Predicting Football Player Transfer Values Using Bagging and Hybrid Machine Learning Approaches

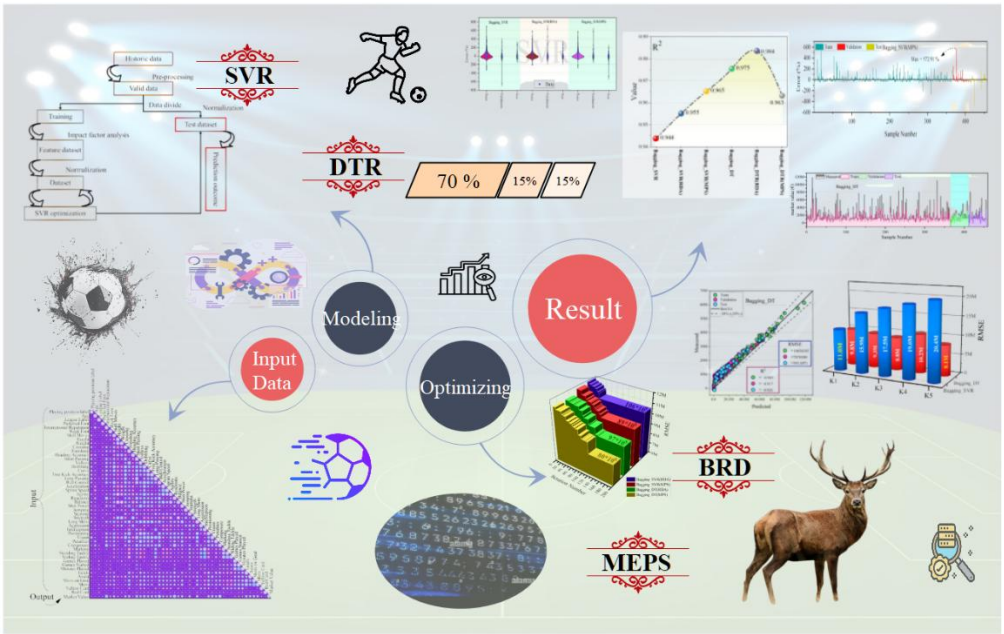
Biao Geng
Department of Physical Education and Military Training, Jiaxing Nanhu University, Jiaxing 314000, Zhejiang, China
E-mail: gbiao6688@126.com

Keywords: machine learning, player value prediction, decision tree regression, bagging regression, fourier amplitude sensitivity test, motion-encoded particle swarm optimization

Received: December 1, 2024

Accurately assessing a football player's market value is essential for enabling informed decision-making by clubs, agents, and investors during player transfers, contract negotiations, and strategic investment planning. In this context, machine learning (ML) algorithms offer a robust framework for analyzing historical data, performance indicators, and market dynamics to produce realistic valuations. These data-driven methods assist in identifying undervalued opportunities and flagging overpriced players, thereby enhancing the overall efficiency of transfer market operations. The dataset employed in this research includes a comprehensive set of player-related features such as age, weight, weak foot rating, preferred foot, and international reputation, among others. These attributes collectively contribute to a detailed profile of each player's capabilities and market relevance. The objective of this study is to develop reliable and accurate predictive models that estimate player market values by leveraging advanced machine learning techniques, thereby improving upon traditional, subjective valuation approaches. Several regression-based models were explored, including Bagging Decision Tree Regression (Bg_DT), and Bagging Support Vector Regression (Bg_SVR). To further enhance model performance, optimization algorithms such as Motion-encoded Particle Swarm Optimization (Motion-encoded PSO) and the Red Deer Algorithm (RDA) were applied for hyperparameter tuning. Among the evaluated models, the Bagging Decision Tree optimized with Motion-encoded PSO (Bg_DT- Motion-encoded PSO) demonstrated superior performance. It achieved the lowest Root Mean Squared Error (RMSE) and the highest coefficient of determination (R^2) across both validation and testing phases. Specifically, the Bg_DT- Motion-encoded PSO model yielded an RMSE of 533×10^5 and an R^2 of 0.962 during validation, indicating strong predictive accuracy and generalization capability. These findings underscore the effectiveness of ensemble learning techniques—particularly Bagging Decision Trees—in conjunction with advanced metaheuristic optimizers like Motion-encoded PSO, for accurately estimating football player market values.

Povzetek: Prispevek predstavlja uporabo strojnega učenja za napovedovanje tržnih vrednosti nogometnih igralcev z uporabo metod, kot sta vrečenje (bagging) in hibridne tehnike.



Graphical Abstract

1 Introduction

The modern football first emerged in Britain in the nineteenth century. Before the medieval period, the processes of industrialization and urbanization had a significant influence on the creation of modern football in Victorian Britain. Association football developed in its early years, between 1863 and 1880, as a result of both rule and play modifications [1]. Midway through the 1900s, the betting landscape was completely changed by the availability of match odds and fan hubs. Internet usage and the late 20th century accelerated betting into a new era.

The transfer of players is one of the most significant arrangements made by managers of a team from a managerial standpoint, so player valuation-related issues, particularly the calculation of Market valuations and transfer fees, are very important. Market valuations and transfer fees are key elements in the financial strategy of football clubs, directly influencing their competitive edge[2]. Football clubs rely heavily on team managers to make strategic decisions, especially when it comes to transfers of players, which significantly affect the team's performance and financial situation. [3]. From a business and athletic standpoint, in professional football, players are the most important investments. [4]. Assessing a player's worth is crucial as it reveals their overall skill set and market value in football. The transfer fee is the amount a player is actually paid by clubs, and it is related to his market value. Thus, evaluation of the market value of a player is an important tool for clubs when estimating their transfer fee. Over the years, the valuation of football players and the determination of a decision to transfer players from one team to another has become a key role of club management [5]. Several researchers [6–8] have tried to find the characteristics that best determine a player's value [9].

Transfermarkt.com represents a website that uses the crowd estimate approach to ascertain the players' market values. With this approach, the website's members assess the values of the players and then select members—referred to as mentors—who calculate the values using the estimates of the other members [10]. Football players were traditionally valued mostly based on the subjective evaluations made by scouts, managers, and agents. These assessments mostly depend on the author's intuition, knowledge, and experience. While this approach yielded insightful results, it was frequently hampered by personal prejudices and irregularities. The sport's quick development and rising financial stakes have made more objective and data-driven methods necessary. Even while they are helpful, traditional analytical techniques are not always able to handle the volume and complexity of current football data [11].

Through training and experience, ML is an artificial intelligence (AI) approach that enhances computer systems' performance on a given job. Rather than being exactly told what to do, ML algorithms [12–14] are trained to generate predictions based on observations and data. AI and ML [15–18] have become increasingly important in various aspects of daily life, including sports. AI mimics human thought processes, enabling big data analysis in sports. ML has transformed football data into actionable insights for clubs and coaches over the past 20 years, particularly in fields like sports [19] [20].

The following papers share related concepts with this research and therefore, can shed more light on this research process. For instance, Majewski (2016) [21] looked at the impact of several aspects on forward players' value to identify the most important factors. Another study by Müller et al., 2017 [3] employed a multilevel regression approach in assessing data analytics' suitability for the calculation of the market values of professional football players. Lamba, in 2019 [22] estimated the factors that determine every player's market value and used them in predicting every player's worth. Apart from the requirements, the work also utilized measures of crowdsourcing, popularity, and statistics of the previous years to predict that the declared goal of automatically detecting the relevant attributes for different player groups, depending on their positions, raised the accuracy and reliability of the market value estimates. Their approach consists of adding position-specific changes and performance data to improve prediction models.

Li et al. (2022) [23] evaluated football players using ML models based on on-field performance metrics analysis. The present study, therefore adopted an improved modeling approach with an ensemble technique such as Random Forest to better the accuracy of the prediction. Behravan and Razavi 2020 [10] Proposed a new Hybrid ML approach in order to estimate the market values of football players. This method used the optimized hybrid of PSO and SVR. The goal of the model is to extract automatically in a relevant way the attributes for different player groups depending on their positions with an aim to further improve the accuracy and reliability of the market value estimates.

Table 1 reports a summary of the existing articles in the study field.

Table 1: Summary of the previous studies.

Authors	Ref	Techniques/Models Used	Dataset Used	Limitations / SOTA Shortcomings
Majewski	[21]	Statistical analysis of influencing factors	Forward players' market data	Did not employ ML or optimization; focused only on forward players, lacks generalizability
Müller et al.	[3]	Multilevel regression	Professional football players	Limited to regression models; prone to overfitting; lacks adaptive optimization
Lamba	[22]	Regression models + crowdsourcing & performance stats	Historical market value data	No metaheuristic optimization; position-based tailoring manually defined
Gadekallu et al.	[24]	Metaheuristic optimization algorithms (general)	Not football-specific	Conceptual study; lacks applied validation in sports/football datasets
Li et al.	[23]	ML (e.g., RF, ensemble methods) based on performance	On-field football performance data	Did not integrate position-specific attributes; optimization technique not elaborated
Behravan & Razavi	[10]	Hybrid PSO + SVR	Market value dataset (unspecified)	Dataset details limited; overfitting risk due to SVR; no ensemble techniques like bagging explored

As highlighted earlier, Müller et al. (2017) [3] employed a method of multilevel regression that may easily face overfitting problems since the predictions of single models usually have a high variance without the application of bagging-like in the paper at hand-models such as SVR or DTR depend on a single dataset and a single model prediction. Due to this dependence, one of the consequences may be overfitting when the model performs well on the training data but not so well on the unknown data.

The primary objective of this study is to develop a robust, accurate, and interpretable machine learning framework for predicting the market values of professional football players. To achieve this, three core models—Bagging Decision Tree Regression (Bg_DT), and Bagging Support Vector Regression (Bg_SVR)—are employed due to their complementary strengths. Bagging Regression is chosen for its capability to reduce model variance and combat overfitting by aggregating multiple base models, thus increasing overall stability and accuracy. Bg_DT is selected for its interpretability and ability to capture non-linear relationships, while Bg_SVR is applied for its effectiveness in handling high-dimensional data and robustness against outliers. To further improve prediction performance and convergence reliability, the models are hybridized with two nature-inspired metaheuristic algorithms: Motion-encoded Particle Swarm Optimization (MPS) and the Red Deer Algorithm (RDA). MPS is incorporated due to its efficient exploration–exploitation balance and ability to dynamically encode complex motion patterns, making it suitable for fine-tuning model hyperparameters. RDA is adopted for its adaptive behavior inspired by the social dynamics of red deer, which helps avoid local minima and enhances global optimization in nonlinear regression settings. Additionally, the Fourier Amplitude Sensitivity Test (FAST) is implemented as a global sensitivity analysis tool to identify and rank the influence of input features on the predicted market values. FAST is selected for its computational efficiency and ability to detect both linear and non-linear interactions among features, which is crucial in a domain as complex and multifactorial as sports analytics.

2 Methodology

2.1 Support vector regression (SVR) based prediction approach

SVR is an ML that estimates functions based on a given data set [25]. $G = \{(x_i, y_i)\}^n$, where n is the ultimate number of data point, y_i is the value of the output, and x_i is the input vector. An SVR model performs a regression first by an ε-sensitive loss. Schoellkopf created the $\varepsilon -$ SVR model and suggested the v − SVR model, which is an adaptation of the $\varepsilon -$ SVR model. It automatically reduces ε and modifies the level of accuracy based on the available data. The expression for the v-SVR model is as follows:

$$min\ R_{SVR}(C) = \left\| = C(v\varepsilon + \frac{1}{n}\sum_{i=l}^{n}(\xi_i + \xi_i^*)) + \frac{1}{2}\|\omega\|^2 \right. \quad (1)$$

Subject to:

$$\begin{aligned} &((\omega \cdot x_i) + b)\_y_i \le \mathcal{E} + \xi_i \\ &y_i - ((\omega \cdot x_i) + b) \le \mathcal{E} + \xi_i \\ &\xi^{(*)} \ge 0, \mathcal{E} \ge 0, v \ge 0 \\ &i = 1, \ldots, n \end{aligned} \quad (2)$$

$\|\omega\|^2/_2$ indicates the Euclidean norm

C : a cost function measuring the empirical risk

R_{SVR} and R_{emp} : the regression and empirical risks,

ω: Weight vector.

b: Bias term.

