Journal of Network Intelligence
Taiwan Ubiquitous Information
 ISSN 2414-8105
Volume **8**, Number **2**, May **2023**

Research on an Intelligent Evaluation System for Used Car Prices Based on Improved Machine Learning Algorithms

Xianning Lin*

School of Information Technology
Guangdong Technology College
Zhaoqing 526100, P.R. China
32028920@qq.com

Salwa Hashim

Faculty of Engineering, Computing and Science
Swinburne University of Technology
Kuching 93350, Malaysia
HashimSiti@yahoo.com

*Corresponding author: Xianning Lin
Received January 8, 2023, revised February 22, 2023, accepted April 7, 2023.

Abstract. *The advent of the big data era offers great opportunities for the development of online trading platforms for used cars. However, problems such as lack of appraisal data, asymmetric appraisal information and high cost of the appraisal process can limit the use of the three traditional used car price appraisal methods. Therefore, it is important to design an accurate and reasonable used car valuation method in the current immature market. As an advanced machine learning model, the Random Forest algorithm has strong advantages in terms of prediction accuracy, missing value handling and noise tolerance. Therefore, this paper introduces the Random Forest model into the study of used car value evaluation and improves the Random Forest model. Firstly, based on the characteristic price theory, 16 indicators were selected as characteristic variables among the factors influencing the value of used cars, and divided into 7 physical variables, 6 functional variables and 3 market variables. Secondly, aiming at the problems of poor training accuracy and relatively poor voting ability, an Inertia Bacterial Foraging Optimization Random Forest algorithm (IBFO-RF) is proposed. A multiplicative weight is set for all decision trees to match their training accuracy, and then the basic parameters of the Random Forest algorithm are optimally selected iteratively using inertia weighted bacterial foraging optimization to improve the output correctness. Finally, the data is imported into the random forest model, and according to the average decreasing degree of mean square error, an intelligent evaluation system for evaluating the value of used cars is established. It was validated against other algorithms on four datasets in the UCI dataset. The experimental results show that the validation of the proposed IBFO-RF algorithm can effectively improve the accuracy of the final classification and has significant advantages in handling the classification problem. The proposed intelligent used car price evaluation system works well with a good fit of 92.21%. The error between the assessed and actual values is mainly within 5%.*
Keywords: Random Forest algorithm; machine learning; decision tree; bacterial foraging optimization algorithm; used car evaluation

1. **Introduction.** In the last decade, China has created a world economic miracle, with an average annual economic growth rate of over 9%. As the level of income generated by consumers has increased, automobiles are becoming an important consumer product for people to purchase. By 2021, 302 million vehicles will have been put into service in China, with sales revenue of $440,823 billion [1]. While new cars are the primary purchasing preference of consumers, there has been a rise in the number of consumers who tend to buy used cars. Used cars are more cost effective, especially for low and middle income people. With the increase in consumer demand for used vehicles, many online trading platforms offer quality testing of used vehicles, valuation and after-sales services for transactions based on quality used vehicle sources [2]. However, the current second-hand trading market still suffers from problems such as an inadequate valuation system, which prevents consumers from making a reasonable evaluation of the value of second-hand goods.

On a macro level, the used car market demand has been growing steadily year by year, and the systematic methods of valuing used cars in the market are mainly divided into the cost method, the market method and the income method [3,4]. However, on a micro level, the used car market lacks a recognised and standardised valuation system. For example, when using the market approach to value, it is necessary to obtain accurate references as well as detailed reference standards. When using the cost approach, it is necessary to rely on the appraiser's knowledge of the overall performance of the vehicle for scoring purposes, which requires a high level of expertise on the part of the appraiser. In addition, the cost method is labour intensive when dealing with large volumes of appraisals. With the growing demand for second-hand transactions, it is vital to find an efficient, fast and accurate means of intelligent appraisal. With the spread of big data as well as intelligent algorithms, algorithms based on machine learning models are widely used in valuation practice. Unlike traditional appraisal methods, computer operation reduces the dependence on the subjective thinking of the appraiser on the one hand, and enables the processing of large volumes of appraisal objects with the help of algorithms on the other, offering greater advantages in terms of saving appraisal labour costs and improving the efficiency of appraisal accuracy. Currently, big data classification is an important part of data mining, which is widely applied to unlabelled data mining under various data platforms [5,6,7]. After various kinds of differentiated structured data enter the data platform, how to extract valuable data in these large-scale data becomes the key to data mining. Machine learning techniques provide an effective means for the realization of big data classification mining and provide effective technical support for the deep development of various industries. Among them, commonly used machine learning techniques include [8,9,10]: Support Vector Machine (SVM), decision tree, AdaBoost algorithm, neural network algorithm, plain Bayesian classification, logistic linear regression, k-Nearest Neighbor (KNN) classification algorithm and Random Forest algorithm Random Forest (RF), etc.

As an advanced machine learning model, the Random Forest algorithm has strong advantages in terms of prediction accuracy, missing value handling and noise tolerance [11,12]. In this paper, we propose to construct a used car valuation system through the Random Forest model to effectively circumvent the various limitations of the market approach, cost approach and income approach in valuation practice, so as to establish a valuation method that is more in line with the current situation of used car transactions. The Random Forest algorithm is a classification model proposed by Breiman in 2001. The essence of Random Forest is a hybrid algorithm combining the Bootstrap Aggregating algorithm [13] and the Random Subspace algorithm [14]. The Random Forest algorithm processes the classification results of multiple decision trees by adopting a voting selection

mechanism to determine the final classification result. Since its introduction, the Random Forest algorithm has been widely used in data mining and classification problems, and many other scholars have later made improvements to the Random Forest model [15,16]. The advantages of the Random Forest algorithm are that it does not require variable selection, has a high noise tolerance, and therefore can omit the tedious work of data pre-processing. Unfortunately, several decision trees with low precision may have similar voting ability due to the vote selection process in the model, which lowers the voting precision. Additionally, the final classification outcomes are frequently greatly influenced by the choice of additional Random Forest model parameters [17,18].

The aim of this work is to assess the value of used cars in a way that avoids the shortcomings of the three traditional valuation methods, reduces the subjectivity of the evaluation, improves the objectivity of the evaluation and makes the assessed price more realistic and scientific. Therefore, in response to the above objectivity requirements, this work proposes a more market-appropriate intelligent appraisal method that not only improves appraisal efficiency, but also reduces appraisal costs. This work uses an improved Random Forest model to determine a system of characteristic variables suitable for the actual situation, thus establishing an intelligent appraisal system for used car values based on a system of variable-related characteristics. The values of the model parameters obtained by analysing the characteristic variable data are used to demonstrate the applicability and accuracy of the model in the used car valuation process.

