**Geomechanics and Engineering**, *Vol. 36, No. 3 (2024) 259-276*
https://doi.org/10.12989/gae.2024.36.3.259


# Estimating pile setup parameter using XGBoost-based optimized models

Xigang Du[1], Ximeng Ma*[1], Chenxi Dong[1] and Mehrdad Sattari Nikkhoo[2]

[1]*Department of Transportation Engineering, Hebei University of Water Resources and Electric Engineering, Hebei Cangzhou,061000*
[2]*Department of Civil Engineering, University of Mohaghegh Ardabili, Ardabil, Iran*

(*Received November 5, 2023, Revised January 4, 2024, Accepted January 7, 2024*)

**Abstract.** The undrained shear strength is widely acknowledged as a fundamental mechanical property of soil and is considered a critical engineering parameter. In recent years, researchers have employed various methodologies to evaluate the shear strength of soil under undrained conditions. These methods encompass both numerical analyses and empirical techniques, such as the cone penetration test ($CPT$), to gain insights into the properties and behavior of soil. However, several of these methods rely on correlation assumptions, which can lead to inconsistent accuracy and precision. The study involved the development of innovative methods using extreme gradient boosting ($XGB$) to predict the pile set-up component "$A$" based on two distinct data sets. The first data set includes average modified cone point bearing capacity ($q_t$), average wall friction ($f_s$), and effective vertical stress ($\sigma_{vo}$), while the second data set comprises plasticity index (PI), soil undrained shear cohesion ($S_u$), and the over consolidation ratio (OCR). These data sets were utilized to develop $XGBoost$-based methods for predicting the pile set-up component "$A$". To optimize the internal hyperparameters of the $XGBoost$ model, four optimization algorithms were employed: Particle Swarm Optimization ($PSO$), Social Spider Optimization ($SSO$), Arithmetic Optimization Algorithm ($AOA$), and Sine Cosine Optimization Algorithm ($SCOA$). The results from the first data set indicate that the $XGBoost$ model optimized using the Arithmetic Optimization Algorithm ($XGB-AOA$) achieved the highest accuracy, with $R2$ values of 0.9962 for the training part and 0.9807 for the testing part. The performance of the developed models was further evaluated using the $RMSE$, $MAE$, and $VAF$ indices. The results revealed that the $XGBoost$ model optimized using $XGBoost-AOA$ outperformed other models in terms of accuracy, with $RMSE$, $MAE$, and $VAF$ values of 0.0078, 0.0015, and 99.6189 for the training part and 0.0141, 0.0112, and 98.0394 for the testing part, respectively. These findings suggest that $XGBoost-AOA$ is the most accurate model for predicting the pile set-up component.

**Keywords:** arithmetic optimization algorithm; cone penetration test; extreme gradient boosting model; particle swarm optimization; pile set-up parameter $A$; social spider optimization

## 1. Introduction

Conducting laboratory studies is directly related to progress in the field of geotechnical engineering (Dehghani and Larijani 2023, Esmaeili-choobar *et al*. 2013, Reihanisaransari *et al*. 2022, Rashidi Nasab and Elzarka 2023). Regardless of whether the soil is granular or non-granular, piles become stronger over time. The soil around the piles is disturbed radially and there is a lot of change in its shape. The extent of this disturbed region is proportional to the amount of soil that has been moved. The soil's effective shear power is also reduced by a substantial increase in excess pore water pressure ($EPWP$) (Esmaeili-Falak and Hajialilue-Bonab 2012, Esmaeili-Falak *et al*. 2018, Malmir *et al.* 2019, Jamil *et al.* 2023). So, the resistance of the piles diminishes. In the long run, disturbed soil's excess pore water pressure gradually breaks down after pile steer, and after that, stabilization and molding takes place again. It also improves the strength of the soil, thereby increasing the pile's resistance. The idiom "pile set-up" shows the increase in resistance of the piles in a variable state with time (Bullock 2008). It depends on several things, including: the specifications of the piles used, the soil characteristics in the desired region, and so forth, how long it will take to tune the piles. Vice versa, sometimes for high over-stabilized soils, negative pore water pressure could be continued where the surrounding soil is affected by the diminishing the general resistance causes to less pile resistance, which is named as “Relaxation” (Richardson 2011, Akbarzadeh *et al*. 2023).