$C > 0$: Regularization parameter controlling the trade-off between model complexity and training error.

v: Parameter that determines the fraction of support vectors and margin errors.

ε: Insensitivity zone (learned from data).

$\xi^{(*)}$: Slack variables representing the deviation from the ε -tube.

x_i, y_i: Input vectors and corresponding target values.

n: Number of training samples.

The SVR-based prediction method, which is based on the v − SVR model, comprises the following five steps:

Step One: Data sampling. Data can be gathered from various sources and in various formats. Additionally, there are several gaps and inconsistencies in the market. Thus, The most reliable data should be selected.

Step Two: Preparing the data. It might be a logarithmic transformation that must be applied, difference, or other techniques to the chosen data in order to place it within a specific acceptable range for network learning. The training and testing sets come next and should be separated from the data set.

Step Three: Education. The training set is used to learn the SVM's parameters.

Step Four: Evaluation. The testing set is used to validate the SVM, and then a final network design for the SVM is determined. s

Step Five: Projecting. Using the scenarios, the SVR-based predicting technique can be used to predict the time series' future values.

Fig. 1 represents the SVR's flowchart.

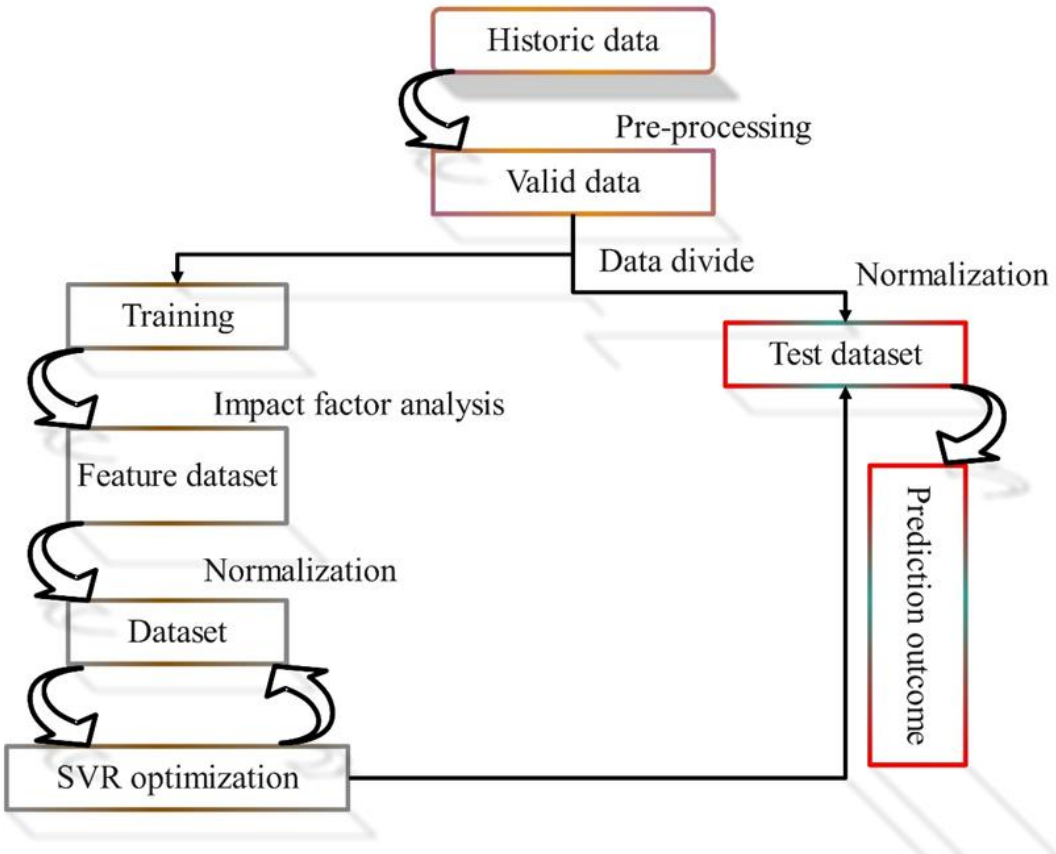


Figure 1: The flowchart of the SVR

2.2 Decision tree regression (DT)

DT, rooted in ML theory, is an effective instrument for addressing both classification and regression challenges. In contrast to other classification approaches that rely on a combination of features for immediate categorization, the DT employs a multi-tiered, hierarchical decision-making process with a structure akin to a tree. Unlike other classification techniques that utilize a single feature set for rapid data categorization, the DT

adopts a hierarchical or multi-level evaluation process to create a structure that resembles a tree.

To enable soft classification, a regression tree is assigned to every class. In regression trees, the known class proportions of a pixel, referred to as soft reference data, act as the target variable or vector, while pixel intensity values from various bands are used as predictor variables or feature vectors. After processing the intensity values for each regression tree, the script outputs the estimated class proportions. The algorithm for building regression trees using the training dataset is also discussed.

1. As a predication, use pixel intensity data from various bands.
2. Utilize class o's known percentage within a pixel as the target variable.
3. Create a regression tree for class o.
4. Repeat steps 1-4 for class o, with values ranging from 1 to n.

Rescaling the outcomes of soft classification to a pixel-by-pixel limit between 0 and 1 typically uncovers the class proportions within the ground pixel area. Therefore, the following process is used to normalize the predicted class proportions from each tree, which are represented as $DT(o)$ for $i = 1,2,3\ldots,n$ [26].

$$M(o) = \frac{DT(o)}{\sum DT(o)}, o = 1,2,3 \ldots, n \quad (3)$$

Using $DT(o)$ which is a function of o, where $M(o)$ is again a function of the natural numbers o.

2.3 Bagging approach

A technique called bagging was put forth by Breiman. It can be applied to a variety of regression and classification techniques to lower prediction variance and enhance the prediction process. It is a straightforward concept from the provided data, several bootstrap samples are chosen, each of which is subjected to a prediction method. The bootstrap sample results are then combined to create an overall prediction that lowers variance, using simple voting for regression and classification [27] [28].

- **Motivation for the method**

To comprehend the logic behind bagging's effectiveness and ascertain the scenarios where significant enhancements can be anticipated through bagging, it could be beneficial to examine the issue of predicting the response variable's numerical value, Yx, that arises from or is associated using a group of inputs, x. Assume that $\phi(x)$ represents the prediction obtained by applying a specific technique, like OLS or CART regression, along with a recommended approach for selecting a model (e.g., choosing a linear model from the set of all models that can be built using just terms of the first and second order created from the input variables) using Mallows' C_p. Using $\mu\phi$ to represent $E(\phi(\mathrm{x}))$, it can be seen that the prediction is related to the distribution that underlies the sample of learning. It can be observed that $\phi(\mathrm{x})$ is a learning sample's function, which is a high-dimensional random variable when considered as a random variable, rather than x (This is assumed to be] fixed:

$$\mathrm{E}\,([Y_x - \phi(x)]^2) = \mathrm{E}\,([Y_x - \mu\phi]^2) + \mathrm{Var}(\phi(\mathrm{x})) \quad (4)$$

In the example above, the learning sample-based predictor, $\phi(x)$, and the future response, Y_x, are employed independently. Since not every random sample that may be used as a learning sample provides the sample value needed to make a prediction, the variance of the predictor $\phi(x)$ is positive in nontrivial scenarios, which means that the prior inequality is stringent. This conclusion indicates that if $\mu\phi = E(\phi(\mathrm{x}))$ could be employed, it would have a lower mean squared prediction error as a predictor than does $\phi(x)$.

Naturally, in most cases, $\mu\phi$ cannot act as a predictor because it is unknown what data is required to determine the value of $E(\phi(\mathbf{x}))$. What is sometimes called the real bagging estimate of $E(\phi(\mathbf{x}))$ is derived from the prediction based on the empirical distribution corresponding to the learning sample.

Although this value is theoretically achievable, in practice, it is usually too challenging to attain reasonably; therefore, the bagged forecast of Y_x, is considered to be:

$$\frac{1}{\mathrm{B}}\sum_{\mathrm{b=1}}^{\mathrm{B}} \phi^*_{\mathrm{b}\,(\mathrm{X}),} \quad (5)$$

where the prediction is made by applying the base regression method (e.g., CART) to the bth bootstrap sample that was taken (with replacement) from the original learning sample is represented by the symbol $\phi^*_{b\,(X)}$. That is, one selects a regression method (also referred to as the base technique) that uses bagging to predict Y_x in a regression scenario by applying the approach to B bootstrap samples extracted from the learning sample. To get the final prediction, the B projected values are then averaged.

2.4 (RDA) based prediction approach

A subspecies of red deer found in the British Isles is the Scottish red deer, primarily in Scotland. They are divided into hinds and stags, with males roaring during breeding. Females prefer males with high roaring rates, possibly due to selective pressure or availability. The strongest males form a harem, with the hinds protected by a commander. The harem engages in predictable conflict, with mature stags becoming enraged in mid-September.

The RDA [6] is a meta-heuristic that assigns a harem to a select group of male RDs who roar first. These RDs are divided into commanders and stags, who engage in combat to control their harems. The number of hinds in a harem is directly proportional to their roaring and combat abilities. The RDA process considers the exploration and exploitation phase with user-adjustable parameters. Male stags' roaring serves as a local search in solution space, while battles between stags and commanders are considered local searches. Harems are established and distributed among commanders based on their power, enhancing exploitation features. The algorithm's exploration phase involves a harem's commander mate with hinds from both harems, enhancing exploration

qualities. Stags should mate with the nearest hind during the breeding season, considering the harem's limitations. This stage also addresses exploration and exploitation, producing RD offspring. The algorithm's next generation offers mediocre solutions, falling under evolutionary algorithms.

Finding a solution that is almost optimum or global in relation to the problem's variables is the aim of optimization. To optimize, a range of values for the variables are formed. For instance, in Georgia, this array is referred to as "chromosome," whereas in the RDA, it is named "Red Deer." Keep in mind that a "Red Deer" in the solution space refers to a workable solution X. Red deer is, therefore, the opposite of a solution. This solution X has Nvar dimensionality. One of the red deer is, hence, a 1× Nvar array in an Nvar ~ dimensional optimization problem. This array's definition is given by:

The process begins by creating the starting population of size Npop. The remaining RDs are then assigned to Nhind (Nhind = Npop - Nmale), while a subset of the best RDs is allocated to Nmale. It is important to note that the number of males reflects the elitist criteria of the algorithm. From another perspective, Nmale preserves the intensification features of the algorithm, while Nhind contributes to its diversification stage.

Two distinct approaches have been used to choose the following generation. All the male RDs, the commander, and all the stags are retained in the first one, or some of the best overall solutions are. The remaining members of the following generation are the subject of the second strategy. Using a fitness tournament or roulette wheel mechanism, hinds are selected from among all hinds and progeny generated during the mating process based on their fitness value.

2.5 Motion-encoded particle swarm optimization

- **Particle Swarm Optimization**

PSO is a population-based stochastic method for addressing optimization problems that were inspired by the social behavior of flocking birds. A swarm of randomly positioned and accelerated particles is first created in PSO [29]. Then, to find the global optimum, each particle travels and evolves with other particles cognitively. Its best position, $L_k$and the swarm's optimal position, G_k, determine those motions. Let x_k and v_k represent a particle's location and speed at generation k, respectively. The following generation's movement of that particle is determined by:

$$v_{k_{+1}} \leftarrow \omega v_k + \varphi_1 r_1 (L_k - \mathcal{X}_k) + \varphi_2 r_2 (G_k - \mathcal{X}_k) \quad (6)$$

$$\mathcal{X}_{k_{+1}} \leftarrow \mathcal{X}_k \quad (7)$$

where ω is the inertial weight, r_1, r_2 are random sequences generated from a uniform probability distribution in the interval [0,1], φ_1 is the cognitive coefficient, φ_2 is the social coefficient, and so on. A particle can move in one of three directions: it can follow its path, travel in the direction of its ideal position, or travel in the direction of the swarm's ideal position. The values of ω, φ_1, and φ_2 define the ratio between those components.