1.1. **Related Work.** At a time when mathematical models are widely used, more and more scholars have started to apply mathematical models to valuation studies. In terms of valuation models, Tan et al. [19] used the replacement-cost approach to appraise pre-owned pure electric cars and used a weight matrix based on hierarchical analysis to adjust the coefficients of relevant factors to improve the efficiency of used vehicle valuation. Brahimi [20] conducted a detailed study on used vehicle price evaluation using the replacement cost method and eventually proved that the factors affecting the value of used vehicles were mainly related to the fuel economy and service life of the used car.

In order to overcome the subjectivity of the traditional feature price method in the selection of feature variables, some scholars have tried to apply data mining algorithms to used car valuation in recent years, providing new ideas for used vehicle value evaluation. Shi et al. [21] used artificial neural network to build a used car valuation model and analysed the influencing factors of used car prices from both macro and micro perspectives. Twelve data factors were finally selected as input data for the used vehicle price valuation model. Verbelen et al. [22] used factor analysis to relate the value of used vehicles to the characteristic variables and extracted three common factors from 15 characteristic variables, confirming that the physical factors of used vehicles had the greatest influence on their valuation values. All of the above research works have effectively improved the accuracy of used car valuation by applying machine learning algorithms to multi-dimensional sample data and screening the importance of relevant characteristic variables based on the characteristic price method. The above research work also reflects the feasibility of the Random Forest algorithm in used car research. The Random Forest algorithm is a general classification technique based on the basic principle of multiple classifiers. The structure of the Random Forest algorithm is clear and relatively easy to implement, and it is able to obtain a good final output. However, there are still many areas for improvement in typical Random Forest algorithms, and researchers have therefore proposed a variety of corresponding improvement schemes. For example, Chen et al. [23] proposed a Random Forest-based feature selection algorithm that uses both sequential backward selection and generalised sequential backward selection methods for feature selection optimisation.

However, the method still suffers from the problem of decision tree voting processing, resulting in a limited degree of improvement in classification accuracy. To address the above problem, Janizadeh et al. [24] proposed the Conditional InferenceRandom Forest (CIRF) algorithm, which attempts to improve the decision tree voting method by distinguishing strong classifiers from weak classifiers. Although the weighted improvement optimizes the decision tree voting mechanism to a certain extent, the problem of selecting the weight parameters arises again. Therefore, how to choose the optimal weight assignment becomes the key to the weighted Random Forest algorithm.

The decision tree voting mechanism in the traditional RF algorithm has drawbacks. Therefore, in order to obtain relatively suitable or optimal parameters, a Random Forest algorithm (IBFO-RF) based on Inertia Bacterial Foraging Optimization (IBFO) is proposed in this paper to improve the final output correctness. The experimental results show that the proposed BFO-RF model exhibits better overall classification performance compared to the RF and the CIRF.

1.2. **Motivation and contribution.** For those decision trees with suboptimal training accuracy and relatively poor voting ability, this paper basically identifies the reasons for the inadequate performance of the traditional Random Forest algorithm by conducting detailed experiments and analysis of the traditional Random Forest algorithm. It is possible for certain decision trees with poor training accuracy to have the same voting power as others due to the Random Forest vote selection process, which has a substantial influence on the accuracy of the Random Forest's final classification findings. The final categorization findings of the Random Forest are significantly less accurate as a result of this. It is also possible that the highest number of votes in multiple categories may be the same while classifying, making it difficult to classify.

The main innovations and contributions of this study are shown below.

(1) To address the classification difficulties arising from low precision decisions and high vote competition, this work proposes a IBFO based Random Forest algorithm (IBFO-RF). Each decision tree is multiplied by a weight proportional to its training accuracy at the time of voting. To address the problem of difficult parameter selection, the Inertia Weighted Colony Optimisation algorithm is used to iteratively improve the parameters affecting the new model.

(2) To address the problems of high cost, low efficiency and subjectivity of the current traditional manual used car valuation model, this work uses a Random Forest model to design an intelligent valuation system for used cars that can be effectively implemented in practice. The proposed system can quickly and efficiently process a large amount of input information and automatically filter the importance of characteristic variables so as to objectively select those variables that have a significant impact on the value of a used car. The proposed system enables an accurate evaluation of used car values in the presence of insufficient data and imperfect information.

2. Traditional used car valuation methods.

2.1. **The Cost Approach.** The cost approach, the most commonly applied of the traditional valuation methods, is also known as the replacement cost approach. First, the replacement cost of a used vehicle is assessed. Then, the depreciation of the used vehicle is assessed by analysing various physical, functional and economic factors. Secondly, by deducting the depreciated cost of the used car each time, the final method of valuation of the used car's value is determined.

The two main valuation model expressions for the cost approach are as follows.

$$P_e = P_r - P_p - P_f - P_c \tag{1}$$

$$P_e = P_r * r_d \tag{2}$$

Where P_e is the assessed value of the vehicle being valued, P_r represents the replacement cost, P_p is the physical depreciation of the vehicle, P_f is the functional depreciation of the vehicle, P_c is the economic depreciation of the vehicle and r_d represents the vehicle's newness rate.

When determining the replacement cost of a used vehicle, two forms are generally included, namely the restoration replacement cost and the replacement cost of a replacement. The determination of replacement cost is generally based on the current cost of restoring the appraised object, while replacement cost refers to the cost of acquiring a brand new vehicle at current prices using new technology.

2.2. **The market approach.** The principle of the market approach is to select a number of similar or identical vehicles for appraisal, in an open-ended manner, in relation to the various factors that affect the value of a used vehicle. The price of a used vehicle will be assessed by adjusting the reference price to determine its value.

Calculation method of the market approach: In the actual valuation of vehicles, it is not possible to find vehicles on the open market that are identical to those being valued, but only to find references that are similar to those being valued and make price adjustments based on the corresponding differences.

$$P_e = P_0 * (1 \pm k) \tag{3}$$

where P_0 is the reference transaction price and k is the adjustment factor.

2.3. **The income approach.** The present value of earnings method involves estimating the expected future earnings of the asset being valued and discounting them to their present value at a certain discount rate.

If a used vehicle has a small expected return, the appraised value of the vehicle will be low, and conversely the appraised value of the vehicle will be higher. The appraised value of the vehicle being valued is equal to the sum of the present value of the benefits over the remaining life of the vehicle and the basic formula is

$$P_e = \sum_{t=1}^{n} \frac{A_t}{(1+i)^t} = \frac{A_1}{(1+i)} + \frac{A_2}{(1+i)^2} + \ldots + \frac{A_n}{(1+i)^n} \tag{4}$$

Where A_t refers to the expected return in the tth future return period. n refers to the number of years of return. i is the discount rate (which generally includes the risk-free rate, the risk-reward rate and the inflation rate). t is the return period of the subject of the appraisal, i.e. the time over which the return can be sustained.

