Developing a Reliable Hybrid Machine Learning Model for Objective Soccer Player Valuation

Hongtao Yu[1], Jialiang Li[2]*
Department of Physical Education, Changchun Institute of Technology, Changchun 130012, Jilin, China[1]
College of Physical Education, Yanching Institute of Technology, Langfang 065201, Hebei, China[2]

***Abstract*—Football is both a popular sport and a big business. Managers are concerned about the important decisions that team managers make when it comes to player transfers, player valuation issues, and particularly the determination of market values and transfer fees. Market values are important because they can be thought of as estimates of transfer fees or prices that could be paid for a player on the transfer market. Football specialists have historically estimated the market. However, expert opinions are opaque and imprecise. Thus, data analytics may offer a reliable substitute or supplement to expert-based market value estimates. This paper suggests a quantitative, objective approach to value football players on the market. The technique is based on applying machine learning algorithms to football player performance data. To achieve this objective, Decision Tree Regression (DTR) was employed to predict the market value of football players. Additionally, two novel metaheuristic algorithms, Honey Badger Algorithm (HBA) and Jellyfish Search Optimizer (JSO), were utilized to enhance the performance of the DTR model. The experiment made use of FIFA 20 game data that was gathered from sofifa.com. In addition, it aims to examine the information and pinpoint the key elements influencing market value assessment. The trial results showed that the DTJS hybrid model performed better in predicting the participants' market pricing than other algorithms. With an R^2 value of 0.984 and the lowest error ratio when compared to the baseline, it gets the highest accuracy score. Lastly, it is thought that these findings may be crucial in the discussions that occur between football teams and the agents of players. This strategy may be used as a springboard to expedite the negotiation process and provide a quantifiable, objective assessment of a player's market worth.**

Keywords*—Market value; machine learning; soccer player; decision tree regression; Honey Badger Algorithm; Jellyfish Search ptimizer

I. INTRODUCTION

A. Background

Regarding players and viewers, football is the most popular sport in the world [1]. $27 billion was estimated to have been made by European football teams alone in 2017 [2]. Therefore, it becomes a key contributor to the world economy [3]. The market for football players has grown significantly over the last several decades, and their worth currently exceeds $\$100\ M$ [4]. These rates are much greater than historical trade numbers when contrasted with the average rate of inflation [5].

Choosing players is the most important management choice football teams have to make. Player transfers have a big influence on a team's chances of winning. As a result, scholars from various fields have investigated the variables influencing transfer fees [6]. Researchers' attention has recently been focused on player market pricing. The player's market value is the amount a club might demand to transfer a player's contract to another team [7]. Market values play a significant part in transfer discussions because they provide estimates of transfer fees, even if transfer fees represent the real prices paid in the market [8]. Market prices have always been important to football experts like team managers and sports finalists, but in recent years, crowdsourcing websites like Transfermarkt ($www.transfermarkt.com$) have shown to help assess market values [9].

Nonetheless, there is a lack of widespread use of data-driven techniques for determining market value in football [10]. The literature has given a detailed account of the difficulty in identifying the critical elements that influence football players' market value [11–14]. Numerous variables were discovered in the literature, and these indications are divided into three groups: player attributes, player effectiveness, and player popularity. According to certain research, the dependent variable (market value) and certain of these variables, like age, have nonlinear relationships [15,16], and [17]. Over the last 20 years, machine learning has become a critical component in turning football data into actionable insights that teams and coaches can use to assess opponents and make better judgments at the moment [18]. There hasn't been much research done on football analytics using machine learning methods. The main reason for this is that there isn't a complete player dataset, which is problematic since teams with significant financial resources may be the only ones able to compile such detailed player data [19].

Video games such as FIFA and Football Manager (FM) are regarded as additional data sources in football analytics. Clubs and academics have been using video games as alternative data sources since 2014 [20]. Shin and Robert forecasted the outcomes of the matches using data from the FIFA video game. They discovered that machine learning programs using this data can produce highly accurate predictions [21, 22]. This study presents an efficient machine-learning technique that was created with the $FIFA$ 20 dataset. This collection contains the different performance ratings of almost 17,000 players [23]. The shooting, passing, and dribbling scores of players are displayed through their attributes in this dataset. It is possible to assess the players' performances from the previous season by using this dataset. As far as awareness goes, the employment of linear regression models has ignored the fact that certain factors have nonlinear relationships with player values. This suggests that nonlinear regression techniques ($such\ as\ decision\ trees$)

may perform better than the conventional strategy that has been documented in the literature.

B. Literature Review

Al-Asadi and Tasdemir [24] proposed an objective and quantitative method for determining football players' market values by applying machine learning algorithms to players' performance data from FIFA 20 video game data collected from sofifa.com. Four regression models—linear regression, multiple linear regression, decision trees, and random forests—were utilized to estimate market values and analyze the data to identify influential factors. The experimental results indicated that the random forest algorithm outperformed other models, achieving the highest accuracy score and lowest error ratio compared to baseline methods. This study demonstrated the effectiveness of the proposed methods in valuing football players, surpassing previous works in this area. Additionally, the findings suggested implications for negotiations between football clubs and players' agents, as the proposed model could simplify the negotiation process and provide an objective quantitative estimate of a player's market value. Herm et al. [25] investigated the evaluation process within a community, assessing the accuracy of its estimated market values and determining the most influential attributes for market-value evaluations. By demonstrating the community's ability to predict actual transfer fees, the study revealed that these evaluations can be largely explained by an econometric model consisting of two blocks of determinants: variables directly linked to players' talent and variables resulting from judgments by external sources, such as journalists. By reorganizing variables used in previous studies into these two blocks, the research offered a more nuanced perspective on the popularity of players compared to recent literature on the "superstar phenomenon." Behravan and Razavi [26] proposed a novel method for estimating football players' market values using the FIFA 20 dataset. It comprised two phases: automatic clustering of the dataset into position-based clusters, and the use of a hybrid regression model combining particle swarm optimization (PSO) with support vector regression (SVR) to predict market values for each cluster. The results demonstrated the effectiveness of the method, achieving a 74% accuracy rate. PSO outperformed other metaheuristics, indicating its superiority in this context. This approach contributed to advancing data-driven player valuation methods, offering potential improvements in accuracy for football market assessments.

C. Objective

This study delves into the critical task of predicting the market value of soccer players using Decision Tree Regression (DTR). Recognizing the dual impact of player market value on both the economic and cultural fabric of teams and society, accurate valuation metrics are imperative for informed decision-making in player acquisitions. To augment the predictive capabilities of the DTR model, we introduce two innovative optimizers: the Jellyfish Search Optimizer (JSO) and the Honey Badger Algorithm (HBA). In this study, we propose the development of hybrid models by integrating DTR with each optimizer, resulting in the creation of the DTJS (Decision Tree + JSO) and DTHB (Decision Tree + HBA) models. These hybrid approaches aim to leverage the complementary strengths of DTR and the respective optimizers, thereby enhancing the accuracy and robustness of player valuation predictions. To rigorously evaluate the effectiveness of these hybrid models, comprehensive performance evaluations are conducted. These evaluations encompass a range of metrics and analyses to assess predictive accuracy, model stability, and generalization capabilities across diverse datasets.

DTR was chosen for its interpretability, simplicity, and effectiveness in handling both numerical and categorical data, making it suitable for modeling the complex, non-linear relationships in soccer player market values. To enhance the DTR model's performance, the HBA and JSO were employed. HBA, inspired by the strategic foraging behavior of honey badgers, optimizes the model by effectively exploring the parameter space and avoiding local optima. JSO, simulating jellyfish movement patterns, fine-tunes the model parameters to achieve an optimal balance between bias and variance. Integrating these optimizers with DTR aims to create hybrid models (DTJS and DTHB) that combine the decision tree's robustness with the advanced optimization capabilities of HBA and JSO, resulting in enhanced accuracy and reliability in predicting soccer players' market values.