### *1.1 Mechanism*

A re-building area has progressed around the pile while it’s being steered in the soil; this area allows the pile steering with the least amount of resistance, like when the volume of soil undergoes this. After re-building, a transferred area with new soil characteristics is created. There is notable additional pore water pressure in the re-building area (Haque 2015, Aghakhani *et al.* 2023, Kiannejad Amiri *et al.* 2023, Sadeghi *et al*. 2023). The additional $PWP$ vanishes over time and integrates the remolded area. The first method that will increase the power of the piles or help to install them is to re-assemble the molding area (Paikowsky and Chernauskas 1992). Due to soil ageing, the pile install does not end even after the entire

*Corresponding author, Ph.D.
E-mail: dsgmxm@126.com



ISSN: 2005-307X (Print), 2092-6219 (Online)

extended EPWP waste is gathered (Focht and Vijayvergiya 1972, Schmertmann 1991, Komurka *et al*. 2003). Adjusting the piles has three steps (Komurka et al. 2003).

Step 1: EPWP waste stage that is in a time and non-linear way,

Step 2: EPWP waste stage is in a time and linear way and

Step 3: Soil aging

In spite of this, soils at different depths along the pile shaft may go through different stages of building up due to uneven treatment.

### *1.2 Empirical models*

In addition to experimental and analytical approaches, numerical techniques are also used to assess the launching of piles (Ng *et al.* 2013, Rosti and Abu-Farsakh 2015). To make an estimate of the adjusted value, an experimental technique was developed (Wang and Reese 1989). This formula was according to a DLT analysis result of clay soil given from Shanghai,Pei, China. By applying this formula, an approximate calculation can be made as follows for pile power after time t (day of launching)

$$\frac{R_t}{R_{t0}} = 0.263[1 + \log(t)]R_{max} \quad (1)$$

In this formula, $t$ expresses the interval after setting up and $R_t$ expresses the initial power of the piles.

In the littoral zone of East China, 70 non-granular soils of test piles were analyzed for 20 different construction operations in order to evaluate the sensitivity $S_t$ related to pile adjustment (Guang-Yu 1988). He pointed out that growth of sensitivity $S_t$ increases the pile resistance re-take.

The Eq. (2) shows the outcomes appraisal of SLT found in this essay.

$$\frac{R_t}{R_{14}} = 0.375S_t + 1 \quad (2)$$

In spite of that, Gwizdała and Więcławski developed the formulas that are often used among the experimental formulas (Gwizdała and Więcławski 2013). They present a semi-logarithmic technique corresponding to data from four German and Danish file records. According to three kinds of soil (chalk, clay, and sand), the Eq. (3) was found.

$$\frac{R_t}{R_{t0}} = 1 + A\log_{10}\frac{t}{t_0} \quad (3)$$

According to the formula above, $R_t$ expresses the power of pile at a time t, $R_{t0}$ expresses the power of the pile at the primary time t0, t expresses passed time from the finishing time of steering, t0 expresses the reference time, and A expresses the adjust parameter.

In the early time, t0, is often known as the time after the steer has finished at that logarithmically linear pile adjust rate. So because of its relation to the size of the piles and the type of soil, it is difficult to assess. Generally, the huge diameter of a pile causes a larger t0 to come (Camp III and Parmar 1999). To accurately measure the value of t0, which is very difficult in reality, it is necessary to obtain the ultimate resistance of the pile at close and different distances. In consequence, to gain t0, the researchers used a returned calculation, made an assumption related to saved data, or maybe used an experimental formula. For PSC and H-piles, t0 was obtained as two days (Camp III and Parmar 1999). Although, they said that one day is a suitable amount. In non-granular soil, a similar t0 =1day was obtained for *PSC* piles (Axelsson 1998). Also, t0 =1 or 2 days were used (Svinkin *et al.* 1994). A lot of studies have also shown that t0 =1 day is appropriate (Bullock 1999, McVay *et al.* 1999). The pile adjust variable A used in eq. (3) related to ingredients like pile material, pile size, pile type, pile resistance, and soil type (Svinkin *et al*. 1994, Camp III and Parmar 1999, Haque *et al.* 2014). A = 0.2 and 0.6 was recommended for clay and sand (Skov and Denver 1988). The amount of clay had t0 including one day, while sand had a value of 0.5 days. That predicted value is apart from both the depth and excess PWP waste, though (Bullock 1999, McVay *et al*. 1999). The researchers had to perform repeated calculations using field data and experimental formulas to obtain the variable A and gain an independent hypothesis. The range rate 0.27-0.75 (Chow *et al.* 1998), within 0.2-0.8 (Axelsson 1998), and average value of 0.20 (Komurka *et al.* 2003) was suggested for A using fully accurate literature research.