The replacement cost method is applicable to most used car valuations. However, due to a certain degree of information asymmetry in the used trade market, consumers do not have accurate access to the precise information required by the market approach. The income approach, on the other hand, is mainly applied to for-profit vehicle valuations. The calculation of the newness rate in the cost approach is also subject to the subjective factors of the appraiser. As a result, all three of these valuation methods present certain difficulties in practice.

With the development of artificial intelligence technology, machine learning, the most mainstream method of implementing artificial intelligence today, covers probability theory, statistics, approximation theory and knowledge of complex algorithms. Machine learning allows the use of computer tools to simulate human learning styles in real time. The Random Forest algorithm, as an integrated learning technique, has strong advantages in terms of prediction accuracy, missing value handling and noise tolerance. The Random Forest algorithm is not only able to process multi-dimensional and complex used car data,

but is also able to objectively select the factors affecting the value of used cars, overcoming the shortcomings of traditional valuation methods to a certain extent.

3. **Theory related to Random Forests.**

3.1. **Introduction to the decision tree generation algorithm.** The training process of decision trees does not require any prior knowledge of the samples and relies only on the multidimensional feature data of the samples to complete the training.
A decision tree is more similar in shape to a multinomial tree [25] and contains three main parts - a root node, an intermediate node and a leaf node. Any of the leaf nodes represents a sample attribution category. Intermediate nodes represent different attributes, and when bifurcating downwards from an intermediate node, it is essentially classifying its corresponding attribute. Each time a sample passes through an intermediate node, it is classified according to the attribute corresponding to that intermediate node, and the number of attributes is reduced by one. The root node, on the other hand, contains all the attributes, and each sample entering the decision tree needs to be classified one step down through the root node and eventually reach the leaf nodes. The basic structure of a decision tree is shown in Figure 1.

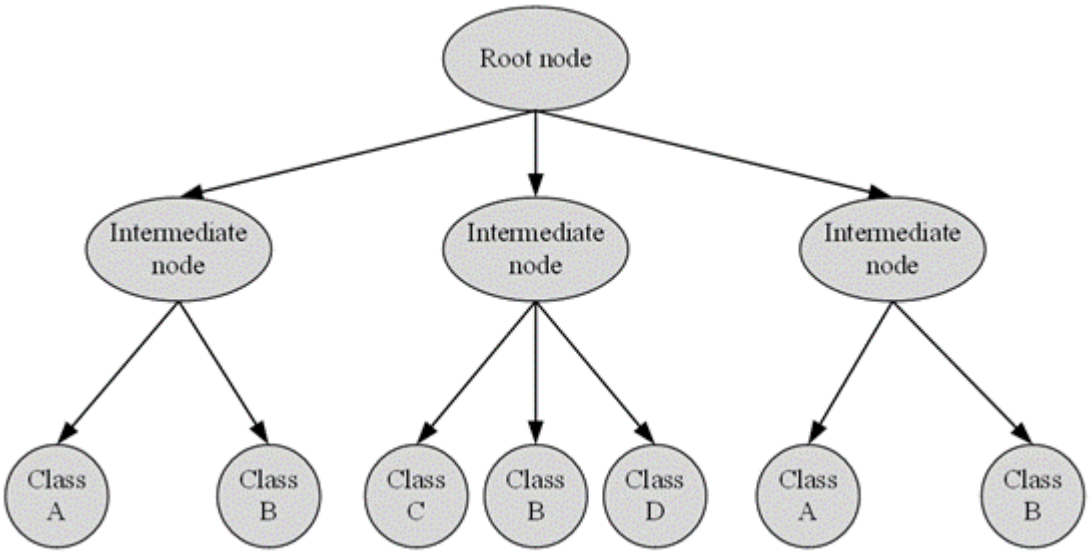

Figure 1. Basic structure of a decision tree.

Decision trees are generally generated using a recursive call approach, working from the top down. The most typical current decision tree classification algorithms are the ID3 and C4.5 algorithms. In the C4.5 algorithms, let S and M denote the number of training data sets and sub-data sets respectively, the "split information" for the attribute A is shown as follows:

$$SplitInfo_A(S) = -\sum_{j=1}^{m} \frac{|S_j|}{} \log_2 \frac{|S_j|}{|S|} \tag{5}$$

Where $|S_j|$ is the number of samples in the jth sub-dataset and $|S|$ is the total number of samples in the dataset before the division. Assuming that $E()$ denotes the information entropy calculation, the information gain of the sample set can be shown as follow:

$$InfoGain(S, A) = E(S) - E_A(S) \tag{6}$$

Where $E_A()$ represents the information entropy of A. The information gain rate of the sample set can be shown as follow:

$$InfoGainRation(S, A) = \frac{InfoGain(S, A)}{SplitInfo(S)} \tag{7}$$

C4.5 algorithm uses the commonly used pruning method on subtrees., with the following implementation conditions:

$$ErrorMean \ + \ ErrorSTD \ \geq \ ErrorMean' \tag{8}$$

Where $ErrorMean$ and $ErrorSTD$ are the average number and standard deviation of pruning errors, respectively, and $ErrorMean'$ indicates the average number of pruning errors for that leaf node.

4. **Decision tree pruning algorithm.** In practice, data is often collected with Gaussian white noise, which results in data that is not completely accurate. Therefore, splitting the sample completely to produce a complete decision tree may lead to overfitting, making the decision tree less capable of identifying new samples. To solve this problem, scholars have proposed the idea of pruning the decision tree. The post-pruning algorithm is to set certain thresholds after the complete generation of the decision tree, and then prune the subtrees that do not satisfy the threshold conditions and define them as leaf nodes. Using this method, it is usually possible to avoid getting trapped in a local optimum solution. There are three common post-pruning algorithms as follows [26].

(1) REP pruning method.

The REP pruning method is a bottom-up pruning method. A portion of the training sample is taken as a separate pruning set and no training is performed to generate the algorithm. The parent nodes of all leaf nodes are then cyclically determined. The decision tree is continually pruned until the error rate of change exceeds a threshold.

(2) PEP pruning method.

The PEP pruning method also calculates the classification error rate, but differs in that the pruning process is top-down. Unlike REP, PEP does not extract a separate pruning set, but directly replaces it with training samples. Starting from the root node and traversing down to each node, the node is defined directly as a leaf node if the following conditions are met

$$E(Node) \leq E\,(Tree) + SE\,(Tree) \tag{9}$$

Where *E(Node)* represents the number of classification errors when the currently visited node is a leaf node, *E(Tree)* represents the number of classification errors when the currently visited node is a subtree, and *SE(Tree)* represents the standard error.

$$SE(X) = \sqrt{\frac{X \times (N - X)}{N}} \tag{10}$$

Where X represents the number of classification errors in that subtree and N represents the total number of samples within that subtree.

