Predicting The California Bearing Ratio Applying The Automated Framework Of Regression Model

Pan Hu*, Jing Jin, and Yu Yun

Civil Engineering and Transportation Engineering, Yellow River Conservancy Technical Institute, Kaifeng 475004, China

Corresponding author. E-mail: hupanyrcti@126.com

Received: Oct. 29, 2023; Accepted: Aug. 10, 2024

The construction of flexible pavement on expansive soil subgrade necessitates the precise determination of the California Bearing Ratio (CBR) value, a crucial aspect of flexible pavement design. However, the conventional laboratory determination of CBR often demands considerable human resources and time. As a result, there is a need to explore alternative methods, such as developing dependable models to estimate the CBR of modified expansive soil subgrade. In this research, a machine learning (ML) model, specifically a Random Forest (RF) machine model, was developed to forecast the CBR of an expansive soil subgrade mixed with sawdust ash, ordinary Portland cement, and quarry dust. The models' performance was assessed using several error indices, and the findings revealed that the RFAO model exhibited superior predictive capability when compared to the RFDA and RFSM machine models. Specifically, the R2 values for the training and testing data for the RFAO model were 0.9952 and 0.9988, respectively. In addition, RFAO obtained the most suitable RMSE equal to 0.4878. The RFAO model generally indicated an acceptable predictive ability and more desirable generalization ability than the other developed models.

Keywords: California bearing ratio, Random Forest, Dynamic Arithmetic Optimization Algorithm, Slime Mould Algorithm, Aquila Optimizer.

http://dx.doi.org/10.6180/jase.202507_28(7).0005

1. Introduction

The CBR is a measure employed to evaluate the strength of the soil subgrade, a critical aspect in the design of flexible pavements [1]. Such an assessment holds considerable significance in this field of engineering. This investigation looks at the link between the force needed to plunge a standard plunger into the soil and the force needed to plunge the same plunger into a standardized material using a penetration test [2, 3]. The CBR value assumes a pivotal role in ascertaining the thickness of a pavement layer essential for the accommodation of the expected traffic volume of a specific design. It is imperative to augment the depth of the pavement stratum to withstand the envisaged load as the CBR value escalates adequately. It can be difficult and expensive to determine the CBR value, especially in a laboratory setting [4]. Because of the aforementioned phenomena, scientists are now investigating different approaches, such as using ML models, to more accurately forecast a soil subgrade's CBR value. The CBR is a metric that may be described as the force per unit area required to pierce a soil mass using a circular plunger with a diameter of 50 mm , compared to the same penetration in a standard material carried out at a constant pace of 1.25 mm per minute. In order to evaluate penetrations, the ratio of 2.5 to 5 mm is typically used, as stated in the references [5, 6]. The utilization of the ratio at 5 mm is consistently higher than that at 2 . There is a choice as to whether to use undisturbed or reconstituted compacted samples for the laboratory test, which may be carried out under water-

saturated or unsaturated circumstances. A recent study has demonstrated that the state of the material at the time of testing has a significant influence on the values of *CBR*, which is worth highlighting [7–9] The endeavor to develop algorithms imbued with the ability to acquire knowledge from datasets and enhance their performance over time is regularly denoted as ML, a subdomain of artificial intelligence. *ML* presents a significant advantage due to its ability to manage large and complex datasets proficiently. The capacity above facilitates the discernment of fundamental trends and the generation of meticulous prognostications characterized by a noteworthy level of precision. A great deal of documentation has been done on the rise of several *ML* approaches in the academic setting. These include deep learning, reinforcement learning, supervised and unsupervised learning, and many more. *ML* gained considerable recognition and has been widely implemented in a wide range of sectors, including healthcare, transportation, manufacturing, and finance [10–13]. The utilization of *ML* in the healthcare industry holds promise for analyzing medical images and detecting potential anomalies. Within finance, the application of *ML* holds promising potential for identifying fraudulent activities and mitigating the associated risks inherent in financial pursuits. This premise has been extensively explored in pertinent scholarly works [14–16].

The RF model in ML represents an exemplary approach within the domain of supervised learning algorithms that are adaptable for performing regression as well as classification tasks. The randomized forest algorithm produces a heterogeneous collection of decision trees and consolidates their outputs to attain a final prediction. Each decision tree is assembled by applying a randomized subset sourced from the training data, while the model utilizes the bagging methodology to amalgamate the individual prognostications produced by each decision tree [17, 18]. The RF [19] algorithm offers several advantages compared to alternative ML methods. The model demonstrates resiliency to overfitting, thereby suggesting its proficiency in mitigating the concern of over-learned training data and the incapability to adjust to new data. The algorithm demonstrates its ability to effectively handle datasets containing missing values and noise while providing valuable metrics for measuring variable significance. These metrics play an essential role in identifying the significant characteristics that hold crucial sway over the forecasted outcome of the dependent variable. The prognostication of the CBR estimation for a soil subgrade may be achieved by implementing an RF model and employing an extensive dataset [20, 21]. The dataset must encompass a wide spectrum of soil properties, such as the Atterberg limits and compaction characteristics, in addition to the precise composition and quantity of stabilizing agents employed. The model discussed above relies on the available dataset to develop a predictive function that can accurately estimate the *CBR* attribute of the subgrade soil based solely on the given features. It is possible to assess the effectiveness of the RF model utilizing various error metrics, such as the mean squared error (MSE) and the coefficient of determination (R^2). The evaluation results can serve as a means of determining the accuracy and reliability of the model. The RF model presents a robust tool for predicting the *CBR* of a soil subgrade in an academic context. The ability to adeptly handle missing, noisy, and variable pertinent information underscores the credibility of the methodology as a reliable means of generating accurate predictions in empirical investigations. By utilizing various ML techniques, such as the *RF* algorithm, it is possible to mitigate the time and cost implications associated with determining the *CBR* of soil subgrades [22]. This process's outcome optimizes both efficiency and cost-effectiveness in pavement design.

This article employs *ML* techniques to address complex systems and the processing of multiple parameters to predict *CBR*. Specifically, the approach utilized in this study involves using the *RF* method. Additionally, meta-heuristic algorithms enhance the output and increase precision while minimizing errors within the relevant model to generate results that mimic those obtained through laboratory experimentation. The ensuing section will expound on diverse optimization techniques, including the Dynamic Arithmetic Optimization Algorithm (DAOA), Slime Mould Algorithm (SMA), and Aquila Optimizer (AO), which belong to the category of novel algorithms. The DAOA was chosen for its adaptability to different problem characteristics, efficient convergence properties, and versatility across diverse optimization tasks. SMA, inspired by natural foraging behavior, offers robustness against uncertainties and a balanced exploration-exploitation approach. Meanwhile, AO excels in global search capability, efficiency in resource consumption, and scalability to handle complex, multidimensional optimization spaces. By leveraging these algorithms, the study aims to enhance model accuracy, efficiency, and robustness while accommodating the complexities inherent in predicting CBR from complex systems and multiple parameters using ML methods. It should be mentioned that, in this study, Python was utilized as the primary programming language for building and deploying the ML models, including the RF method and meta-heuristic algorithms such as the DAOA, SMA, and AO.