II. Datasets and Methodology

A. Data Gathering

The dataset for predicting soccer players' market value is sourced from sofifa.com. This study's dataset comes from (https://www.openml.org/search?type=data&status=active&id=43604), which includes real-world statistical records and the FIFA 19 video game database. This large dataset required data engineering to make it acceptable for evaluating the market worth of players with diverse playing positions in well-known football leagues. Originally, it contained 53 attributes for 491 sampled players. Fig. 1 shows the comprehensive player attributes and performance metrics available in the FIFA 20 game data. In the selection process, active players in the FIFA 20 game are considered, ensuring a broad representation of player positions to capture diverse playing styles. The dataset incorporates various input variables, including demographic information such as age and international reputation, technical skills like weak foot, skill moves, short pass, finishing, heading accuracy, crossing, volleys, curve, dribbling free kick accuracy, long pass, and ball control, as well as physical attributes such as height, weight, sprint speed, acceleration, agility, stamina, shot power, balance, jumping, reactions, and strength.

Additionally, key performance metrics like goals, assists, and shots on goal, yellow cards, and red cards are included, along with potential and overall ratings. The dataset also encompasses mental attributes, covering aggression, interception, positioning, vision, penalties, composure, and marking, standing tackle, and sliding tackle. The data collection process involves systematic extraction from the sofifa.com database using web scraping techniques, ensuring accuracy and consistency. Subsequently, the dataset undergoes a meticulous cleaning process to address missing values, outliers, and inconsistencies, enhancing its integrity and reliability for subsequent analyses.

The choice of utilizing FIFA 20 game data from sofifa.com holds particular relevance to this study for several reasons.

Firstly, FIFA 20 is one of the most widely played and recognized football simulation video games globally, capturing a vast array of player attributes and performance metrics. By leveraging this extensive dataset, which is regularly updated to reflect real-world player performances and transfers, we ensure the inclusion of current and comprehensive player data in our analysis. Furthermore, sofifa.com serves as a reputable and reliable source for FIFA player data, providing structured and standardized information that facilitates systematic analysis and comparison across players. The accessibility and completeness of the data available on sofifa.com enable researchers to construct robust predictive models and conduct rigorous evaluations of player valuation methodologies.

B. Decision Tree Regression (DTR)

One kind of tree – based structure used to forecast the dependent variable's numerical results is decision tree regression. An implementation of Quinlan's $M5$ algorithm is also referred to as the $M5P$ algorithm [27]. $M5P$ is a tree-based structure similar to $CART$ (classification and regression tree); however, it has multivariate linear models instead of regression trees with values at the leaves like in CART. Furthermore, the M5P method typically produces smaller model trees than the CART algorithm's tree. The following describes how decision tree regression operates.

First, a tree is constructed using a traditional decision-tree approach. This decision tree uses a splitting criterion that lower the intra-subset volatility in the class values of instances that descend each branch. The root node is determined by selecting the property that maximizes the projected reduction in error. Eq. (1)'s formula is used to compute the standard deviation decrease.

$$SDR = sd(T) - \sum_i \frac{T_i}{|T|} \times sd(T_i) \quad (1)$$

The tree is then trimmed back to just a few leaves. Ultimately, a smoothing process is employed to mitigate the abrupt changes in slope that will unavoidably transpire among neighboring linear models at the tree's leaves after pruning [28].

C. Jellyfish Search Optimizer (JSO)

The JFS optimizer is controlled by three pillars and takes its cues from the movements of jellyfish. The first pillar states that the jellyfish can travel either within their swarm or toward the ocean current [29]. By alternating between these two forms, the temporal control (TC) mechanism can regulate the movements of the jellyfish. The jellyfish are lured to their locations when there is an adequate supply of food, which is the second pillar. The third pillar is that the quantitative objective function is used to characterize the amount of food [30]. The jellyfish population is randomly initialized during chaotic logistic mapping, and it can be expressed as follows:

$$X_i(t+1) = 4P_0(1 - X_i), 0 \le P_0 \le 1 \quad (2)$$

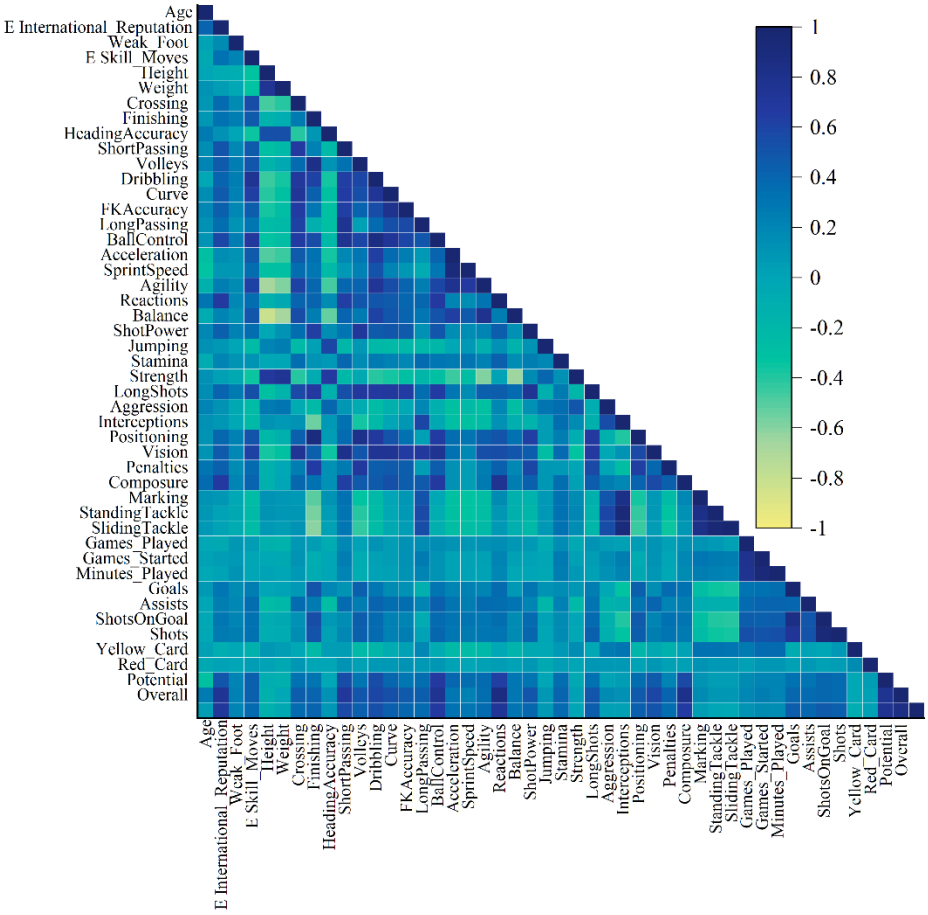

Fig. 1. Correlation matrix to analyze the relationships between input and output variables.

where, P_o indicates the starting jellyfish population, which may produce a value of $P_0 \in (0,1)$; $P_0 \notin \{0.0, 0.25, 0.75, 0.5, 1.0\}$, and X_i reflects the i_{th} jellyfish logistic chaotic value.

The time control function CF(t) in comparison to a constant CO_0 is one of the two key components of the TC [31]. Here is how the time control function is computed:

$$CF(t) = \left|(1 - \frac{t}{Max_{iter}}) \times (2 \times rand(0,1) - 1)\right| \quad (3)$$

where t and Max_{iter} stand for the number of iterations and the maximum number of iterations, respectively.

