]. Techniques such as Shapley Additive Explanations (SHAP) enable researchers to quantify the contribution of each feature, such as popularity metrics, performance indicators, and demographic variables, to a model's predictions, offering actionable insights. Recent studies have applied SHAP as an explainability technique to assess the influence of individual features on predictions in various fields related to sports, such as performance analytics in professional basketball, focusing on the varying influence of key performance indicators on match outcomes [35] or to assess the influence of individual features on match-specific score predictions in football [36, 37]. Plakias et al. [38] developed an explainable machine learning model identifying factors crucial for securing a top-three position in French Ligue 1, ensuring UEFA Champions League qualification. Also, Moustakidis et al. [34] identified key team-level performance indicators influencing football match outcomes using explainable machine learning techniques. By analyzing team-specific features such as ball possession and pass behaviors, the pipeline incorporates data preprocessing, feature selection, model training, and SHAP-based explainability. Furthermore, the integration of explainable machine learning allows the development of improved, interpretable ML tools that bridge the gap between predictive accuracy and usability.

It is rare in the existing literature to use real-world and virtual football simulation game datasets for predicting the market value of players based on their attributes, performance metrics, popularity scores, and transfer market values for various seasons. Also, feature selections by decision-making and the development of explainable machine learning models and ensemble predictions are the technical gaps in the literature. In this investigation, the aim is to determine the most relevant features among those extracted for players with various popularity levels from five top-tier European leagues on their market value and examine the prediction capability of different machine learning models. At the conclusion of this research, the following questions will be addressed: Which variables are most influential in predicting players' market values, and are they primarily related to player information, physical and performance attributes, or game statistics? Which optimization algorithm performs best in optimizing players' market value predictions? What insights can be drawn from explainable machine learning models, and what are the practical interpretations of these predictions? How can clubs utilize estimated market values during player negotiations?

The organization of the research is as follows:

In the "Dataset Exploration" section, the employed data source (FIFA19 in Sofifa.com) is introduced as a prominent football-related data platform, which includes player attributes, performance metrics, popularity, and transfer market values. Then, the dataset went under preprocessing, and the most imperative features were selected. The popularity index of players is one of the important features selected as the base of feature engineering. In the "Machine Learning Algorithms" section, regression boosting methods (Adaptive Boosting (ADA) and Extreme Gradient Boosting (XGB)) for structured data were reviewed, and four metaheuristic optimization algorithms, including Ali Baba and Forty Thieves, Crystal Structure Algorithm, Henry Gas Solubility Optimization, and Mayfly Optimization Algorithm, were introduced for hybrid model development. In the "Statistical Evaluation Metrics" section, the optimized hyperparameters of two base models are reported, and the iterative optimization procedure is illustrated. Then, the prediction performance of developed models was evaluated through statistical parameters and comparative analysis conducted by various figures. Also, error values in estimating the market value of sample players with different levels of international reputation were presented to examine the accuracy of predictions in detail. In the "Discussion of Results" section, the novel SHAP method for explaining the sensitivity of predicted values is presented to assess the generalization performance of estimations and give valuable insights for future works. Finally, concluded results and real-world applications are presented. Framework of the research is presented in Fig. 1.

Dataset Exploration

Data Collection

The dataset of the study was extracted from [39], for which https://sofifa.com/?r=190075&set=true was the original reference. Sofifa.com is a prominent online database for football-related data, particularly focusing on player statistics within virtual football simulation games. The platform offers a wealth of information, including player attributes, performance metrics, popularity, and transfer market value

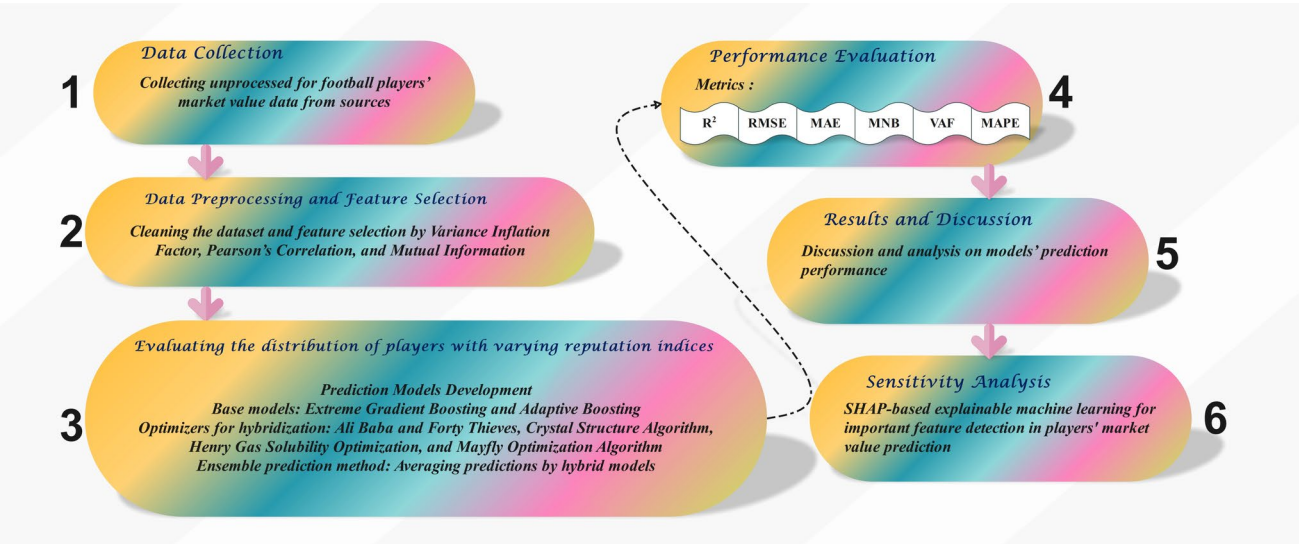

Fig. 1 Research framework

across various seasons, as is clear in Fig. 2 for Cristiano Ronaldo and Kylian Mbappé in FIFA 19.

Enthusiasts, analysts, and researchers have widely utilized this dataset. In research focused on predicting football player market values, Kirschstein and Liebscher [40] employed machine learning techniques, utilizing data from the FIFA16 video game. Their model estimated player market values based on skill variables in FIFA, aligning this information with actual market values from the German First Division and Second Division sourced from transfermarkt.com. Results indicated a significant impact of a club's reputation on player market values. Behravan and Razavi [10] utilized the FIFA20 dataset to predict football player market values, emphasizing the importance of considering different player positions and overall ratings in their research.

Data Preprocessing and Feature Selection

The FIFA19 dataset used in this study comprised 53 features for 491 sampled players. Data engineering was necessary to estimate market values accurately, accounting for different reputation indexes in well-known European football leagues. To clean the dataset and exert necessary modifications on the dataset, preparing for the regression-based prediction task, the following steps were conducted:

1. During the initial preprocessing phase, seven samples were excluded due to incomplete feature values.
2. Following this, the dataset was enriched by extracting corresponding leagues and playing positions for each player based on club and player names. This led to the addition of two new columns, "league name" and "playing position."
3. To ensure the stability of the target value and prevent excessive variation, 27 players from less professional leagues with lower-value players were excluded from the dataset. A substantial majority of the 457 remaining players were affiliated with defenders, forwards, midfielders, and players with dual professional positions (forward and midfielder) in Serie A, Premier League, League 1, La Liga, and Bundesliga.
4. Certain feature columns lacking analytical significance, such as player nationality, were omitted, resulting in a final set of 47 features (Specified in Table 1).
5. Nominal features like footedness (preferred foot), weak foot, and league name were converted to numerical labels, enhancing their suitability for machine-learning regression tasks.

Feature selection demonstrates its effectiveness in diminishing dimensionality, eliminating irrelevant data, boosting learning accuracy, and refining the comprehensibility of the obtained results. Filter, wrapper, and embedded represent the three primary categories of feature selection methods employed in learning contexts. The filter method is the most prevalent, involving the selection of features without the use of a machine learning algorithm. Essentially, this method filters out irrelevant features through various selection principles. Filter methods employ selection criteria to

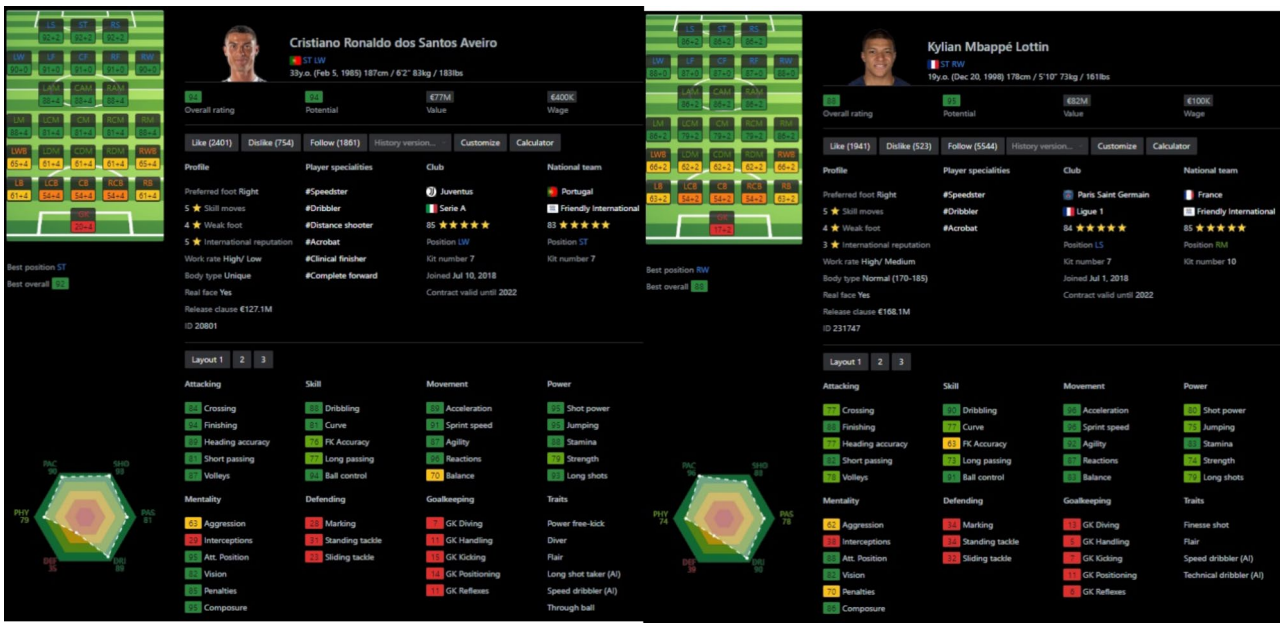

Fig. 2 Presented data for Cristiano Ronaldo and Kylian Mbappé on Sofifa.com

Table 1 Feature selection–based ranking of the features by decision-making between Pearson's correlation coefficients and Mutual Information scores

Features	*Pearson's correlation coefficient*	*Mutual information score*	*TOP-ranked features (TOPSIS)*
Playing Position	8.05E-02	3.48E-02	-
Age	3.93E-01	3.14E-02	-
League	9.06E-01	1.55E-02	-
Preferred Foot	1.19E-01	4.55E-02	-
International Reputation	4.02E-47	2.36E-01	8
Weak Foot	5.56E-03	7.74E-03	-
Skill Moves	9.45E-10	1.00E-01	-
Height	1.09E-02	0.00E+00	-
Weight	2.67E-01	0.00E+00	-
Crossing	1.79E-10	4.86E-02	-
Finishing	1.01E-10	1.41E-01	14
Heading Accuracy	2.24E-01	3.52E-02	-
Short Passing	2.36E-29	3.69E-01	4
Volleys	4.53E-13	1.27E-01	16
Dribbling	1.45E-20	2.92E-01	6
Curve	4.30E-15	2.23E-01	9
Free Kick Accuracy	3.60E-15	1.59E-01	12
Long Passing	2.74E-13	1.95E-01	10
Ball Control	3.93E-39	4.85E-01	2
Acceleration	5.28E-07	8.24E-02	20
Sprint Speed	4.42E-05	1.21E-02	-
Agility	3.39E-10	3.86E-02	-
Reactions	9.62E-49	4.96E-01	1
Balance	1.66E-07	8.81E-02	-
Shot Power	1.14E-11	1.59E-01	13
Jumping	8.10E-01	2.82E-02	-
Stamina	1.32E-09	1.11E-01	18
Strength	6.56E-01	0.00E+00	-
Long Shots	1.18E-13	1.90E-01	11
Aggression	6.52E-01	4.25E-02	-
Interception	8.57E-01	2.84E-02	-
Positioning	1.24E-11	2.38E-01	7
Vision	1.02E-20	3.98E-01	3
Penalties	1.83E-10	1.08E-01	19
Composure	8.94E-38	3.60E-01	5
Marking	4.56E-01	9.47E-02	-
Standing Tackle	6.15E-01	7.74E-02	-
Sliding Tackle	4.35E-01	6.43E-02	-
Games Played	1.46E-01	1.91E-02	-
Games Started	1.47E-02	2.74E-02	-
Minutes Played	6.05E-02	4.93E-02	-
Goals	1.33E-12	3.75E-02	-
Assist	2.22E-13	3.69E-02	-
Shots on Goal	5.17E-14	1.22E-01	-
Shots	3.61E-11	1.39E-01	17
Yellow Card	1.44E-01	0.00E+00	15
Red Card	5.34E-01	7.07E-03	-

assign scores to features in the training dataset, followed by a ranker search method that ranks each feature based on computed scores (Tang et al., 2014). Features with higher informativeness receive elevated scores, while less informative ones receive lower scores. The resultant complete set of features, ranked according to computed scores, is then presented to the end user for subset selection. Diverse filter-based feature selection methods, such as Variance Inflation Factor (VIF)–based, Pearson's correlation–based, and mutual-information-based feature selection, exist based on the selection principles applied.