2. Materials and methodology

2.1. Data gathering

The experiment carried out by Ikeagwuani provided the intake-output dataset used in this investigation [23]. 2 different experimental techniques are used in the mix ratio: the response surface approach and the Taguchi array technique. The stabilizers and modified expansive soil qualities shown in Table 1 are correlated with the CBR value, which is the output variable and has inherent geotechnical characteristics. The modified physical properties of the soil include the Atterberg limits and compaction attributes of the altered expansive soil. The Atterberg limits comprise specific parameters, namely the liquid limit (LL), plastic limit (PL), and plasticity index (PI).

Additionally, the compaction characteristics comprise 2 critical attributes: maximum dry density (MDD) and optimum content (OMC). Furthermore, the intake parameters encompass the SDA, OPC, and QD stabilizers. Overall, the study comprised 8 intake parameters and a solitary outcome parameter. Table 1 shows that there are 109 sampled points in the data set used for this investigation.

2.2. Random Forest

2.2.1. The RF basis

An RF framework comprises a tree-shaped classifier collection denoted as $\{\mathrm{s}\,(\mathrm{x}, \aleph_1)\,, \mathrm{m} = 1, \ldots\}$. Every tree contributes a single option for establishing the predominant group for a specific intake x. In this context, the $\{\aleph_1\}$ stand for random vectors that exhibit independence and identical distribution characteristics.

An RF is composed of numerous tree-shaped classifiers created by a training demo collection and an arbitrary variable, denoted as $\{\aleph_1\}$, for the m-th tree within Breiman's framework [24]. These arbitrary ones exhibit independence and identical distribution across any 2 trees, thereby giving rise to a classifier $s\,(x, \aleph_l)$, in which, x stands for the input vector. A sequence of classifiers $\{s_1(x), s_2(x), \ldots, s_m(x)\}$ is created, and this sequence can be harnessed to construct several categorization frameworks.

$$S_{(x)} = \mathrm{argmax}_m \sum_{i=1}^{m} F\,(s_i(x) = W) \tag{1}$$

Every tree possesses the prerogative to exercise its voting power in favor of the aptest classification outcome of a specified intake variable. The discrete decision tree forms conglomeration is symbolically represented using $S(x)$. The outcome would be W, and the index function is shown by F(.) [25]. The procedure for choosing the suitable classification result is shown in Fig. 1.

2.2.2. RF Characteristics

The function of margin that is utilized in the RF algorithm [25] to evaluate the degree to which the mean number of correct class votes exceeds that of incorrect class votes at X, W may be expressed as follows:

$$dz(X, W) = av_l F\,(s_l(X) = W) - max_{j \neq W} av_l F\,(s_l(X) = j) \tag{2}$$

The margin function serves as a metric that gauges the precision and assurance of the classification forecast, wherein an elevated value indicates an augmented level of accuracy and confidence. The classifier's generalization error can be formulated as follows:

$$EQ^* = E_{X,W}(dz(X, F) < 0) \tag{3}$$

Leo Breiman found that as the number of decision trees in an RF classifier becomes large enough, the arbitrary variable $s_m(X) = s\,(x, \aleph_m)$, adheres to the Strong Law of Large Numbers. With an increasing decision tree number, $EQ*$ congregate to a specific amount for nearly all of $\aleph_1$ sequences. Additionally, Breiman showed that RFs are resilient against overfitting and have the capacity to produce a limit value for error of generalizing.

$$E\left(E_\theta\,(s_m(x, \theta) = W) - \max_{j \neq W} E_\theta(s(x, \theta) = j) < 0\right) \tag{4}$$

Leo Breiman also drew the conclusion that there exists an upper bound for the generalization error:

$$EQ^* \leq \bar{\beta}\left(1 - c^2\right)/c^2 \tag{5}$$

The error of generalizing an RF is affected by 2 key factors: every forest tree's strength, which is distributed as (c), and the level of correlation among the trees, indicated using the mean correlation value $\bar{\beta}$. A small correlation amount signifies diminished reliance on the trees, leading to enhanced performance of the RF [26].

2.2.3. Out-of-bag estimation

Constructing an RF model entails generating a tree structure on a fresh training dataset that employs a random selection of features. The present study introduces a novel training set that employs bagging techniques, extracting random demos of the primary training collection. Bagging was employed for 2 fundamental purposes. The efficacy of bagging in enhancing accuracy, specifically in incorporating random features, has emerged as a pertinent observation. Another aspect to consider is that bagging generates out-of-bag data and is able to offer ongoing assessments

Table 1. The numerical characteristics of output and intake parameters.

Numerical Features	Dataset Components								
	PL	**LL**	**MDD**	**PI**	**OMC**	**QD (%)**	**SDA (%)**	**CBR (%)**	**OPC (%)**
Min	17.9	21.2	1.365	2.1	18.9	0	0	19.69	2
Max	37.2	52.1	1.777	19.5	29.5	20	20	66.75	8
St. Dev.	4.2812	6.1538	0.0883	4.1149	2.426	8.1961	7.1546	10.866	2.3798
Mean	26.683	35.845	1.4929	9.1623	24.143	10.642	10.660	39.959	4.9449

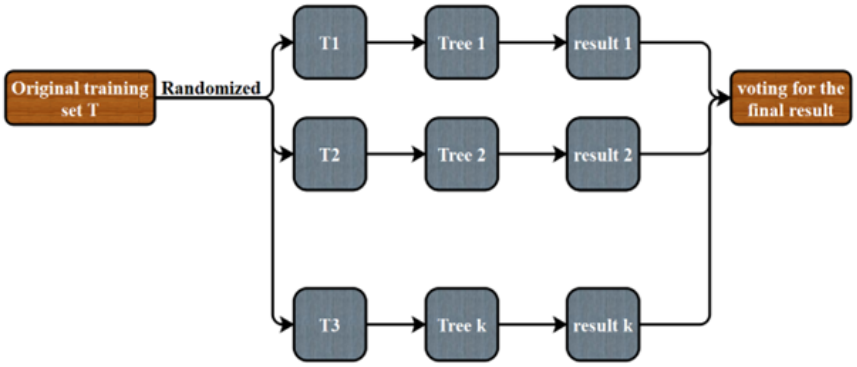

Fig. 1. Schematic of RF.

for the out-of-sample $EQ*$ of the RF , as well as correlation and strength predictions.