While the value of CO_0 is fixed at 0.5, the value of $CF(t)$ varies with time, ranging from 0 to 1. When the CF value is greater than the COo value, the jellyfish will migrate in the direction of the ocean current [32]. The average of each jellyfish's vectors to the optimal jellyfish site is used to determine the direction of this current. Therefore, each jellyfish's new location is determined using the formula shown in Eq. (4):

$$\begin{aligned} X_i(t+1) &= R \times (X^* - 3 \times R \times \mu) + X_i(t) \\ X_i(t+1) &= R \times (X^* - 3 \times R \times \mu c) + X_i(t) \\ \mu c &= \frac{\sum_{i=1}^{rr} X_i(t)}{rr} \end{aligned} \quad (4)$$

where, R is a random quantity within the range $[0-1]$, and the optimal jellyfish position at that precise instant is shown by X^*, whereas the parameter (m) indicates the mean of all jellyfish locations in the swarm.

The jellyfish will go into the swarm when CF is less than CO_0. Two types of mobility within a swarm are covered: passive $(Type\ A)$ and active $(Type\ B)$. The majority of jellyfish in (Type A) are moving around their own positions, as shown by Eq. (5), with each jellyfish's position being updated:

$$X_i(t+1) = 0.1 \times R \times (U_b - L_b) + X_i(t) \quad (5)$$

where, the search spaces' $upper$ and $lower$ bounds are indicated, individually, by U_b and L_b.

A vector that extends from the jellyfish of interest (i) to the randomly selected jellyfish (j) of $type\ B$ which is not the one of interest determines the direction of movement. This kind of effective local search space exploitation is seen in Eq. (6), where the selected jellyfish's updated position is mimicked.

$$X_i(t+1) = \begin{cases} X_i(t) + R \times \left(X_j(t) - X_i(t)\right) if\ f(X_i) \geq f(X_j) \\ X_i(t) + R \times \left(X_i(t) - X_j(t)\right) if\ f(X_i) < f(X_j) \end{cases} \quad (6)$$

where, f stands for the jellyfish location $X's$ objective function value.

Types A or B are chosen based on the TC mechanism. When comparing the term (1 -CF(t)) with a random number in the range of [0-1], it is important to keep this in mind. If this is more than the calculated value of $(1 - CF(t))$, type A motion is shown by the JSO. Conversely, jellyfish travel in a $type\ B$ motion in the case that the random number is less than the computed result. To be explicit, type B motion is favored over time, while $type\ A$ motion is selected at the start condition when the TC function quickly decreases from 1 to 0 over time.

A jellyfish will return to the reverse limit if it goes past the search zone's boundaries as stated in Eq. (7).

$$\begin{cases} X'_{i,d} = (X_{i,d} - U_{b,d}) + L_b(d)\ if\ X_{i,d} > U_{b,d} \\ X'^{i,d}_{,} = (X_{i,d} - L_{b,d}) + U_b(d)\ if\ X_{i,d} < L_{b,d} \end{cases} \quad (7)$$

where, $X_{i,d}$ represents the location of the i_{th} jellyfish in the dth dimension, which is updated following a study of the limit constraints. Fig. 2 presents the flowchart of the JSO.

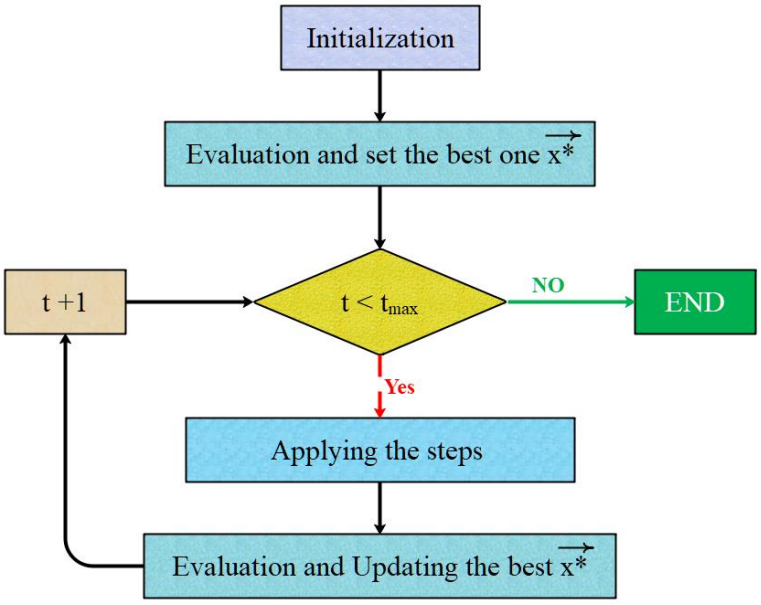

Fig. 2. The JSO's flow chart.

D. Honey Badger Algorithm (HBA)

The properties of the HBA are given in detail in this section [33]. The way honey badgers forage impacted the design of the HBA. The honey badger locates its food primarily via smell, although it also employs digging as a backup strategy. To find and enter the hives, the honey badger depends on honey-guide birds [34]. The first strategy was called the "digging phase," while the second strategy was called the "honey phase," after the people who created the algorithm. Movement is controlled by the honey badger's sensitivity to smell; a strong fragrance will cause it to move more quickly, and vice versa [33]. The following are the primary phases of the HBA and the associated equations:

The issue space's upper (HU) and lower (HL) limits are used to identify the first possible solution during the initialization procedure [35]. Consequently, the initial solutions are stochastic sets, which can be generated using the subsequent procedure in accordance with Eq. (8) [33].

$$H_i = HL + r_1(1, D) \times (HU - HL), i = 1,2, \dots, N \quad (8)$$

where, N is the number of solution providers (honey badgers), H is the total number of possible solutions, and D is the dimension of the solution.

Position updates: At this stage, the candidates' coordinates are updated for H_{new}. This could entail, for example, using a method that employs the digging or honey stages.

Digging phase: The strength of the predator's scent and the distance between the honey badger (agent) and the prey (P) affect the possible search subjects' movements during this phase. The polarized honey badger excavates in a circular region [36]. The stated formula for its motion is as follows:

$$H_{new} = P + Fg \times \beta \times In \times P + Fg \times r_3 \times (P - H_i) \times (cos2\pi r_4) \times (1 - cos2\pi r_5) \quad (9)$$

where, β is the capacity of an insect to gather food. According to [33], there is a maximum value of 6 for β. The r_3, r_4, and r_5 are random variables with a range of 0 to 1, chosen from a uniform distribution, and the intensity is In. The following process yields the Fg, an indication of the search direction:

$$Fg = \begin{cases} 1 & if\ r_6 \le 0.5 \\ -1 & if\ else \end{cases} \quad (10)$$

Honey phase: Honey badgers use the honey phase to move in relation to the honey lead bird when searching for beehives. The study in [33] used the following formula to calculate the honey phase:

$$H_{new} = P + Fg \times r_7 \times \sigma \times (P - H_i) \quad (11)$$

where r_7 is a random number with values ranging from 0 to 1, and P is the best answer found thus far.

Intensity modeling since the honey badger's behavior is determined by its perception of insect scent, [33] developed the next formula for each candidate's scent intensity In_i of the prey.

$$In_i = r_2 \times \frac{(H_i - H_{i+1})^2}{4\pi(P - H_i)} \quad (12)$$

where, r_2 is a random value in the interval [0, 1] and P represents the prey's location.

Modeling density parameter ($sigma$): Hashim et al. state that the sigma value controls transmission between the local and global search phases [33]. According to the hypothesis put forth by Hashim et al. [33], beta is represented across the iterations as follows:

$$\sigma = C \times \exp(\frac{-IT}{IT_{max}}) \quad (13)$$

where, IT and IT_{max} stand for total iterations and current iterations, respectively. It was suggested that the value of the constant C have a value of 2. Fig. 3 presents the flowchart of the HBA.