Various improvements and alterations have been made to the PSO algorithm, contingent on its intended use. Still, it is a difficult challenge to apply PSO for online dynamic target searching in a complicated environment, especially within a short time frame. The goal of the search problem is to encode the particle positions so that the particles can progressively approach the global optimum. Defining a position as a multi-dimensional vector that represents a potential search path is a frequent technique:

$$x_k \sim O_k = (o_{k,1,\ldots,}o_{k,N}), \quad (8)$$

The search map node is associated with a search map node, but this technique has limitations, such as not accounting for neighboring dynamic behavior in path nodes. To address this, discrete PSO can be used, but local maxima can occur due to the lack of particle momentum preservation. Indirect methods like priority-based encoding PSO and angle-encoded PSO may be viable, but they require phase angles to fall within [-pi/2, +pi/2] for their mapping functions to operate, reducing search capacity in large dimensions.

The Motion-encoded PSO equations can be expressed as follows, where U_k represents the location of each particle.

$$\Delta U_{k+1} \leftarrow \omega U_k + \varphi_1 r_1 (L_k - U_k) + \varphi_1 r_1 (G_k - U_k) \quad (9)$$

$$U_{k+1} \leftarrow U_k + \Delta U_{k+1} \quad (10)$$

Additionally, mapping U_k to a direct path O_k throughout the search is necessary to enable the evaluation of the costs related to U_k. One way to start the mapping process is to limit the UAV's movements to one of its eight neighbors for each time step. After that, it is possible to normalize the motion magnitude p_t and quantize the motion angle a_t as follows:

$$p_t{}^* = 1 \quad (11)$$

$$a_t^* = 45° \lfloor a_t / 45° \rceil, \quad (12)$$

where the operator to round to the closest integer is represented by $\lfloor a_t / 45° \rceil$. Next, the position of the UAV in Cartesian space, denoted as node ok,t+1, is obtained as follows:

$$o_{k,t+1} = o_{k,t+} u_{k,t}^* \quad (13)$$

where:

$$u_{k,t}^* = (\lfloor cos a_t^* \rceil, \lfloor sin a_t^* \rceil \quad (14)$$

The objective function may evaluate the cost value from the decoded path O_k, and the local and global best can then be calculated as follows:

$$L_k = \begin{cases} U_k & if\ J(O_k) > J(L^*_{k-1} \\ L_{k-1} & otherwise \end{cases} \quad (15)$$

$$G_k = argmaxJ(O_k), L_k \quad (16)$$

where Lk is the route that has been deciphered, from the mapping process discretizes the motion to one of eight potential directions.

3 Description of data

The dataset which is derived from public source [30] is designed to enhance the prediction of football players' market values by incorporating a comprehensive set of features that capture various aspects of a player's profile and performance. The information utilized in this study includes a number of characteristics and parameters such as age, weight, weak foot, preferred foot, international reputation, etc., pertaining to market valuations, player demographics, and football performance.

The correlation matrix below helps in identifying the strength and direction of relationships between variables. For instance, the relationship between the goals of a player (The number of goals scored by the player) with market value is determined by observing the blue dots (and other dots in different colors showing the strength between variables) and the correlation that connects these variables. The relationship between the age of a player (which affects both the experience and future potential of players) with their international reputation is shown with a small blue top in the figure, which determines the perfect positive correlation between these two variables. On the contrary, the relationship between age and sprint speed (Maximum velocity a player can achieve during a full-out sprint) is a perfect negative correlation and is determined with a small pink dot (-0.4). The correlation matrix represented in Fig. 2 is essential for preliminary data analysis. This step is particularly important for building robust and accurate predictive models. Table 2, which outlines the input parameters and factors affecting the value of football players, provides a detailed overview of the variables considered in the analysis of player valuation.

While the dataset offers valuable insights for predicting football players' market values, it is important to acknowledge the potential ethical implications of automated player valuations. One key concern is the bias introduced by certain features, particularly subjective or culturally influenced ones like International Reputation.

- **International reputation:** This feature, which reflects a player's global recognition, can lead to biases in valuation, as players from well-known leagues or countries might receive inflated market values, regardless of their actual performance or potential. This introduces an implicit preference for players with higher visibility or from certain countries, perpetuating inequalities and under-valuing players from less recognized leagues or nations.
- **Age:** The correlation between age and market value may also create biases, as older players could be undervalued due to assumptions about their future performance potential, even if they possess considerable experience and skill.
- **Performance metrics:** While metrics such as goals scored or assists are often reliable, these can also be influenced by the quality of a player's teammates or the team's overall performance, which could lead to unintentional favoritism toward players in high-performing teams.

Table 2: Input parameters and factors affecting the value of football players.

Parameter	Description
Age	The age of the player affects both the experience and future potential of players.
Preferred Foot	Dominant or more proficient foot that a soccer player uses for shooting, passing, and dribbling.
International Reputation	Globally perceived measure of trust, esteem, and recognition of a player shaped by achievements.
Weak Foot	The weakness of the player in using both legs in football reflects the level of inflexibility in the players.
Skill Moves	Techniques performed by players to outmaneuver opponents involve intricate ball control, dribbling, and feints.
Height	Height of the player affects the likelihood of scoring or preventing a goal.
Weight	The weight of the player affects the movement skills of the players.
Crossing	Technique where a player delivers the ball into the penalty area from the flanks
Finishing	Player's ability to successfully score goals
Heading Accuracy	Player's proficiency in directing the ball with their head.
Short Passing	The number of passes to other players and the accuracy of passing
Volleys	The technique is where a player strikes the ball while it is in the air without allowing it to touch the ground.
Dribbling	Skill that involves a player using controlled touches to maneuver the ball while on the move
Curve	Bending or swerving trajectory applied to the ball by the player during a shot or a pass.
FK Accuracy	Accuracy in taking free kicks.
Long Passing	The number of passes delivering the ball over a significant distance to a teammate

Ball Control	Player's skill in skillfully receiving, trapping, and manipulating the ball using various body parts.
Acceleration	How quickly a player can reach their top speed
Sprint Speed	Maximum velocity a player can achieve during a full-out sprint.
Agility	The player's ability to change direction rapidly.
Reactions	Player's quick responses to the movement of the ball, changes in the game situation, or the actions of opponents or teammates.
Balance	Player's ability to maintain stability and control their body position during various movements.
Shot Power	Strength with which a player strikes the ball during a shot on goal.
Jumping	Jumping ability of the player
Stamina	The player's overall ability to sustain physical effort and performance over an extended period of time.
Strength	The player's physical power and ability to exert force against resistance.
Long Shots	The number of successful shots from a considerable distance away from the goal, often outside the penalty area.
Aggression	Player's assertiveness and determination in challenging for the ball.
Interception	Successfully blocks a pass or a ball played by the opposing team.
Positioning	Playing the position of the player.
Vision	Player's ability to perceive and understand the unfolding dynamics of the game.
Penalties	The number and accuracy of penalty kicks by a player
Composure	Player's ability to maintain calmness, control, and mental focus in high-pressure situations during a match
Marking	The tactic of closely tracking and guarding an opponent to prevent them from receiving or playing the ball effectively.
Standing Tackle	Performance of player in standing tackles.
Games Played	The number of games played by the player
Games Started	The number of games started by the player
Minutes played	Playing time (minutes) for players.
Goals	The number of goals scored by the player
Assist	Helping other players score a goal.
Shots on Goal	The number of shots of a player toward the goal.
Shots	The total number of shots by a player
Yellow Cards	The number of yellow cards received by a player.
Red Card	The number of red cards received by a player.

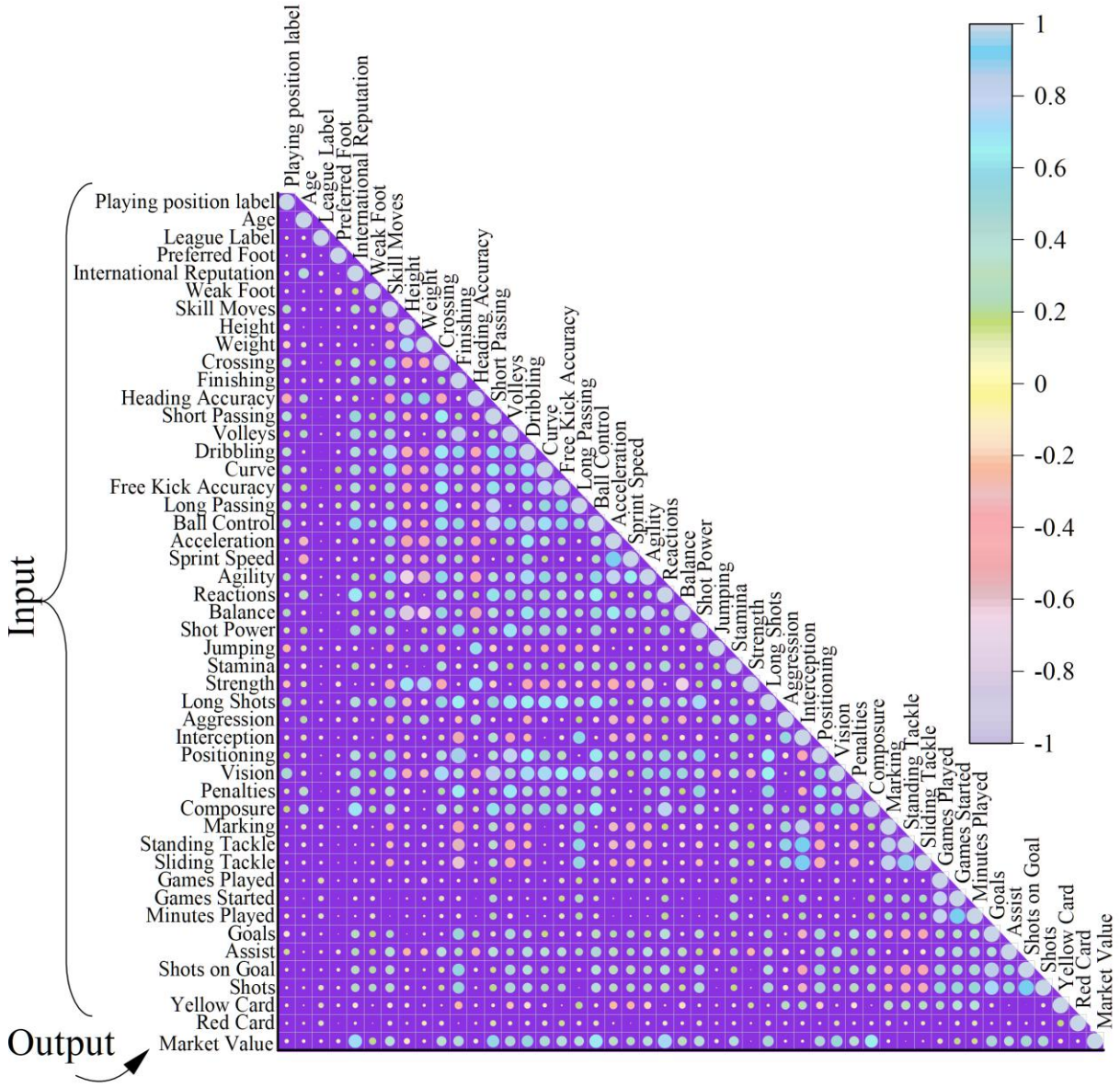


Figure 2: The relationships between input and output variables.

System configuration

The experiments were conducted on a system powered by an Intel® Core™ i7-3770K CPU running at 3.50 GHz, supported by 16 GB of RAM to ensure smooth multitasking and computational efficiency. The machine operates on a 64-bit Windows 11 Pro platform with an x64-based architecture. For handling graphics-related tasks, an NVIDIA GeForce GT 640 GPU is utilized, providing stable and responsive visual performance. Data storage is managed by a 1 TB hard drive, offering sufficient capacity for storing datasets and project files.

Software environment

The implementation was carried out using Python as the primary programming language. Machine learning models were developed and evaluated using the scikit-learn library. For data manipulation and analysis, Pandas and NumPy were employed, while Matplotlib was used for visualizing results and presenting analytical insights effectively.