### *1.3 Developed prediction models*

Nowadays, more than before, The artificial intelligence approach has been used in the field of civil engineering (Dehghan *et al*. 2023, Joushideh *et al*. 2023a, Khorshidi *et al*. 2023, Shi *et al*. 2023, Dawei *et al*. 2023), particular on the field of geotechnical engineering (Johari *et al.* 2011a, b, 2016, Nazoktabar *et al*. 2014, Anbari *et al*. 2020, Hashemi 2022, Sadeghi *et al*. 2022, Sarkhani Benemaran *et al*. 2022, Hashemi *et al*. 2023). Hence, by applying nonlinear artificial neural regressions, we may approximate the fundamental model of various materials at a considerable reduction in time and cost (Dehghani *et al*. 2020, Mahmoudabadi 2020, Faramarzi *et al*. 2021, Shakouri Mahmoudabadi 2021, Moqadam *et al.* 2022, Joushideh *et al.* 2023b, Larijani and Dehghani 2023, Moshtaghi Largani and Lee 2023). As well as, a number of studies have recorded using of artificially based techniques in many kind of engineering majors, such as additives in concrete, chloride diffusion in cement mortar (Stone 1974, Hoang *et al.* 2017, Motie *et al.* 2018, Ahmadi *et al.* 2020, Roudini *et al.* 2020, Murdoch *et al.* 2021, 2023b, a, Iraji *et al.* 2023), triaxial compressive strength and Young's modulus of frozen sand (Esmaeili-Falak *et al.* 2019), so on. According to the descriptions, types of artificial neural networks (ANNs) are used to assess soil attributes adapted from CPT, such as pending ANNs to measure soil not-drained shear power (Abu-Farsakh and Mojumder 2020), and final pile capacity and pile adjust variable A (Mojumder 2020). Many other hybrid techniques, like the random forest model with optimization methods and Harris hawk optimization (HHO) for synchronized energy management (Abdelsalam *et al.* 2021), creep index forecast according to PSO and RF(Zhang *et al*. 2020), landslide susceptibility mapping based on RF hyper-parameter optimization using bayes technique

(Sun *et al*. 2020), and grasshopper optimization algorithm (GOA) -based proceed forward to classify epileptic EEG signals (Singh *et al.* 2019), landslide susceptibility mapping (Sun *et al.* 2021a, b), permanence assessment of GFRPrebar (Iqbal *et al.* 2022), for heart diseases diagnosis (Asadi *et al.* 2021), prediction of proton-exchange membrane fuel cell (Huo *et al.* 2021), UHPFRC under ductility requirements for seismic retrofitting applications (Abellán-García and Guzmán-Guzmán 2021), smart meter data classification (Zakariazadeh 2022), rapid chloride permeability of self-compacting concrete (Ge *et al*. 2022), so on, have also been effectively and realistic used. In the most related studies, pile setup parameter has been predicted in the three papers. In the first one, optimized $GOA$ -$RF$ model has been successfully employed for this purpose, so that the value of $R^2$ obtained 0.918 (Zhao *et al.* 2022). In another study, $WOA$-$RF$ and $HHO$-$RF$ approaches have been developed to predict the pile setup parameter so that $R^2$ of the mentioned models obtained 0.87 and 0.89, respectively (Duan *et al.* 2022). In the last one which is the most recent study in this area, $PSO-RF$ has been utilized to forecast the pile setup parameter. The results showed the developed $PSO-RF$ model leads to the most accurate approach with $R^2$ of 0.964 (Dawei *et al.* 2023). Investigating the literature review demonstrates that, often conventional models like $ANN$ and $RF$ have been employed in this area, while in the present study, optimized $XGB$ approach has been extended to predict the pile setup parameter.

### *1.4 Main objective of study*

The main objective of the present study is to develop a novel hybrid XGB − based approach such as XGB − SCA, XGB − PSO, $XGB-SSA$ and $XGB-AOA$ to predict the pile-setup variable (A) using CPT variables in which 2 various kinds of data for training and testing datasets. Dataset type I includes average modified cone point bearing capacity ($q_t$), average wall friction ($f_s$), and effective vertical stress ($\sigma_{vo}$), and dataset type II consists of plasticity-Index ($PI$), undrained shear cohesion of soil ($S_u$), and the ratio of over-consolidation ($OCR$). The novelty perspective of the present study is that 4 XGB- based hybrid approaches were primarily proposed as an approach to estimate pile-setup parameters from field test results. For this purpose, cone penetration test dataset and corresponding borehole data were collected from seventy multiple positions from the Louisiana state. Statistical assessment diversity is considered while assessment of the successfulness of the developed approaches.