VIF-Based Feature Selection

A popular measure for identifying multicollinearity between independent variables in a dataset is VIF. Interpreting the link between the characteristics and the target variable can be challenging due to multicollinearity, which can lead to instability in the calculated regression coefficients. By calculating the extent to which the correlation with other characteristics inflates the variance of a feature's coefficient, VIF measures the degree of multicollinearity [41, 42]. This equation is used to calculate it:

$$VIF = \frac{1}{1 - R_i^2} \tag{1}$$

where the coefficient of determination for the dataset's ith feature is denoted by R_i^2. In this study, the threshold of 10 has been utilized for selecting the features as illustrated in Fig. 3. Among all 47 features, 8 variables had $VIF \geq 10$ and removed from dataset. So 39 features were in the selected features set showing less mutlicolinearity. Therefore, other feature selection methods utilized to select more decreased number of features for prediction.

Correlation-Based Feature Selection

The underlying concept of correlation-based feature selection is not dependent on specific data transformations; it necessitates a means of quantifying the correlation between any pair of variables. Consequently, this technique is versatile and applicable to various supervised problems, including those involving the prediction of variables. It is an entirely automated algorithm, eliminating the need for users to specify thresholds or the number of

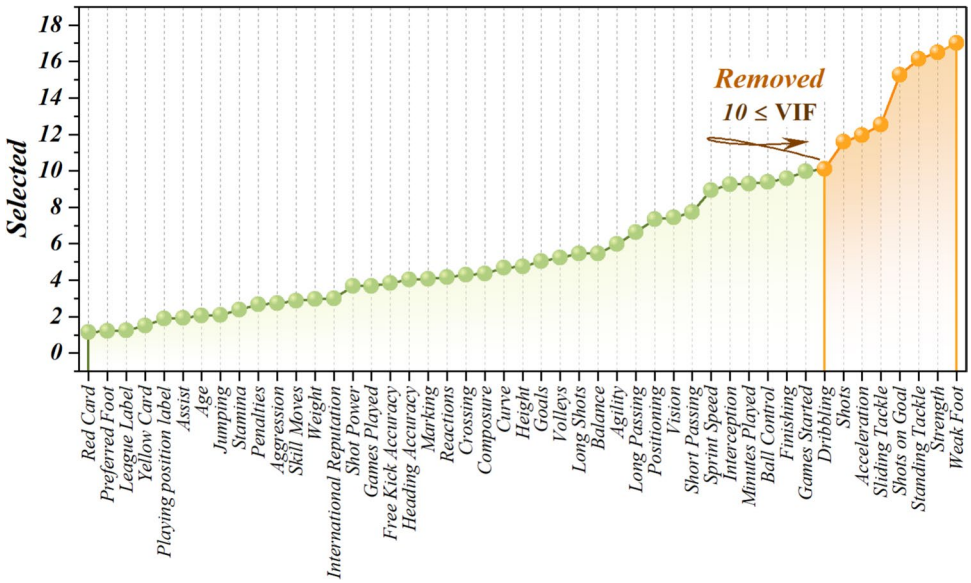

Fig. 3 Variance Inflation Factor results in feature selection

features to be selected, although such parameters can be easily incorporated if desired. Importantly, this method functions as a filter, avoiding the computational costs associated with repetitively employing a learning algorithm [43].

According to this approach, a feature V_i is said to be relevant if there exists some v_i and c for which $p(V_i = v_i) > 0$:

$$p(C = c|V_i = v_i) \neq p(C = c) \tag{2}$$

Empirical findings in the field of feature selection emphasize the necessity of removing not only irrelevant features but also redundant information. A feature is deemed redundant when it exhibits a high correlation with one or more other features. Therefore, an effective feature subset is characterized by features that are highly correlated with the target value yet display low correlations with each other.

If the correlation between each component in a test and an external variable is known, along with the inter-correlation between each pair of components, the correlation between a composite test (formed by summing the components) and the external variable can be anticipated from these parameters.

$$r_{zc} = \frac{k\overline{r_{zi}}}{\sqrt{k + k(k-1)\overline{r_{ii}}}} \tag{3}$$

where r_{zi} represents the correlation between the summed components and the outside variable, k is the number of components, $\overline{r_{zi}}$ is the average of the correlations between the components and the outside variable, and $\overline{r_{ii}}$ is the average inter-correlation between components.

Equation (2) represents Pearson's correlation coefficient, standardized for all variables. It indicates that the correlation between a composite and an external variable depends on the number of component variables, their inter-correlations, and the correlations between components and the external variable.

Higher correlations between components and the external variable elevate the overall composite-external variable correlation. Lower inter-correlations among components are linked to a stronger composite-external variable correlation. Additionally, increasing the number of components in the composite while maintaining consistent intercorrelations with other components and the external variable leads to a heightened correlation between the composite and the external variable.

Mutual Information–Based Feature Selection

In information theory, mutual information $I(X;Y)$ is the amount of uncertainty in X due to the knowledge of Y [44, 45]. Mathematically, mutual information is defined as

$$I(X;Y) = \sum_{x,y} p(x,y) log \frac{p(x,y)}{p(x)p(y)} \tag{4}$$

where $p(x, y)$ is the joint probability distribution function of X and Y, and $p(x)$ and $p(y)$ are the marginal probability distribution functions for X and Y. We can also say

$$H(X;Y) = H(X) - H(X|Y) \tag{5}$$

In this framework, where $H(X)$ represents marginal entropy, $H(X|Y)$ is conditional entropy and $H(X;Y)$ is joint entropy, mutual information serves as a measure of information gain between features and class attributes. Employing a greedy strategy, the method selects features that provide maximum information about the class attribute with minimal redundancy from the remaining set.

Then, TOPSIS (Order Priority Technique by Similarities to Ideal Solution) [46], as a simple ranking method, attempted to choose alternatives that simultaneously have the lowest Pearson's correlation coefficient and highest Mutual Information score. The optimal number of features (20 top features) was selected based on reports in Table 1.

Feature Engineering

In football, a handful of players labeled as superstars significantly impact a club beyond their on-field performance, influencing their transfer market values [47]. Clubs invest in popular players for global commercial appeal, enhancing profitability through various channels. Record-breaking transfers, like Cristiano Ronaldo's €94 million move in 2009, and ongoing record-breaking transfers, including those of Ousmane Dembélé (€140 million), Kylian Mbappé (€145 million plus €35 million commission), and Neymar Jr (€222 million), prompt examination into the factors influencing a player's market value.

To be able to answer the question: "*How is the distribution of players with a reputation index of 1 (least popular) to 5 (most popular) in five top-tier European leagues, and what effect has it had on the average market value of players in these leagues?*" useful reports are provided in Fig. 4 and Table 2.

Premier League had the largest number of players with a reputation index of 3 and 4, and the average market value of its players with a reputation index of 4 was higher than other leagues. Moreover, in other groups, players with marginal differences with La Liga were the second most expensive players. League 1 also had one expensive superstar, but dominant players in this league were less popular ones with the lowest market value. La Liga paid the highest range of salaries to less popular players. All these contributions led to conclude that managers of clubs in League 1 and Serie A were more concentrated on attracting superstars to earn more from ticket sales and reach a high level of performance based on their experience. Nevertheless, managers of Premier League clubs are more concentrated on players with low reputations but batter international reputations in the future and pay high amounts of money for those players. However, in La Liga and Bundesliga, managers preferred to invest in less popular players with low market values and concentrated on the experience of coaches in training them well to create new superstars.

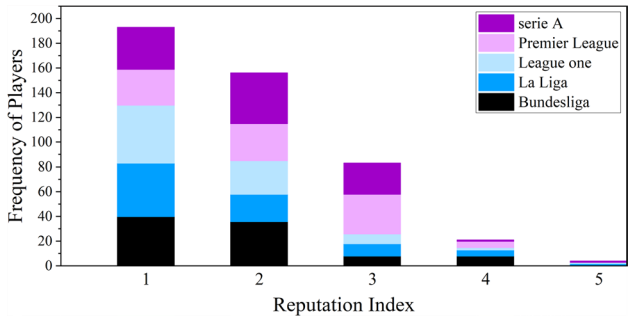

Fig. 4 Frequency of five top-tier European League football players in each category of reputation index

Machine Learning Algorithms

Extreme Gradient Boosting (XGB)

XGB, a highly effective gradient boosting machine (GBM) method, is well-known for its versatility in tackling supervised learning problems, including regression and classification. It is popular among data scientists because of its fast execution speed, which is due to its out-of-core processing capabilities [48]. XGBoost uses a set of equations to create predicted outputs $\hat{y}_i$ for ensemble tree models on a dataset DS with n instances and m features $DS = \{(X_i, y_i) : i = 1 \ldots n, x_i \in R^m, y_i \in R\}$

$$A_{.i} = \varnothing(X_i) = \sum_{k=1}^{K} f_k(X_i), f_k \in F \tag{6}$$

The variable K defines the number of trees, whereas f_k indicates the K_{th} tree.

To solve the given equation, the key is to minimize both the loss and regularization objectives by identifying the optimal set of functions.

Table 2 Average market values (€) of football players in each category of the reputation index

League	*Reputation index*				
	1	*2*	*3*	*4*	*5*
Serie A	7,602,206	12,490,854	31,640,000	57,000,000	77,000,000
Premier League	9,293,103	17,420,000	37,156,250	65,000,000	-
League 1	6,586,170	10,744,444	34,312,500	44,750,000	118,500,000
La Liga	11,479,069	18,704,545	38,300,000	52,000,000	95,250,000
Bundesliga	8,951,250	15,125,000	19,012,500	40,875,000	-

$$\varphi(\varnothing) = \sum_i 1\left(y_i, A_{\cdot i}\right) + \sum_k \Omega\left(f_k\right) \tag{7}$$

The loss function, written as 1, is defined by the difference between the expected output $\hat{y}_i$ and the real output y_i

O serves as a metric assessing the model's complexity to mitigate overfitting, and its calculation is governed by Eq. (7):

$$\Omega\left(f_k\right) = \gamma T + \frac{1}{2}\lambda\|w\|^2 \tag{8}$$

The weight assigned to each leaf is signified by W, and the whole number of leaves is symbolized by T.

During model training, decision trees are boosted to reduce the objective value iteratively. As the model begins to learn, it incorporates a new function (tree) symbolized by Eqs. (9) to (12):

$$\varphi^{(t)} = \sum_{i=1}^{n} 1(y_i, A_i^{\cdot(t-1)} + f_t(X_i)) + \Omega(f) \tag{9}$$

$$\Omega_{\text{split}} = \frac{1}{2}\left[\frac{\left(\sum_{i\in I_L} g_i\right)^2}{\sum_{i\in I_L} h_i + \lambda} + \frac{\left(\sum_{i\in I_R} g_i\right)^2}{\sum_{i\in I_R} h_i + \lambda} - \frac{\left(\sum_{i\in I} g_i\right)^2}{\sum_{i\in I} h_i + \lambda}\right] \tag{10}$$

$$\mathrm{g_i} = \partial_{A^{t-1}} 1(y_i, A.^{(t-1)}) \tag{11}$$

$$h_{\mathrm{i}} = \partial^2{}_{A^{t-1}} 1(y_i, A.^{(t-1)}) \tag{12}$$

Figure 5 depicts the schematic representation of XGB.

Adaptive Boosting (ADA)

Schapire [49] first proposed boosting techniques in 1990 to overcome the apparent weakness of decision trees when employed alone. Despite its limited capabilities, decision trees can be successively merged to produce a strong learner. The boosting technique iteratively adds new tree models to the ensemble, with each addition resulting in the replacement of the weakest tree. This guarantees that only the strongest tree contributes to the ensemble, gradually increasing the overall model efficiency, as shown in Eq. (12). However, issues developed after creating the initial basic tree model, with some samples in the dataset properly categorized and others misclassified. The approach improves the model's performance over time by doing repetitive computations and gradually building up tree models.

$$G_n(x) = G_{n-1}(x) + argmin_h \sum_{i=1}^{n} L(y_i, G_{n-1}(x) + T\left(x_i\right)) \tag{13}$$

The notation $T\left(x_i\right)$ refers to the newly added tree, $G_{n-1}(x)$ to the overall model produced in the previous round, and $G_n(x)$ to the overall model obtained after adding the new tree. The prediction result of the ith tree is indicated by $\mathrm{y_i}$.