4 Results

4.1 Evaluation metrics

Several ML models and optimizers were employed in this research. To improve the accuracy and reliability of football player market value prediction, hybrid models that combine the Bagging Regression, DTR, and SVR models with Motion-encoded PSO and the Red Deer algorithm (RDA) were utilized.

The assessment uses RMSE, R-squared (R^2), U95 uncertainty, SI, and a bespoke N10_ index. Root Mean Square Error (RMSE) is a widely used metric in ML and statistics to assess the accuracy of a prediction model. A statistical metric called R-squared (R^2) is used to quantify how well a regression model fits data. It shows the percentage that the independent variable(s) accounts for in explaining the variance in the dependent variable. The expanded uncertainty at a 95% confidence level is represented by U95.

The N10_Index is a bespoke accuracy metric introduced in this study. It represents the percentage of predicted values that fall within ±10% of the corresponding actual (measured) values. This index provides a direct, interpretable measure of how often the model predictions are acceptably close to reality, which is particularly useful in practical decision-making contexts such as sports analytics. A higher N10 value indicates stronger predictive reliability.

These performance evaluation metrics are presented in Table 3. Where the metrics are presented by measured values (M_i), predicted values by models (P_i), average measured and predicted values ($\bar{M}$ and $\bar{P}$), and the total number of studied samples (n), the following metrics utilized for evaluation of the estimation performance of the proposed models.

Table 3: Performance evaluation metrics.

$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(M_i - P_i)^2}$	root mean square error (RMSE)
$R^2 = \left(\frac{\sum_{i=1}^{n}(M_i - \bar{M})(P_i - \bar{P})}{\sqrt{[\sum_{i=1}^{n}(M_i - \bar{P})^2][\sum_{i=1}^{n}(P_i - \bar{P})^2]}} \right)^2$	Coefficient Correlation (R^2)
$U_{95} = \sqrt{\sum_{i=1}^{n}(P_i - \bar{P})^2/(n*(n-1))}$	Uncertainty Index
$SI = \frac{RMSE}{M_i}$	Scatter Index
$n10 - index = \frac{n10}{n}$	n10-index

4.2 Results of K-Fold cross validation

K-fold cross-validation is a widely adopted technique for model evaluation and selection, particularly in classification and regression tasks. The method partitions the dataset into k equal subsets; in each iteration, one subset is held out for testing while the remaining k−1 subsets are used for training. This process is repeated k times, ensuring that each subset is used exactly once as test data. In this study, a 5-fold cross-validation (k = 5) was employed to robustly assess and enhance the generalization performance of the proposed models by rotating the training and testing sets. As shown in Fig. 3, the Decision Tree (DT) model achieved its best performance in Fold 5, with the highest R^2 value of 0.96 and the lowest RMSE of 8.1 million. For the Support Vector Regression (SVR) model, Fold 1 yielded the most accurate predictions, attaining an R^2 of 0.944 and an RMSE of 11.0 million.

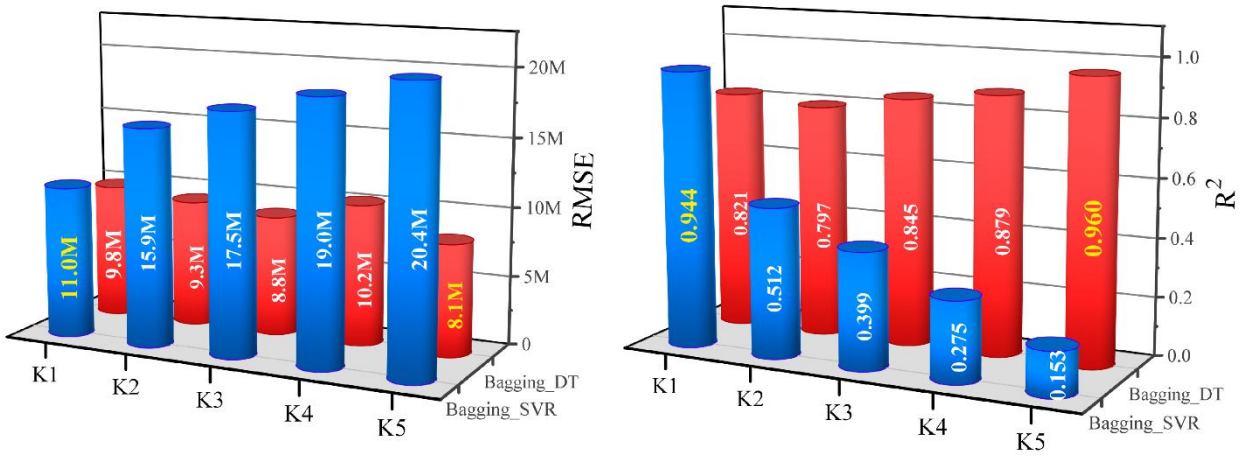


Figure 3: The results of 5-Fold cross validation

4.3 Results of hyperparameters

In machine learning, hyperparameters are critical predefined settings that control the learning process of a model. Unlike model parameters, which are learned during training, hyperparameters are set beforehand and have a substantial impact on the model's performance. To achieve optimal accuracy and efficiency, hyperparameter tuning is necessary, and one of the most common techniques for this task is random search. In this study, random search was utilized to optimize the hyperparameters of the proposed hybrid models. The optimized hyperparameter values for each model are presented in Table 4. For the Bg_SVR model, the key hyperparameters include n_estimators, n_jobs, and random_state. For the Bg_SVR(RDA) model, the most important hyperparameters were n_estimators (52), n_jobs (63), and random_state (49). Other models, such as Bg_SVR(MPS) and Bg_DT, also had their hyperparameters fine-tuned to enhance predictive performance and maintain computational efficiency. For example, Bg_DT(RDA) used hyperparameter values of n_estimators (61), n_jobs (29), and random_state (50), while Bg_DT(MPS) had n_estimators set to 10.

Table 4: The result Hyperparameters for hybrid models.

Models	Hyperparameters		
	N-estimators	N-jobs	Random-state
Bg_{SVR}	10	None	None
$Bg_{SVR(RDA)}$	52	63	49
$Bg_{SVR(MPS)}$	7	18	18
Bg_{DT}	34	73	47
$Bg_{DT(RDA)}$	61	29	50
$Bg_{DT(MPS)}$	10	None	None

4.4 Convergence curves

Fig. 4 illustrates the convergence behavior of four hybrid machine learning models throughout 200 optimization iterations. The y-axis represents the RMSE in units of market value, which measures the average magnitude of the prediction error. A lower RMSE value indicates higher model accuracy. The x-axis shows the number of iterations during the optimization process.

Initial performance: At the beginning of training, models typically start with a higher error rate or lower accuracy, indicating poor performance.

Learning phase: As training progresses, the models' performance improves, indicated by a downward trend in the error rate or an upward trend in accuracy.

Convergence point: The point where the curve starts to flatten indicates the model's convergence. Beyond this point, additional training provides minimal improvements.

By comparing the convergence curves, we can see which model converges faster and performs better. A steeper curve implies a higher learning rate, but a lower convergence point indicates better end performance. The best model appears to be Bagging_DT (MPS) (yellow line), as it is closer to the center, indicating lower error

values. The weakest model seems to be Bagging_SVR (RDA) (magenta line), as it is farther from the center, indicating higher error values. The Bagging_DT (MPS) model shows the lowest values around 8 * 10^4. The Bagging_SVR (RDA) model shows the highest values, exceeding 106*10^5.

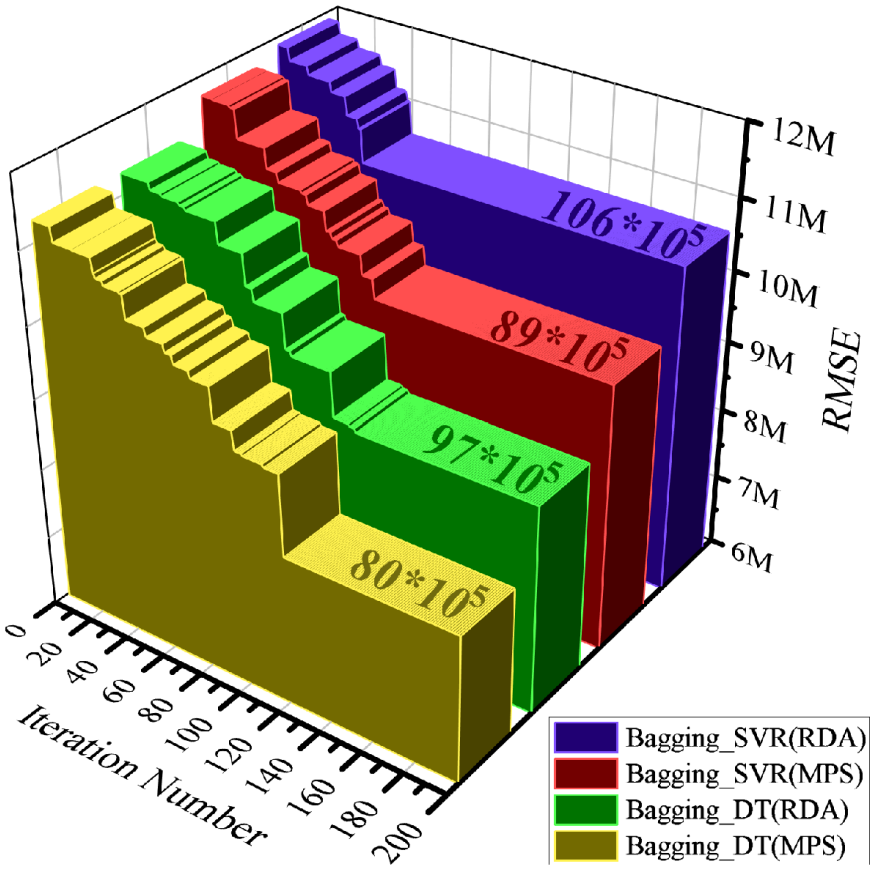


Figure 4: The convergence curve of the four hybrid models

4.5 Results of the evaluation metrics

Table 5 compares the performance metrics of several bagging models, notably Bagging Support Vector Regression (Bg_SVR) and Bagging Decision Tree (Bg_DT), as well as their modifications employing Red Deer Algorithm (RDA) and Motion PSO (MPS). It is a widely used metric in uncertainty analysis that expresses the range of values that a measured quantity's true value is most likely to lie within. The Scatter Index is a normalized measure of error that represents the percentage of error relative to the mean observation. It is a measure of how consistent the error is, with lower values indicating better model performance. During training, Bg_SVR had the minimum RMSE of 114*10^5 and maximum R^2 of 0.950, which already justifies excellent predictive capability. This is actually the best model in all respects since it has attained the top rating in every category, adding up to an overall ranking score of five.

In comparison, the Bg_DT models, especially the Bg_DT(MPS), showed significant improvements and achieved an RMSE of 846×10^4 and R^2 of 0.986 but with a higher-ranking score in general of 30 due to lower ranks on other indices. These patterns in performances by these models are further defined through validation and testing phases. The Bg_SVR keeps dominating in the validation phase with an RMSE of 763×10^5 and R^2 of 0.887, making its weak point over all the parameters. On the other side, Bg_DT(MPS) has a minimum RMSE of 533*10^4 in the validation phase and 763*10^4 in the test phase, with a maximum R^2 value of 0.962 and 0.980, indicating strong performance over the two stages. These findings also point out the usefulness of the bagging approaches, in particular Bg_DT(MPS), for enhanced model performance during many assessment phases.