Around 36.8% of the demos in dataset T are excluded from the training collection of m th, S_l, obtained from the initial training collection S by employing bagging with an exchange. S_l comprises N demos, in which, N represents the whole demos amount in S. The possibility of any particular demo not being a part of S_l is given by $(1-1/N)^N$, and these approaches c^{-1} as N grows. These demos, which are absent from S_l are commonly referred to as out-of-bag (OOB) info [27].

The OOB prediction approach applies a method that leverages out-of-bag info to approximate the classifying efficiency of a given model. This technique acquires an error estimate of every tree constituent of the RF algorithm. Determining the RF's generalization error involves the derivation of the mean of all tree error approximations within the RF. Tibshirani, Macready, and Wolpert proposed integrating the OOB prediction in the generalization error estimation due to its expedited computational nature and reduced bias compared to cross-validation. Moreover, the OOB estimate exhibits higher precision in comparison to the cross-validation approach. The utilization of the OOB error estimate obviates the requirement for allocating a distinct test framework. As Breiman showed, the OOB precision estimate is commensurate with utilizing a test framework that matches the training framework size. Furthermore, the utilization of the OOB technique has the potential to furnish an internal appraisal of the potency and interdependence of the analysis, thereby facilitating the comprehension of the classification precision and the identification of prospective enhancements.

2.3. Dynamic Arithmetic Optimization Algorithm (DAOA)

The dynamic version of optimization methods employs solution candidates with adaptive characteristics and evaluates domains that enable the utilization of exploitation and exploration approaches. A salient feature of $DAOA$ is its innate capacity to forego the need for pre-defined parameter adjustments in the context of commonly employed meta-heuristics.

The cruciality of dynamic acceleration functions in the exploration phase of dynamics is evident in Arithmetic Optimization Algorithms. The alignment with the recent downward function necessitates that AOA implements modifications to the upper and lower thresholds of the initial data for the acceleration capability. Utilizing an algorithm devoid of modifiable internal parameters is highly recommended due to the substitution of the dynamically accelerated function. The improvement factor of the algorithm is characterized by the following terms:

$$R = \left(\frac{\text{Max}_{\text{iter}}}{R} \right)^{Z} \tag{6}$$

Balancing exploitation and exploration is an essential factor in the effectiveness of metaheuristic methods, given

that the features are their fundamental components [28]. As a part of the optimization approach that places equal emphasis on both exploration and exploitation, each solution adapts its location dynamically regarding the optimal solution achieved in the present iteration, as described below:

$$x_{i,j}(q+1) = \begin{cases} \text{best}\left(x_j\right) \div (S+z) \times \left(\left(vh_j - ch_j\right) \times U + ch_j\right) \\ , z_1 < 0.5 \\ \text{best}\left(x_j\right) \times S \times \left(\left(vh_j - ch_j\right) \times U + ch_j\right) \\ , \text{ otherwise} \end{cases} \tag{7}$$

$$x_{i,j}(q+1) = \begin{cases} \text{best}\left(x_j\right) - (S+z) \times \left(\left(vh_j - ch_j\right) \times U + ch_j\right) \\ , z_2 < 0.5 \\ \text{best}\left(x_j\right) + S \times \left(\left(vh_j - ch_j\right) \times U + ch_j\right) \\ , \text{ otherwise} \end{cases} \tag{8}$$

During the progression of iterations, the value decreases, while a dynamic function (S) for candidate solutions is defined based on the decreasing influence of the proportion of available resolutions, such as:

$$S(0) = 1 - \sqrt{\frac{R}{\text{Max}_{\text{iter}}}} \tag{9}$$

$$S(q+1) = S(q) \times 0.99 \tag{10}$$

Incorporating candidate solutions within the framework of DAOA holds the potential to enhance the AOA's convergence rate. This is chiefly due to the many exploration agents and iterative processes involved. Additionally, the improvements augment the quality level of the solution. "Parameter-free functioning is a commonly acclaimed attribute of these algorithms, regarded as a prominent strength. The use of dynamic functions differs between $DAOA$ and AOA [29]. The DAOA algorithm possesses a significant advantage by demonstrating an inherent capability to acclimate to parameters, thus necessitating minimal modifications. Unlike numerous methods that necessitate parameter modification to align with diverse problem scenarios, the DAOA algorithm distinguishes itself. One potential drawback of $DAOA$ is its reliance on the number of iterations for its adaptive mechanism rather than improving overall fitness. The pseudocode for DAOA is presented through algorithm 1.

Algorithm 1. The dynamic form of AOA

Initialize the parameters for $DAOA$
Generate random values for primary positions.
while ($q <$ maximu number of iterations) ***do***
Calculate the fitness values for the given solution representation.
Discover the optimal solution
Revise the S value using Eq. (6)
Modify the R-value using Eq. (9)

for i=1:
for j=1:

Generate random values ranging from 0 to 1 for z_1, z_2, and z_3
if Z_1 >S

if Z_2>q
By employing the first mentioned formula in Eq. (7)
Else
By making use of the second formula mentioned in Eq. (7)
end if

if Z_1<S
if Z_2>0.5

By applying the first mentioned formula in Eq. (8)
Else
By employing the second stated formula in Eq. (8))
end if
end if
end if
q=q+1
end while
Return the optimal solution

2.4. Slime Mould Algorithm (SMA)

This particular optimization approach is centered on an organism known as Physarum Polycephalum, a variety

of slime mold. The Plasmodium stage constitutes the predominant metabolic phase of the mold, characterized by its highly active and dynamic nature. During this stage [30], the organic material in the slime mold actively seeks nourishment, envelops it, and then secretes various enzymes to facilitate its decomposition and digestion. These organisms possess distinctive traits and patterns that enable them to create a network of veins connecting with multiple food sources simultaneously. The slime creature is able to choose the best path for connecting to food sources by means of both positive and negative feedback processes [31]. As a result, using both mathematical modeling and real-world investigations, academics have investigated the application of slime molds in the fields of graph theory and route networks. This segment will comprehensively account for the suggested mathematical model and approach [32].