The AdaBoost technique overcomes restrictions by gradually enhancing the model's classification skills through continual training. It starts by generating an initial weak classifier from training samples and then combining misclassified

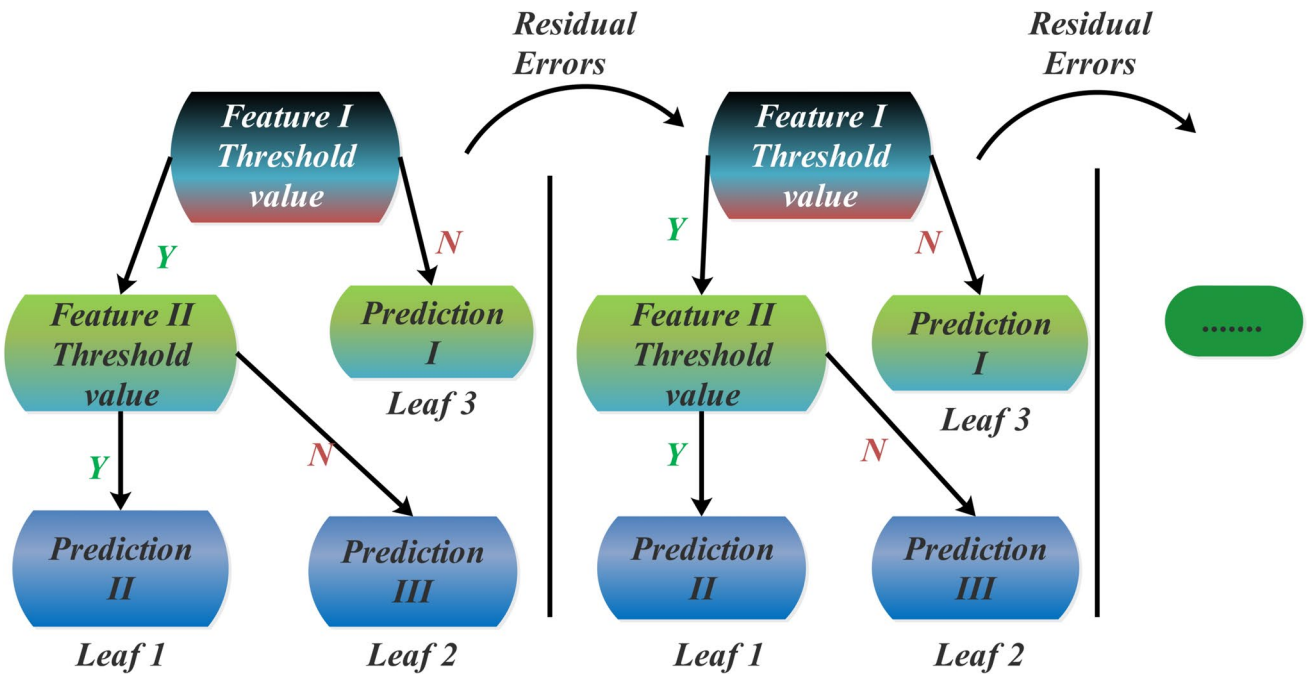

Fig. 5 Schematic representation of the XGB tree

examples with untrained data. The learning process is then used to generate further weak classifiers, and this iterative process is done several times. Each cycle consists of merging misclassified samples with untrained data to generate a new training sample [50]. This incremental technique improves the model's overall performance as a weak classification method. The AdaBoost method generates a robust classifier by combining numerous weak classifiers. It improves proper categorization by allocating different weights to samples. Correctly categorized samples are given lower weights, while incorrectly classified samples are given higher weights. This method pushes the model to prioritize misclassified data, increasing its capacity to accurately identify them in subsequent rounds [51].

Figure 6 depicts the computation process for the AdaBoost technique, in which each basic tree model is trained by modifying the weight distribution of each dataset sample. This results in varying training outcomes for each training dataset, and the final results are calculated by adding all the individual results [52].

The provided pseudocode outlines the AdaBoost algorithm [54].

Algorithm 1: Pseudocode outlining the AdaBoost algorithm.

Input: a set S, of m labeled samples: $S = ((x_i, y_i), i = (1, 2, \ldots, m))$, with labels in Y

Learn

A constant L

Initialize for all i: $w_j(i) = 1/m$ // initialize the wights

For $j = 1$ to L **do**

For all i

Calculate normalized weights $P_i(i) = \frac{w(i)}{\sum_i^m w(i)}$

$h_j = (S, P_j)$ // call weak Learn with normalized weights

Compute the error of h_j $\varepsilon_j = \sum_i p_j(i)\,[h_j(x_i \neq y_i)]$

If $\varepsilon_j > \frac{1}{2}$ then

L=j-1 $\beta_j = \frac{\varepsilon_j}{1-\varepsilon_j}$

For all i:

Estimate new weights $w_{j+1}(i) = w_j(i)\beta_j^{1-[h_j(x_i-y_i)]}$

End For

Output: $h_{final(x)} = \sum_{y \in Y}^{\arg\max} \sum_{j=1}^{L} (\log \frac{1}{\beta_h})\,[h_j(x = y)]$

Ali Baba and Forty Thieves Algorithm (AFT)

Malik et al. invented the AFT algorithm, which was inspired by Ali Baba and the Forty Thieves [55]. The algorithm reflects the repetitive aspect of the story, as a band of robbers pursues Ali Baba. Countermeasures in the novel, carried out by the main character, Marjane, are consistent with adaptive techniques in the algorithm. The village where Ali Baba dwells represents the algorithm's search space. Clever techniques from the narrative inspire the algorithm's exploration efficiency, serving as the foundation for mathematical models in the AFT algorithm's construction and development.

The fundamental goal of this project is to employ an optimization approach that uses Ali Baba and the Forty Thieves story as a unified model for human social interaction. The study connects aspects of the narrative, such as the thieves' coordinated attempts to find Ali Baba's residence, their range of travel, and Marjane's ingenious strategies to deceive them, to an objective function for optimization. This

approach leads to the creation of a novel metaheuristic technique, which is discussed in the following sections [56].

The AFT algorithm initializes by randomly positioning individuals (thieves) in a multidimensional search space, with each location representing a potential solution within defined boundaries. Fitness values, derived from a user-defined objective function, guide the search process, where thieves update their positions if a better solution is found. The search mimics the pursuit of Ali Baba, employing strategies like refining positions using shared knowledge, random exploration when misled, and balancing exploration and exploitation through global and individual best solutions. These adaptive strategies enhance the algorithm's optimization capability in complex problem spaces.

The AFT algorithm's pseudocode can be concisely outlined through the iterative steps provided in Algorithm 2.

Algorithm 2: A pseudocode description of the AFT.

u_i and l_j represent the upper and lower bounds of the j_{th} measurement, respectively. x denotes the position, and n indicates the number of thieves.

Prepare the location for the best $best_t^i$ and the global best $gbest$ location.

Assess the fitness function.

Set $t \leftarrow 1$

While $(t < T)$ **do**

$\mathrm{T}d_t = 1 \times e^{-2(\frac{t}{T})^2}$

$\mathrm{P}p_t = 0.1 \times \log\left(2\left(\frac{t}{T}\right)^{0.1}\right)$

For $i = 1, 2, \ldots, n$ **do**

If $(rand \geq 0.5)$ **then**

If $(rand \geq P_{pt})$ **then**

$$x_{t+1}^i = gbest_t + \left[Td_t\left(best_t^i - y_t^i\right)r_1 + Td_t\left(y_t^i - m_t^{a(i)}\right)r_2\right]sgn(rand - 0.5)$$

Else

$$x_{t+1}^i = Td_t\left[\left(u_j - l_j\right)rand + l_j\right]$$

Else If

Else

$$x_{t+1}^i = gbest_t - \left[Td_t\left(best_t^i - y_t^i\right)r_1 + Td_t\left(y_t^i - m_t^{a(i)}\right)r_2\right]sgn(rand - 0.5)$$

End If

End For

For $i = 1, 2, \ldots, n$ **do**

Inspect, assess, and revise the new positions.

Assess and revise the solutions for $m_t^{a(i)}$, $best_t^i$, $gbest_t$

End For

$t = t + 1$

End While

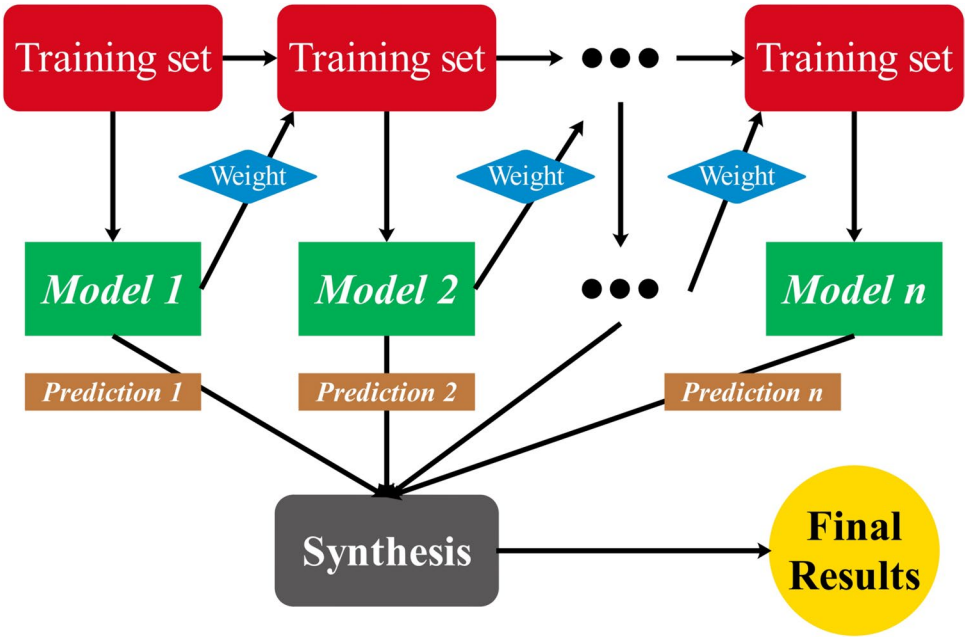

Fig. 6 Process of AdaBoost calculation [53]

Crystal Structure Algorithm (CSA)

Crystals are solid minerals with atoms and particles structured in a periodic crystalline shape. The name is derived from the Greek word for *Cold-Frozen*. Discovered by Kepler, Hooke, and Hogens in the seventeenth century [57], crystals exhibit a cyclical arrangement of atoms that forms a lattice, influencing the overall structure. The crystal lattice, characterized by infinite geometric forms, gives rise to diverse shapes [58]. The structure is discontinuous, defined by an endless lattice form, with each lattice point connected to its position. Crystals vary in size and form, and their properties might be anisotropic or isotropic [59].

The CSA optimizes solutions by simulating crystallographic processes, where crystal positions are initialized randomly within defined bounds. Iterative updates refine these positions using four improvement strategies (basic, optimal, average, and combined cubic), guided by the best crystal configurations and their mean merit. The algorithm balances exploration and exploitation through dynamic adjustments, ensuring effective convergence. A boundary flag maintains constraints on variable solutions, and optimization ends after a set number of iterations, signaling the completion of the search process. Figure 7 illustrates the technique using a flow chart.

Henry Gas Solubility Optimization (HGSO)

The Henry Gas Solubility Optimization algorithm, inspired by Henry's gas law, is a novel physics-based optimization method introduced by Hashim et al. It utilizes the principles of gas solubility in liquids, with a focus on low-solubility gases [60]. The main contributing elements are temperature and pressure, with higher temperatures typically increasing solid solubility and decreasing gas solubility. The program uses these insights to optimize operations across a variety of applications [61].

The HGSO algorithm improves gas solubility in liquids by simulating the effects of increased pressure through an eight-step optimization process. It begins by initializing gas populations, positions, Henry's constants, and partial pressures. Gases are grouped by type, and the optimal gas in each group is identified based on equilibrium positions. The algorithm iteratively updates Henry's constant using temperature-dependent calculations, which influence solubility values for each gas. Positions of gases are dynamically updated, leveraging solubility and fitness evaluations to guide the search for optimal solutions. To avoid local optima, gas positions are adjusted based on interactions and fitness, ensuring exploration and exploitation. Finally, the weakest agents are ranked and repositioned, enhancing the algorithm's convergence and overall performance.

The steps of the HGSO algorithm are detailed in Algorithm 3.

Algorithm 3: Pseudocode definition of the HGSO.

Starting: $X_i(1 = 1, 2, \ldots, N)$, number of gas types $i, H_j, P_i, j, C_j, l_1, l_2\ and\ l_3$.

Separate the population of agents into groups based on the gas types, each group having the same value for Henry's constant (H_j).