Compared to REP, PEP has a faster running speed because the top-down process avoids repeatedly judging all subtrees within the prunable subtree. However, this PEP pruning method also has the potential to fall into a local optimum solution.

(3) MEP pruning method.

The MEP pruning method also compares the error rate of each node with its subtree from the bottom up.

$$STE(Node) - DYE(Tree) \leq \varepsilon \tag{11}$$

Where *STE()* represents the static error rate of the intermediate node, *DYE()* represents the dynamic error rate of the corresponding subtree of the intermediate node, and ϵ is a freely set threshold to guarantee the correctness of the decision tree.
The decision tree algorithm in this work is mainly applied to build out a Random Forest, as the voting strategy of the Random Forest will further improve the correct classification rate, so a single decision tree does not require a particularly high training accuracy. Since the Random Forest has a large number of decision trees, the training time is a key parameter. For this reason, the MEP pruning algorithm is chosen as the pruning algorithm for decision trees in this paper.

4.1. **Random Forest Algorithm.** The weak classifier used in the Random Forest model is the decision tree, which is essentially a modified version of the Bagging algorithm.
The Random Forest algorithm uses a no-relaxation extraction operation for attributes in the random subspace algorithm, which improves global search capabilities and therefore better classification performance. Any one decision tree of Random Forest is different and can be applied to different types of samples. For different samples, the Random Forest algorithm is more adaptable than Bagging and random subspace algorithms alone.
All weak classifiers will be randomly selected N times. The probability that a given sample is not drawn any of the N *times* is P.

$$P = \left(1 - \frac{1}{N}\right)^N \tag{12}$$

Where N is the total number of training samples, all weak classifiers have their own training samples. Instead of using all training samples, sampling training is used to reduce the probability of incorrect sample extraction due to noise. The weights determined by this method have some impact on the fairness of the voting when they are stacked.
In addition, another major component of constructing a Random Forest is the random subspace algorithm, which mainly accomplishes the selection of the number of features m selected at random each time, a value usually chosen empirically and calculated as follows [27].

$$m = \lfloor \log_2(M+1) \rfloor \tag{13}$$

The Random Forest model's voting technique involves combining the outputs of many weak classifiers. The classification result of the Random Forest algorithm for the sample to be tested during the voting process is $f_{RF}(x)$.

$$f_{RF}(x) = \underset{i=1,2,\ldots,c}{\arg\max} \{I\left(f_l(x) = i\right)\} \tag{14}$$

Where i denotes one of all types, *I()* denotes the number of expressions satisfying the brackets, c is the number of types in the entire Random Forest, and $f_l(x) = i$ denotes that the output of the lth decision tree is i.

5. An intelligent system for valuing used cars based on improved weighted Random Forest.

5.1. **Disadvantages of the traditional RF.** The traditional RF model not only solves the drawbacks in the original decision tree classification algorithm, but also improves the accuracy of the classification. However, there are two problems with the decision tree voting mechanism in the traditional RF algorithm.
Problem 1: In the final voting strategy of the traditional Random Forest algorithm, each decision tree can cast a vote for its own classification result. Therefore, this work introduces the principle of 'winner-take-all' by assigning a larger weight to trees that grow well and a smaller weight to trees that do not grow well, thus allowing them to be

weighted fairly for voting. This would solve the previous problem to some extent.
Problem 2: The traditional voting process does not take into account the occurrence of tie votes. This is because decision trees that are not trained with sufficient accuracy will cast the wrong number of votes. The reason for this problem is that all decision trees have the same weight when voting.

Table 1. All feature variables.

Variable type	Variable name
Entity variables	Brand, product range, class, number of seats, date of registration, mileage, condition damage
Functional variables	Power, gearbox type, drive type, fuel consumption type, safety features, engine displacement
Market Variables	New car prices, emission standards, number of transmissions

5.2. **Weighting improvements in voting methods.** As mentioned above, the traditional RF algorithm performs with some decision trees that classify well and others that may be relatively poor.
Therefore, the classification ability is used on this paper to set the weights that match it in order to solve the problem that the classification ability of each decision tree is different, which requires the calculation of the correct classification rate of the first l decision tree w_l.

$$w_l = \frac{X_l^{correct}}{X_L} \tag{15}$$

Where X_L indicates the number of Pre-test samples,$X_l^{correct}$ indicates the amount of samples accurately categorised by l-th decision tree, and L indicates the number of decision trees.
Each decision tree is given the same weight when voting, but there is no guarantee that each decision tree will end up with the same classification accuracy. As a result, certain decision trees having relatively low precision will always make the incorrect decisions, thus affecting the classification ability of the whole Random Forest. To solve such problems and reduce the impact of decision trees with relatively low training accuracy on the overall model, therefore, to obtain relatively suitable or optimal parameters (pruning threshold, number of decision trees and Pre-test sample rate), this work proposes an improved weighted Random Forest algorithm based on IBFO (IBFO-RF) to improve the final output correctness, as shown in Figure 2.

Setting w_l as the weights to match the decision tree, the output of the RF algorithm is:

$$f_l^{\text{tree}}(x) = i \tag{16}$$

$$f'_{RF}(x) = \underset{i=1,2,\ldots,c}{\arg\max} \left\{ \sum_{l \in L, f_l^{\text{tree}}(x)=i} w_i \right\} \tag{17}$$

Where x is the sample to be tested in the weighted Random Forest.
To facilitate the optimization of weights, the Pre-test sample rate $X_L^{'}$ is used instead of the number of Pre-test samples in the specific implementation of the weighted RF algorithm, which is calculated as follows:

$$X_L^{'} = \frac{X_L}{\text{Total number of training samples}} \tag{18}$$

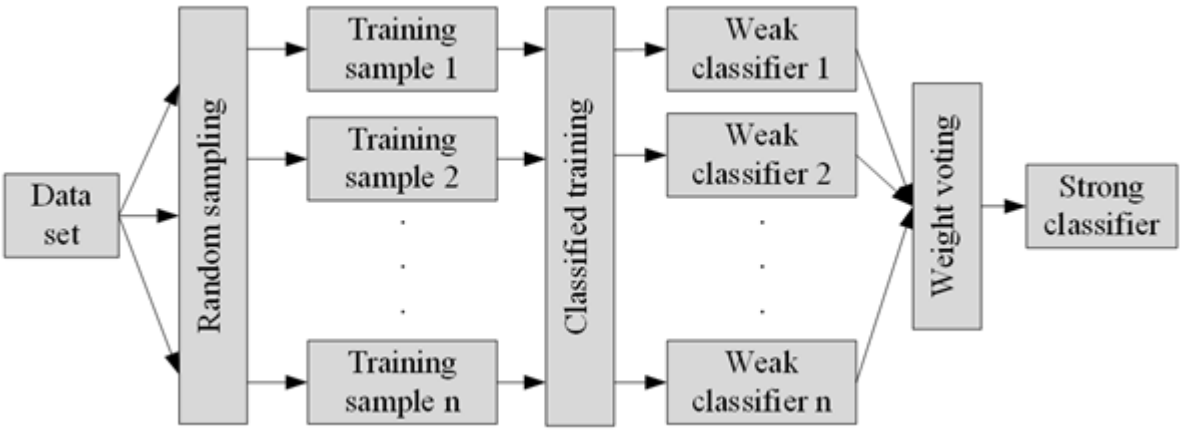

Figure 2. The principle of IBFO-RF.