## 2. Methodology

### *2.1 Dataset description*

To estimate the pile adjustment variable from the CPT, 70 CPT data sets from various Louisiana positions were studied, as shown in Fig. 1 (Abu-Farsakh and Mojumder 2020, Mojumder 2020). These studies done the early experimental experiments on samples of soil like grain size distribution (ASTM D422-63 2007), water content (ASTM D422-63 2007), Atterberg limits (ASTM D4318-17e1 2018), and unit weight (ASTM D7263-21 2021) and in-situ CPT testing at mentioned positions in Louisiana.

Undistributed Shelby tube samples corrected from various borehole depths at every position were part of the experimental experiment plan. Triaxial unconsolidated undrained experiments were done in correspondence to ASTM-D 2850-03a to assess the soil's not-drained shear power (ASTM D2850-03 2017). CPT (ASTM D3441-16 2018) and also RCPT In the first place were used in the in-situ testing that were done surrounding the boreholes that were bored using 10 cm2 and 15 cm2 piezocone penetrometers. If the RCPT also assessed excess pore pressure behind the cone (u2) besides the prior defined factors, the cone penetration experiment may assess cone tip resistance (qc) and cone sleeve friction (fs). At a constant space of 2 cm in-depth, recordings were given whereas the cone was being excavated in a constant infiltrate level of 2cm/sec.

The parameters used in the input data of the model have a considerable impact on creating the desired model. In order to optimize the model's performance, it is important to have the right variables selected. Increasing the number of variables will increase the size of the network, which will affect its speed. A look at the effect of different parameters on variable A has been conducted in this article by using two types of parameters that are correlated with the literature. Chosen variables for data type 1 were mean modified cone tip resistanceqt, mean skin friction ((fs)=, and efficient overburden pressure (σvo), and for data type 2 were plasticity index (PI), not-drained soil's shear power (Su), and over consolidation ratio (OCR) for train models to recognize the network that produces the huge results. In this study, the pile adjusted variable A was the result, with values for data types 1 and 2 ranging from 0.07 to 0.53 and from 0.12 to 0.53, respectively. The power of the various layers was assessed according to the strain gauge measurements of the load diffusion while the static load experiments. Then, the adjust space variable A was calculated using Eq. (3) where the resistance of the pile (Rt) at certain time (t) and the resistance of the pile (Rt0) at the beginning time (t0) had already been measured (Skov and Denver 1988). The value of A in this research is seriously related to the factors of the soil. CPT technique, which is very efficient and effective, has been used to measure soil resistance during research. Having determined the data set to be used, these input data are evaluated and refined to determine the best set of data to be used to predict variable A.

Randomizing the order of the data can help reduce bias in the model. If the data is not randomized, the model might learn patterns based on the order of the data, leading to biased results. Randomizing the data ensures that the model is exposed to a diverse range of examples during training. This helps the model generalize better to new, unseen data. Randomizing the data can prevent the model from overfitting to specific patterns in the data. Overfitting occurs when the model learns to perform well on the


Fig. 1 Distribution of CPT locations in the Louisiana (Abu-Farsakh and Mojumder 2020)

Table 1 The statistical characteristics of the inputs and outputs variables

| Category | Property | Data type 1 | | | | Data type 2 | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | *Inputs* | | | Output | *Inputs* | | | Output |
| | | $q_t$ | $f_s$ | $\sigma_{vo}$ | $A$ | $PI$ | $S_u$ | $OCR$ | $A$ |
| | | $tsf$* | $tsf$ | $tsf$ | – | % | $tsf$ | – | – |
| Train data | $Min$ | 1.52 | 0.04 | 0.68 | 0.07 | 4 | 0.07 | 1 | 0.12 |
| | $Max$ | 173.33 | 3.74 | 9.15 | 0.53 | 84 | 1.57 | 3 | 0.53 |
| | $St.deviation$ | 28.690 | 0.616 | 2.099 | 0.128 | 20.952 | 0.461 | 0.651 | 0.124 |
| | $Median$ | 14.89 | 0.46 | 2.915 | 0.34 | 34 | 0.54 | 1 | 0.29 |
| | $Average$ | 23.647 | 0.653 | 3.583 | 0.303 | 38.481 | 0.716 | 1.477 | 0.289 |
| | $Skewness$ | 3.529 | 2.766 | 0.726 | -0.162 | 0.463 | 0.293 | 1.066 | 0.205 |
| Test data | $Min$ | 5.64 | 0.1 | 0.55 | 0.08 | 16 | 0.07 | 1 | 0.15 |
| | $Max$ | 174.62 | 3.74 | 10.27 | 0.48 | 84 | 1.45 | 3.05 | 0.53 |
| | $St.deviation$ | 44.646 | 0.892 | 2.651 | 0.103 | 18.180 | 0.389 | 0.845 | 0.095 |
| | $Median$ | 23.6 | 1.12 | 2.21 | 0.25 | 37 | 0.79 | 1.77 | 0.27 |
| | $Average$ | 39.506 | 1.120 | 3.24 | 0.245 | 39.412 | 0.785 | 1.889 | 0.288 |
| | $Skewness$ | 2.163 | 1.665 | 1.509 | 0.694 | 0.951 | -0.082 | 0.266 | 0.861 |