To imitate the approach behavior of slime molds, which can detect food through scent in the air, the subsequent equations are suggested:

$$\overrightarrow{X(h+1)} = \begin{cases} \overrightarrow{X_p(h)} + \overrightarrow{wp} \cdot \left(\vec{C} \cdot \overrightarrow{X_Z(h)} - \overrightarrow{X_B(h)}\right), s < q \\ \overrightarrow{wZ} \cdot \overrightarrow{X(h)}, s \geq q \end{cases} \tag{11}$$

The parameter $\overrightarrow{wp}$ has a range of $-e$ to e, $\overrightarrow{wZ}$ decreases linearly from one to zero, and X reflects the position of the slime mold. $\overrightarrow{X_p}$ indicates the location of the individual with the greatest odor concentration currently observed. The variables $\overrightarrow{X_Z}$ and $\overrightarrow{X_B}$ represent 2 randomly chosen slime mold individuals, whereas h indicates the current iteration and $\vec{C}$ indicates the weight of the slime mold. Below is the formula for q :

$$q = \tanh|G(i) - BF| \tag{12}$$

For variables $i \in 1, 2, \ldots, n$, the BF shows the optimal fitness attained throughout all iterations, while $G(i)$ shows the fitness of X. The following is the formula for $\overrightarrow{wp}$:

$$\overrightarrow{wp} = [-e, e] \tag{13}$$

$$e = \operatorname{arctanh}\left(-\left(\frac{h}{\max\_ h}\right) + 1\right) \tag{14}$$

The formula for $\vec{C}$ is presented below:

$$\overrightarrow{C(\text{ SmellIndex }(\imath))} = \begin{cases} 1 + s \cdot \log\left(\frac{pF - G(i)}{pF - mF} + 1\right), & \text{condition} \\ 1 - s \cdot \log\left(\frac{pF - G(i)}{pF - mF} + 1\right), & \text{other} \end{cases} \tag{15}$$

$$\text{SmellIndex } = \text{sort}(G) \tag{16}$$

The following formula defines $\vec{C}$: A random value between 0 and 1 is represented by s; the best fitness value obtained in the current iteration is shown by pF; the worst fitness value obtained in the current iterative procedure is shown by mF; and the fitness values pattern, which is arranged in ascending order for value minimization, is represented by SmellIndex. Furthermore, $G(i)$ falls into the first half of the population according to condion.

Eq. (17) is a mathematical model that simulates the contraction of the vascular structure of slime molds during their feeding search. A larger vein is indicative of a more robust wave emerging from the quicker cytoplasm and flow bio-oscillator, and the model relies on this interplay between the vein's diameter and the concentration of food. Variables are added to Eq. (15) to account for the uncertainty in the venous tissue contraction mode. The component log's inclusion helps to stabilize the contraction frequency's numerical values. Based on the nutritional value of the food, the variable condition replicates the adaptable search pattern of the slime mold. In other words, the weight of the region grows in high food concentration places and reduces in low food concentration parts, leading to the exploration of other regions.

The following mathematical expression may be stated by updating the location of slime molds using the aforementioned principle:

$$\overrightarrow{X^*} = \begin{cases} \text{rand}. \ (CR - LR) + LR, & \text{rand } < o \\ \overrightarrow{X_p(h)} + \overrightarrow{wp} \cdot \left(C \cdot \overrightarrow{X_Z(h)} - \overrightarrow{X_B(h)}\right), & s < q \\ \overrightarrow{wz} \cdot \overrightarrow{X(h)}, & s \geq q \end{cases} \tag{17}$$

LR and CR, respectively, stand for the lower and upper limits of the search range. Random values inside the interval $[0, 1]$ are represented by rand and s. The SMA pseudocode is shown by algorithm 2.

Algorithm 2. SMA's pseudocode

Initialize the parameters pop size, Max_iteraition;
Initialize the positions of the slime mold $X_i (i = 1, 2, \ldots, n)$;
While (h Max_iteraition)
  Calculate the fitness of all slime mold;
  update bestFitness, X_p
  Calculate the C by Eq. (15);
  For each search portion
    update q, wp, wz;
    update positions by Eq. (17);
End For
h = h + 1 ;
End While
Return *bestFitness*, X_p;

2.5. Aquila Optimizer (AO)

In 2021, Abualigah et al. introduced a new swarm intelligence algorithm called *AO*. The algorithm is inspired by Aquila's 4 distinct hunting behaviors, which are tailored for different types of prey. Aquila can effortlessly switch between hunting strategies to adapt to prey, utilizing its agility, powerful feet, and claws to launch swift attacks. A concise explanation of the mathematical model is presented below [33].

1 : $The first step, Expanded Exploration\ (X_1)$

It entails Aquila flying at a high altitude to thoroughly explore a vast search space, followed by a vertical dive towards the prey's location. The corresponding mathematical model is expressed as follows:

$$X_1(h+1) = X_{\text{best}}(h) \times \left(1 - \frac{h}{H}\right) + (X_W(h) - X_{\text{best}}(h) \times z_1) \tag{18}$$

$$X_W(h) = \frac{1}{U}\sum_{i=1}^{U} X_i(h) \tag{19}$$

Whereas, $X_W(h)$ represents the average position of all Aquilas in the present version., and $X_{\text{best}}(h)$ denotes the best position achieved up to that point. Here, h denotes the current iteration, and H represents the maximum number of iterations. U corresponds to the size of the population, and z_1 is a random number ranging from 0 to 1 .

2 : $Narrowed Exploration\ (X_2)$: *Contour Flight with Short Glide Attack*

Aquila's primary hunting strategy involves descending within a selected area and flying around the prey, followed by a short gliding attack [34]. The corresponding formula for updating the position is as follows:

$$X_2(h+1) = X_{\text{best}}(h) \times LF(B) + X_S(h) + (y - x) \times z_2) \tag{20}$$

The formula for updating the position is given by: where $X_S(h)$ denotes a random position of Aquila, B represents the size of the dimension, and z_2 is a random number between 0 and 1 . Moreover, $LF(B)$ represents the Levy fight function, which is defined as follows:

$$LF(B) = c \times \frac{v \times \sigma}{|\vartheta|^{\frac{1}{\alpha}}} \tag{21}$$

$$\sigma = \left(\frac{\Gamma(1+\alpha) \times \sin\left(\frac{\pi\alpha}{2}\right)}{\Gamma\left(\frac{1+\alpha}{2}\right) \times \alpha \times 2^{\left(\frac{\alpha-1}{2}\right)}}\right) \tag{22}$$

The constant parameters c and α are set to 0.01 and 1.5 , respectively. v and ϑ are random numbers within the range of 0 to 1 . The variables y and x are utilized to generate a spiral trajectory during the search process, and their calculation is as follows:

$$\begin{cases} x = z \times \sin(\theta) \\ y = z \times \cos(\theta) \\ z = z_3 + 0.00565 \times B_1 \\ \theta = -\varphi \times B_1 + \frac{3\times\pi}{2} \end{cases} \tag{23}$$

Here, z_3 is an integer variable representing the number of search cycles, ranging from 1 to 20.B_1 is a set of integers from 1 to the dimension size (B), while φ is assigned a value of 0.005 .