5.3. **Operation of the Bacterial Foraging Optimization algorithm.** Based on the Darwinian theory of biological evolution, the Bacterial Foraging Optimization (BFO) algorithm [28] is an intelligent bionic optimization algorithm that mimics the instinctive foraging search behaviour of an E. coli population.

BFO is able to solve efficient optimisation problems in a number of real-world applications. In the BFO algorithm, let the population size of bacteria be S and $\theta^i(j,k,l)$ denote the position vector of bacteria i after multiple operations, the operations of the BFO algorithm are divided into 4 types.

(1) Convergent operation.

The location of the bacteria after this procedure is updated as follow:

$$\theta^i(j+,k,l) = \theta^i(j,k,l) + C(i)\mathbf{\Phi}(j) \tag{19}$$

Where $\mathbf{\Phi}(j)$ is the unit random direction vector when the bacteria are flipped, j is the number of copy operations, k is the number of migration operations, l is the number of convergent operations, and $C(i)$ is the unit step size of the convergent operations.

(2) Aggregate operations.

The mathematical expression for the aggregative operation is shown as follow [29]:

$$\begin{aligned} J_{cc}(\theta, P(j,k,l)) &= \sum_{i=1}^{s} J_{cc}^{i}\left(\theta, \theta^i(j,k,l)\right) \\ &= \sum_{i=1}^{s}\left[-h_{\text{attart}}\ \exp\left(-\omega_{\text{attarct}}\ \sum_{m=1}^{P}\left(\theta_m - \theta_m^i\right)^2\right)\right] + \sum_{i=1}^{s}\left[h_{\text{repellant}}\ \exp\left(-\omega_{\text{repellant}}\ \sum_{m=1}^{P}\left(\theta_m - \theta_m^i\right)^2\right)\right. \end{aligned} \tag{20}$$

Where $\omega_{\text{repellant}}$ and ω_{attarct} represent the repulsive width and gravitational width values respectively, and $h_{\text{repellant}}$ and h_{attart} represent the repulsive height and gravitational height values respectively.

(3) Copy operations.

In the replication operation, the bacteria are eliminated and the number of bacteria eliminated is:

$$S_r = \frac{S}{2} \tag{21}$$

(4) Migration operations.
The migration operation occurs with a certain probability P and thus determines whether a bacterium dies or produces a new individual at a random location.

5.4. **Convergence step improvement for inertia weights.** In the BFO algorithm is more sensitive to the unit step size *C(i)* parameter of the convergence operation, which directly determines the strength of its local search and global search capability.
In the traditional BFO algorithm, the parameter *C(i)* is fixed [30], which is not conducive to a balance between local search for excellence and global search capability, so it is improved by using the inertial weight factor approach. The proposed IBFO algorithm uses a linearly decreasing convergence step, which balances between local search and global search capabilities while reducing the complexity of the algorithm.

$$C(i,j) = C_{MAX} - \frac{(C_{MAX} - C_{MIN})}{N_c} \times j \tag{22}$$

Where j and N_c denote the current and maximum number of iterations of the convergence operation, respectively, and C_{MAX} and C_{MIN} denote the maximum and minimum values of the convergence step, respectively. Taking $C_{MAX} = 0.4, C_{MIN} = 0.1$ and $N_c = 100$ as examples, the trend of inertial convergence step is shown in Figure 3.

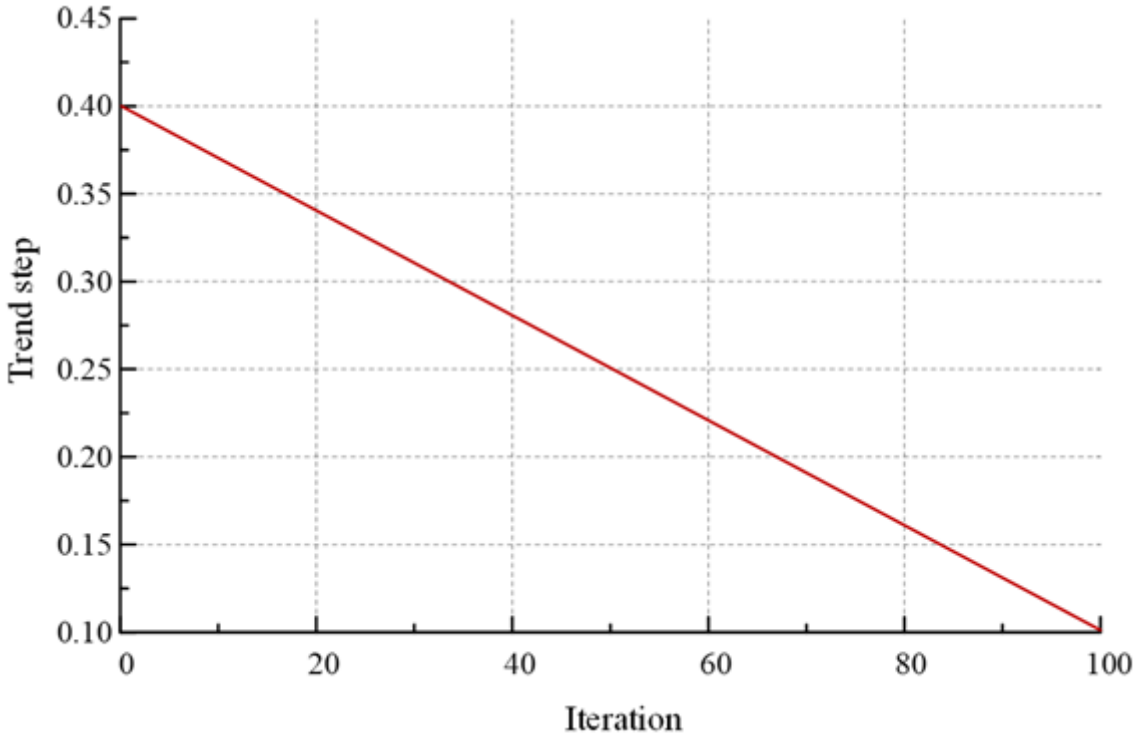

Figure 3. Trend of inertial trending step size.