Assess each cluster j

Become the best gas $X_i, best$ in each cluster, and the finest search agent X_{best}

While t <maximum number of iterations **do**

For each quest agent **do**

Update the locations of all search agents using:

$$X_{i,j}(t+1) = X_{i,j}(t) + F \times r \times \gamma \times (X_{i,best}(t) - X_{i,j}(t)) + F \times r \times a \times (S_{i,j}(t) \times X_{i,best}(t) - X_{i,j}(t))$$

and

$$\gamma = \beta \times \exp\left(-\frac{F_{best}(t) + \varepsilon}{F_{i,j}(t) + \varepsilon}\right), \qquad \varepsilon = 0.05$$

End For

Update Henry's coefficient of each gas type using:

$$H_j(t+1) = H_j(t) \times e^{(-c_j(1/T(t)) - (1/T^{\theta}))}, T(t) = e^{(-t/iter)}$$

Update the solubility of each gas using:

$$S_{i,j}(t) = K \times H_j(t+1) \times P_{i,j}(t)$$

Rank and choose the number of worst agents using:

$$N_w = N * (rand\ ((C_2 - C_1) + C_1), C_1;\ C_1 = 0.1, C_2 = 0.2$$

Update the situation of the worst agents using:

$$G_{i,j} = G_{Min(i,j)} + r \times (G_{Max(i,j)} - G_{Min(i,j)})$$

Update the best gas $X_i, best$, and the best search agent X_{best}

End While

$\boldsymbol{t = t + 1}$

Return X_{best}

Mayfly Optimization Algorithm (MOA)

Mayflies, categorized under Palaeoptera in the order Ephemeroptera, are insects that remarkably emerge in May in the UK. Immature mayflies undergo several years of growth as aquatic nymphs before transitioning to adult mayflies. Male adults typically gather in swarms a few meters above the water surface to attract females. Mating is a brief process lasting only a few seconds, after which eggs are deposited into the water, continuing the cycle. Allan and Flecker [62] and Barbara, Peckarsky et al. [63] provide detailed information on this behavior. Zervoudakis and Tsafarakis [64] introduced MOA as an innovative method for addressing problems. This hybrid method merges the strengths of traditional optimization methods like PSO [65], GA [66], and FA. The elements of MOA are outlined as follows:

The motion of male mayflies in the algorithm is determined by their position and velocity, which are updated based on their previous states and the influence of cognitive

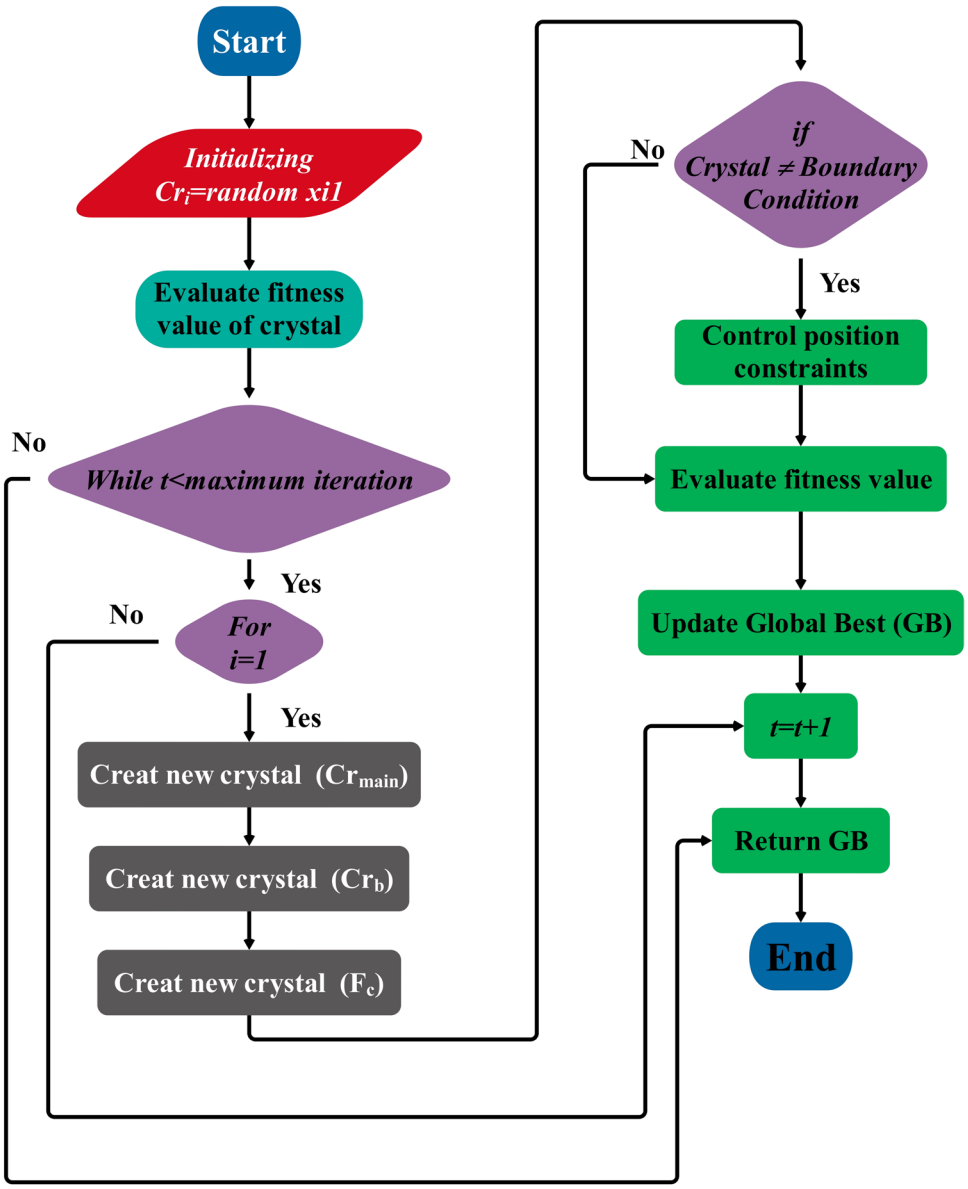

Fig. 7 Flow chart of the Crystal Structure Algorithm (CSA)

and social factors. Male mayflies maintain high speeds and adjust their movements according to the best solutions they have visited (personal best) and the global best, with attraction and visibility coefficients influencing these updates. The nuptial dance introduces a stochastic element to enhance exploration, where coefficients gradually decrease with iterations. Female mayflies are attracted to high-performing males, with their movements influenced by the fitness of their solutions and the Cartesian distance to the males. If a female's solution is not attracted to a male, a random walk component is applied. The algorithm incorporates interbreeding, where the top-performing male and female mayflies are paired to produce offspring with traits inherited from both parents. These offspring undergo mutation, introducing variability and enhancing the exploration of the solution space.

Algorithm 4 signifies pseudocode Mayfly Optimization Algorithm (MA) [67].

Algorithm 4: Pseudocode delineation of the MOA.

Objective function $f(x), x = (x_1, \ldots, x_d)^T$

Adjust the male mayfly populace $x_i (i = 1, 2, \ldots, N)$ and velocities v_{mi}

Adjust the female mayfly populace $y_i (i = 1, 2, \ldots, M)$ and velocities v_{fi}

Estimate solution

Discovery global best

While stopping criteria are not met

Update velocities and solutions of males and females

Estimate solutions

Rank the mayflies

Mate the mayflies

Estimate offspring

Separate offspring to male and female randomly

Change the best solutions with the best new ones

Update $pbest$ and $gbest$

End While

Postprocess results and visualization

Statistical Evaluation Metrics

The performance evaluation of the analyzed models is conducted employing five specified metrics designated as R^2, RMSE, MAE, MNB, VAF, and MAPE.

R^2 (Coefficient of Determination):

R^2 is a statistical measure denoting the quality of the model fit by quantifying how much variability in the dependent variable is accounted for by the independent variable. The scale ranges from 0 to 1, where a higher R^2 signifies a better model fit and an R^2 of 1 indicates a complete explanation of the dependent variable's variability.

RMSE (Root Mean Square Error):

RMSE serves as a metric to measure the average magnitude of disparities between predicted and observed values in a model. A lower RMSE reflects heightened predictive accuracy, signifying minimized discrepancies between the model's predictions and the actual observed values.

MAE (Mean Absolute Error):

MAE assesses model performance by measuring the average absolute differences between predicted and actual values, offering a direct gauge of accuracy by conveying the average magnitude of prediction errors.

MNB (Mean Normalized Bias):

MNB is a statistical metric applied in model evaluation, specifically in regression analysis. It gauges the average deviation between predicted and actual values, considering normalization to provide a comprehensive assessment of predictive performance.

VAF (Variance Accounted For):

Variance Accounted For (VAF) is a metric used in regression analysis or predictive modeling to measure the proportion of variance in a dependent variable explained by an independent variable or a statistical model.

MAPE (Mean Absolute Percentage Error):

Mean Absolute Percentage Error (MAPE) is a widely used metric for forecast accuracy due to its scale-independence and ease of interpretation.

Equations (14) to (19) serve as the mathematical representations of the metrics mentioned above.

$$R^2 = \left(\frac{\sum_{i=1}^{n}(A_i - \overline{A})(P_i - \overline{P})}{\sqrt{\left[\sum_{i=1}^{n}(A_i - \overline{P})^2\right]\left[\sum_{i=1}^{n}(P_i - \overline{P})^2\right]}} \right)^2 \quad (14)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(P_i - A_i)^2}{n}} \quad (15)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n} \|P_i - A_i\| \quad (16)$$

$$MNB = \frac{1}{n}\sum_{i=1}^{N} \frac{P_i - A_i}{R} \quad (17)$$

$$VAF = \left(1 - \frac{var(b_i - \overline{b})}{var(b_i)}\right) \times 100 \quad (18)$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{N} \frac{A_i - P_i}{A_i} \quad (19)$$

Symbols A_i and P_i denote the actual observed values and corresponding predicted values, respectively. $\overline{A}$ and $\overline{P}$ represent the mean outcomes in testing and predicting, with n indicating the total sample count in the analyzed dataset. The symbol R serves as a scaling factor specific to each data point, commonly employed for normalizing bias.

Discussion of Results

Cross-Validation

To assess the model's performance, the dataset is divided into subsets based on the K-fold cross-validation. The models are trained on all but one of these subsets (the "training fold") and evaluated on the remaining subset (the "test fold"). This procedure is repeated, with each subset being used as the test fold exactly once [68]. The final performance metrics are reported in Table 3. One of the key advantages of cross-validation is that it helps mitigate overfitting, providing a more reliable estimate of the model's generalizability by ensuring that the model is tested on different subsets of the data.

Hyperparameter Optimization and Convergence Analysis

The learning process of an algorithm is influenced by hyperparameters, which represent specific values or weights. XGB and ADA, as highlighted earlier, provide an extensive range of hyperparameters that can be fine-tuned to maximize accuracy. Automated tuning of these learnable parameters enables XGB and ADA to effectively identify patterns and regularities within datasets. In tree-based models like XGB and ADA, these parameters include decision variables at each node. Given the complexity of XGB and ADA, with its numerous design choices, the primary challenge lies in optimizing the selection of multiple hyperparameters. This can be addressed through efficient hyperparameter tuning methods. In this study, the hyperparameters for XGB and ADA were selected based on previous research on detecting efficient hyperparameters [69, 70]. These hyperparameters are n_estimators, max_depth, learning_rate, colsample_bytree, and *subsample*. The results of adjusting hyperparameters for XGB and ADA models are elaborated in Tables 4 and 5, respectively.

Convergence graphs are invaluable instruments for visualizing the dynamics of iterative processes across diverse domains. Through a deep examination of trends, rates, and minimum values, profound insights can be gleaned into the efficacy and constraints inherent in the applied processes or algorithms [71–74].

Table 4 The results of hyperparameter optimization for XGB

Hyperparameter	Models			
	XGAF	XGCS	XGMO	XGHG
n_estimators	102	104	134	576
max_depth	76	4	56	248
learning_rate	0.0779797	0.435474	0.130039	0.336626
colsample_bytree	0.334355	0.358836	0.763059	0.13099
subsample	0.769512	0.420914	0.302734	0.586162

Table 3 Cross-validation results

Model	Metric	Metric results through folds (test)				
		K1	K2	K3	K4	K5
XGB	R^2	0.806	0.805	0.968	0.882	0.808
	RMSE	8,650,976	6,280,439	1,821,063	3,332,494	7,628,000
ADA	R^2	0.916	0.794	0.927	0.946	0.875
	RMSE	6,084,973	4,013,015	6,914,160	2,540,025	8,272,643

Table 5 The results of hyperparameter optimization for ADA

Hyperparameter	Models			
	ADAF	ADCS	ADMO	ADHG
n_estimators	54	60	86	36
learning_rate	0.15145	0.397159	0.252832	0.904226

Figure 8 illustrates the convergence process of hybrid models associated with XGB and ADA. The graph demonstrates a consistent decrease in RMSE with extended training periods, indicating improved model fit and successful learning of underlying data patterns. The RMSE reduction rate is especially swift in the initial periods, followed by a gradual stabilization in the later stages.

The XGAF, XGCS, XGMO, and XGHG models demonstrated lower RMSE value (approximately €5 million) in the initial iteration compared to the ADAF, ADCS, ADMO, and ADHG models. The XGAF model displayed superior performance with the lowest RMSE value of €1.91 million at the 170th iteration. Additionally, the ADAF model had effectively converged, displaying a lower RMSE than other ADA-based hybrid models, specifically at €2.69 million in 160 iterations.

Prediction Performance of Developed Models

Tables 6 and 7 demonstrate the results obtained from both the XGB- and ADA-based models, along with the corresponding hybrid models for comprehensive examination.