* $tsf$ = Tons Force per Square Foot

training data but fails to generalize to new data. By randomizing the data, the model becomes more robust and less sensitive to the order or structure of the input data. Randomizing data is essential for proper cross-validation and performance evaluation of the model. It ensures that each fold of the data contains a representative sample of the entire dataset. Overall, randomizing data before training AI models helps improve model performance, reduce bias, and enhance generalization to new data. For this purpose, $Rand\_perm$ code has been employed in MATLAB software.

Training data (used to expand) and testing data (used to determine when the training should be stopped) are collected while the divide step is taking place. Various data set ratios in between these sub-groups were used (Stone 1974, Hammerstrom 1993, Looney 1996). In this paper, the data collected from the Louisiana region were divided into two parts where 25% of them were grouped for testing and

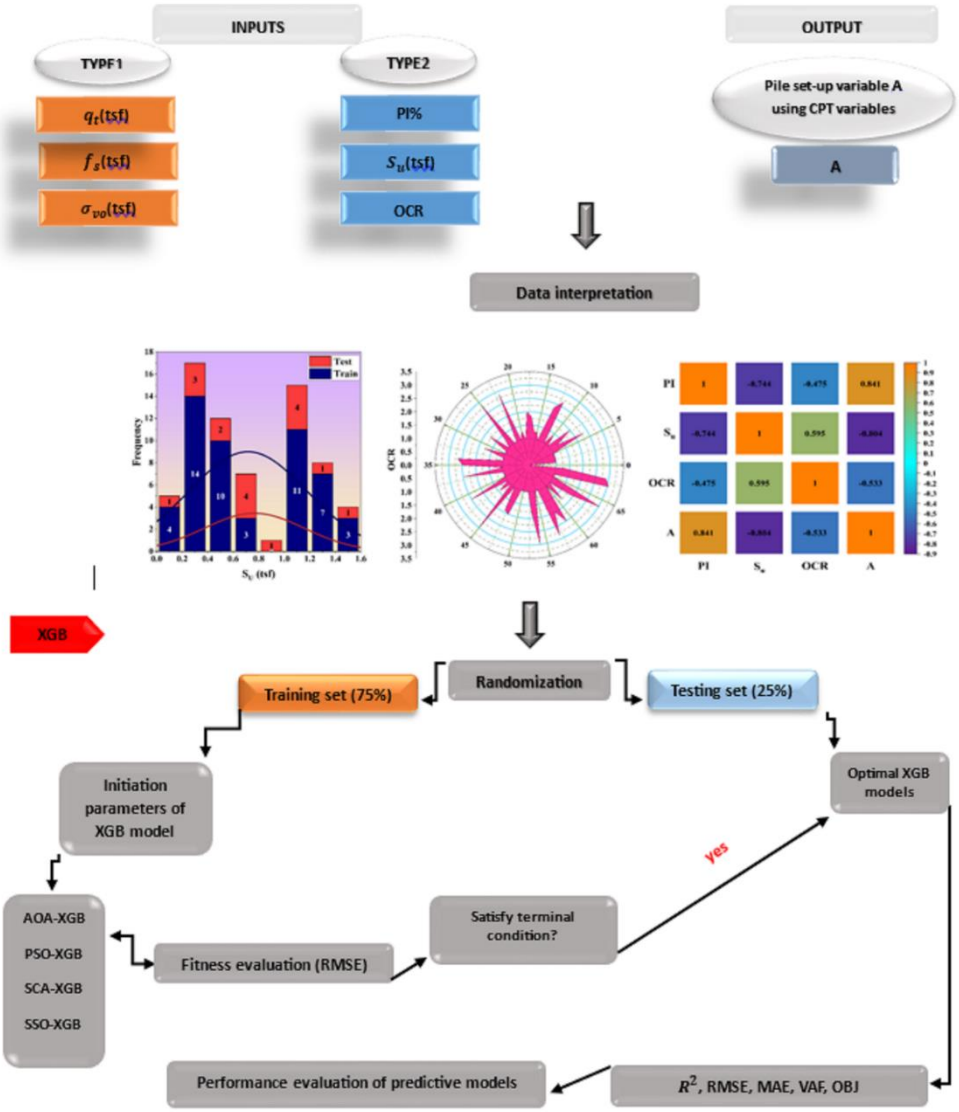



Fig. 2 Research methodology flowchart

75% for training. The statistical value of input variables as well as output variables for training (75%) and testing (25%) data phases are presented in Table (Zhang et al. 2020).