5.5. **Optimal parameter selection.** To obtain relatively suitable or optimal parameters, this work uses three variables to form a space vector as a bacterium in the IBFO algorithm.
This work takes the pruning threshold, the number of decision trees and the Pre-test sample rate and forms a spatial vector (ε, L, X'_L) as a bacterium in the IBFO algorithm. The IBFO iterative process is used to search for the best bacteria within the range of

values taken in order to arrive at the optimal parameters. the fitness function for the IBFO optimisation operation is formulated as

$$J(i, j+1, k, l) = J(i, j, k, l) + J_{cc}(\theta, P) \tag{23}$$

Where $J(i, j, k, l)$ denotes the value of the fitness function for bacteria i at the time of the pre-trending operation.

6. **Experimental results and analysis.**

6.1. **Experimental environment.** A comparative analysis with the RF algorithm and the CIRF algorithm was carried out on four datasets in the UCI database, and several metrics were used to validate the performance of the proposed classification algorithm. Quantitative evaluation was carried out through 3 metrics: accuracy, integrated F1, and kappa coefficient of consistency test. The simulation environment for all algorithms in this work is Matlab 7.0. The experimental platform is a Windows 7 64-bit operating system with an Inter(R) Core(TM) I7-4790K @2.4GHz 2.39GHz CPU and 8GB RAM.

6.2. **Experimental datasets.** The relevant parameters of the four datasets from the UCI database used for the experiments are shown in Table 2.
The relevant parameter settings for the IBFO-RF algorithm implementation are shown in Table 3. The number of training samples was all approximately two-thirds of the number of test samples. Each algorithm was run nearly 2000 times on each dataset.

Table 2. Experimental data set.

Serial number	**Data sets**	**Number of samples**	**Number of attributes**	**Number of categories**
1	Balance-scale	625	4	3
2	Car Evaluation	1728	6	4
3	Wine Quality	178	13	3
4	Yeast	101	16	7

Table 3. Experimental parameters of IBFO-RF

Parameters	**Numerical values**
C_{MAX}	0.4
C_{MIN}	0.1
L	100
N_c	100

6.3. **Evaluation metrics.** In this paper, a quantitative evaluation was performed by three metrics: accuracy, composite F1, and kappa coefficient [31].

(1) Accuracy (accuracy) is calculated using the following formula.

$$acc = \frac{1}{n} \sum_{i=1}^{n} I\left(f(x_i) = y_i\right) \tag{24}$$

Where $I(\cdot)$ indicates the number of expressions that satisfy the brackets, n indicates the total number of samples tested, and $f(x_i) = y_i$ indicates the cases where the samples are correctly classified.

(2) The formula for F1-measure is as follows.

$$F_1 = \frac{2P \cdot R}{R + R} \tag{25}$$

Where P is the accuracy rate and R is the completeness rate.

(3) The kappa coefficient is calculated as follows.

$$\kappa = \frac{P_1 - P_2}{1 - P_2} \tag{26}$$

6.4. **Performance analysis of the IBFO-RF algorithm.** The experimental results of the RF algorithm, CIRF algorithm and IBFO-RF algorithm on the four test datasets are shown in Table 4, Table 5 and Table 6 respectively.

Table 4. Comparison of accuracy rate indicators

	RF	**CIRF**	**IBFO-RF**
Balance-scale	0.841	0.8421	0.8518
Car Evaluation	0.9902	0.9915	0.9924
Wine Quality	0.9772	0.9734	0.9871
Yeast	0.9562	0.9601	0.9738

Table 5. Comparison of comprehensive F1 indicators

	RF	**CIRF**	**IBFO-RF**
Balance-scale	0.581	0.5903	0.604
Car Evaluation	0.9686	0.9705	0.9761
Wine Quality	0.9985	0.9991	1.0023
Yeast	0.9074	0.9157	0.9245

Table 6. Comparison of Kappa coefficient indicators

	RF	**CIRF**	**IBFO-RF**
Balance-scale	0.7213	0.7235	0.7402
Car Evaluation	0.9824	0.9717	0.9812
Wine Quality	0.9811	0.9912	1.0051
Yeast	0.9527	0.9627	0.9678

It can be seen that the proposed IBFO-RF algorithm outperforms both the RF algorithm and the CIRF algorithm in three metrics: accuracy, integrated F1, and kappa coefficient. In addition, the accuracy comparison of the 3 algorithms on the 4 data sets is shown in Figure 4, using Table 4 as an example.

It can be seen that the IBFO-RF algorithm performs better in terms of accuracy on any dataset. Similarly, the results of the F1-measure, and the other two metrics of the kappa coefficient are consistent, which means that the IBFO-RF algorithm has the highest classification accuracy on the dataset used in the experiments and obtains more satisfactory results.

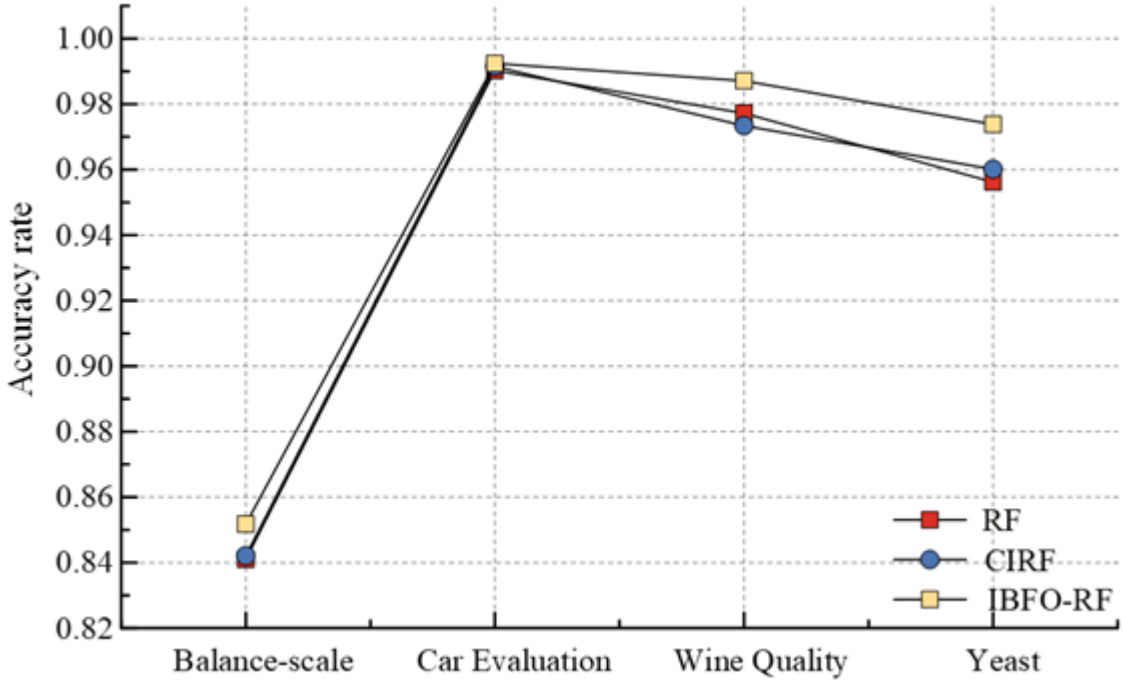

Figure 4. Comparison of the accuracy of the three algorithms on four data sets.