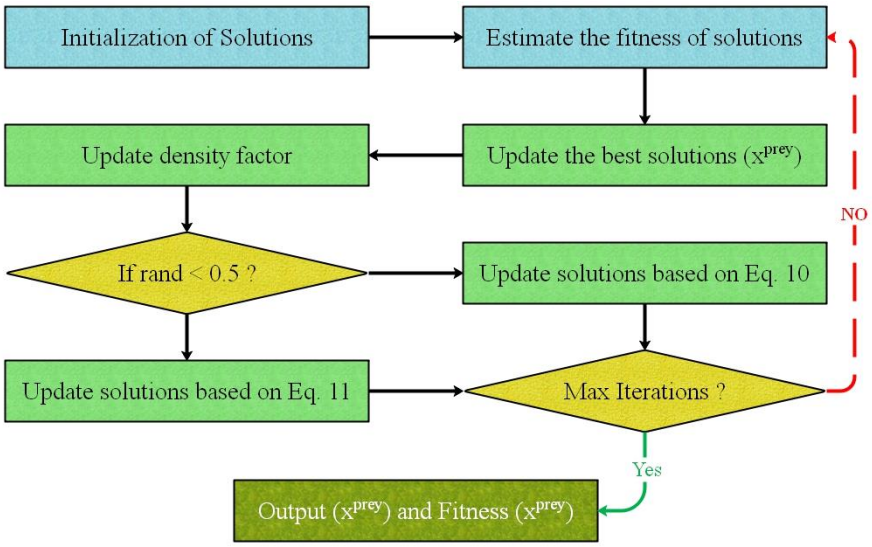

Fig. 3. The flowchart of the HBA.

E. Performance Evaluators

Various measures are outlined in this section to evaluate the performance of hybrid models, including correlations and error levels. Mean Square Error (MSE), Root Mean Square Error ($RMSE$), U95, Prediction Interval (PI), and Coefficient Correlation (R^2) are among the metrics that are being examined. Below is a list of the matching formulas for each of these measurements.

$$R^2 = \left(\frac{\sum_{i=1}^{n}(b_i - \bar{b})(m_i - \bar{m})}{\sqrt{[\sum_{i=1}^{n}(b_i - \bar{b})^2][\sum_{i=1}^{n}(m_i - \bar{m})^2]}} \right)^2 \quad (14)$$

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

$$U95 = \frac{1.96}{n}\sqrt{\sum_{i=1}^{n}(m_i - b_i)^2 + \sum_{j=1}^{n}(m_j - b_j)^2} \quad (16)$$

$$\text{MSE} = \frac{1}{n}\sum_{j=1}^{n}(m_i - b_i)^2 \quad (17)$$

$$PI = \pm t \times SE \times \sqrt{(1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\Sigma(x_i - \bar{x})^2})} \quad (18)$$

Alternatively, the variables can be represented in the following manner:

- The sample size is denoted by n.
- The predicted value is represented by b_i.
- $\bar{m}$ and $\bar{b}$, respectively stand for the measured and mean predicted values.
- The measured value is denoted by m_i.
- The critical value from the $t-$distribution is based on the desired level of confidence and the degrees of freedom denoted by t.
- SE is the Standard Error of the Estimate, a measure of the variability of the model's predictions.
- The value of the predictor variable for which the prediction is being made is represented by x^*.
- The mean of the predictor variable in the dataset is represented by $\bar{x}$.

III. Result and Discussion

In this segment, the outcomes from the created models are examined and compared, employing visual representations to gauge their accuracy and precision. The evaluation of the hybrid DTR+HBA (DTHB), DTR+JSO (DTJS), and DTR single-mode models took place across three sections: training, validation, and testing.

A. Convergence Curve

The convergence curve in Fig. 4 graphically depicts the evolution of an iterative optimization method over time. It shows how the algorithm's objective function value changes with each iteration, showing whether it is approaching the optimal answer. In the context of optimization problems, an algorithm approaches convergence when it continuously minimizes or maximizes the objective function until it reaches a stage where further iterations only yield small improvements.

As evident from Fig. 4, the DTJS model achieved optimal performance significantly faster than the DTHB model. The DTJS model exhibited a steady decline in error rate from the outset, reaching an optimal level with minimal error. In contrast, the DTHB model commenced with a high error rate and remained consistently elevated throughout training.

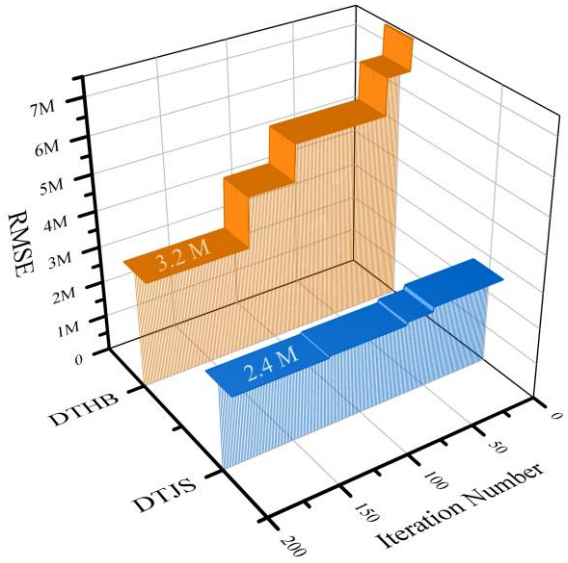

Fig. 4. Convergence curve of hybrid models.

B. Models Comparison

Table I displays the results of the developed models that are on display. Five distinct metric values and three distinct sections have been used to compare the models. Train, Validation, and Test comprise the sections. The metric values are U95, PI, R^2, RMSE, and MSE. The best-performing model is indicated by values in RMSE, MSE, U95, and PI that approach zero, while the highest-performing model is indicated by values in R^2 value that approach one. For instance, the DTJS model performs better than the other two models in the Train segment at the RMSE metric value, while the DT model is the poorest.

The DTJS model also performs flawlessly in the validation stage. The DTJS model performs best in the test section at R^2 value, while the DTHB model is the second-best model. The DTHB model is the weakest in the validation stage at the MSE value. The DTHB model is the second-best in the Test section based on the U95 value. The DTJS model has the best performance in the Train section's PI value.

TABLE I. THE OUTCOME OF THE SHOWCASED DEVELOPED MODELS

Section	Model	Metric values				
		RMSE	***R^2***	***MSE***	***U95***	***PI***
Train	DT	4E+06	0.958	1.58E+13	1.10E+07	0.104
	DTHB	3E+06	0.971	1.13E+13	9.30E+06	0.087
	DTJS	2E+06	0.984	6.11E+12	6.85E+06	0.064
Validation	DT	3E+06	0.956	7.89E+12	7.78E+06	0.097
	DTHB	3E+06	0.965	8.39E+12	7.45E+06	0.100
	DTJS	2E+06	0.973	4.85E+12	6.09E+06	0.076
Test	DT	4E+06	0.924	1.43E+13	1.04E+07	0.184
	DTHB	3E+06	0.954	1.04E+13	8.31E+06	0.155
	DTJS	2E+06	0.971	5.36E+12	6.10E+06	0.111
All	DT	4E+06	0.955	1.44E+13	1.05E+07	0.111
	DTHB	3E+06	0.969	1.07E+13	9.02E+06	0.095
	DTJS	2E+06	0.982	5.81E+12	6.67E+06	0.070