Research methodology flowchart is presented in Fig. 2. All of the records available techniques are placed in the training data because it was given in this way. When a large amount of training data is removed, which causes unrelated states in the test data, instead of checking the model's interpolation power, check its extrapolation capacity. The statistical analysis of the input and output data used for the expansion of the model is shown in Table 1. In addition, Figs. 3 and 4 illustrate the histograms of inputs and outputs by their normal distributions.

Figs. 3(a) to 3(d) demonstrate the histogram and distribution curve for $q_t$, $f_s$, $\sigma_{vo}$ and $A$, respectively, in which the first 3 parameters are related to the input data and the fourth parameter is related to the output data. As can be seen there is no considerable discreteness of data series I as well as reasonable dividing between training and testing dataset. Distribution line of each parameter in the histograms are similar to a normal distribution line which may lead well-convergence of the developed models. Likewise Figs. 4(a) to 4(d) illustrate the $PI$, $S_u$, $OCR$ and $A$, respectively, in which the first three parameters are related to the input data and the fourth parameter is related to the output data. Distribution line of each variable in the histograms are equivalent to a normal distribution line which may lead well-convergence of the developed models.

In order to understand and accurately interpret the distribution of data, the 360-degree distribution circular diagrams for data type I are shown in Figs. 5(a) to 5(d). Correspondingly, the 360-degree distribution circular diagrams for data type II are shown in Figs. 6(a) to 6(d). 360-degree distribution circular diagrams show that each parameter have a reasonable distribution which may lead well-convergence of the developed models.

Then, the relation coefficients among the variables under consideration were calculated and compiled in Figs. 6(a) and 6(b) for data types I and II, respectively. A large

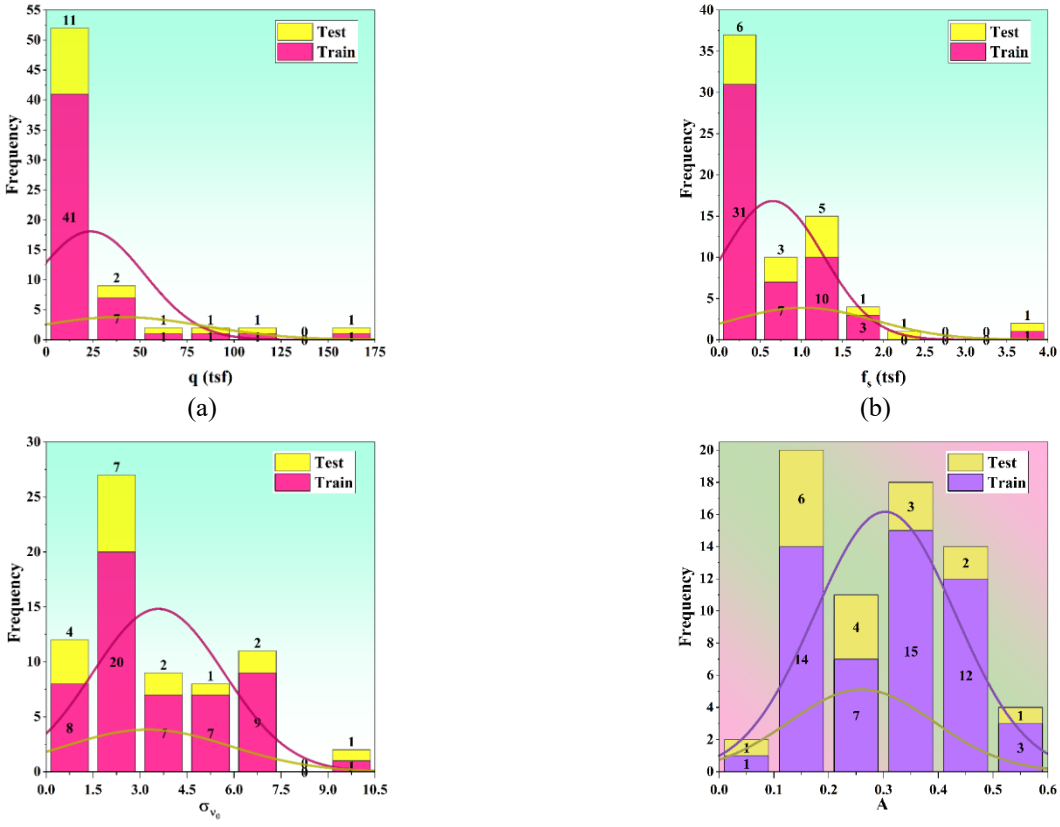



Fig. 3 Train and test histogram of data type I; (a) $q_t$, (b) $f_s$, (c) $\sigma_{vo}$ and (d) $A$ (with their normal distribution)