6.5. **Case validation of an intelligent valuation system for used cars.** The used car value evaluation results obtained by running the IBFO-RF algorithm were compared with the actual values. The test set was a sample of 214 used cars.
Four metrics, namely goodness of fit, mean relative error MRE, mean absolute error MAE and root mean square error RMSE, were used to measure the accuracy of the value evaluation, and the results are shown in Table 7. The results are shown in Table 7. It can be seen that the proposed intelligent used car price evaluation system works well, with a goodness-of-fit of 92.21%. The error between the assessed value and the actual value is mainly only 4.19%, which is far below the permitted error range of 20%. The relative

Table 7. Error Analysis of Used Car Intelligent Valuation Systems

R^2	MRE	MAE (million yuan/vehicle)	RMSE
92.21%	4.19%	0.8942	0.9770

errors of the 214 used cars were analysed based on their predicted and actual sold values and the results are shown in Figure 5.

It can be seen that 207 used cars, or 96.7% of the test dataset (214 cars), had an error rate of 20% or less. 122 of the 207 used cars had an error of less than 5%. Of the 214 vehicles in the sample, 7 had an error of more than 20%. The reason for the large error is that the initial data sample was obtained relatively small and therefore anomalies may not have been detected when the data was obtained or processed. In addition, there may be important factors that were missed in this work in addition to the 16 characteristic variables selected. Whichever perspective is used to explain the errors in the evaluation results, it can be confirmed that the IBFO-RF algorithm has some resistance to missing data. When there are missing or abnormal in the initial data, the IBFO-RF based intelligent evaluation system for used cars still works and evaluates well.

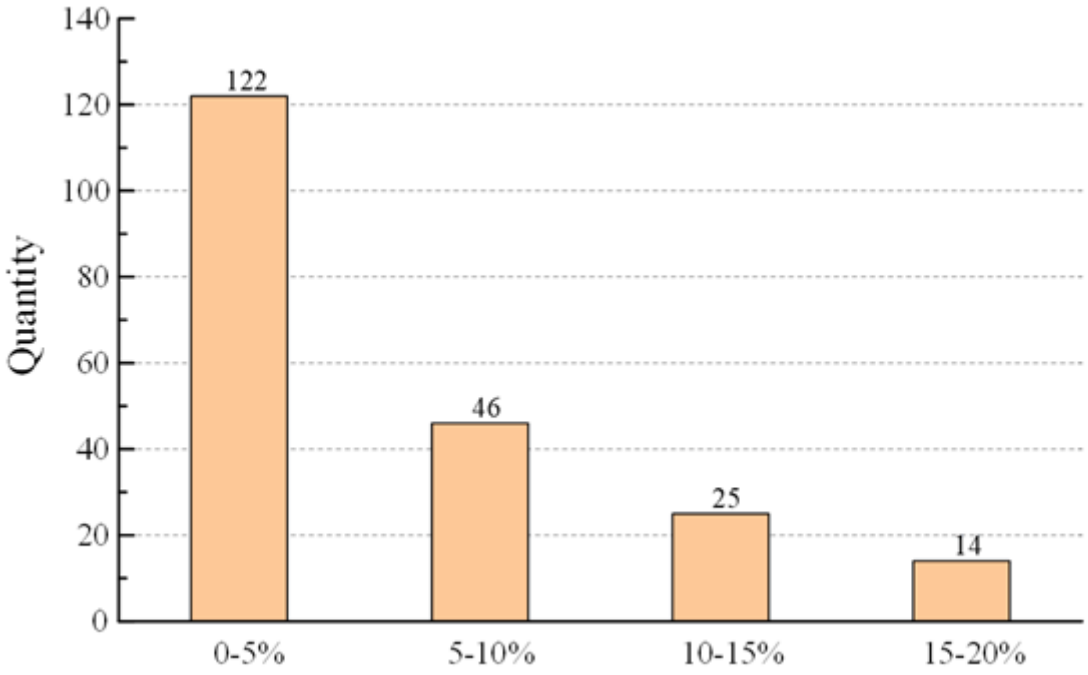

Figure 5. Error histogram of the used car intelligent valuation system.

7. **Conclusion.** This work presents an optimally weighted Random Forest (IBFO-RF) algorithm based on IBFO. The algorithm improves on the traditional voting selection mechanism where decision trees have the same voting power by setting a multiplicative weight for all decision trees that matches their training accuracy. Secondly, the traditional colony optimisation algorithm is improved by proposing a convergence step for the inertia weights, which enables optimal iterative selection of the basic parameters of the Random Forest algorithm. Experiments show that the proposed BFO-RF improves the final output correctness and exhibits better overall classification performance than the RF and CIRF algorithms. 214 used cars were analysed in the case study, showing that the IBFO-RF-based intelligent used car evaluation system achieved a goodness-of-fit of 92.21% and the error between the evaluation value and the actual value was within 5%. This indicates that the proposed system is well suited to the evaluation of a large number of used car values. Subsequent research will be conducted to investigate the diversity of the Random Forest classification algorithm in order to enhance its classification capability in a comprehensive manner.

Funding Statement. This work supported by the project "2021 Guangdong Higher Education Teaching Reform Project, NO.644" and "The First Class Course of Guangdong Institute of Technology in 2021, NO. YLKC202107".

REFERENCES

[1] A. Gavazza, A. Lizzeri, and N. Roketskiy, "A Quantitative Analysis of the Used-Car Market," *American Economic Review*, vol. 104, no. 11, pp. 3668-3700, 2014.

[2] C.-M. Chen, S. Lv, J. Ning, and J. M.-T. Wu, "A Genetic Algorithm for the Waitable Time-Varying Multi-Depot Green Vehicle Routing Problem," *Symmetry*, vol. 15, no. 1, 124, 2023.

[3] T.-Y. Wu, J. Lin, Y. Zhang, and C.-H. Chen, "A Grid-Based Swarm Intelligence Algorithm for Privacy-Preserving Data Mining," *Applied Sciences*, vol. 9, no. 4, 774, 2019.