In Table 6, a thorough investigation reveals that the XGB single model displaying R^2 values of 0.9586 and an RMSE of €4.55 million demonstrated comparatively lower predictive efficiency. Conversely, the XGAF model outperformed others, revealing superior performance with R^2 and RMSE values of 0.9905 and €1.91 million, respectively. This error value is under 10% of average market values reported in Table 2 for players with various levels of popularity. Notably, the XGAF model attained a VAF of 98.8, signifying a higher percentage of explained variance and superior explanatory power compared to alternative models. Following, the XGAMCH ensemble model secured the second position with $R^2 = 0.9889$ and RMSE = 2.03 million euros.

The results presented in Table 7 underscore the superior performance of the ADAF model, attaining values of 0.9847 and €2.69 million for the R^2 and RMSE metrics, respectively. Remarkably, the ADAF model exhibited the highest VAF value, equal to 977, and an MNB value of − 0.269, indicative of a smaller average bias between predicted and actual values. The ADAMCH ensemble model secured the second position with values of R^2=0.9742 and RMSE = 3.23 million euros.

Also, Fig. 9 illustrates that based on all metric values results (except for MNB), XGAF was the best performer, and its superiority was more visible in the case of RMSE (€2.6 million lower than XGB as the worst model) and MAE error values.

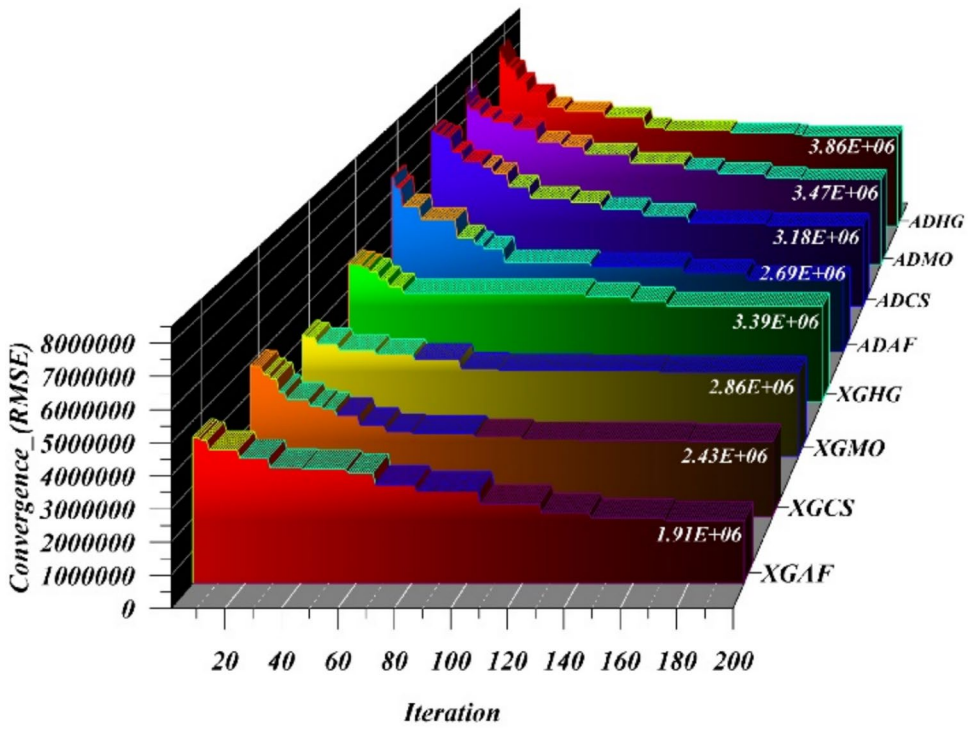

Fig. 8 The convergence of optimization procedures related to hybrid models

Table 6 Metric reports related to XGB-based prediction models' performance in players' market value prediction

Model	Part	Indicator values					
		RMSE	R^2	MAE	MNB	VAF	MAPE
XGB	*Train*	4.62E+06	0.9615	2.31E+06	−7.58E-02	9.12E+01	1.34E+01
	Validation	5.36E+06	0.9448	2.86E+06	−6.11E-02	8.76E+01	1.53E+01
	Test	3.11E+06	0.9676	1.82E+06	−4.93E-02	9.52E+01	1.13E+01
	Total	4.55E+06	0.9586	2.32E+06	−6.96E-02	9.12E+01	1.34E+01
XGAF	*Train*	1.79E+06	0.9920	6.79E+05	−4.41E-02	9.90E+01	5.42E+00
	Validation	2.31E+06	0.9881	1.34E+06	−5.13E-02	9.82E+01	8.88E+00
	Test	2.01E+06	0.9850	1.31E+06	−3.98E-02	9.82E+01	7.47E+00
	Total	1.91E+06	0.9905	8.74E+05	−4.45E-02	9.88E+01	6.24E+00
XGCS	*Train*	2.37E+06	0.9838	1.85E+06	−2.60E-02	9.84E+01	4.12E+01
	Validation	2.33E+06	0.9858	1.51E+06	7.02E-03	9.83E+01	1.42E+01
	Test	2.76E+06	0.9697	1.76E+06	−2.73E-02	9.69E+01	9.51E+00
	Total	2.43E+06	0.9821	1.78E+06	−2.12E-02	9.82E+01	3.24E+01
XGMO	*Train*	2.74E+06	0.9784	1.36E+06	−2.96E-02	9.77E+01	1.58E+01
	Validation	3.14E+06	0.9721	2.09E+06	−3.11E-02	9.70E+01	1.59E+01
	Test	3.12E+06	0.9611	2.14E+06	−3.50E-02	9.60E+01	1.34E+01
	Total	2.86E+06	0.9753	1.59E+06	−3.06E-02	9.74E+01	1.54E+01
XGHG	*Train*	3.36E+06	0.9710	9.40E+05	−2.02E-02	9.62E+01	4.74E+00
	Validation	4.24E+06	0.9592	2.22E+06	−3.74E-02	9.33E+01	1.17E+01
	Test	2.46E+06	0.9757	1.78E+06	−3.45E-02	9.75E+01	1.16E+01
	Total	3.39E+06	0.9687	1.26E+06	−2.50E-02	9.59E+01	6.81E+00
XGAMCH	*Train*	1.84E+06	0.9914	9.65E+05	−3.00E-02	9.89E+01	7.83E+00
	Validation	2.53E+06	0.9857	1.53E+06	−2.82E-02	9.79E+01	9.18E+00
	Test	2.25E+06	0.9798	1.47E+06	−3.42E-02	9.79E+01	8.51E+00
	Total	2.03E+06	0.9889	1.13E+06	−3.04E-02	9.86E+01	8.13E+00

Figure 10 illustrates scatter plots for XGB and ADA single models, the best-performing XGAF and ADAF hybrid models, and the XGAMCH and ADAMCH ensemble models. The scatter plot is the prevalent form of data visualization, which employs the method of data representation for depicting bivariate data (x_i, y_i), aiming to show associations between the x and y values in each respective pair i [75]. Detailed dispersion points, primarily controlled by RMSE and R^2, ensure that higher density corresponds to lower RMSE values. A superior fit of the line to the data is indicated by a higher R^2 value. Within the plots, three lines are featured: the best-fit line ($X = Y$) and two dashed lines representing a 15% underestimation and overestimation.

The XGB and ADA models feature RMSE values at 45.5% and 43.1% and the lowest R^2 values of 0.9586 and 0.9532 displayed notable dispersion compared to the Best Fit line, suggesting reduced accuracy in predicting the market value of football players. Contrastingly, the XGAF and ADAF hybrid models displayed outstanding performance, displaying significant reductions in RMSE values compared to the XGB and ADA models. Particularly, the XGAMCH and ADAMCH ensemble models secured a commendable second position in the overall performance ranking. The R^2 values for the XGAF and ADAF models stand at 0.9905 and 0.9847, respectively.

Figure 11 shows the Taylor diagram for the developed models. The Taylor diagram compares estimated values by models with actual values. It visually represents the standard deviation ratio, correlation coefficient, and centered root mean squared error of each model compared to the reference dataset. This allows for easy identification of models that perform better in prediction [76].

The XGAF and XGAMCH models exhibited superior performance, as evidenced by the highest correlation coefficients and smallest standard deviation, indicating closer alignment with the actual market values. Following closely in performance assessment is the ADAF model with higher RMSE and lowest R^2 value. Conversely, the single ADA and XGB models displayed the lowest efficiency and performance, rendering the models unsuitable for reliable predictions of players' market values.

Table 7 Metric reports related to ADA-based prediction models' performance in players' market value prediction

Model	Part	Indicator values					
		RMSE	R^2	MAE	MNB	VAF	MAPE
ADA	*Train*	4.37E+06	0.9553	3.54E+06	−4.06E-01	9.33E+01	2.46E+01
	Validation	4.55E+06	0.9503	3.11E+06	−2.23E-01	9.24E+01	2.01E+01
	Test	3.80E+06	0.9462	2.54E+06	−1.57E-01	9.31E+01	1.45E+01
	Total	4.31E+06	0.9532	3.32E+06	−3.41E-01	9.31E+01	2.24E+01
ADAF	*Train*	2.60E+06	0.9877	2.20E+06	−3.09E-01	9.80E+01	1.88E+01
	Validation	3.09E+06	0.9790	2.22E+06	−2.03E-01	9.69E+01	1.67E+01
	Test	2.71E+06	0.9741	1.95E+06	−1.49E-01	9.69E+01	1.25E+01
	Total	2.69E+06	0.9847	2.17E+06	−2.69E-01	9.77E+01	1.75E+01
ADCS	*Train*	3.14E+06	0.9772	2.52E+06	−2.95E-01	9.67E+01	1.95E+01
	Validation	3.63E+06	0.9638	2.31E+06	−1.64E-01	9.59E+01	1.59E+01
	Test	2.86E+06	0.9695	1.98E+06	−1.18E-01	9.63E+01	1.23E+01
	Total	3.18E+06	0.9739	2.41E+06	−2.49E-01	9.65E+01	1.79E+01
ADMO	*Train*	3.49E+06	0.9704	2.76E+06	−3.03E-01	9.59E+01	2.07E+01
	Validation	3.73E+06	0.9647	2.46E+06	−1.66E-01	9.52E+01	1.69E+01
	Test	3.09E+06	0.9631	2.15E+06	−1.18E-01	9.57E+01	1.29E+01
	Total	3.47E+06	0.9685	2.62E+06	−2.54E-01	9.57E+01	1.89E+01
ADHG	*Train*	3.84E+06	0.9655	3.17E+06	−3.65E-01	9.50E+01	2.30E+01
	Validation	4.22E+06	0.9519	2.89E+06	−2.33E-01	9.43E+01	1.99E+01
	Test	3.54E+06	0.9529	2.43E+06	−1.51E-01	9.41E+01	1.45E+01
	Total	3.86E+06	0.9616	3.02E+06	−3.13E-01	9.48E+01	2.13E+01
ADAMCH	*Train*	3.20E+06	0.9773	2.62E+06	−3.18E-01	9.66E+01	2.03E+01
	Validation	3.60E+06	0.9666	2.38E+06	−1.91E-01	9.58E+01	1.71E+01
	Test	2.99E+06	0.9667	2.10E+06	−1.34E-01	9.60E+01	1.28E+01
	Total	3.23E+06	0.9742	2.50E+06	−2.71E-01	9.64E+01 s	1.87E+01

Prediction of Market Value for Players with Different Reputation Index

To examine the generalization capability of XGAF and XGAMCH models in estimating the market value of players within various datasets, their effectiveness in the prediction of player's value in each group of reputation index (least popular with an index of 1 to most popular with an index of 5) is presented in Fig. 12. In the case of players with a reputation index of 5, due to the low number of samples, there is not a comprehensive representation of error values, but in the case of superstars with a reputation index of 4, XGAF was the most accurate predictor with most errors lower than 10%. The estimation performance of XGAF in the remaining two groups of samples became less accurate, with five to more than ten times higher error values. In contrast to XGAF, XGAMCH demonstrated maximum error in the case of the market value of samples with a reputation index of 4. However, still, its prediction error in the case of most of the players with higher popularity was lower than those with least popularity, indicating that introduced estimation models are reliable market value predictors, especially useful for clubs planning to concentrate on popular and expensive players with a high risk of decision making in the transfer market.

To discuss more on the obtained results, firstly, the Wilcoxon signed-rank test was conducted to compare models' prediction accuracy in pairs. Then, the prediction performance of the developed best prediction models and their performance metrics are compared with those of models in the literature.