3. $Expanded\ Exploitation\ (X_3)$: *Low Flight with a Slow Attack*

The third hunting strategy involves Aquila descending vertically for a preliminary attack once the prey's location is approximately determined. *AO* capitalizes on the chosen area to close in on and attack its prey. The corresponding behavior is expressed as follows:

$$\begin{aligned} X_3(h+1) = {} & (X_{\text{best}}(h) - X_W(h)) \times \gamma - z_4 + \\ & ((VC - LC) \times Z_5 + LC) \times \delta \end{aligned} \tag{24}$$

Where $X_{\text{best}}(h)$ represents the optimal position attained so far and $X_W(h)$ denotes the average position of the present positions. The parameters γ and δ are the exploitation adjustment parameters, which are set to a value of 0.1.VC and LC correspond to the upper and lower bounds of the problem, whereas Z_4 and Z_5 are random numbers ranging from 0 to 1.

2.6. Methods for assessing performance

As previously noted, current research employs a range of metrics for assessing the models, including the correlation coefficient $\left(\mathrm{R}^2\right)$, median absolute percentage error (MDAPE), root mean square error (RMSE), and mean square error (MSE), and n10-index. These metrics are calculated using Eqs. (25) to (29):

$$R^2 = \left(\frac{\sum_{i=1}^{u}(z_i - \bar{z})(e_i - \bar{e})}{\sqrt{\left[\sum_{i=1}^{W}(z_i - z)^2\right]\left[\sum_{i=1}^{u}(e_i - \bar{e})^2\right]}}\right)^2 \tag{25}$$

$$RMSE = \sqrt{\frac{1}{U}\sum_{i=1}^{u}(e_i - z_i)^2} \tag{26}$$

$$MSE = \frac{1}{U}\sum_{i=1}^{u} h_i^2 \tag{27}$$

$$\text{MDAPE} = 100 \times \text{median}\left(\frac{|z_i - \bar{z}|}{|e_i - \bar{e}|}\right) \tag{28}$$

$$n10 - \text{index} = \frac{n10}{n} \tag{29}$$

Here, z_i stands for the predicted values, while e_i corresponds to the experiential amounts. The forecasted and observed samples' average values are referred to as z and e, respectively. Alternatively, U represents the quantity of samples under consideration.

3. Results and discussion

3.1. Comparative analysis based on evaluators' results

In this section, the models are discussed based on the criterion. The dataset was randomly split into a training dataset and a test dataset. 30% of the data set of the test was utilized to assess its dependability after it was constructed using 70% of the training data set. Establish reasonable connections between the reasons. Based on Table 2, the maximum estimated value of R^2 was reported to the $RFAO_{\text{test}}$ and equal to 0.9988 ; the lowest value was reported to the $\text{RFSM}_{\text{train}}$ and equal to 0.9776 . The RMSE component also showed that $\text{RFSM}_{\text{train}} = 1.7210$ had the greatest value and $\text{RFAO}_{\text{test}} = 0.4878$ had the lowest. When it comes to MSE, the RFAO has the best reported result (0.238), while the $\text{RFSM}_{\text{test}}$ has the worst reported value (3.664). In MDAPE, $\text{RFAO}_{\text{test}} = 0.5979$ had the best performance, although the $\text{RFSM}_{\text{test}} = 2.4504$ had the worst performance. In the end, the $\text{RFAO}_{\text{test}}$ yielded the most appropriate result for T_{estate} , whereas the $\text{RFSM}_{\text{test}}$ produced the lowest value. Since the test data has been decreased, with the exception of a few instances, the model's findings show that all of the models have been appropriately trained. In general, the parameters of the RFSM model decrease, but since the parameters had high values from the beginning, it is not a suitable prediction model, while the RFAO model has a slight increase except for R^2, and the rest of the parameters decrease. As a result, the model is suitable and with It is highly accurate for prediction.

CBR prediction is essential for assessing geotechnical risks associated with subgrade soils. Understanding CBR values allows engineers to evaluate soil stability, susceptibility to deformation, and potential for bearing capacity failures, enabling proactive risk management and mitigation measures during project planning and execution.

Fig. 2 depicts a scatter plot comparing the estimated against the real amounts of 3 hybrid ones: RFAO, RFDA, and RFSM. 2 linear fits plus a center line, which stand in for the distinct training and testing frameworks, make up the scatter plot. Given that the anticipated and observed numbers have a significant positive correlation, the results show that all 3 frameworks are capable of producing accurate predictions. However, based on the scatter plot, RFAO exhibits the densest clustering of data points around the linear fit lines among the 3 models, suggesting higher accuracy. Although the data points show more dispersion, there is a substantial link between RFDA and RFSM. Both models have identical slopes and intercepts for their linear regression lines, indicating equivalent expected capabilities.

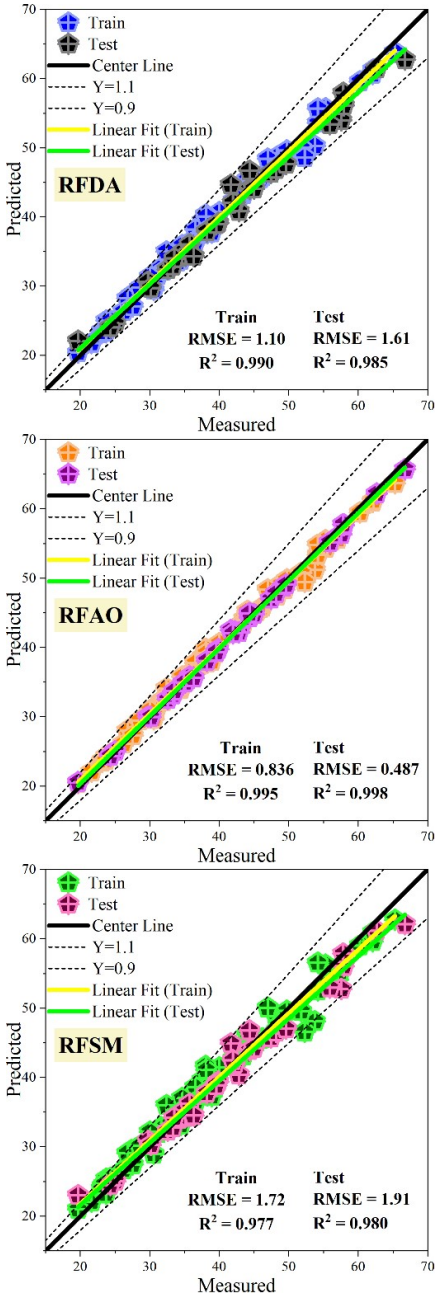

Fig. 2. The scatter diagram representing the hybrid improved models.