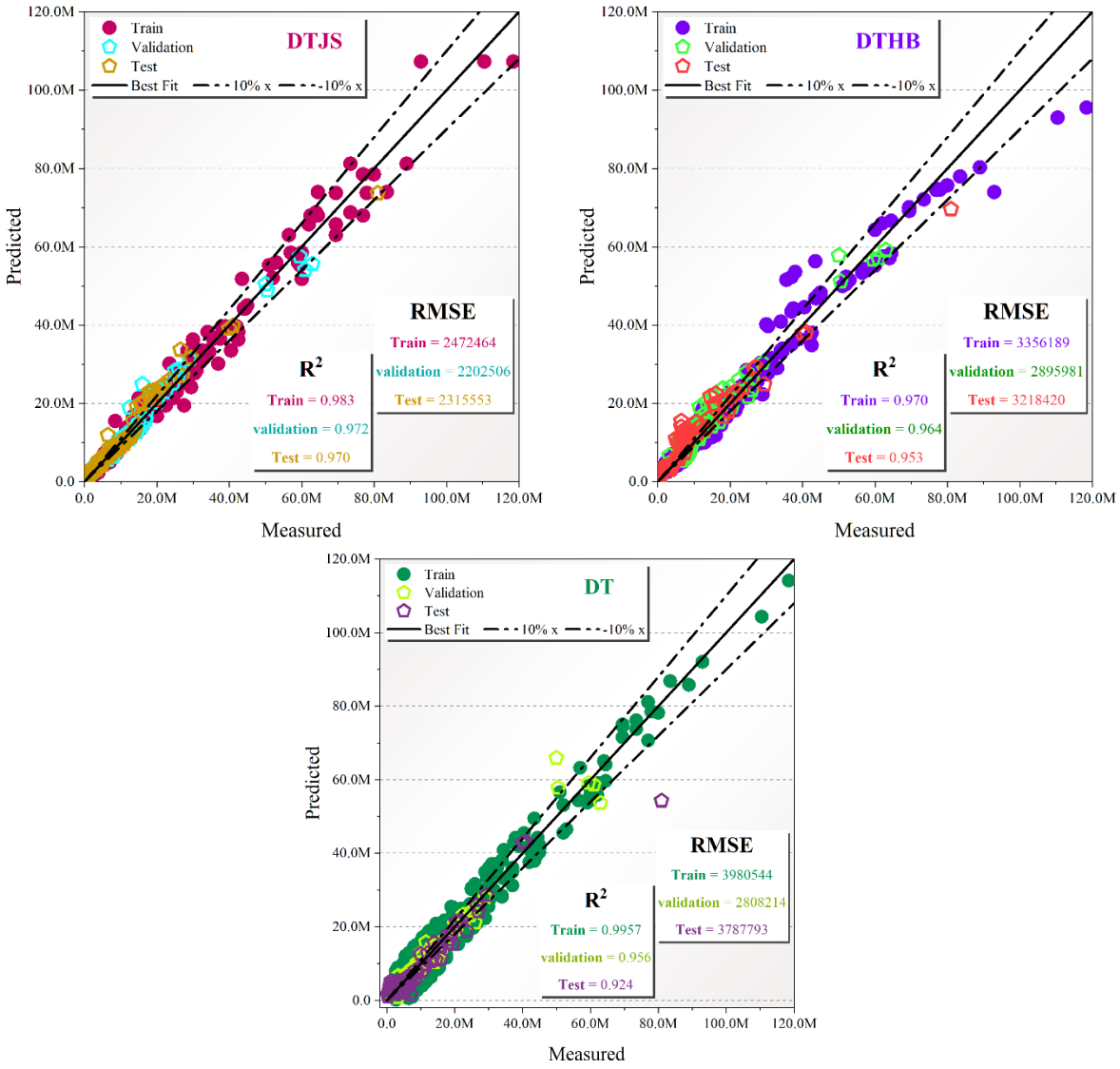

Fig. 5. The scatter plot of the dispersion of evolved hybrid models.

The scatter plot of the dispersion of evolving hybrid models is displayed in Fig. 5. The Y-axis indicates the anticipated value, and the X-axis displays the measured value. The population surrounding the central line, representing the R^2 value, fills up to show that the model with the best performance is in the center. The portions are color-coded for clarity, as shown in Fig. 5. An underestimation is shown when the population is below the center line, and an overestimation is indicated when the population is above the center line. The strong performance of the model is shown if the linear line is in close alignment with the center line and does not exhibit any discernible angle between them. It is clear from Fig. 5 comparison of the diagrams for these three models that the DTJS model performs flawlessly in contrast to the DTHB and DT models.

A comparison of the measured and predicted values is presented in Fig. 6. A visual representation of the model's prediction accuracy is provided by the congruence between the measured circle and the forecasted line. A high degree of accuracy is shown by a close fit between the measured circle and the anticipated line; deviations, on the other hand, point to a lower level of performance. For instance, the DTJS model scored well in the Train part since only a small percentage of the measured circles had a distance greater than the predicted line. The validation component of this model performs much better than the test section. The forecast of the DTJS model closely matches the observed data.

This model performed worse because fewer predicted lines had a distance with measured circles in the Train part of the DTHB model than the DTJS model. Both in the test and validation sections, the DTHB model performs admirably. When compared to the DTJS and DTHB models, the DT model performs the worst in the test segment. There is too much space between the measured circle and the anticipated line. However, this model performs satisfactorily in the Validation and Train part.

The error percentage of the models based on the column plot is displayed in Fig. 7. When the error rate is close to zero, the model is performing admirably. For example, in the DTJS model, the maximum error rate in the Train portion is 80%, although the error rate varies between (-40) and (80). Compared to the other two models, this one has the lowest maximum error rate (84.37%) and performs the best.

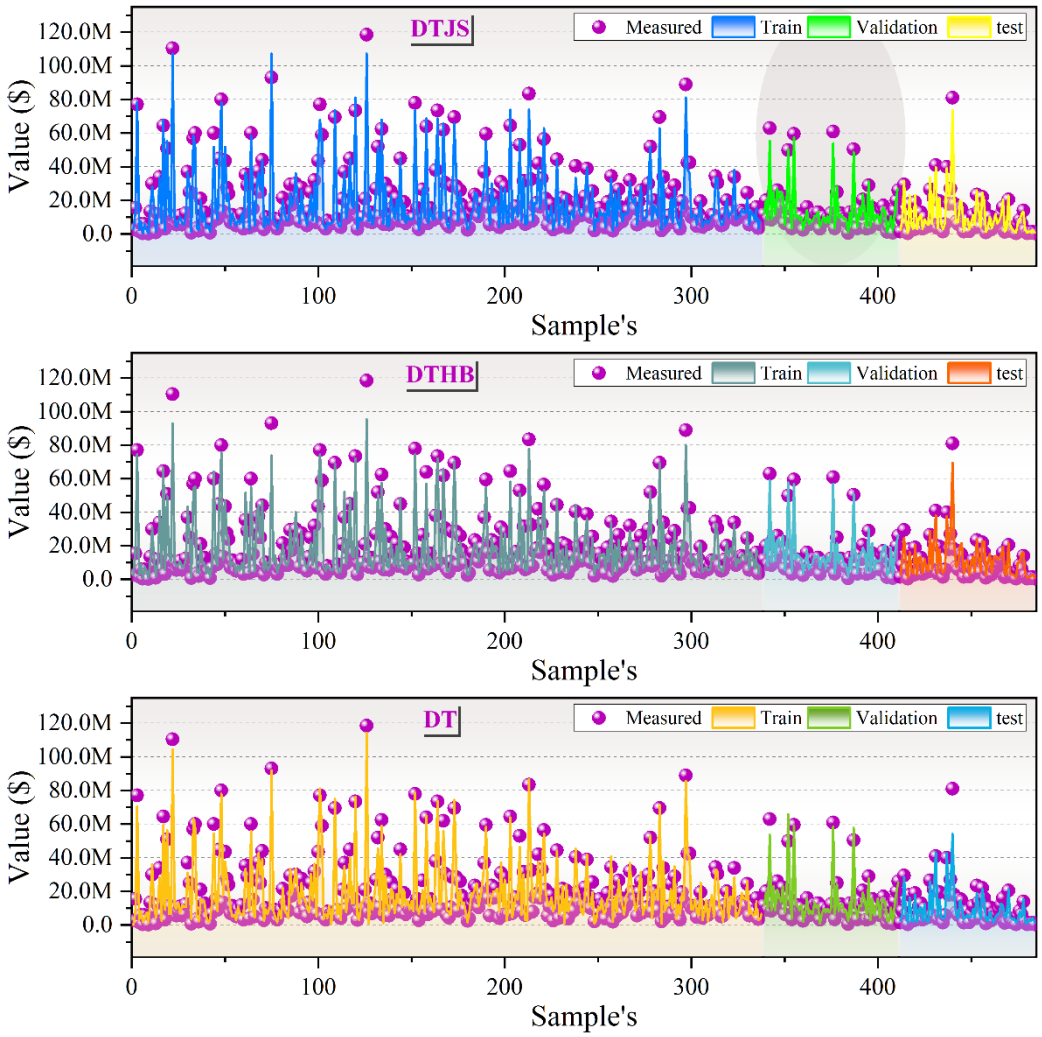

Fig. 6. The comparison of predicted and measured values.