Statistical Testing

This study provides an in-depth analysis of model performance by conducting a pairwise comparison of prediction models using the Wilcoxon signed-rank test, as proposed by Demšar (2006) [77]. A total of 66 pairwise comparisons were made, with Bonferroni-adjusted p-values below 0.05, indicating significant performance differences between the models [78, 79]. The results, as reported in Table 8, show that the combination of ADCS_ADMO exhibits the most notable performance difference, achieving the highest value of 0.8246. Additionally, no clear preference emerges

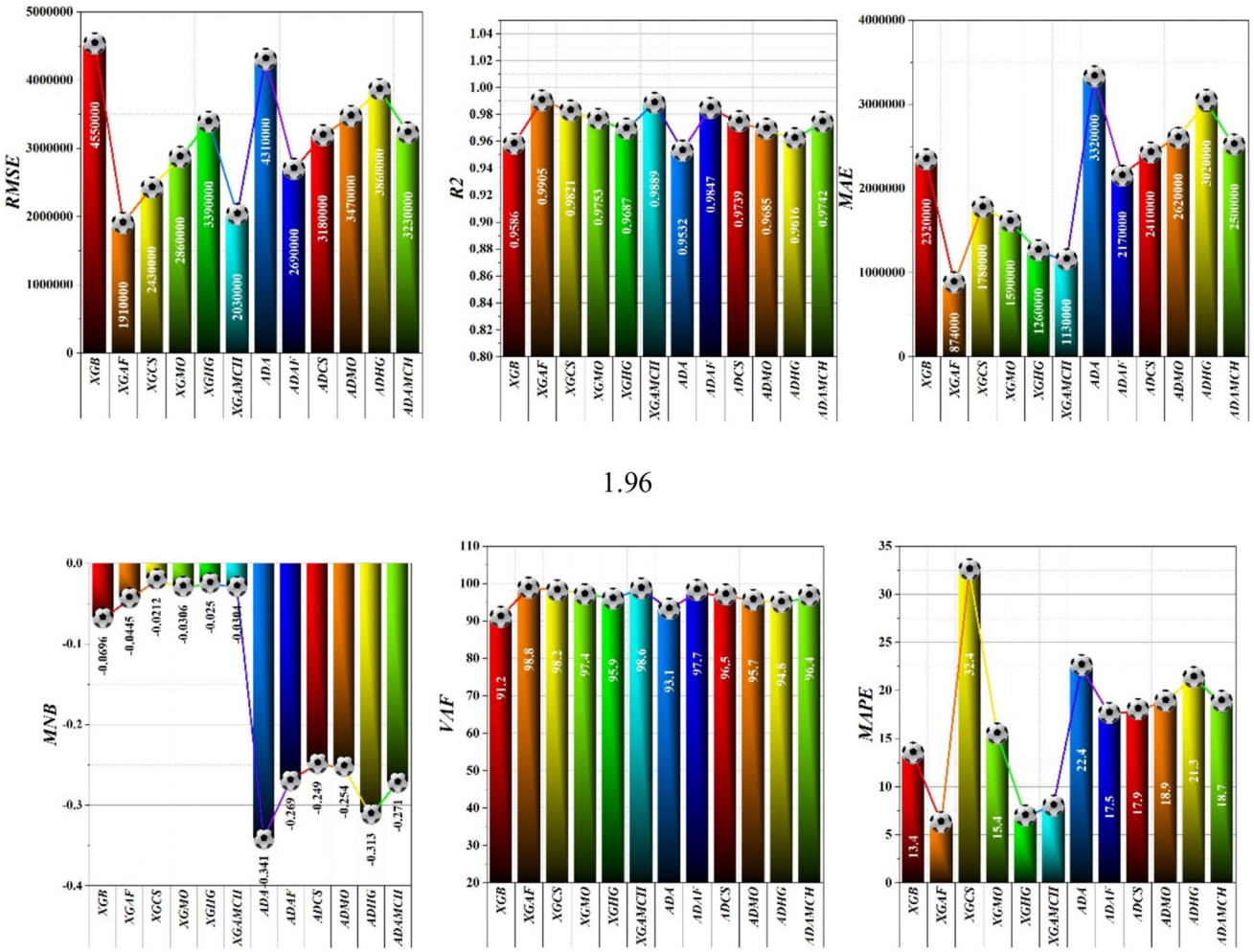

Fig. 9 Comparison of developed models' performance in predicting the market value of players based on performance metric results

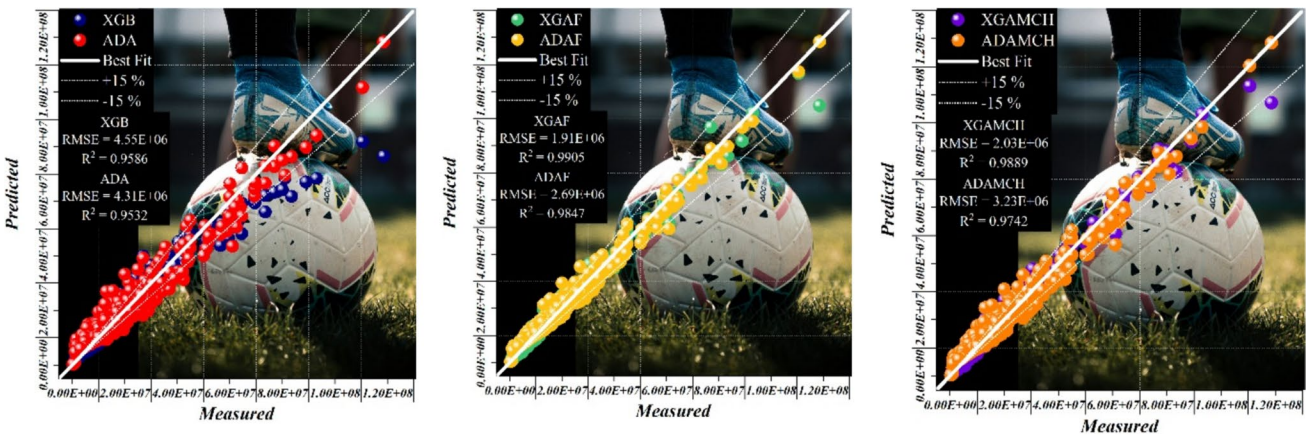

Fig. 10 The scatter diagram illustrates the formulated models

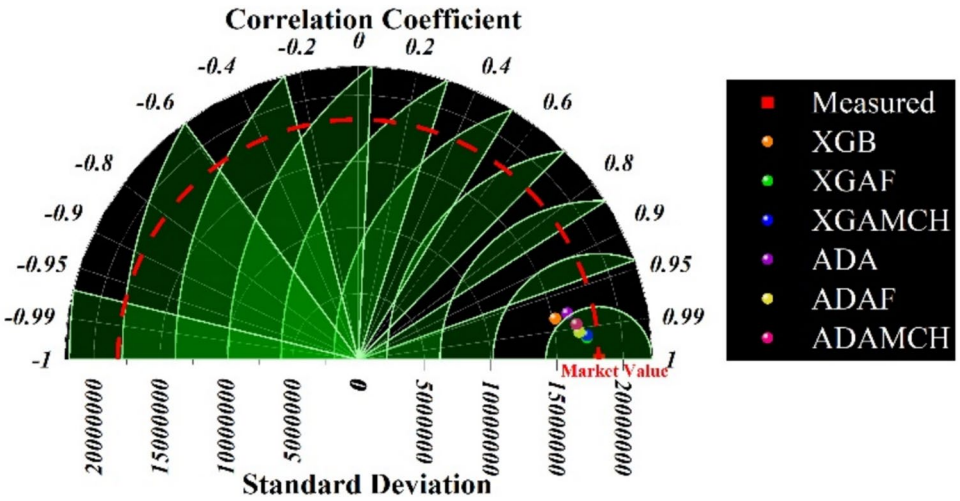

Fig. 11 The Taylor diagram of selected models

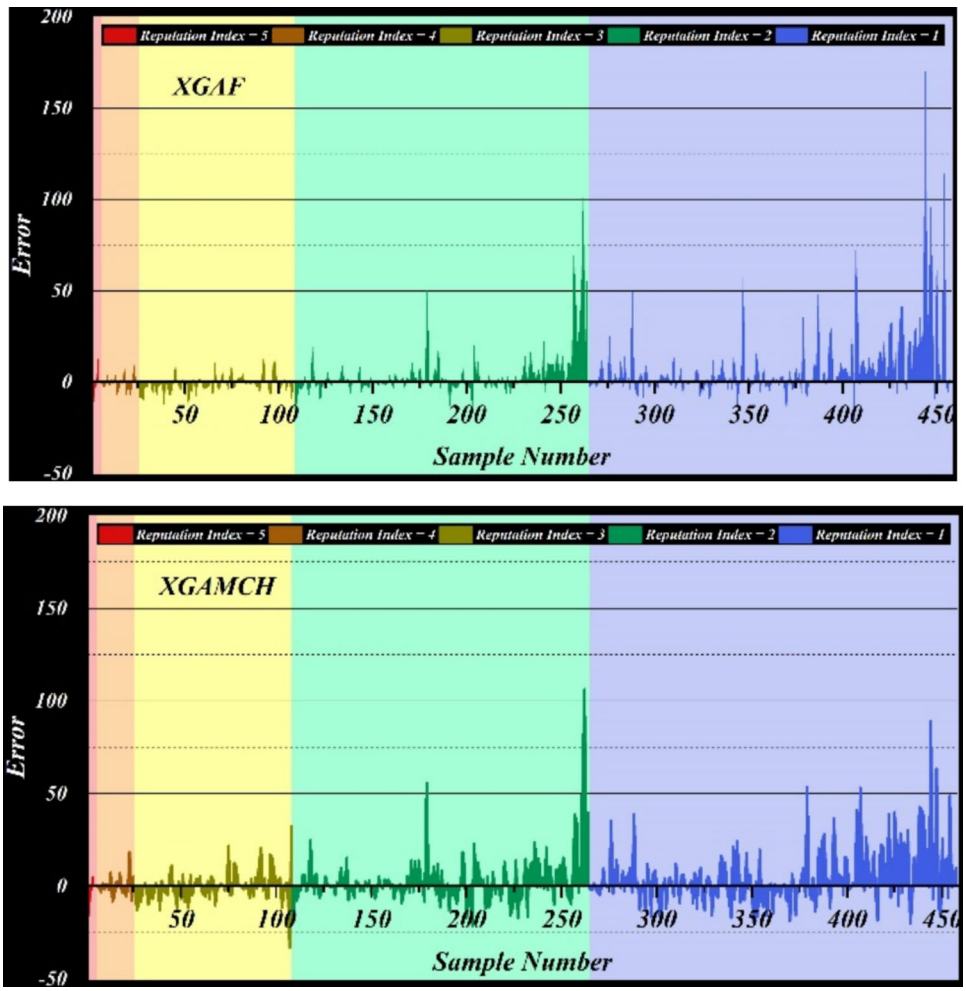

Fig. 12 Comparison between real market values of players with different reputation index and those predicted by best models

between several other model pairings, such as XGMO_ XGHG, ADAF_ADHG, and XGAF_ XGAMCH, which show similar performance levels. However, certain combinations, like XGAF_ XGAMCH (p-value = 0.6846) and ADAF_ADHG (p-value = 0.6710), consistently perform well, demonstrating significant differences when compared to others, particularly outperforming lower-performing pairs like XGB_XGMO (p-value = 0.1505) and ADCS_ADHG (p-value = 6.81E-06).

Table 8 Result of Wilcoxon test

Difference of models	Parameter *p*_value	Difference of models	Parameter *p*_value
XGB_XGAF	0.012242	XGMO_ADAF	5.82E-19
XGB_XGCS	0.072716	XGMO_ADCS	2.56E-08
XGB_XGMO	0.150498	XGMO_ADMO	1.21E-06
XGB_XGHG	0.145147	XGMO_ADHG	5.98E-10
XGB_XGAMCH	0.056635	XGMO_ADAMCH	1.33E-10
XGB_ADA	7.84E-20	XGHG_XGAMCH	0.255295
XGB_ADAF	3.97E-28	XGHG_ADA	1.45E-11
XGB_ADCS	4.92E-17	XGHG_ADAF	9.26E-20
XGB_ADMO	1.67E-14	XGHG_ADCS	1.17E-09
XGB_ADHG	1.80E-17	XGHG_ADMO	4.97E-08
XGB_ADAMCH	5.86E-20	XGHG_ADHG	1.29E-10
XGAF_XGCS	0.615935	XGHG_ADAMCH	1.03E-11
XGAF_XGMO	0.114656	XGAMCH_ADA	9.52E-10
XGAF_XGHG	0.10491	XGAMCH_ADAF	5.90E-21
XGAF_XGAMCH	0.684584	XGAMCH_ADCS	1.99E-08
XGAF_ADA	1.99E-09	XGAMCH_ADMO	7.47E-07
XGAF_ADAF	1.27E-21	XGAMCH_ADHG	8.33E-10
XGAF_ADCS	1.75E-08	XGAMCH_ADAMCH	7.11E-11
XGAF_ADMO	8.00E-07	ADA_ADAF	0.606313
XGAF_ADHG	5.69E-10	ADA_ADCS	3.10E-10
XGAF_ADAMCH	3.26E-11	ADA_ADMO	1.28E-14
XGCS_XGMO	0.289442	ADA_ADHG	0.12244
XGCS_XGHG	0.208925	ADA_ADAMCH	1.66E-05
XGCS_XGAMCH	0.342312	ADAF_ADCS	3.42E-17
XGCS_ADA	6.81E-06	ADAF_ADMO	4.09E-13
XGCS_ADAF	3.00E-09	ADAF_ADHG	0.670999
XGCS_ADCS	0.000965	ADAF_ADAMCH	1.85E-08
XGCS_ADMO	0.002831	ADCS_ADMO	0.824608
XGCS_ADHG	1.04E-05	ADCS_ADHG	1.46E-10
XGCS_ADAMCH	3.06E-05	ADCS_ADAMCH	1.69E-13
XGMO_XGHG	0.648296	ADMO_ADHG	9.67E-16
XGMO_XGAMCH	0.155277	ADMO_ADAMCH	7.49E-16
XGMO_ADA	5.25E-10	ADHG_ADAMCH	5.95E-06

Comparison with Existing Literature

Following are the comparative descriptions:

- Al-Asadi and Tasdemir [80] proposed a machine learning–based method to predict football players' market values using FIFA 20 data. Four regression models (Linear Regression, Multiple Linear Regression, Regression Tree, and Random Forest Regression) were tested, with Random Forests achieving the best accuracy and lowest error with R^2 and RMSE values of 0.95 and 1,649,921. These results are comparable with the prediction performance of single models in this study, but it is weaker compared with hybrid and ensemble prediction methods.
- Behravan and Razavi [10] used a hybrid regression model combining Particle Swarm Optimization (PSO) and Support Vector Regression (SVR) to predict player values, with PSO optimizing feature selection and parameter tuning for SVR. The results showed that the method achieved R^2 of 0.74 in value estimation which has considerable difference with high accuracy obtained by prediction models in this study.