Table 5: The result of developed models based on Bagging (SVR, DT)

Phase	**Model**	Index values					Score of the predicted models (1 for the worst and 6 for the best.)					*Total Ranking Score*
		RMSE	R^2	U95	SI	N10_index	RMSE	R^2	U95	SI	N10_index	
Train	Bg_SVR	$114*10^5$	0.950	$308*10^5$	0.630	0.145	1	1	1	1	1	5
	Bg_SVR(RDA)	$110*10^5$	0.961	$296*10^5$	0.606	0.167	2	2	2	2	2	10
	Bg_SVR(MPS)	$101*10^5$	0.971	$271*10^5$	0.555	0.197	4	4	4	4	4	20
	Bg_DT	$106*10^5$	0.970	$287*10^5$	0.585	0.158	3	3	3	3	3	15
	Bg_DT(RDA)	$925*10^4$	0.979	$249*10^5$	0.508	0.191	5	5	5	5	5	25
	Bg_DT(MPS)	$846*10^4$	0.986	$228*10^5$	0.465	0.216	6	6	6	6	6	30
Validation	Bg_SVR	$763*10^4$	0.887	$200*10^5$	0.463	0.130	1	1	1	1	1	5
	Bg_SVR(RDA)	$705*10^4$	0.915	$188*10^5$	0.428	0.196	2	2	2	2	2	10
	Bg_SVR(MPS)	$695*10^4$	0.923	$184*10^5$	0.422	0.239	4	4	4	4	4	20
	Bg_DT	$707*10^4$	0.917	$188*10^5$	0.430	0.196	3	3	3	3	3	15
	Bg_DT(RDA)	$636*10^4$	0.942	$169*10^5$	0.386	0.217	5	5	5	5	5	25
	Bg_DT(MPS)	$533*10^4$	0.962	$141*10^5$	0.324	0.196	6	6	6	6	6	30
Test	Bg_SVR	$102*10^5$	0.923	$274*10^5$	0.586	0.178	1	1	1	1	1	5
	Bg_SVR(RDA)	$104*10^5$	0.918	$277*10^5$	0.595	0.244	2	2	2	2	2	10
	Bg_SVR(MPS)	$95*10^5$	0.932	$253*10^5$	0.545	0.200	4	4	4	4	4	20
	Bg_DT	$95*10^5$	0.926	$256*10^5$	0.546	0.178	3	3	3	3	3	15
	Bg_DT(RDA)	$87*10^5$	0.954	$233*10^5$	0.500	0.178	5	5	5	5	5	25
	Bg_DT(MPS)	$73*10^5$	0.980	$196*10^5$	0.422	0.222	6	6	6	6	6	30

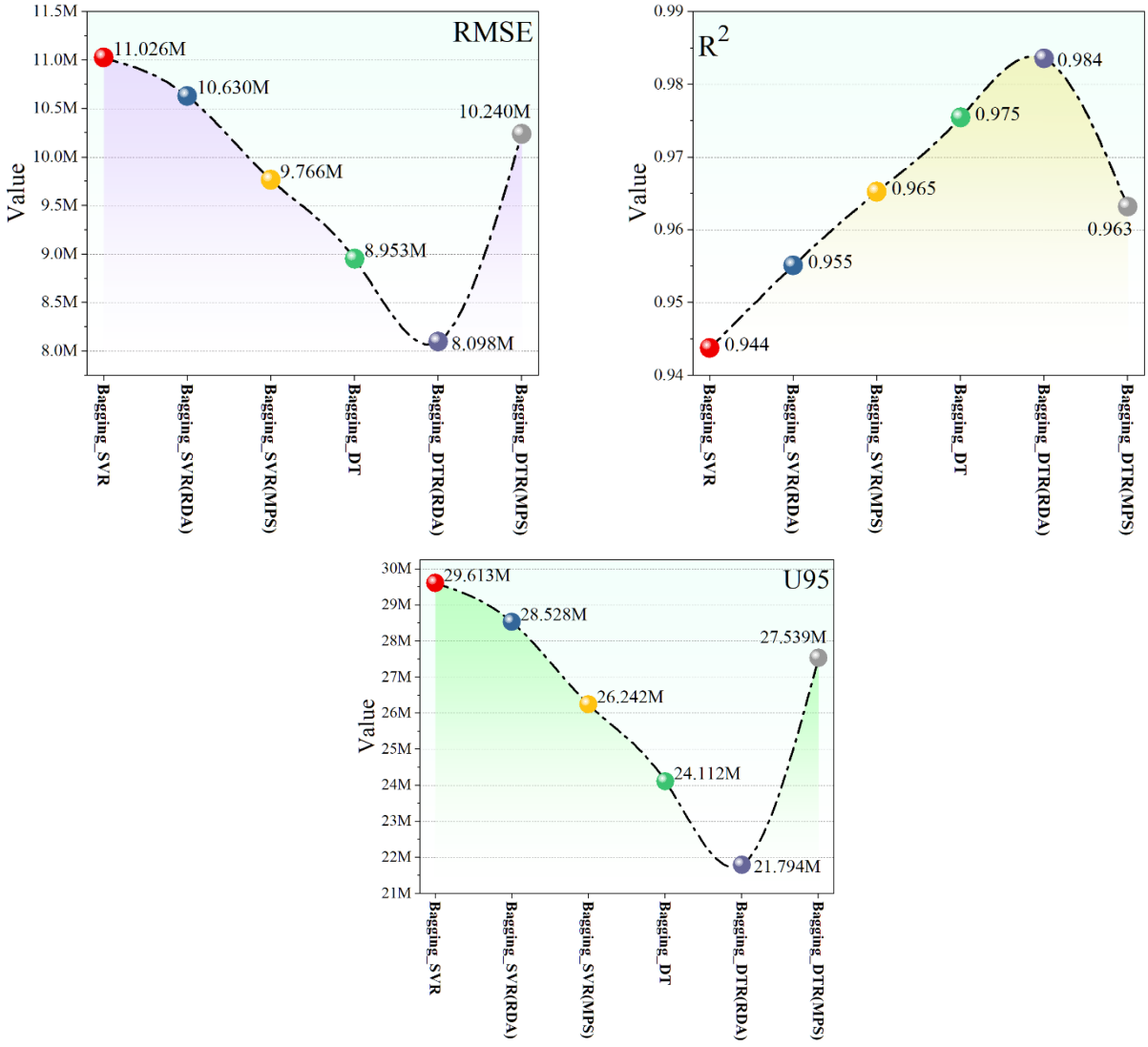
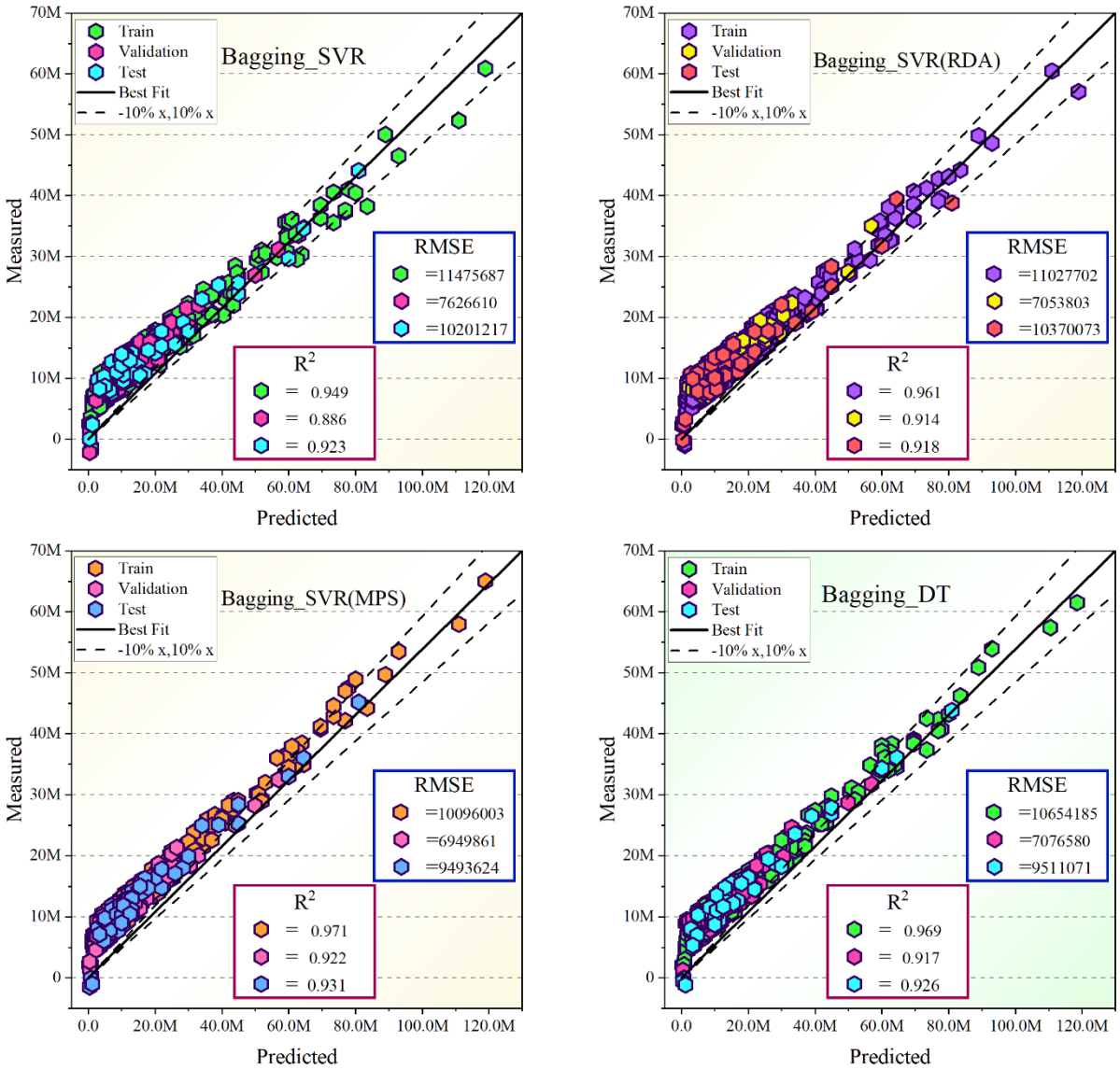


Figure 5: Results of evaluation metrics for models.

Fig. 6 shows the dispersion or variability of different evolved hybrid models that have been developed and tested. The term "evolved" in this context would mean that such models have been optimized through some iterative improvement process, including genetic algorithms or other evolutionary strategies. Dispersion in this regard shows the variance of different models from each other in terms of performance or characteristics and hence gives an insight into the models concerning stability and robustness. A small dispersion means the performance of the various models is fairly consistent across different conditions and, hence, might imply robustness and reliability. In contrast, a large dispersion indicates that the performances of the model vary greatly fact that may result either from the model's sensitivity to particular parameters or datasets. The understanding of this dispersion would, therefore, help one in selecting the most stable and reliable models for practical applications. The best model, which is Bg_DT(MPS), follows the highest performance curve at each step, and its data points are very closely lying on the central line, which means a minimal error with high prediction accuracy. The poorest performance is indicated by the weakest model of Bagging_SVR, which shows a broader dispersion of the data point from the central line, meaning significant inaccuracies in the prediction with the worst performance.