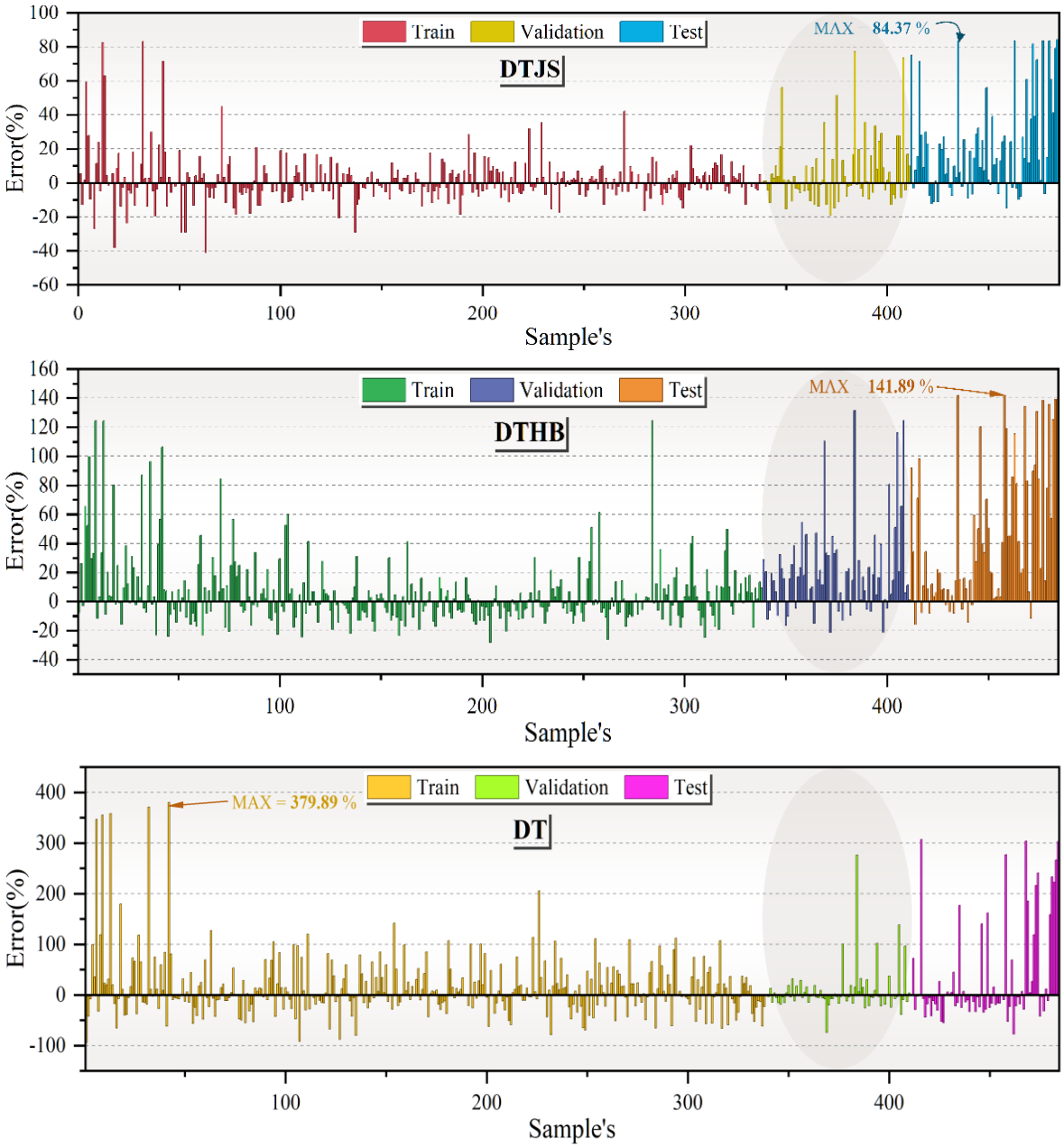

Fig. 7. The error percentage of the models is based on the column plot.

The highest error rate in the DTHB model is 141.89%, which is more than the error rate in the DTJS model. The Test part of the DTHB model contains the maximum error rate. The error rate varies from (-22) to (130) in the validation section. In this investigation, this model performs the second-best. Compared to the DTJS and DTHB models, the maximum error rate in the DT model is 390%, which is the highest mistake rate. The Train section of this model has the highest inaccuracy rate. The Validation part of this model has the lowest error rate.

The suggested models' distribution plot errors are displayed in Fig. 8. The x-axis denotes errors, while the y −axis represents their corresponding frequency. When numbers on the x-axis come close to zero, the model's error rate is diminished. The vertical line that appears exactly above zero signifies that a well-defined, sharp, conical shape, a feature of a normal distribution, emerges as values go closer to zero.

The conical form denotes left skewness if it extends to the left of this vertical line and right skewness if so. A conical shape that is sharper indicates that the model performs better than other models. For instance, it is clear from the Train section that, when compared to the DTHB and the DT model, the DTJS model has the best acute conical shape. Every section of the DTJS model is perfectly shaped like a sharp conical.