A line-symbol plot comparing the estimated and observed demonstrations of 3 hybrid ones-RFAO, RFDA, and RFSM—is shown in Fig. 3. The chart demonstrates the effectiveness of the models by assessing the degree of agreement

Table 2. Developed assessment results of models by evaluators.

Hybrid Models	Sections	Evaluators				
		RMSE	**R^2**	**MSE**	**MDAPE**	**n10_index**
RFDA	**Train**	1.1014	0.9907	1.2132	1.4874	1
	Test	1.6107	0.9858	2.5947	2.0118	0.9697
RFAO	**Train**	0.8365	0.9952	0.6998	1.3268	1
	Test	0.4878	0.9988	0.238	0.5979	1
RFSM	**Train**	1.7210	0.9776	2.9621	2.3613	0.9737
	Test	1.9142	0.9809	3.6644	2.4504	0.9697

between the predicted and actual values. As seen by the strong agreement between the observed and predicted values throughout the dataset, the findings show that RFAO achieves great accuracy. Although RFDA and RFSM also show a significant connection between the projected and observed results, there is a slightly bigger degree of departures from the empirical data. This implies that, despite their effectiveness, RFDA and RFSM might not provide as high of an accuracy as RFAO.

Fig. 4 displays the distribution of the improved ones' error percentages. There is a vertical axis for incidence frequency and a horizontal axis for mistake rate. According to the results, RFAO has the fewest mistakes, with most error percentages ranging between 0 and 10%. Conversely, RFSM and RFDA have a larger range of error percentage distribution and a higher frequency of results across the 10% threshold. Furthermore, an extremely small subset of data points exhibits notably greater error percentage levels in both RFDA and RFSM, indicating a right-skewed distribution. Given the graph's indication that the testing framework's error percentages are comparatively lower than those of the training framework, overfitting of the training data is a possibility. Overall, the graph does a good job of showing how the upgraded ones' error percentages are distributed, emphasizing how precise the RFAO approach is. Reliable CBR prediction contributes to cost-effective project management by optimizing material usage, reducing over-design, and minimizing construction delays and unforeseen repairs. Improved efficiency in pavement design and maintenance translates into substantial cost savings for infrastructure stakeholders and taxpayers.

Fig. 5 displays a boxed plot illustrating the error percentages of the models represented. During the training phase, RFAO had a mean error rate of 0%, with a normal distribution and minimal dispersion. Favorable values fell below the 10% threshold in the error distribution. With a more uniform and symmetrical normal distribution, RFDA, on the other hand, showed dispersion in both stages. The error percentage of the model was limited to 10%, nevertheless. It is unusual in statistical analysis for a single outlier datum to account for more than 10% of the dataset during the evaluation stage, yet RFSM revealed the most significant and varied discrepancies. Compared to the other 2 models, the RFDA's Gaussian distribution had higher dispersion due to a lower frequency of incidence near zero. In general, all the models performed satisfactorily, although RFAO produced better results. CBR prediction supports environmentally sustainable practices by facilitating the selection of eco-friendly construction materials and techniques. By optimizing subgrade soil characteristics based on CBR values, engineers can minimize environmental footprint and enhance the long-term sustainability of infrastructure projects.

3.2. Wilcoxon Test

The Wilcoxon test was utilized to evaluate the comparative performance of 3 models: RFDA, RFAO, and RFSM. The test results, including p-values and statistics computed for each pair of models, offer valuable insights into their statistical significance and relative performance. Detailed outcomes of the Wilcoxon test are presented in Table 3, providing a comprehensive overview of the model comparisons and their significance levels. The results of the Wilcoxon test indicate that there is no statistically significant difference in performance between RFAO and RFSM (p-value = 0.9698, Statistic = 2985), as well as between RFDA and RFSM (p-value = 0.0680, Statistic = 2394), suggesting comparable model pairs. However, the comparison between RFDA and RFAO reveals a marginally significant difference (p-value = 0.0268, Statistic = 2265). Although this difference does not reach conventional levels of significance, it suggests a potential difference that may require further investigation or consideration. In summary, according to the Wilcoxon test, RFAO performs comparably to both RFDA and RFSM, while the comparison between RFDA and RFAO shows a marginally significant difference, highlighting the importance of cautious interpretation and potential further exploration.

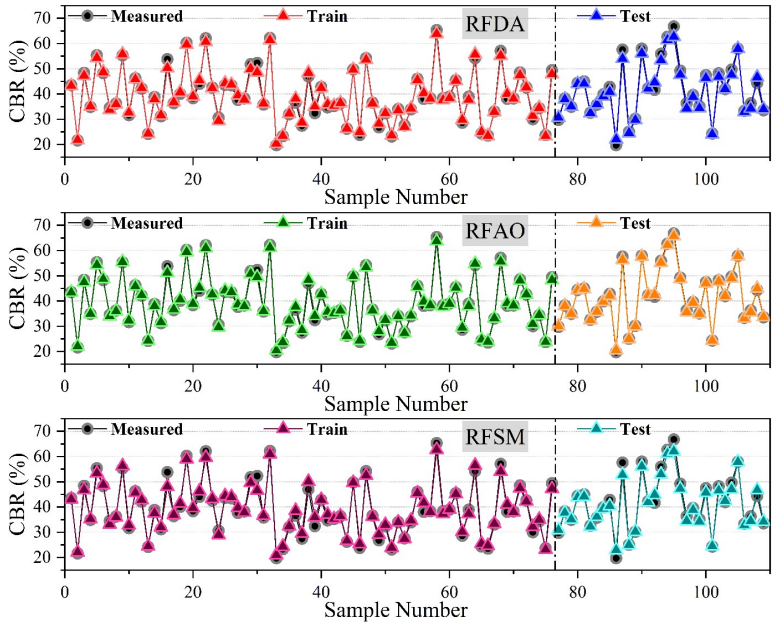

Fig. 3. The juxtaposition of forecasted and observed samples.