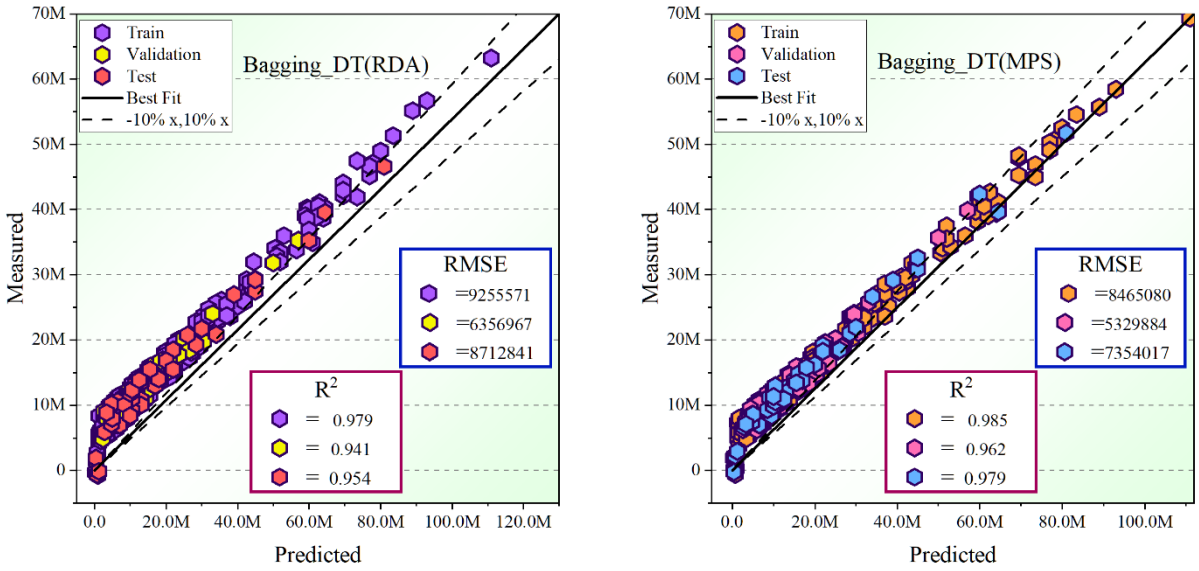
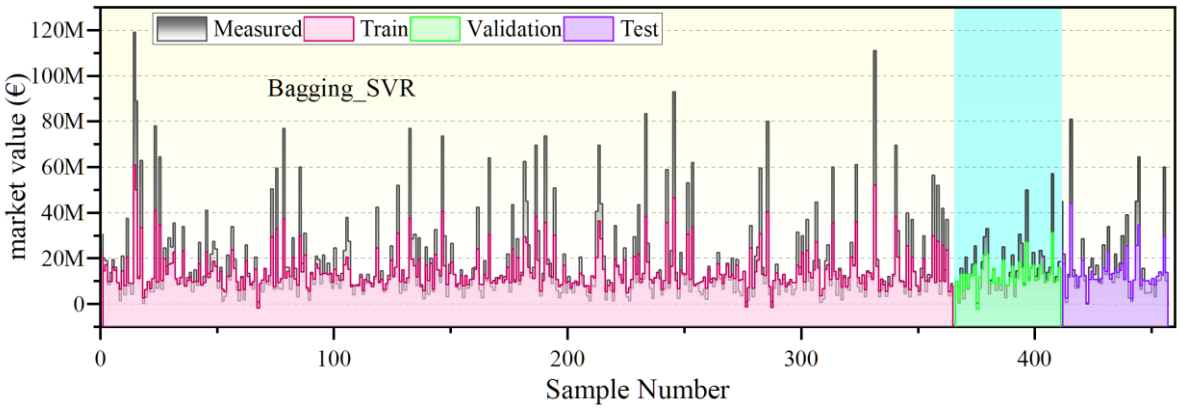
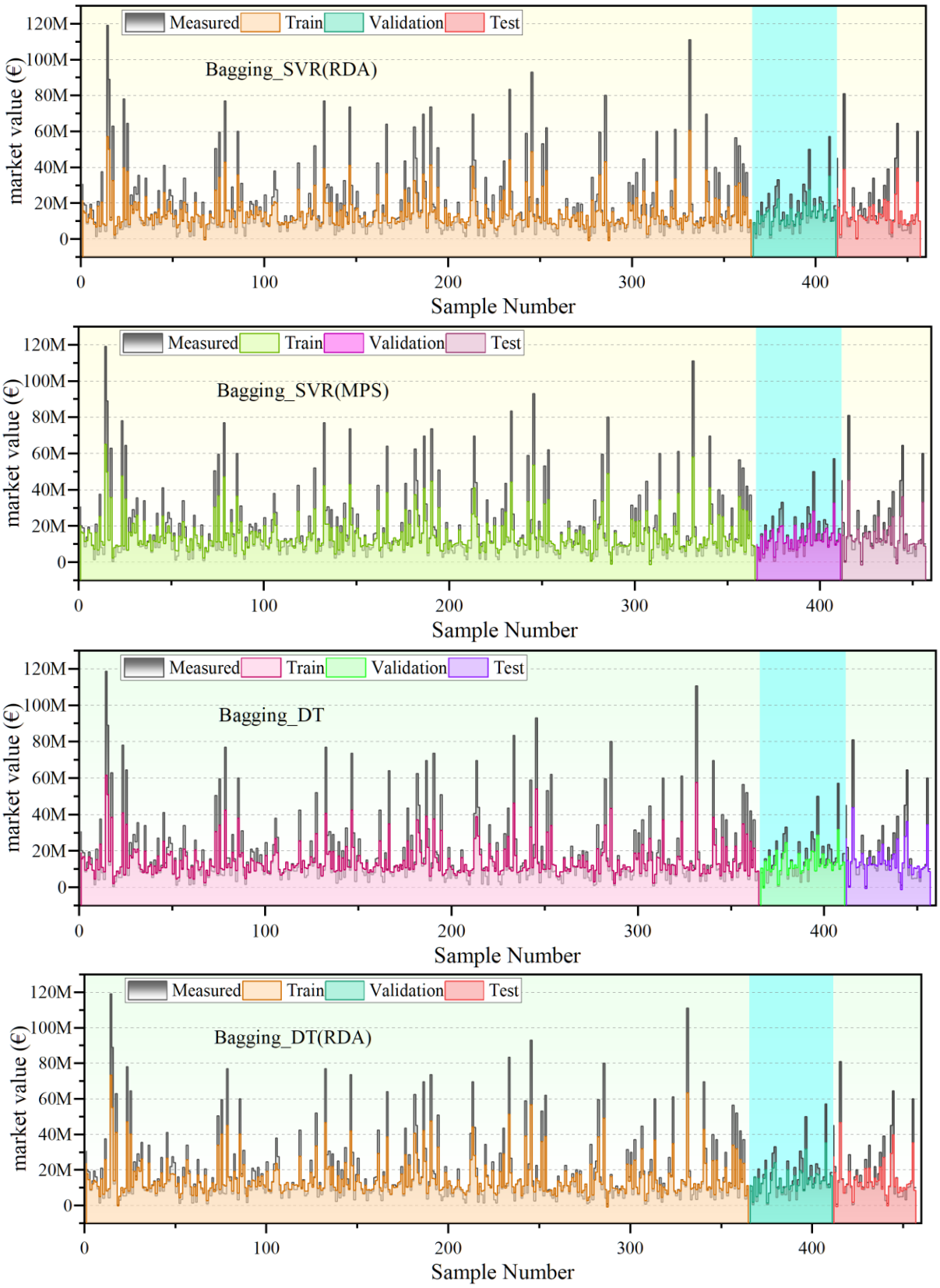


Figure 6: scatter plot of developed models

Figs. 7 and 8 present critical visual data comparisons-expressing the predicted versus measured values and the error percentages of various models, respectively. Fig. 7 serves as a graph of the predicted versus measured values, using color differentiation between the two. Usually, this would be useful for immediate discernment of how close the model's prediction comes to the actual measurement. Each model is color-coded in order to make the performance scenario comparison straightforward. The following visualization is important in establishing how well each of the models being tested performs correctly. Fig. 8 represents a column plot of the error percentages of these models. Different colors in this figure represent the error distribution of different models. Hence, it is more convenient to find out which one provides better or worse with respect to each other concerning their performance in prediction accuracy. This plot will be useful in finding out which model has the lowest error percentage that will indicate the most efficient model. Fig. 7 best model, Bagging Decision Tree with Motion PSO (Bg_DT(MPS)), shows that the predicted and measured values are almost aligned in a single trend line.

While this happened, the poorest performance recorded was from the Bagging_SVR model, which had a great deviation from the actual values and hence carried the least predictability. Fig. 8 quantifies this performance by the error percentage shown in a column plot. The Bg_DT(MPS) model has the smallest error percentage to confirm that it is the most accurate and reliable. In contrast, the highest error percentage is contributed by the Bagging_SVR model, which means this model remains the weakest among all.




Measured
Train
Validation
Test
Bagging_SVR(RDA)
market value (€)
120M
100M
80M
60M
40M
20M
0
0
100
200
300
400
Sample Number
Measured
Train
Validation
Test
Bagging_SVR(MPS)
market value (€)
Sample Number
Measured
Train
Validation
Test
Bagging_DT
market value (€)
Sample Number
Measured
Train
Validation
Test
Bagging_DT(RDA)
market value (€)
Sample Number

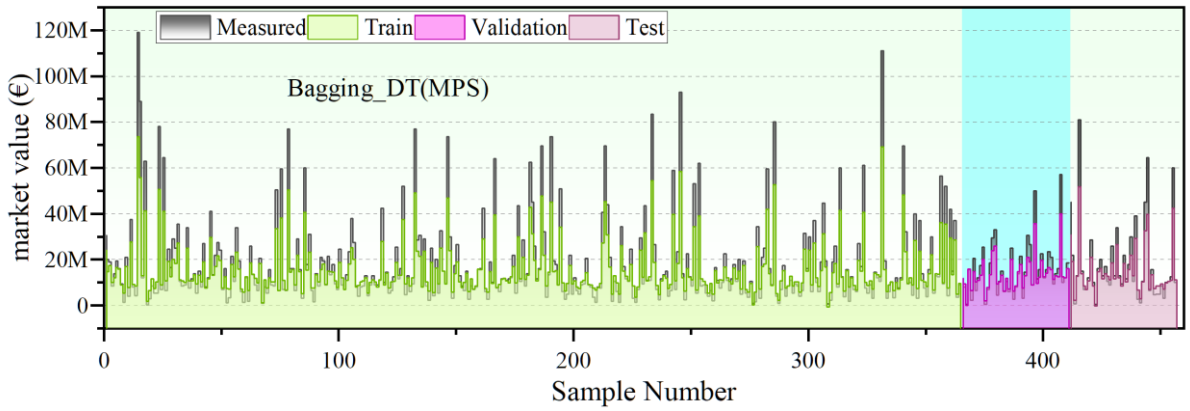
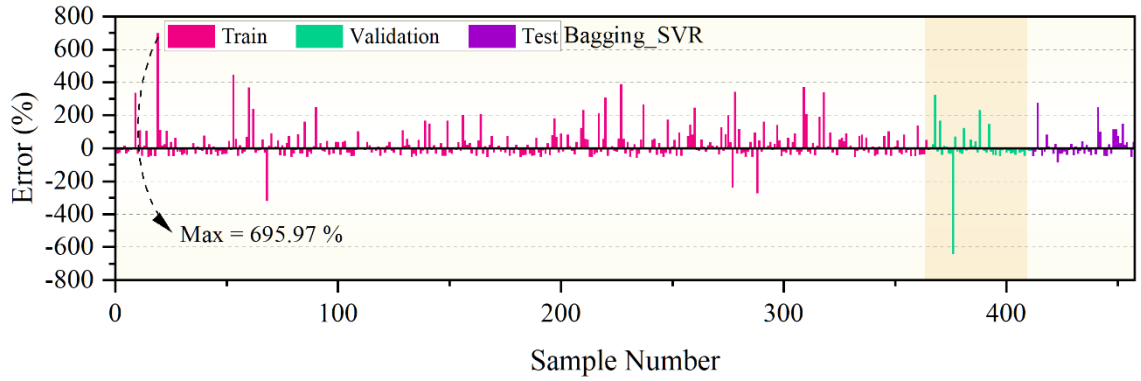
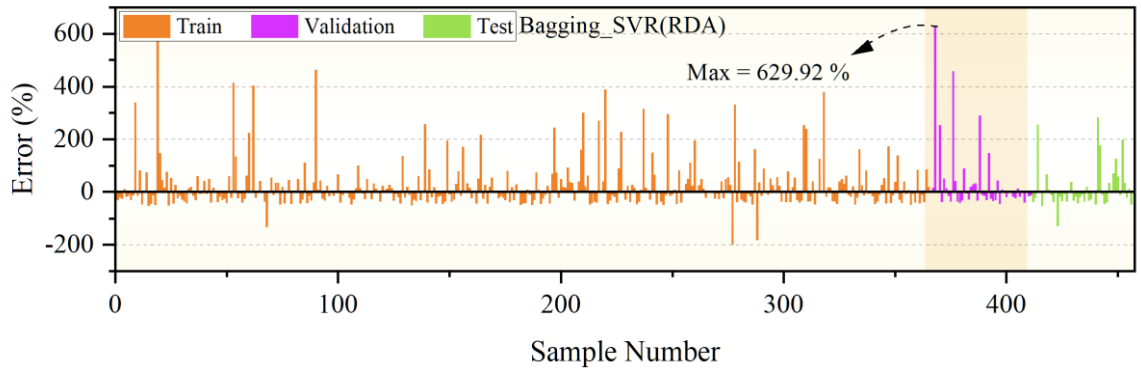
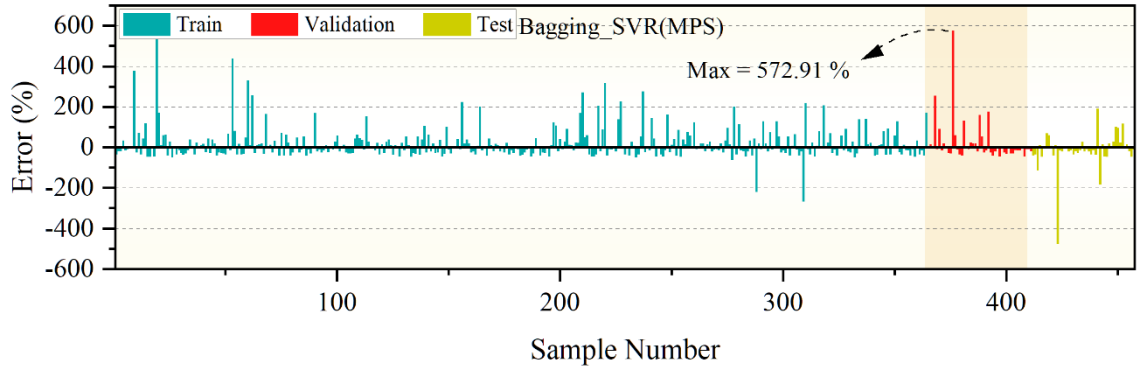


Figure 7: The comparison between the predicted and actual values of the Market Value.