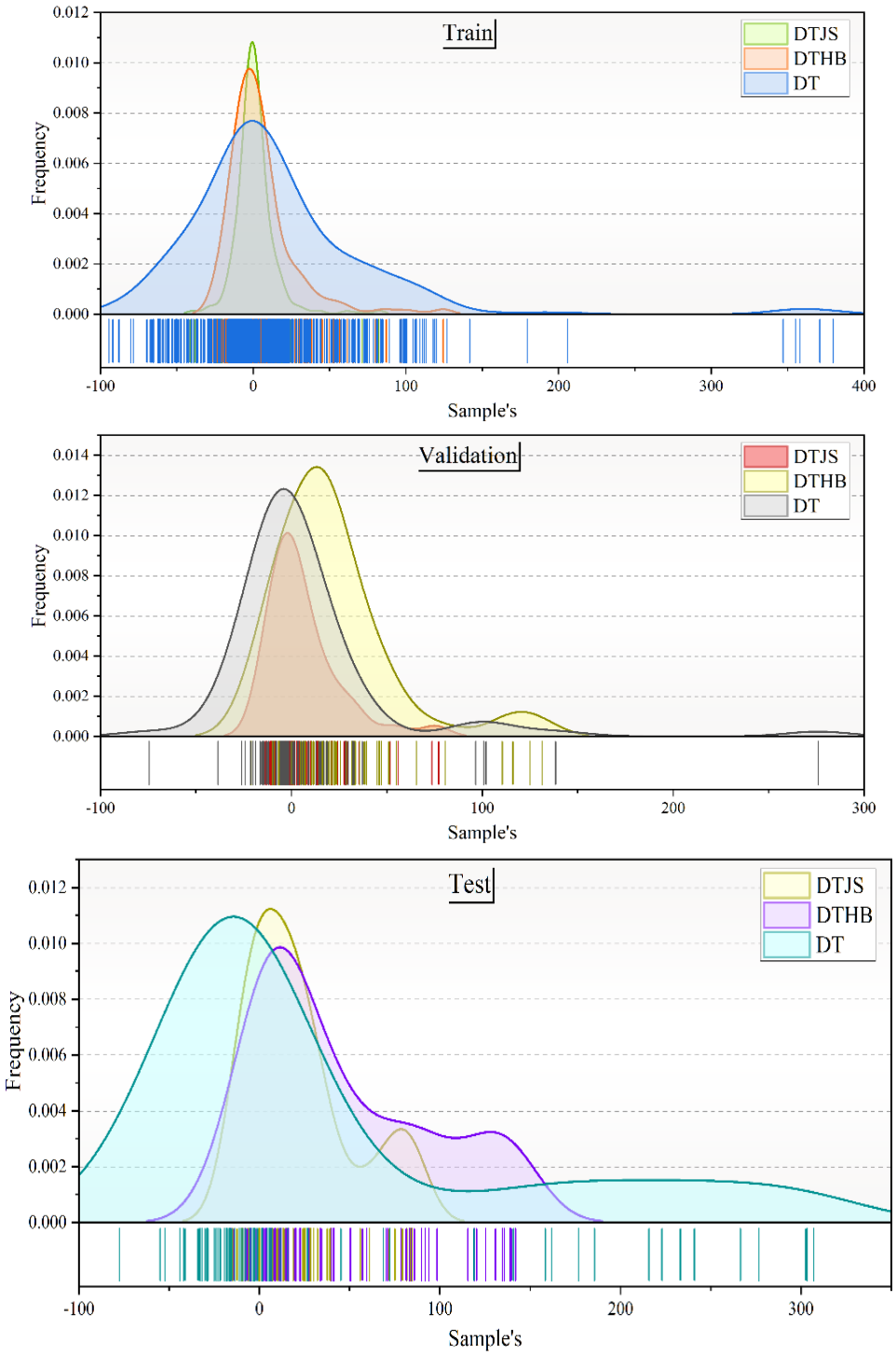

Fig. 8. The distribution plot errors of proposed models.

C. Attributes Analysis

The attribute analysis is shown in Fig. 9, where different inputs show how much of an impact each has on soccer players' market worth. For example, when a soccer player has poor ball control, it has less effect on their market value, but when they have good ball control, their market value is greatly affected. Interestingly, Fig. 9 shows that age is a significant factor that influences a player's market worth in a noticeable way. A player's market worth will decrease if their interception abilities deteriorate. Lower priority qualities add very little to market worth, such as heading accuracy, dribbling, crossing, and leaping. On the other hand, the most significant variables impacting soccer players' market worth are interceptions, age, ball control, and responses.

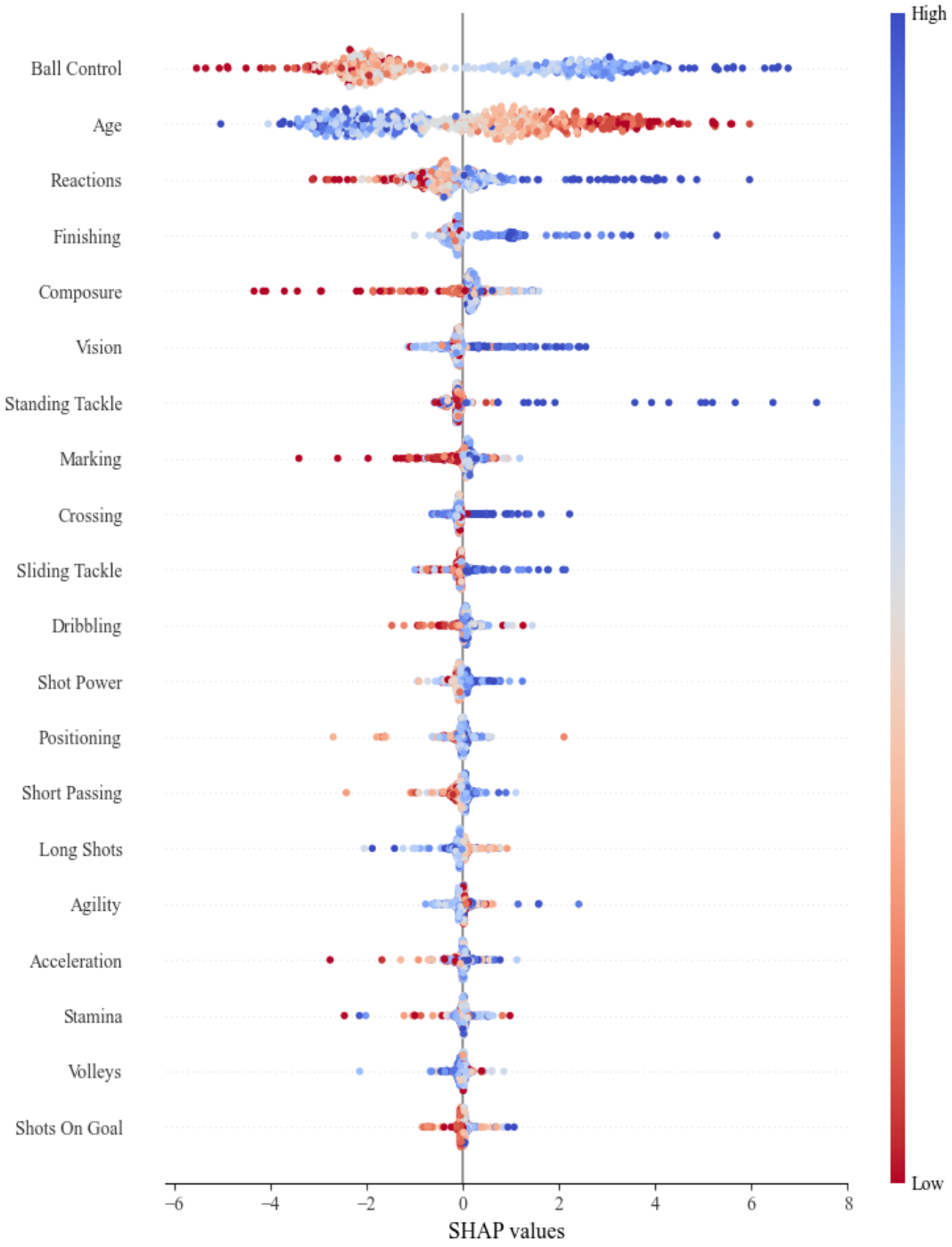

Fig. 9. The SHAP sensitivity analysis of the best models.

D. Comparing with Published Papers

Table II presents a comparative analysis of the best prediction models from various published papers against the model proposed in the present study. The table highlights the R^2 values, which indicate the proportion of variance in the market value of soccer players that is predictable from the models.

The comparative analysis clearly demonstrates that the DTJS model proposed in the present study outperforms the models from the referenced papers in terms of predictive accuracy. The R^2 value of 0.982 indicates that the DTJS model explains a higher proportion of the variance in soccer player market values, underscoring its effectiveness and potential application in real-world scenarios. This highlights the value of incorporating advanced optimization techniques such as JSO into traditional regression models to significantly enhance their performance.