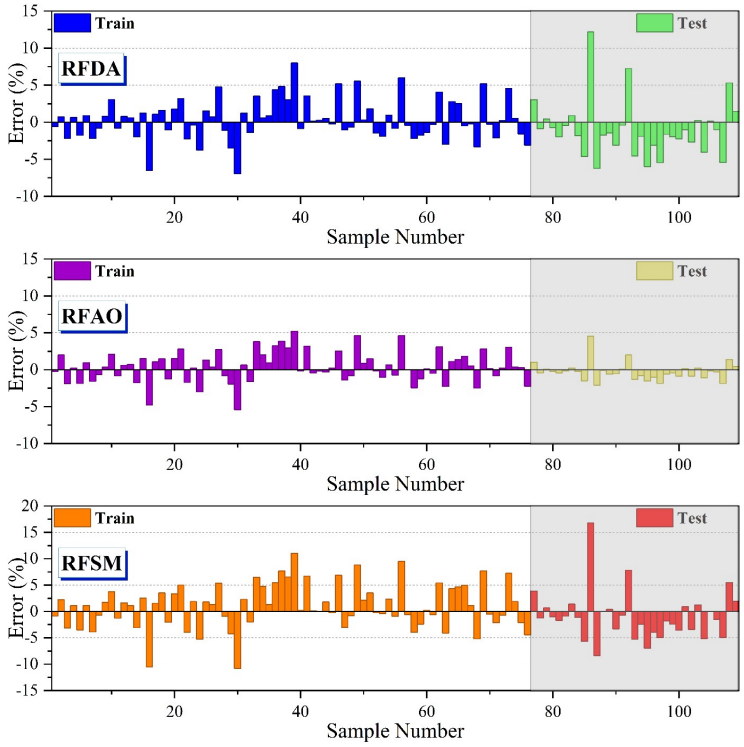

Fig. 4. The percentage of error rates for the models on display.

4. Conclusion

To anticipate R^2 and RSME, the construction and implementation of durable and adaptable pavements frequently require the use of reputable methods for evaluating or verifying the CBR. Regrettably, the customary testing methodology for determining the subgrade's CBR is often challenging because of the substantial time commitments necessitated by the methodology utilized for the test. Therefore, the necessity to examine substitute approaches, specifically the creation of prognostic models, has arisen as a way to

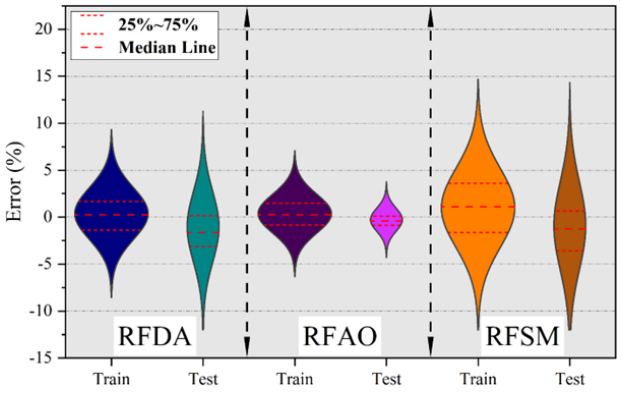

Fig. 5. The distribution of errors among the hybrid models is illustrated by the violin plot depicted in the diagram.

Table 3. Developed assessment results of models by evaluators.

Difference of models	Parameter	
	p_value	statistic
Def. between RFDA and RFAO	0.0268	2265
Dif. between RFDA and RFSM	0.0680	2394
Dif. between RFAO and RFSM	0.9698	2985

develop a method for estimating the CBR of expansive soil subgrades. Due to this rationale, using learning machines has superseded human experiments. This article aims to introduce the RF model, an ML method utilized for predicting CBR.

In addition, the relevant model was combined with 3 meta-heuristic algorithms-the DAOA, SMA, and AO-to improve accuracy and reduce mistakes, creating a hybrid model. Subsequently, various performance metrics, namely R^2, RMSE, MSE, MDAPE, and n10_index, were utilized to evaluate the efficiency of these models in this research article. The blended models' achievements were evaluated based on distinct criteria employed for model selection. Therefore, it can be inferred that the AO model demonstrated superior performance compared to the other 2 models when combined with RF.

Identifying study limitations is essential for improving research and guiding future endeavors. This study's limitations include challenges related to the generalizability of RF combined with meta-heuristic algorithms, necessitating validation across diverse soil types and environments to enhance reliability. Data quality issues, such as missing values or outliers, may affect model effectiveness, emphasizing the critical reliance on high-quality data. Exploration of optimal algorithm selection, tuning, and evaluation metrics is warranted to fully capture model performance and utility. External validation through independent datasets is crucial to confirm the approach's effectiveness beyond experimental settings. Future research should explore different soil conditions, integrate additional data sources, conduct comparative studies with alternative methods, enhance model interpretability, deploy real-time monitoring systems, validate models through field studies, and develop user-friendly tools. These efforts aim to advance predictive modeling for CBR estimation, enhancing accuracy and applicability in pavement design and infrastructure management.

References

[1] B. Yildirim and O. Gunaydin, (2011) *"Estimation of California bearing ratio by using soft computing systems"* **Expert Systems with Applications** ***38***: 6381–6391. DOI: 10.1016/j.eswa.2010.12.054.

[2] T. Taskiran, (2010) *"Prediction of California bearing ratio (CBR) of fine grained soils by AI methods"* **Advances in Engineering Software** ***41***: 886–892. DOI: 10.1016/j.advengsoft.2010.01.003.

[3] S. Alzabeebee, S. A. Mohamad, and R. K. S. Al-Hamd, (2022) *"Surrogate models to predict maximum dry unit weight, optimum moisture content and California bearing ratio form grain size distribution curve"* **Road Materials and Pavement Design** ***23***: 2733–2750. DOI: 10.1080/14680629.2021.1995471.

[4] C. Khasnabis, K. H. Motsch, K. Achu, K. Al Jubah, S. Brodtkorb, P. Chervin, P. Coleridge, M. Davies, S. Deepak, K. Eklindh, et al. "About the CBR guidelines". In: *Community-Based Rehabilitation: CBR Guidelines*. World Health Organization, 2010.

[5] S. Bhatt, P. K. Jain, and M. Pradesh, (2014) *"Prediction of California bearing ratio of soils using artificial neural network"* **Am. Int. J. Res. Sci. Technol. Eng. Math** *8*: 156–161.

[6] M. Ahmed, S. AlQadhi, J. Mallick, N. B. Kahla, H. A. Le, C. K. Singh, and H. T. Hang, (2022) *"Artificial neural networks for sustainable development of the construction industry"* **Sustainability** ***14***: 14738. DOI: 10.3390/su142214738.

[7] A. M. Ebid, (2021) *"35 Years of (AI) in geotechnical engineering: state of the art"* **Geotechnical and Geological Engineering** ***39***: 637–690. DOI: 10.1007/s10706-020-01536-7.