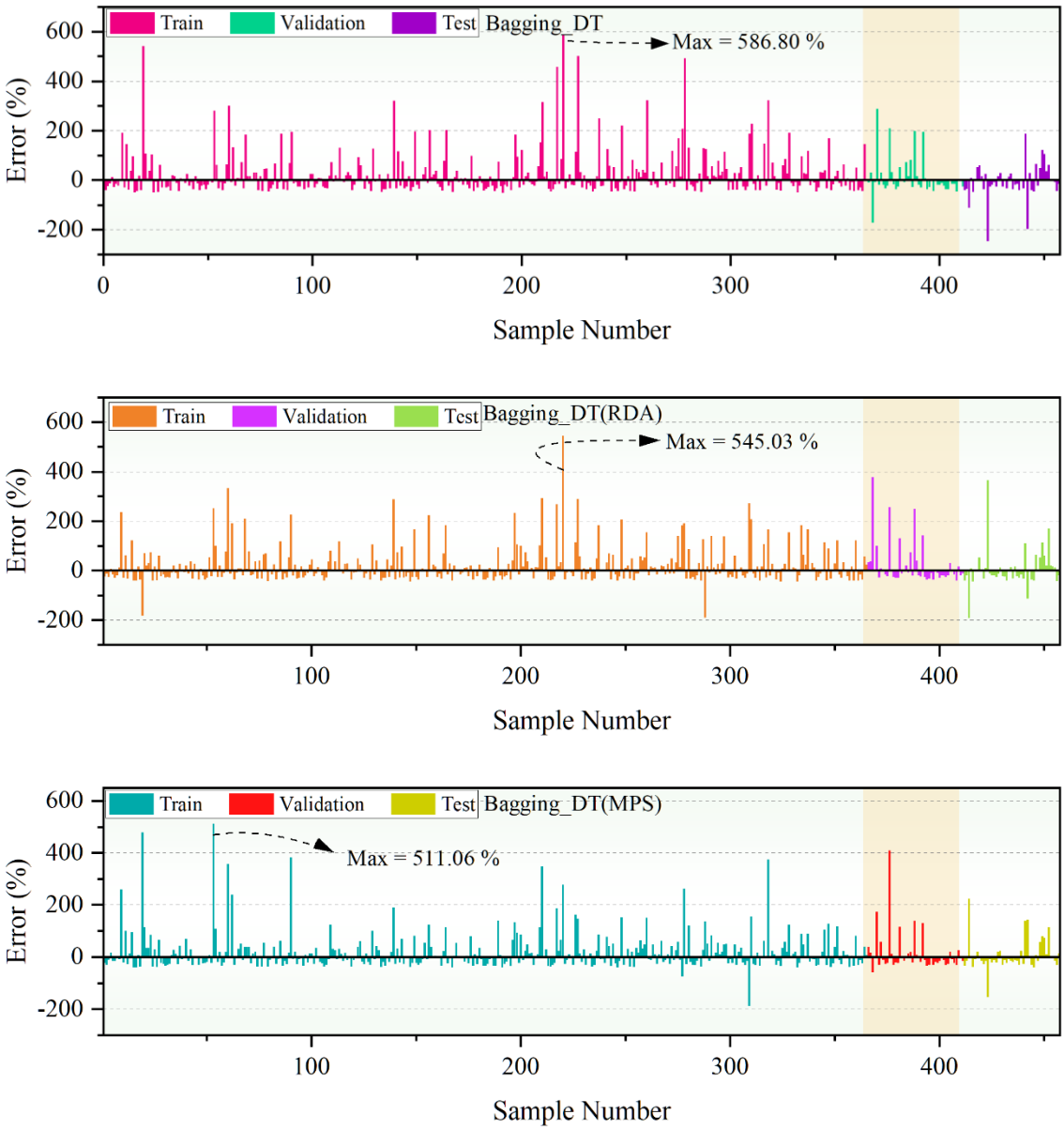


Figure 8: The histogram plots for illustrating the models' error.

Fig. 9 illustrates the error percentages of various models using a violin plot, allowing for a clear comparison of their performance in terms of prediction accuracy. The best-performing model is the Bagging Decision Tree with Motion PSO (MPS), Bagging_DT shows the best performance overall. During training, it displays a wide error range from -400% to 800%, but with a high concentration around the median, indicating some overfitting. However, during validation and testing, Bagging_DT(MPS) exhibits a much tighter error distribution, with values mostly within -100% to 100% and median errors close to 0%, indicating good generalization and consistency. The reduced variability in errors across validation and test datasets compared to the other models demonstrates Bagging_DT(MPS)'s superior ability to maintain accuracy and robustness.

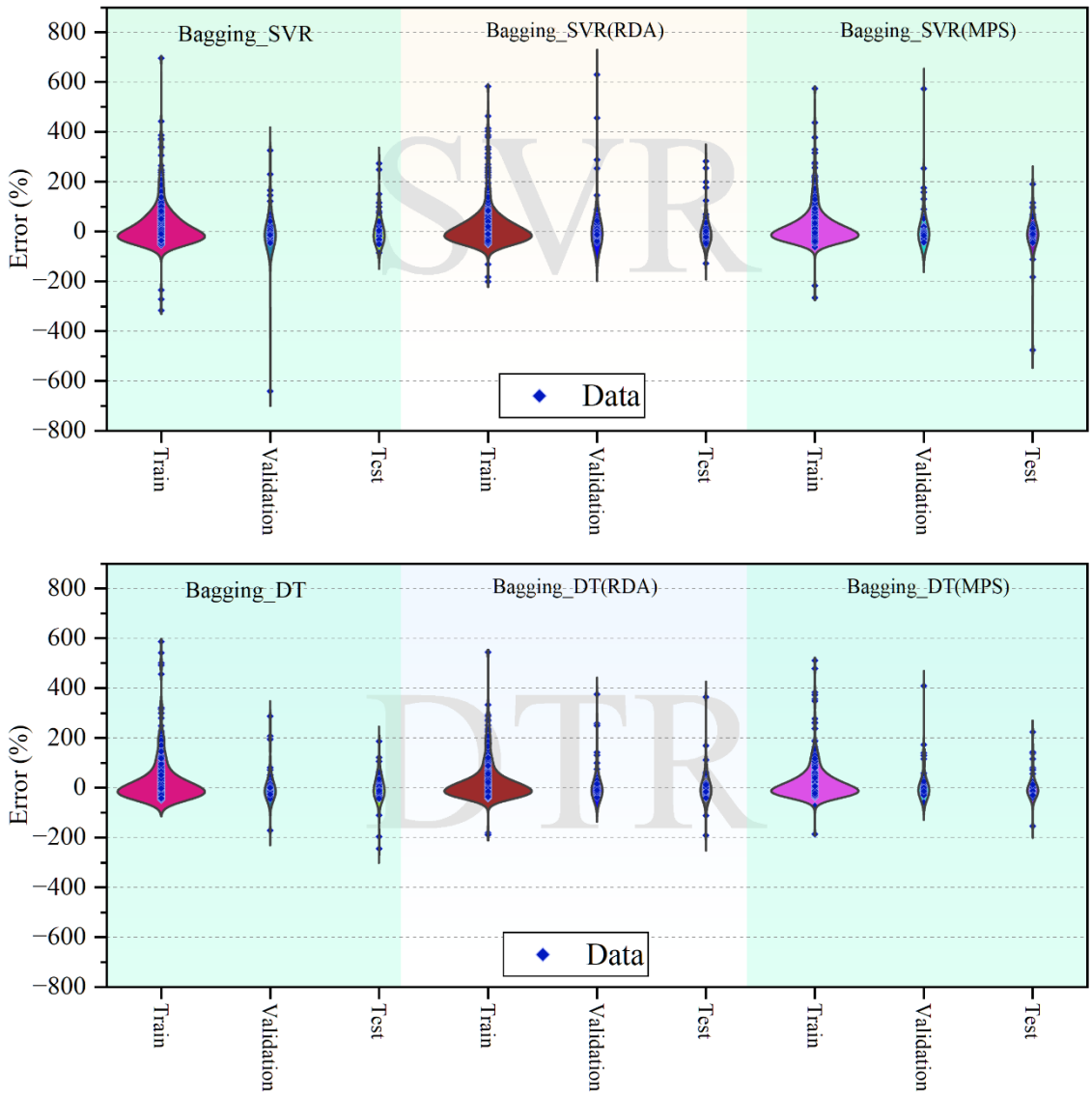


Figure 9: The violin plot errors of proposed models.

5 Fourier amplitude sensitivity test (FAST)

Fourier Amplitude Sensitivity Test (FAST) [31] is a widely adopted global sensitivity analysis method designed to evaluate how uncertainty in each input parameter influences the variability in model outputs. FAST is particularly suited for nonlinear, complex systems and is frequently used in model validation, simplification, and interpretability.

FAST provides two key indices:

- **First-order sensitivity index (S1)**: Quantifies the direct contribution of each input parameter to the output variance, ignoring interactions with other inputs. An S1 value close to 1 indicates that a variable independently accounts for a large portion of the output variance, whereas a value near 0 indicates minimal individual influence.
- **Total-order sensitivity index (ST)**: Captures the combined effect of a parameter, including both its direct impact and its interactions with other variables.

In this study, FAST is applied to assess how different football-related features contribute to the predicted output values. The goal is to determine which features are most influential in shaping model output, thereby guiding model refinement and feature prioritization.

Fig. 10 visually presents the S1 for the input variables used in the prediction model. Each bar in the figure corresponds to a specific feature (e.g., Finishing, Sprint Speed, Age), and the bar height reflects its S1 value. Higher bars indicate a stronger direct impact on the model's predictions, while lower bars suggest limited or negligible individual influence.

Key observations from Fig. 8 include:

- **Finishing (S1 = 0.543)**, **Sprint Speed (S1 = 0.517)**, and **Positioning (S1 = 0.344)** show the highest first-order sensitivity indices, identifying them as **core predictive variables** for the striker role.

- On the other hand, attributes such as **Yellow Card (S1 = 0.000)** and **Red Card (S1 = 0.000)** exhibit no measurable effect, indicating they do not meaningfully contribute to value prediction for this player type.