Sensitivity Analysis

SHAP (Shapley Additive Explanations) [81] is a tool for locally analyzing predictions by breaking them down into individual feature contributions. This method is motivated by scenarios like linear regression with structured data and continuous response, where predictions are expressed as:

$$\widehat{y}_i = b_0 + b_1 x_{1i} + \cdots + b_d x_{di} \tag{20}$$

where $\widehat{y}_i$ denotes the $i-$th estimation result, $\{x_{1i}, \ldots, x_{di}\}$ are the predictors, and $\{b_0, \ldots, b_d\}$ are the predicted regression coefficients. If the predictors are independent, the contribution of the $k-$th predictor to the predicted response $\widehat{y}_i$ can be expressed as $b_k x_{ki}$ for$k = 1, \ldots d$. SHAP represents an extension of this principle to encompass more intricate models within the domain of supervised learning, where F is the entire set of features, and S denotes a subset. $S \cup i$ represents the union of the subset S and feature i. $E[f(X)|X_S = x_S]$ is the conditional expectation of $f(\cdot)$ when a subset S of features are fixed at x (local point).

SHAP value to measure the contribution of the $i-$th feature is defined as follows [82]:

$$\phi_i = \sum_{S \subseteq F/\{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \{E[f(X)|X_{SUi} = x_{SUi}] - E[f(x)|X_S = x_S]\} \tag{21}$$

SHAP has demonstrated its adherence to favorable properties, including fairness and consistency, in assigning importance scores to individual features.

In this study, AHAP is utilized to detect the sensitivity of market value predictions by XGB to selected features. In Fig. 13, the X-axis represents data points, and SHAP values (model's output) are presented on the Y-axis. The explanation for the 12th and 68th samples in the testing dataset is illustrated in Fig. 13. As it is evident, Reactions, Ball control, and Dribbling were the factors that affected predicted market values most, indicating that exact reports of such factors may lead to more accurate estimations of the market value of players.

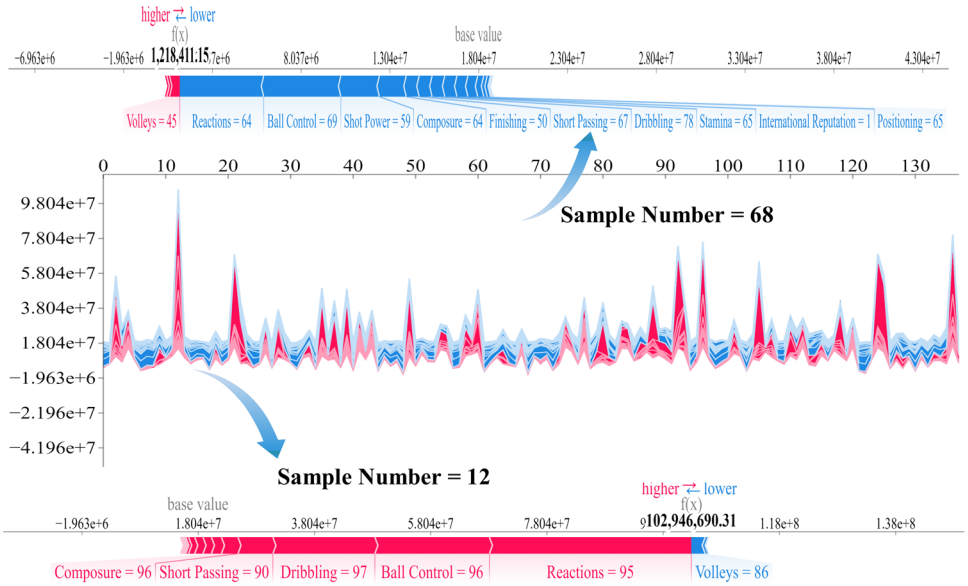

Fig. 13 Explanation generated by SHAP visualized with force plot

Conclusions and Real-World Applications

Various factors, including performance evaluators of football players such as reactions in critical situations of match and ability to control the ball, to the popularity level of players, especially in well-known leagues, influence a football player's value in the competitive transfer market. To overcome this, multidimensionality machine learning estimators are the best solutions for more accurate predictions of players' value. They assist in strategic decision-making in football clubs, allowing them to focus on players with optimal market value predictions for team performance improvement and ensuring future financial benefits. Sofifa.com is one of the most popular football-related data sources utilized for extracting a comprehensive dataset of FIFA19 and real-world statistical sources. This study pursues three main objectives: first, to identify the most relevant features extracted for players across various popularity levels from five top-tier European leagues that influence their market value; second, to evaluate the prediction performance of advanced machine learning methods in forecasting the market value of players with different popularity levels from these leagues; and third, to assess the effectiveness of various variables in predicting market value using SHAP-based explainable machine learning techniques.

The compiled dataset went through detailed preprocessing, feature selection, and engineering. Two filtering-based feature selection methods independently assigned scores to features, allowing the selection of a relevant subset (20 features). Prediction outcomes related to Adaptive Boosting (ADA) and Extreme Gradient Boosting (XGB) base models and their hybrid versions (optimized with Ali Baba and Forty Thieves (AFT), Crystal Structure Algorithm (CSA), Henry Gas Solubility Optimization (HGSO), and Mayfly Optimization Algorithm (MOA)) reported and ensemble outcomes obtained as most reliable predicted values. XGAF was the best predictor among developed models with an RMSE value of €1.9 million misestimation. This error was less than 10% of average market values obtained for players of five well-known European leagues with 1 (less popular) to 5 (superstars with the highest popularity) reputation indexes. Also, sensitivity analysis revealed that Reactions, Ball control, and Dribbling were the factors that affected predicted market values the most.

Football clubs can directly apply the predictions derived from the models to inform decision-making in various operational areas. For instance, the predicted market values of players can aid in transfer negotiations, allowing clubs to assess whether the asking price for a player is realistic based on their predicted value. Clubs could also utilize these predictions to determine whether investing in younger players, who may not have the same market value as superstars but possess significant growth potential, could yield long-term financial returns. Furthermore, knowing the predicted market values can help clubs in contract renewal or termination decisions, ensuring they align with financial expectations. Additionally, clubs can leverage these insights to plan strategic investments, such as focusing on acquiring players with specific skill sets that can generate increased ticket sales and enhance overall team performance.

While this study presents promising results, several limitations should be acknowledged. First, the results of this research primarily focused on five well-known European leagues, which may not necessarily be representative of other global leagues. The transfer dynamics, player evaluations, and economic factors in these leagues could differ

from those in smaller leagues or emerging football markets. As such, while the methods and findings could provide valuable insights for clubs within these top-tier leagues, further studies are needed to assess the applicability of the models in different league contexts, especially in terms of player valuation in markets with distinct economic conditions or player pool characteristics. Additionally, cross-validation methods and data balancing across reputation indices can be used in future studies to further validate the robustness of the models across a broader range of datasets and ensure the fairness of the predictions. Also, other global sensitivity methods and techniques are used to assess the sensitivity of models to input variations. These techniques include variance-based methods, such as Sobol's indices and Morris screening, as well as Fourier-based approaches, like Fourier Amplitude Sensitivity Testing (FAST) and its extended version, eFAST. These methods help identify the individual contributions of input factors and the interactions between them in influencing model outputs.

Acknowledgements N/A

Author Contribution Authors' contribution statement. Dan Cao: Writing-Original draft preparation, Conceptualization, Supervision, Project administration. Wenjing Zhang: Methodology, Software, Validation, Formal analysis.

Data Availability No datasets were generated or analysed during the current study.

Declarations

Human Participants and/or Animals Not applicable.

Ethical Approval Not applicable.

Conflict of Interest The authors declare no competing interests.

References

1. Tamir I, Hilik Limor Y, Galily Y. Sports: faster, higher, stronger, and public relations. Hum Aff. 2015;25(1):93–109. https://doi.org/10.1515/humaff-2015-0008.
2. Seippel Ø, Dalen HB, Sandvik MR, Solstad GM. From political sports to sports politics: on political mobilization of sports issues. Int J Sport policy Polit. 2018;10(4):669–86. https://doi.org/10.1080/19406940.2018.1501404.
3. UEFA. 2021b. [Online]. Available: https://www.uefa.com/uefachampionsleague/history/seasons/2021/. Accessed 20 Oct 2024
4. Bridge T, Hammond T, Tantam L. Deloitte Football Money League 2023: get up, stand up. Deloitte. 2023 Jun [cited 2023 June 21]." [Online]. Available: https://www2.deloitte.com/uk/en/pages/sportsbusiness-group/articles/deloitte-football-money-league.html. Accessed 20 Oct 2024
5. Malagón-Selma P, Debón A, Domenech J. Measuring the popularity of football players with Google Trends. PLoS ONE. 2023;18(8):e0289213. https://doi.org/10.1371/journal.pone.0289213.
6. BBC. 2017b. [Online]. Available: https://www.bbc.com/worklife/article/20170829-how-does-a-football-transfer-work. Accessed 20 Oct 2024
7. BBC. [Online]. Available: neymar: Paris St-Germain sign Barcelona forward for world record%0D222m euros. BBC. 2017a.
8. Walker C. [Online]. Available: Manchester United will recover the costs of the sensational%0Dsigning of Cristiano Ronaldo... with analysts expecting a windfall of £30 million within%0D12 months as sponsors line up to cash in on "a match made in heaven." Daily Mail. 2021.
9. Bifet A et al. Machine learning and knowledge discovery in databases: Machine learning and knowledge discovery in databases. Springer International Publishing AG, European Conference, ECML PKDD 2015, Porto, Portugal, 2015, Proceedings, Part III, vol. 9286. 2015. https://doi.org/10.1007/978-3-319-23461-8
10. Behravan I, Razavi SM. A novel machine learning method for estimating football players' value in the transfer market. Soft Comput. 2021;25(3):2499–511. https://doi.org/10.1007/s00500-020-05319-3.
11. Müller O, Simons A, Weinmann M. Beyond crowd judgments: data-driven estimation of market value in association football. Eur J Oper Res. 2017;263(2):611–24. https://doi.org/10.1016/j.ejor.2017.05.005.
12. Garcia-del-Barrio P, Pujol F. Hidden monopsony rents in winner-take-all markets — sport and economic contribution of Spanish soccer players. Manag Decis Econ. 2007;28:57–70. https://doi.org/10.1002/mde.1313.
13. Herm S, Callsen-Bracker H-M, Kreis H. When the crowd evaluates soccer players' market values: accuracy and evaluation attributes of an online community. Sport Manag Rev. 2014;17(4):484–92. https://doi.org/10.1016/j.smr.2013.12.006.
14. Barbuscak L. What makes a soccer player expensive? Analyzing the transfer activity of the richest soccer. Augsbg Honor Rev. 2018;11(1):5.
15. Carmichael F, Forrest D, Simmons R. The labour market in association football: who gets transferred and for how much? Bull Econ Res. 1999;51(2):125–50. https://doi.org/10.1111/1467-8586.00075.
16. Frick B. The football players' labor market: empirical evidence from the major European leagues. Scott J Polit Econ. 2007;54(3):422–46.
17. Rosen S. The economics of superstars. Am Econ Rev. 1981;71(5):845–58.
18. Adler M. Stardom and talent. Am Econ Rev. 1985;75:208–12.
19. Ferreira MSBP. The impact of performance measures in football players' Transfer market value, (Master's thesis, Universidade NOVA de Lisboa (Portugal)). 2022.
20. Garcia-del-Barrio P, Pujol F. Hidden monopsony rents in winner-take-all markets—sport and economic contribution of Spanish soccer players. Manag Decis Econ. 2007;28(1):57–70. https://doi.org/10.1002/mde.1313.
21. Kiefer S. The impact of the Euro 2012 on popularity and market value of football players. Diskussionspapier des Instituts für Organisationsökonomik (No. 11/2012), 2012, https://hdl.handle.net/10419/67719
22. Hofmann J, Schnittka O, Johnen M, Kottemann P. Talent or popularity: what drives market value and brand image for human brands? J Bus Res. 2021;124:748–58. https://doi.org/10.1016/j.jbusres.2019.03.045.
23. Singh P, Lamba PS. Influence of crowdsourcing, popularity and previous year statistics in market value estimation of football players. J Discret Math Sci Cryptogr. 2019;22(2):113–26. https://doi.org/10.1080/09720529.2019.1576333.
24. Goes FR, et al. Unlocking the potential of big data to support tactical performance analysis in professional soccer: a systematic