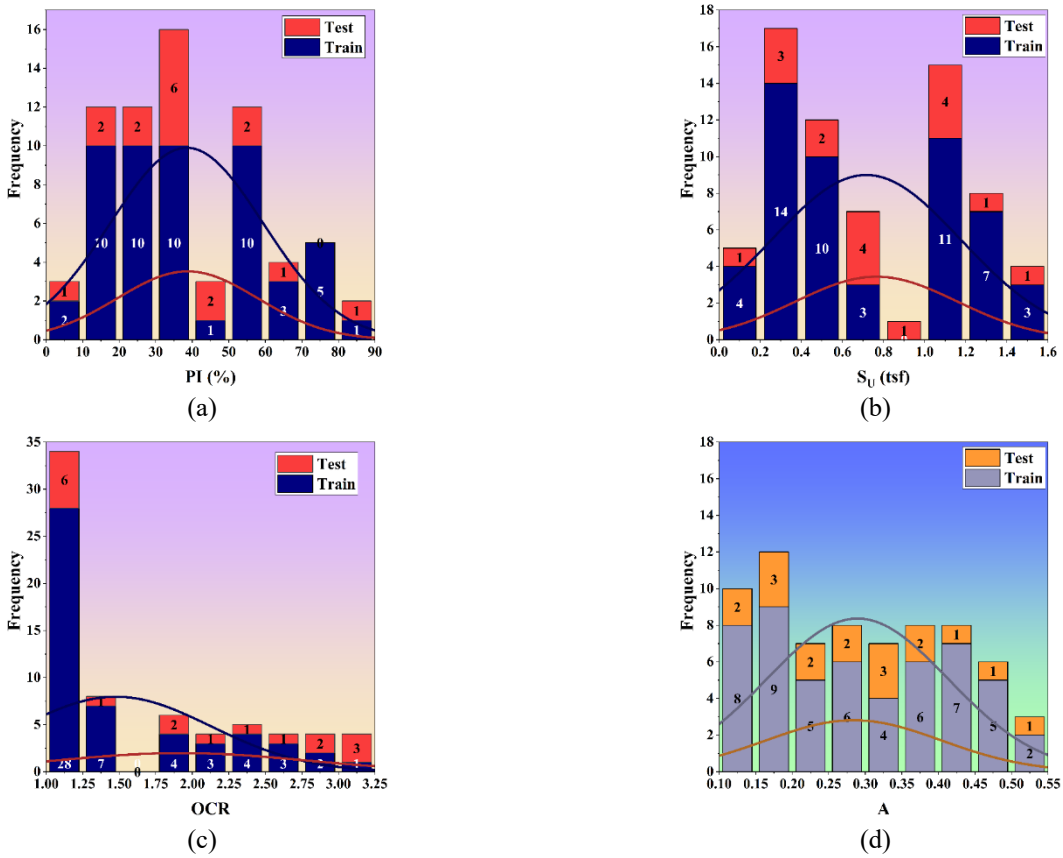
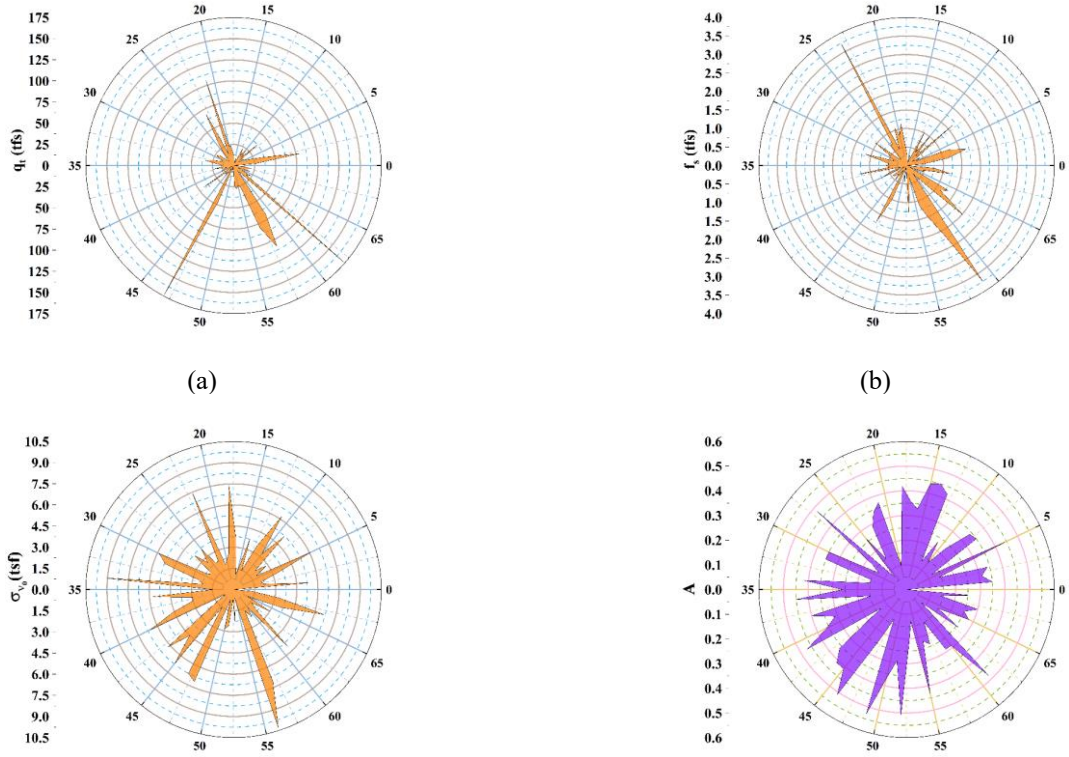