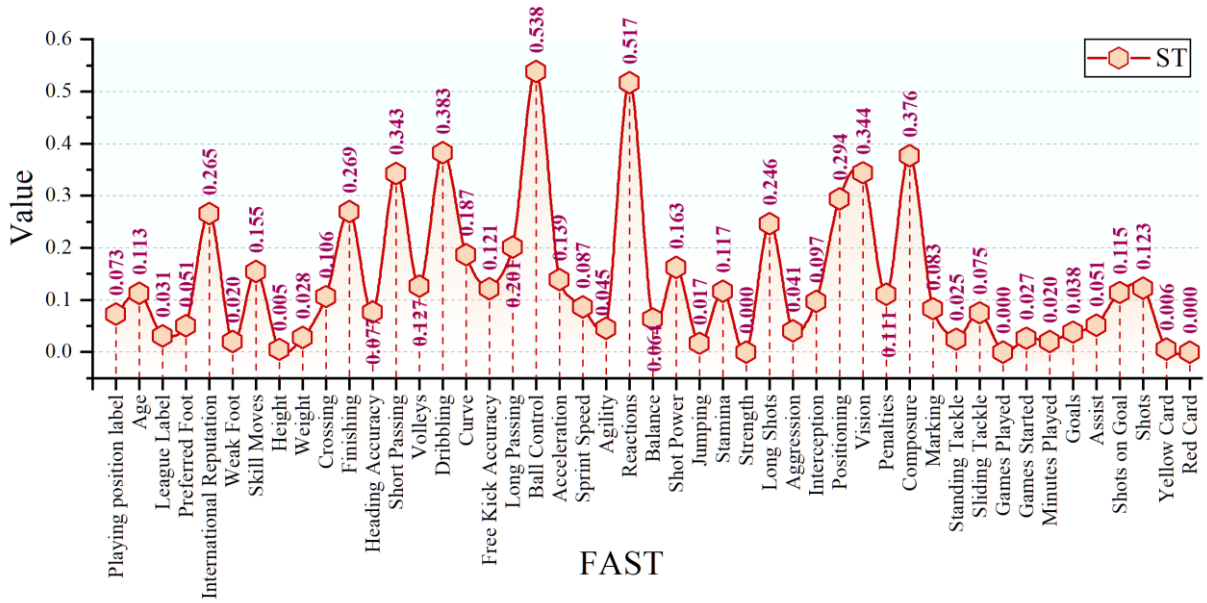


Figure 10: The FAST sensitivity analysis of the best-performed model

6 Discussion

6.1. Limitations of the study

While the findings of this study demonstrate the effectiveness of hybrid machine learning models in predicting the market value of football players, several limitations should be acknowledged to contextualize the results.

Limitations:

1. **Dataset scope and representativeness:** The dataset employed in this research was compiled from previously published sources, focusing primarily on historical data. Although it includes a diverse range of features, it may not fully capture the rapidly changing dynamics of the football market, such as recent transfers, injuries, or market inflation.
2. **Reliance on historical and static features:** Player valuation is influenced by dynamic, real-time factors such as performance in ongoing tournaments, managerial changes, or media influence. However, the current dataset relies on static, pre-existing attributes, limiting the model's responsiveness to real-time fluctuations.
3. **Model generalization:** While the models showed high performance within the training and testing phases, their generalizability to unseen leagues, seasons, or drastically different market conditions remains uncertain.
4. **Computational complexity:** The hybrid models, especially those incorporating metaheuristic optimizers like MPSO and RDA, demand significant computational resources. This may pose challenges for real-time implementation or use in resource-constrained environments.

6.2 Future research directions

1. **Dataset expansion and real-time updating:** Future studies should aim to incorporate more recent and real-time data, including match-by-match statistics, social media sentiment, and dynamic market indicators. Expanding the dataset to include players from lower-tier leagues or different continents could also improve model robustness and applicability.
2. **Integration of temporal and sequential Features:** Incorporating time-series data to track player performance over multiple seasons or transfer windows could enhance the model's predictive power by capturing performance trends and fluctuations.
3. **Exploration of alternative and hybrid optimization techniques:** Further research could explore other metaheuristic or hybrid optimization algorithms, such as Grey Wolf Optimizer, Harris Hawks Optimization, or Multi-Objective Evolutionary Algorithms, to potentially improve convergence speed and prediction accuracy.
4. **Model interpretability and explainability:** Developing interpretable models using techniques like SHAP (SHapley Additive

exPlanations) could help stakeholders in understanding the reasoning behind predictions, making the models more trustworthy and actionable.

6.3 Practical implications

The outcomes of this study hold significant practical relevance for various stakeholders within the football ecosystem, particularly in the domains of talent scouting, player valuation, and strategic financial planning.

1. **Data-driven decision making in player valuation:** The developed hybrid models offer a robust and objective approach to estimating market values of football players. By leveraging machine learning and optimization algorithms, clubs can reduce reliance on subjective assessments, leading to more accurate and transparent valuations.
2. **Enhanced scouting and recruitment efficiency:** Scouting departments can use these models to pre-screen a wide range of players across leagues and markets, identifying undervalued talent or potential high-return investments. This can streamline recruitment efforts and reduce the time and cost associated with manual evaluations.
3. **Financial strategy and contract negotiations:** Club management and financial planners can incorporate the model outputs into contract renewal negotiations or transfer strategies. The quantification of a player's market value with high predictive accuracy supports better budgeting and risk assessment.
4. **Real-Time adaptation to market dynamics:** While the current study utilizes historical data, the models are designed to be adaptable. With real-time data integration in future implementations, clubs could dynamically adjust player valuations based on recent performance, injuries, or other market changes.
5. **Benchmarking and performance analysis:** The ML framework can also serve as a benchmarking tool to compare players across different teams and leagues, helping clubs identify performance gaps or overvalued assets.
6. **Commercial and sponsorship valuation:** Beyond on-field performance, player value has implications for sponsorship and branding. Accurate valuation models provide a foundation for estimating the commercial potential of players, assisting marketing teams in forming profitable partnerships.

7 Conclusion

This study explored ML models to predict the market value of football players using a comprehensive dataset. It was designed to develop more approaches with ML, including Bagging Regression, SVR, and Bagging DTR improved by Motion-encoded PSO and the Red Deer Algorithm. Among those, the Bagging Decision Tree with Motion PSO was the best in both the validation and test phases, with RMSE $533*10^4$ and R^2 0.962 in the validation phase and RMSE $73*10^5$ and R^2 0.980 in the testing. These results confirmed how hybrid models can capture such complexities in the valuations. By far, the Bg_SVR performance was excellent in the training phase, with a minimum RMSE of $114*10^5$ and a maximum R^2 of 0.950, showing that this algorithm is strong. However, since Bg_DT(MPS) had a good performance for both validation and testing, it is ranked as the best overall model despite its lower performance for some indices, such as the N10_ index. The results hereby confirm that ML bears the potential of becoming an increasingly objective and more accurate valuation method for football players than traditional subjective assessments. Advanced optimization algorithms were highly effective in increasing the accuracy and reliability of the models developed in this study. Future research might further refine these models by adding more parameters and testing different optimization techniques. This study represents a great step toward more data-driven decision-making processes in football management and may potentially turn upside down the traditional habits of player evaluations.

References

[1] Kitching G. The Origins of Football: History, Ideology and the Making of 'The People's Game.' History Workshop Journal 2015; 79:127–53. https://doi.org/10.1093/hwj/dbu023.

[2] Dobson S, Gerrard B. The determination of player transfer fees in English professional soccer. Journal of Sport Management 1999; 13:259–79.

[3] Müller O, Simons A, Weinmann M. Beyond crowd judgments: Data-driven estimation of market value in association football. Eur J Oper Res 2017; 263:611–24.

[4] Frick B. The Football Players' Labor Market: Empirical Evidence from The Major European Leagues. Scott J Polit Econ 2007; 54:422–46. https://doi.org/https://doi.org/10.1111/j.1467-9485.2007.00423.x.

[5] 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 Management Review 2014; 17:484–92.

[6] Razali MN, Mustapha A, Mostafa SA, Gunasekaran SS. Football matches outcomes prediction based on gradient boosting algorithms and football rating system. Human Factors in Software and Systems Engineering 2022;61:57.

[7] Li C, Kampakis S, Treleaven P. Machine learning modeling to evaluate the value of football players. ArXiv Preprint ArXiv:220711361 2022.

[8] Laros GGPK. Predicting Transfer Value of Professional Football Players Based on Player Skills and Characteristics Using Multiple Linear

Regression, Support Vector Regression, and Random Forest Regression 2022.

[9] Felipe JL, Fernandez-Luna A, Burillo P, de la Riva LE, Sanchez-Sanchez J, Garcia-Unanue J. Money Talks: Team Variables and Player Positions that Most Influence the Market Value of Professional Male Footballers in Europe. Sustainability 2020;12:1–8.

[10] Behravan I, Razavi SM. A novel machine learning method for estimating football players' value in the transfer market. Soft Comput 2021;25:2499–511.

[11] Peeters T. Testing the Wisdom of Crowds in the field: Transfermarkt valuations and international soccer results. Int J Forecast 2018;34:17–29.

[12] Adeshina AM, Razak SFA, Yogarayan S, Sayeed S. Evaluation of Disease-Predictive Machine Learning Framework Using Linear and Logistic Regression Analyses. Informatica 2024;48.

[13] Kaushal A, Gupta AK, Sehgal VK. Hybrid CatBoost and SVR Model for Earthquake Prediction Using the LANL Earthquake Dataset. Informatica 2025;49.

[14] Sun J. Prediction and estimation of book borrowing in the library: Machine learning. Informatica 2021;45.

[15] Sadaghat B, Ebrahimi SA, Souri O, Yahyavi Niar M, Akbarzadeh MR. Evaluating strength properties of Eco-friendly Seashell-Containing Concrete: Comparative analysis of hybrid and ensemble boosting methods based on environmental effects of seashell usage. Eng Appl Artif Intell 2024;133:108388. https://doi.org/https://doi.org/10.1016/j.engappai.2024.108388.

[16] Li H, Chen J, Zhang W, Zhan H, He C, Yang Z, et al. Machine-learning-aided thermochemical treatment of biomass: a review. Biofuel Research Journal 2023;10:1786–809.

[17] Huang G-B, Zhu Q-Y, Siew C-K. Extreme learning machine: theory and applications. Neurocomputing 2006;70:489–501.

[18] Nsangou JC, Kenfack J, Nzotcha U, Ekam PSN, Voufo J, Tamo TT. Explaining household electricity consumption using quantile regression, decision tree and artificial neural network. Energy 2022;250:123856.

[19] 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.

[20] Memmert D. Data analytics in football: positional data collection, modeling, and analysis. Journal of Sport Management 2019;33:308–2019.

[21] Majewski S. Identification of Factors Determining Market Value of the Most Valuable Football Players. Journal of Management and Business Administration Central Europe 2016;24:91–104. https://doi.org/10.7206/jmba.ce.2450-7814.177.

[22] Singh P, Lamba PS. Influence of crowdsourcing, popularity and previous year statistics in market value estimation of football players. Journal of Discrete Mathematical Sciences and Cryptography 2019;22:113–26.

[23] Li C, Kampakis S, Treleaven P. Machine learning modeling to evaluate the value of football players. ArXiv Preprint ArXiv:220711361 2022.

[24] Talbi E-G. Metaheuristics: from design to implementation. John Wiley & Sons; 2009.

[25] Li D-Y, Xu W, Zhao H, Chen R-Q. A SVR based forecasting approach for real estate price prediction. 2009 International conference on machine learning and cybernetics, vol. 2, IEEE; 2009, p. 970–4.

[26] Xu M, Watanachaturaporn P, Varshney PK, Arora MK. Decision tree regression for soft classification of remote sensing data. Remote Sens Environ 2005;97:322–36.

[27] Sutton CD. 11 - Classification and Regression Trees, Bagging, and Boosting. In: Rao CR, Wegman EJ, Solka JLBT-H of S, editors. Data Mining and Data Visualization, vol. 24, Elsevier; 2005, p. 303–29. https://doi.org/https://doi.org/10.1016/S0169-7161(04)24011-1.

[28] Opitz D, Maclin R. Popular ensemble methods: An empirical study. Journal of Artificial Intelligence Research 1999;11:169–98.

[29] Eberhart R, Kennedy J. A new optimizer using particle swarm theory. MHS'95. Proceedings of the sixth international symposium on micro machine and human science, Ieee; 1995, p. 39–43.

[30] fifa 19 n.d. https://www.kaggle.com/karangadiya/fifa19.

[31] Edmund Ryan, Oliver Wild, Apostolos Voulgarakis and LL. Fast sensitivity analysis methods for computationally expensive models with multi-dimensional output 2018.