[4] T.-Y. Wu, J. C.-W. Lin, U. Yun, C.-H. Chen, G. Srivastava, and X. Lv, "An efficient algorithm for fuzzy frequent itemset mining," *Journal of Intelligent & Fuzzy Systems*, vol. 38, no. 5, pp. 5787–5797, 2020.

[5] A. L. H. P. Shaik, M. K. Manoharan, A. K. Pani, R. R. Avala, and C.-M. Chen, "Gaussian Mutation–Spider Monkey Optimization (GM-SMO) Model for Remote Sensing Scene Classification," *Remote Sensing*, vol. 14, no. 24, 6279, 2022.

[6] G. Atluri, A. Karpatne, and V. Kumar, "Spatio-Temporal Data Mining," *ACM Computing Surveys*, vol. 51, no. 4, pp. 1-41, Aug. 2018.

[7] J. Yang, Y. Li, Q. Liu, L. Li, and A. Feng, "Brief introduction of medical database and data mining technology in big data era," *Journal of Evidence-Based Medicine*, vol. 13, no. 1, pp. 57-69, 2020.

[8] A. Bhardwaj and R. Gupta, "Financial Frauds: Data Mining based Detection - A Comprehensive Survey," *International Journal of Computer Applications*, vol. 156, no. 10, pp. 20-28, 2016.

[9] R. Kumar and R. Tripathi, "DBTP2SF: A deep blockchain-based trustworthy privacy-preserving secured framework in industrial internet of things systems," *Transactions on Emerging Telecommunications Technologies*, vol. 5, no. 8, pp. 109-121, 2021.

[10] A. Ferdowsi and W. Saad, "Deep Learning for Signal Authentication and Security in Massive Internet-of-Things Systems," *IEEE Transactions on Communications*, vol. 67, no. 2, pp. 1371-1387, 2019.

[11] J. L. Speiser, M. E. Miller, J. Tooze, and E. Ip, "A comparison of random forest variable selection methods for classification prediction modeling," *Expert Systems with Applications*, vol. 134, pp. 93-101, 2019.

[12] P. Probst, M. N. Wright, and A. Boulesteix, "Hyperparameters and tuning strategies for random forests," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 9, no. 3, pp. 89-102, 2019.

[13] M. Schonlau and R. Y. Zou, "The random forest algorithm for statistical learning," *The Stata Journal: Promoting communications on statistics and Stata*, vol. 20, no. 1, pp. 3-29, 2020.

[14] A. Paul, D. P. Mukherjee, P. Das, A. Gangopadhyay, A. R. Chintha, and S. Kundu, "Improved Random Forest for Classification," *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 4012-4024, 2018.

[15] R. Couronné, P. Probst, and A.-L. Boulesteix, "Random forest versus logistic regression: a large-scale benchmark experiment," *BMC Bioinformatics*, vol. 19, no. 1, pp. 56-73, 2018.

[16] S. Georganos, T. Grippa, A. Niang Gadiaga, and C. Linard, "Geographical random forests: a spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling," *Geocarto International*, vol. 36, no. 2, pp. 121-136, 2019.

[17] C. Iwendi, A. K. Bashir, A. Peshkar, R. Sujatha, and J. M. Chatterjee, "COVID-19 Patient Health Prediction Using Boosted Random Forest Algorithm," *Frontiers in Public Health*, vol. 8, 357, 2020.

[18] P. A. Rajakumari and P. Parwekar, "Boosting Blockchain Mechanism Using Cryptographic Algorithm in WSN," *Rising Threats in Expert Applications and Solutions*, vol. 3, pp. 509-517, 2022.

[19] Z. P. Tan, Y. Cai, Y. D. Wang, and P. Mao, "Research on the Value Evaluation of Used Pure Electric Car Based on the Replacement Cost Method," *IOP Conference Series: Materials Science and Engineering*, vol. 324, p. 012082, 2018.

[20] B. Brahimi, "Arabic Text Mining for Used Cars and Equipments Price Prediction," *Computación y Sistemas*, vol. 26, no. 2, pp. 34-46, 2022.

[21] H. Shi, J. Qian, N. Zhu, and T. Zhang, "RecNet: A Resource-Constraint Aware Neural Network for Used Car Recommendation," *International Journal of Computational Intelligence Systems*, vol. 15, no. 1, 91, 2022.

[22] R. Verbelen, K. Antonio, and G. Claeskens, "Unravelling the predictive power of telematics data in car insurance pricing," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 67, no. 5, pp. 1275-1304, 2018.

[23] Y. Chen, W. Zheng, W. Li, and Y. Huang, "Large group activity security risk evaluation and risk early warning based on random forest algorithm," *Pattern Recognition Letters*, vol. 144, pp. 1-5, 2021.

[24] S. Janizadeh, S. Chandra Pal, A. Saha, I. Chowdhuri, and K. Ahmadi, "Mapping the spatial and temporal variability of flood hazard affected by climate and land-use changes in the future," *Journal of Environmental Management*, vol. 298, 113551, 2021.

[25] B. Charbuty and A. Abdulazeez, "Classification Based on Decision Tree Algorithm for Machine Learning," *Journal of Applied Science and Technology Trends*, vol. 2, no. 01, pp. 20-28, 2021.

[26] K. Patil, N. Sonawane, E. Patil, K. Kulkarni, and P. Padiya, "Blockchain-Based Security for Super-peer Wireless Sensor Networks," *IC-BCT 2019*, vol. 4, pp. 241-256, 2020.

[27] F. Wang, J. Xu, X. Wang, and S. Cui, "Joint Offloading and Computing Optimization in Wireless Powered Mobile-Edge Computing Systems," *IEEE Transactions on Wireless Communications*, vol. 17, no. 3, pp. 1784-1797, 2018.

[28] B. Niu, H. Wang, J. Wang, and L. Tan, "Multi-objective bacterial foraging optimization," *Neurocomputing*, vol. 116, pp. 336-345, Sep. 2013.

[29] D. Satria, D. Park, and M. Jo, "Recovery for overloaded mobile edge computing," *Future Generation Computer Systems*, vol. 70, pp. 138-147, 2017.

[30] S. Devi and M. Geethanjali, "Application of Modified Bacterial Foraging Optimization algorithm for optimal placement and sizing of Distributed Generation," *Expert Systems with Applications*, vol. 41, no. 6, pp. 2772-2781, 2014.

[31] C. Vallati, A. Virdis, E. Mingozzi, and G. Stea, "Mobile-Edge Computing Come Home Connecting things in future smart homes using LTE device-to-device communications," *IEEE Consumer Electronics Magazine*, vol. 5, no. 4, pp. 77-83, 2016.