TABLE II. COMPARITIVE ANALYSIS BETWEEN THE PRESENT PAPER AND PUBLISHED PAPERS

Paper	Best prediction model	R^2
[37]	SVR-PSO	0.74
[38]	XGB	0.77
[39]	RFR	0.95
Present Paper	DTJS	0.982

IV. CONCLUSION

This study proposes a novel approach to valuing soccer players using machine learning algorithms. The proposed method, the DTJS hybrid model, effectively combines DTR with the JSO, and the DTHB hybrid model combines the DTR with the HBA metaheuristic algorithms to achieve superior prediction accuracy in estimating player market values. The experimental results on FIFA 20 game data demonstrate the effectiveness of the DTJS hybrid model, outperforming other algorithms in terms of performance evaluators, including RMSE, R^2, MSE, U95, and PI. These findings suggest that machine learning holds the capacity to bring about substantial changes in player valuation within the football league. By providing a more objective and quantitative assessment of player worth, machine learning models can potentially lead to more informed transfer negotiations, enhanced decision-making by football teams and player agents, and a more efficient and transparent transfer market overall. As indicated by the results presented in the study, the R^2 value in the training section of the DTJS model stands at 0.984, surpassing both the DT and DTHB models. The DTJS model emerges as the most effective in this study for predicting the market values of players, showcasing exceptional performance in the prediction task. The study demonstrated advancements in predicting soccer players' market values using the DTJS and DTHB models. However, the dataset was limited to FIFA 20 game data from sofifa.com, which may not capture all real-world complexities. The data represented a specific point in time, so the model's predictions might not remain accurate without regular updates. Additionally, relevant features like psychological factors and team dynamics were not included, potentially affecting prediction accuracy. The models showed high performance on the FIFA 20 dataset, but their applicability to other datasets or real-world scenarios requires further validation. Future research should integrate diverse and real-time data sources, including actual player transfer fees and performance statistics from various leagues. Regularly updating the dataset and retraining the models will help maintain accuracy. Expanding the feature set to include psychological assessments, social media presence, and fan base size could enhance predictive capabilities. Cross-dataset validation will help assess robustness and generalizability. Exploring advanced optimization techniques and machine learning methods, such as deep learning and ensemble learning, could further improve model performance. Addressing these areas can lead to more accurate, reliable, and generalizable models for predicting soccer players' market values, benefiting football clubs, agents, and analysts in their decision-making processes.

ACKNOWLEDGMENTS

Project source: Jilin Provincial Sports Bureau Sports Science Research Project, Project Name: Research on the Construction of Intelligent Sports Park in Changchun City, Jilin Province, Project number: 202325.

REFERENCES

[1] Cotta L, de Melo P, Benevenuto F, Loureiro A. Using fifa soccer video game data for soccer analytics. Workshop on large scale sports analytics, 2016.

[2] Vroonen R, Decroos T, Van Haaren J, Davis J. Predicting the potential of professional soccer players. Proceedings of the 4th workshop on machine learning and data mining for sports analytics, vol. 1971, Springer; 2017, p. 1–10.

[3] Asif R, Zaheer MT, Haque SI, Hasan MA. Football (soccer) analytics: A case study on the availability and limitations of data for football analytics research. International Journal of Computer Science and Information Security 2016;14:516.

[4] Li Y. When Moneyball Meets the Beautiful Game: A Predictive Analytics Approach to Exploring Key Drivers for Soccer Player Valuation 2021.

[5] González-Rodenas J, Moreno-Pérez V, López-Del Campo R, Resta R, Del Coso J. Evolution of tactics in professional soccer: An analysis of team formations from 2012 to 2021 in the Spanish LaLiga. J Hum Kinet 2023;87:207.

[6] Félix LGS, Barbosa CM, Carvalho IA, da F. Vieira V, Xavier CR. Forecasting soccer market tendencies using link prediction. Computational Science and Its Applications–ICCSA 2020: 20th International Conference, Cagliari, Italy, July 1–4, 2020, Proceedings, Part I 20, Springer; 2020, p. 663–75.

[7] 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.

[8] Stanojevic R, Gyarmati L. Towards data-driven football player assessment. 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW), IEEE; 2016, p. 167–72.

[9] Yiğit AT, Samak B, Kaya T. Football player value assessment using machine learning techniques. Intelligent and Fuzzy Techniques in Big Data Analytics and Decision Making: Proceedings of the INFUS 2019 Conference, Istanbul, Turkey, July 23-25, 2019, Springer; 2020, p. 289–97.

[10] 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.

[11] Wicker P, Prinz J, Weimar D, Deutscher C, Upmann T. No Pain, No Gain? Effort and Productivity in Professional Soccer. International Journal of Sport Finance 2013;8.

[12] Majewski S. Identification of factors determining market value of the most valuable football players. Central European Management Journal 2016;24:91–104.

[13] Inan T, Cavas L. Estimation of market values of football players through artificial neural network: a model study from the turkish super league. Applied Artificial Intelligence 2021;35:1022–42.

[14] Lee H, Tama BA, Cha M. Prediction of Football Player Value using Bayesian Ensemble Approach. ArXiv Preprint ArXiv:220613246 2022.

[15] 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.

[16] 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.

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

[18] Frenger M, Emrich E, Geber S, Follert F, Pierdzioch C. The influence of performance parameters on market value. Diskussionspapiere des Europäischen Instituts für Sozioökonomie eV; 2019.

[19] Cotta L, de Melo P, Benevenuto F, Loureiro A. Using fifa soccer video game data for soccer analytics. Workshop on large scale sports analytics, 2016.

[20] Liu G, Luo Y, Schulte O, Kharrat T. Deep soccer analytics: learning an action-value function for evaluating soccer players. Data Min Knowl Discov 2020;34:1531–59.

[21] Shin J, Gasparyan R. A novel way to soccer match prediction. Stanford University: Department of Computer Science 2014.

[22] Rodríguez MS, Ortega Alvarez AM, Arango-Vasquez L. Worldwide trends in the scientific production on soccer players market value, a bibliometric analysis using bibliometrix R-tool. Team Performance Management: An International Journal 2022;28:415–40.

[23] Jana A, Hemalatha S. Football player performance analysis using particle swarm optimization and player value calculation using regression. J Phys Conf Ser, vol. 1911, IOP Publishing; 2021, p. 12011.

[24] 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.

[25] 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.

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

[27] Quinlan JR. Learning with continuous classes. 5th Australian joint conference on artificial intelligence, vol. 92, World Scientific; 1992, p. 343–8.

[28] Wang S, Yao X. Using class imbalance learning for software defect prediction. IEEE Trans Reliab 2013;62:434–43.

[29] Rajpurohit J, Sharma TK. Chaotic active swarm motion in jellyfish search optimizer. International Journal of System Assurance Engineering and Management 2022:1–17.

[30] Alam A, Verma P, Tariq M, Sarwar A, Alamri B, Zahra N, et al. Jellyfish search optimization algorithm for mpp tracking of pv system. Sustainability 2021;13:11736.

[31] Manita G, Zermani A. A modified jellyfish search optimizer with orthogonal learning strategy. Procedia Comput Sci 2021;192:697–708.

[32] Farhat M, Kamel S, Atallah AM, Khan B. Optimal power flow solution based on jellyfish search optimization considering uncertainty of renewable energy sources. IEEE Access 2021;9:100911–33.

[33] Hashim FA, Houssein EH, Hussain K, Mabrouk MS, Al-Atabany W. Honey Badger Algorithm: New metaheuristic algorithm for solving optimization problems. Math Comput Simul 2022;192:84–110.

[34] Düzenli T, Onay FK, Aydemir SB. Improved honey badger algorithms for parameter extraction in photovoltaic models. Optik (Stuttg) 2022;268:169731.

[35] Han E, Ghadimi N. Model identification of proton-exchange membrane fuel cells based on a hybrid convolutional neural network and extreme learning machine optimized by improved honey badger algorithm. Sustainable Energy Technologies and Assessments 2022;52:102005.

[36] Abou El Ela AA, El-Sehiemy RA, Shaheen AM, Shalaby AS, Mouafi MT. Reliability constrained dynamic generation expansion planning using honey badger algorithm. Sci Rep 2023;13:16765.

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

[38] McHale IG, Holmes B. Estimating transfer fees of professional footballers using advanced performance metrics and machine learning. Eur J Oper Res 2023;306:389–99. https://doi.org/https://doi.org/10.1016/j.ejor.2022.06.033.

[39] 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.