[8] M. W. Kin, (2006) *"California bearing ratio correlation with soil index properties"* **Master degree Project, Faculty of Civil Engineering, University Technology Malaysia**:

[9] D. K. Talukdar, (2014) *"A study of correlation between California Bearing Ratio (CBR) value with other properties of soil"* **International Journal of Emerging Technology and Advanced Engineering *4***: 559–562.

[10] W. Zhang, X. Gu, L. Tang, Y. Yin, D. Liu, and Y. Zhang, (2022) *"Application of machine learning, deep learning and optimization algorithms in geoengineering and geoscience: Comprehensive review and future challenge"* **Gondwana Research *109***: 1–17. DOI: 10.1016/j.gr.2022.03.015.

[11] C. Li, J. Zhou, D. Dias, and Y. Gui, (2022) *"A kernel extreme learning machine-grey wolf optimizer (KELM-GWO) model to predict uniaxial compressive strength of rock"* **Applied Sciences *12***: 8468. DOI: 10.3390/app12178468.

[12] S. Talamkhani, (2023) *"Machine Learning-Based Prediction of Unconfined Compressive Strength of Sands Treated by Microbially-Induced Calcite Precipitation (MICP): A Gradient Boosting Approach and Correlation Analysis"* **Advances in Civil Engineering *2023***: 3692090. DOI: 10.1155/2023/3692090.

[13] Z.-H. Zhou. *Machine learning*. Springer nature, 2021.

[14] I. G. Farias, W. Araujo, and G. Ruiz, (2018) *"Prediction of California bearing ratio from index properties of soils using parametric and non-parametric models"* **Geotechnical and geological engineering *36***: 3485–3498. DOI: 10.1007/s10706-018-0548-1.

[15] T. F. Kurnaz and Y. Kaya, (2019) *"Prediction of the California bearing ratio (CBR) of compacted soils by using GMDH-type neural network"* **The European Physical Journal Plus *134***: 326. DOI: 10.1140/epjp/i2019-12692-0.

[16] S. M. Kassa and B. Z. Wubineh, (2023) *"Use of machine learning to predict california bearing ratio of soils"* **Advances in Civil Engineering *2023***: 8198648. DOI: 10.1155/2023/8198648.

[17] I. D. Mienye, Y. Sun, and Z. Wang, (2019) *"Prediction performance of improved decision tree-based algorithms: a review"* **Procedia Manufacturing *35***: 698–703. DOI: 10.1016/j.promfg.2019.06.011.

[18] S. B. Kotsiantis, (2013) *"Decision trees: a recent overview"* **Artificial Intelligence Review *39***: 261–283. DOI: 10.1007/s10462-011-9272-4.

[19] W. Zhang, C. Wu, H. Zhong, Y. Li, and L. Wang, (2021) *"Prediction of undrained shear strength using extreme gradient boosting and random forest based on Bayesian optimization"* **Geoscience Frontiers *12***: 469–477. DOI: 10.1016/j.gsf.2020.03.007.

[20] F. Livingston, (2005) *"Implementation of Breiman's random forest machine learning algorithm"* **ECE591Q Machine Learning Journal Paper *1***: 13.

[21] G. Biau and E. Scornet, (2016) *"A random forest guided tour"* **Test *25***: 197–227. DOI: 10.1007/s11749-016-0481-7.

[22] V. Y. Katte, S. M. Mfoyet, B. Manefouet, A. S. L. Wouatong, and L. A. Bezeng, (2019) *"Correlation of California bearing ratio (CBR) value with soil properties of road subgrade soil"* **Geotechnical and Geological Engineering *37***: 217–234. DOI: 10.1007/s10706-018-0604-x.

[23] C. C. Ikeagwuani, (2019) *"Optimisation of additives for expansive soil reinforcement"* **Unpublished PhD thesis**:

[24] G. Biau and E. Scornet, (2016) *"A random forest guided tour"* **Test *25***: 197–227. DOI: 10.1007/s11749-016-0481-7.

[25] W. Lin, Z. Wu, L. Lin, A. Wen, and J. Li, (2017) *"An ensemble random forest algorithm for insurance big data analysis"* **Ieee access *5***: 16568–16575. DOI: 10.1109/ACCESS.2017.2738069.

[26] A. D. Kulkarni and B. Lowe, (2016) *"Random forest algorithm for land cover classification"*:

[27] N. Mohapatra, K. Shreya, and A. Chinmay. "Optimization of the random forest algorithm". In: Springer, 2020, 201–208. DOI: 10.1007/978-981-15-0978-0_19.

[28] N. Khodadadi, V. Snasel, and S. Mirjalili, (2022) *"Dynamic arithmetic optimization algorithm for truss optimization under natural frequency constraints"* **IEEE Access *10***: 16188–16208. DOI: 10.1109/ACCESS.2022.3146374.

[29] L. Abualigah, A. Diabat, S. Mirjalili, M. A. Elaziz, and A. H. Gandomi, (2021) *"The arithmetic optimization algorithm"* **Computer methods in applied mechanics and engineering *376***: 113609. DOI: 10.1016/j.cma.2020.113609.

[30] S. Li, H. Chen, M. Wang, A. A. Heidari, and S. Mirjalili, (2020) *"Slime mould algorithm: A new method for stochastic optimization"* **Future generation computer systems *111***: 300–323. DOI: 10.1016/j.future.2020.03.055.

[31] M. Abdel-Basset, V. Chang, and R. Mohamed, (2020) *"HSMA_WOA: A hybrid novel Slime mould algorithm with whale optimization algorithm for tackling the image segmentation problem of chest X-ray images"* **Applied soft computing *95***: 106642. DOI: 10.1016/j.asoc.2020.106642.

[32] H. Chen, C. Li, M. Mafarja, A. A. Heidari, Y. Chen, and Z. Cai, (2023) *"Slime mould algorithm: a comprehensive review of recent variants and applications"* **International Journal of Systems Science *54***(1): 204–235. DOI: 10.1080/00207721.2022.2153635.

[33] L. Abualigah, D. Yousri, M. A. Elaziz, A. A. Ewees, M. A. A. Al-Qaness, and A. H. Gandomi, (2021) *"Aquila optimizer: a novel meta-heuristic optimization algorithm"* **Computers Industrial Engineering *157***: 107250. DOI: 10.1016/j.cie.2021.107250.

[34] A. M. AlRassas, M. A. A. Al-qaness, A. A. Ewees, S. Ren, M. A. Elaziz, R. Damaševičius, and T. Krilavičius, (2021) *"Optimized ANFIS model using Aquila Optimizer for oil production forecasting"* **Processes *9***: 1194. DOI: 10.3390/pr9071194.