review. Eur J Sport Sci. 2021;21(4):481–96. https://doi.org/10.1080/17461391.2020.1747552.
25. Memmert D, Raabe D. Data analytics in football: positional data collection, modelling and analysis. Routledge. 2018. https://doi.org/10.4324/9781351210164.
26. Teixeira JE, et al. Data mining paths for standard weekly training load in sub-elite young football players: a machine learning approach. J Funct Morphol Kinesiol. 2024;9(3):114. https://doi.org/10.3390/jfmk9030114.
27. Mandadapu P. The evolution of football betting-a machine learning approach to match outcome forecasting and bookmaker odds estimation. 2021. Available from: https://arxiv.org/abs/2403.16282.
28. Yang Y, Koenigstorfer J, Pawlowski T. Predicting transfer fees in professional European football before and during COVID-19 using machine learning. Eur Sport Manag Q. 2024;24(3):603–23. https://doi.org/10.1080/16184742.2022.2153898.
29. El-Kenawy E-SM, Khodadadi N, Mirjalili S, Abdelhamid AA, Eid MM, Ibrahim A. Greylag goose optimization: nature-inspired optimization algorithm. Expert Syst Appl. 2024;238:122147. https://doi.org/10.1016/j.eswa.2023.122147.
30. Ibrahim A, Khodadadi E, Khodadadi E, Dutta PK, Bailek N, Abdelhamid AA. Apple perfection: assessing Apple quality with waterwheel plant algorithm for feature selection and logistic regression for classification. J Artif Intell Eng Pract. 2024;1(1):34–48. https://doi.org/10.21608/jaiep.2024.355003.
31. El-Kenawy ESM, et al. Football optimization algorithm (FbOA): a novel metaheuristic inspired by team strategy dynamics. J Artif Intell Metaheuristics. 2024;8:21–38. https://doi.org/10.54216/JAIM.080103.
32. Morciano G, Zingoni A, Calabrò G. Optimization and comparison of machine learning algorithms for the prediction of the performance of football players. Neural Comput Appl. 2024;36(31):19653–66. https://doi.org/10.1007/s00521-024-10260-9.
33. Samek W, Montavon G, Vedaldi A, Hansen LK, Müller KR, editors. Explainable AI: interpreting, explaining and visualizing deep learning. Springer Nature. 2019. https://doi.org/10.1007/978-3-030-28954-6.
34. Moustakidis S, Plakias S, Kokkotis C, Tsatalas T, Tsaopoulos D. Predicting football team performance with explainable ai: leveraging shap to identify key team-level performance metrics. Futur Internet. 2023;15(5):174. https://doi.org/10.3390/fi15050174.
35. Plakias S, Kokkotis C, Pantazis D, Tsatalas T. Comparative analysis of key performance indicators in Euroleague and national basketball leagues. J Phys Educ Sport. 2024;24(6):1360–72. https://doi.org/10.7752/jpes.2024.06154.
36. Ren Y, Susnjak T. Predicting football match outcomes with explainable machine learning and the Kelly index. 2022. Available from: https://arxiv.org/abs/2211.15734.
37. Geurkink Y, Boone J, Verstockt S, Bourgois JG. Machine learning-based identification of the strongest predictive variables of winning and losing in Belgian professional soccer. Appl Sci. 2021;11(5):2378. https://doi.org/10.3390/app11052378.
38. Plakias S, Kokkotis C, Mitrotasios M, Armatas V, Tsatalas T, Giakas G. Identifying key factors for securing a champions league position in French Ligue 1 using explainable machine learning techniques. Appl Sci. 2024;14(18):8375. https://doi.org/10.3390/app14188375.
39. FootballPlayersDataset. [Online]. Available: https://www.openml.org/search?type=data&status=active&id=43604. Accessed 20 Oct 2024
40. Kirschstein T, Liebscher S. Assessing the market values of soccer players–a robust analysis of data from German 1. and 2. Bundesliga. J Appl Stat. 2019;46(7):1336–49. https://doi.org/10.1080/02664763.2018.1540689.
41. Goswami S, Chakrabarti A. Feature selection: a practitioner view. Int J Inf Technol Comput Sci. 2014;6(11):66. https://doi.org/10.5815/ijitcs.2014.11.10.
42. Akinwande MO, Dikko HG, Samson A. Variance inflation factor: as a condition for the inclusion of suppressor variable (s) in regression analysis. Open J Stat. 2015;5(07):754. https://doi.org/10.4236/ojs.2015.57075.
43. Hall MA. Correlation-based feature selection for machine learning. PhD diss, The University of Waikato; 1999.
44. Swingle B. Renyi entropy, mutual information, and fluctuation properties of Fermi liquids. Phys Rev B. 2010;86. https://doi.org/10.1103/PhysRevB.86.045109.
45. Kraskov A, Stögbauer H, Grassberger P. Estimating mutual information. Phys Rev E Stat Nonlin Soft Matter Phys. 2004;69:66138. https://doi.org/10.1103/PhysRevE.69.066138.
46. Hwang C-L, Yoon K, Hwang C-L, Yoon K. Methods for multiple attribute decision making. Mult Attrib Decis Mak methods Appl A State-of-the-Art Surv. 1981;58–191. https://doi.org/10.1007/978-3-642-48318-9_3.
47. Franck E, Nüesch S. Talent and/or popularity: what does it take to be a superstar? Econ Inq. 2012;50(1):202–16. https://doi.org/10.1111/j.1465-7295.2010.00360.x.
48. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 2016, pp. 785–794. https://doi.org/10.1145/2939672.2939785.
49. Schapire RE. The strength of weak learnability. Mach Learn. 1990;5:197–227. https://doi.org/10.1007/BF00116037.
50. Freund Y, Schapire R, Abe N. A short introduction to boosting. J-Japanese Soc. 1999;14(771–780):1612.
51. Ying C, Qi-Guang M, Jia-Chen L, Lin G. Advance and prospects of AdaBoost algorithm. Acta Autom Sin. 2013;39(6):745–58. https://doi.org/10.1016/S1874-1029(13)60052-X.
52. Schapire RE. Explaining adaboost. Empir. Inference Festschrift Honor Vladimir N. Vapnik, 2013. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 37–52. https://doi.org/10.1007/978-3-642-41136-6_5.
53. Wang C, Xu S, Yang J. Adaboost algorithm in artificial intelligence for optimizing the IRI prediction accuracy of asphalt concrete pavement. Sensors. 2021;21(17):5682. https://doi.org/10.3390/s21175682.
54. Korada NK, Kuma N, Deekshitulu Y. Implementation of naïve Bayesian classifier and ada-boost algorithm using maize expert system. Int J Inf Sci Tech. 2012;2. https://doi.org/10.5121/ijist.2012.2305.
55. Braik M, Ryalat MH, Al-Zoubi H. A novel meta-heuristic algorithm for solving numerical optimization problems: Ali Baba and the Forty Thieves. Neural Comput Appl. 2022;34(1):409–55. https://doi.org/10.1007/s00521-021-06392-x.
56. Rehman KU, et al. Fast tracking of maximum power in a shaded photovoltaic system using Ali Baba and the Forty Thieves (AFT) algorithm. Processes. 2023;11(10):2946. https://doi.org/10.3390/pr11102946.
57. Averill BA, Eldredge P. Chemistry: principles, patterns, and applications. Pearson Benjamin Cummings International ed, 2007.
58. Sareh P. The least symmetric crystallographic derivative of the developable double corrugation surface: computational design using underlying conic and cubic curves. Mater Des. 2019;183:108128. https://doi.org/10.1016/j.matdes.2019.108128.
59. Talatahari S, Azizi M, Tolouei M, Talatahari B, Sareh P. Crystal structure algorithm (CryStAl): a metaheuristic optimization method. IEEE Access. 2021;9:71244–61. https://doi.org/10.1109/ACCESS.2021.3079161.
60. Hashim FA, Houssein EH, Mabrouk MS, Al-Atabany W, Mirjalili S. Henry gas solubility optimization: a novel physics-based

algorithm. Futur Gener Comput Syst. 2019;101:646–67. https://doi.org/10.1016/j.future.2019.07.015.
61. Brown TL, LeMay HE, Bursten BE, Murphy C, Woodward P, Langford S, Sagatys D, George A. Chemistry: The central science. Pearson Higher Education AU, 2013. https://books.google.com/books?id=zSziBAAAQBAJ
62. Allan JD, Flecker AS. The mating biology of a mass-swarming mayfly. Anim Behav. 1989;37:361–71. https://doi.org/10.1016/0003-3472(89)90084-5.
63. Bersier L-F, Banašek-Richter C, Cattin M-F. Quantitative descriptors of food-web matrices. Ecology. 2022;83(9):2394–407. https://doi.org/10.1890/0012-9658(2002)083[2394:QDOFWM]2.0.CO;2.
64. Zervoudakis K, Tsafarakis S. A mayfly optimization algorithm. Comput Ind Eng. 2020;145:106559. https://doi.org/10.1016/j.cie.2020.106559.
65. Eberhart R, Kennedy J. A new optimizer using particle swarm theory. In MHS'95. Proceedings of the sixth international symposium on micro machine and human science. Nagoya, IEEE; 1995. pp. 39–43. https://doi.org/10.1109/MHS.1995.494215.
66. D. E. Goldberg, *Genetic algorithms*. pearson education India, 2013.
67. Mo S, Ye Q, Jiang K, Mo X, Shen G. An improved MPPT method for photovoltaic systems based on mayfly optimization algorithm. Energy Rep. 2022;8:141–50. https://doi.org/10.1016/j.egyr.2022.02.160.
68. Sweet L, Müller C, Anand M, Zscheischler J. Cross-validation strategy impacts the performance and interpretation of machine learning models. Artif Intell Earth Syst. 2023;2(4):e230026. https://doi.org/10.1175/AIES-D-23-0026.1.
69. Putatunda S, Rama K. A comparative analysis of hyperopt as against other approaches for hyper-parameter optimization of XGBoost. In Proceedings of the 2018 international conference on signal processing and machine learning, New York, 2018. pp. 6–10. https://doi.org/10.1145/3297067.3297080.
70. Yang L, Shami A. On hyperparameter optimization of machine learning algorithms: theory and practice. Neurocomputing. 2020;415:295–316. https://doi.org/10.1016/j.neucom.2020.07.061.
71. Schwer LE. Verification and validation in computational solid mechanics and the ASME Standards Committee. WIT Trans Built Environ. 2005;84. https://doi.org/10.2495/FSI050111.
72. Saad Y. Iterative methods for linear systems of equations: a brief historical journey. 75 Years of Mathematics of Computation, arXiv:1908.01083, American Mathematical Society, vol. 754, pp. 197–215, 2020. https://doi.org/10.1090/conm/754/15141.
73. Cai C, Wang Y. Convergence of invariant graph networks. In Proceedings of the 39th international conference on machine learning. PMLR 2022;162:2457–2484. https://doi.org/10.48550/arXiv.2201.10129.
74. Awasthi P, Das A, Gollapudi S. A convergence analysis of gradient descent on graph neural networks. Adv Neural Inf Process Syst. 2021;34:20385–97.
75. Bergstrom CT,West JD. Why scatter plots suggest causality, and what we can do about it. 2018. Available from: https://arxiv.org/abs/1809.09328.
76. Sáenz J, Carreno-Madinabeitia S, Esnaola G, González-Rojí S, Ibarra-Berastegi G, Ulazia A. The sailor diagram. An extension of taylor's diagram to two-dimensional vector data. Geosci Model Dev Discuss. 2019;13:1–24. https://doi.org/10.5194/gmd-2019-289.
77. Demšar J. Statistical comparisons of classifiers over multiple data sets. J Mach Learn Res. 2006;7:1–30.
78. Golafshani EM, Kashani A, Kim T, Arashpour M. Concrete chloride diffusion modelling using marine creatures-based metaheuristic artificial intelligence. J Clean Prod. 2022;374:134021. https://doi.org/10.1016/j.jclepro.2022.134021.
79. Moghaddas SA, Nekoei M, Golafshani EM, Behnood A, Arashpour M. Application of artificial bee colony programming techniques for predicting the compressive strength of recycled aggregate concrete. Appl Soft Comput. 2022;130:109641. https://doi.org/10.1016/j.asoc.2022.109641.
80. Al-Asadi MA, Tasdemır S. Predict the value of football players using FIFA video game data and machine learning techniques. IEEE Access. 2022;10:22631–45. https://doi.org/10.1109/ACCESS.2022.3154767.
81. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst. 2017;30. https://doi.org/10.48550/arXiv.1705.07874. (**Focus to learn more**).
82. Lundberg SM, Erion GG, Lee S-I. Consistent individualized feature attribution for tree ensembles. 2018. Available from: https://arxiv.org/abs/1802.03888.