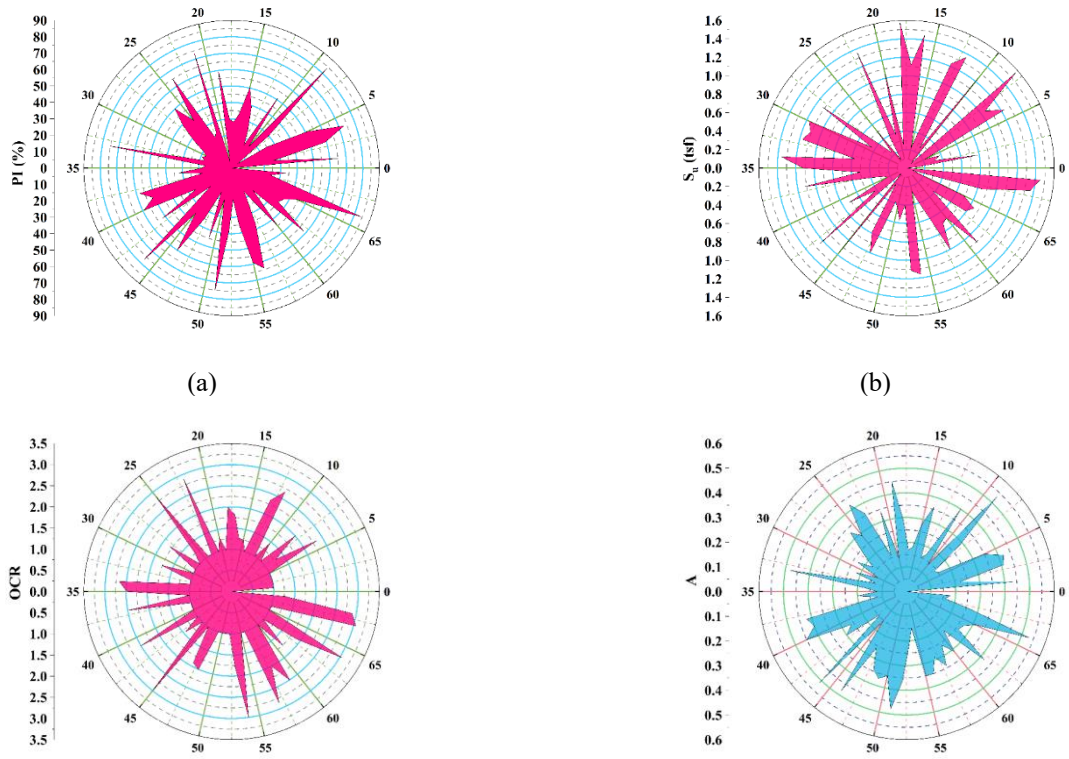

Fig. 4 Train and test histogram of data type II; (a) $PI$, (b) $S_u$, (c) $OCR$ and (d) $A$ (with their normal distribution)




(a) (b)

(c) (d)

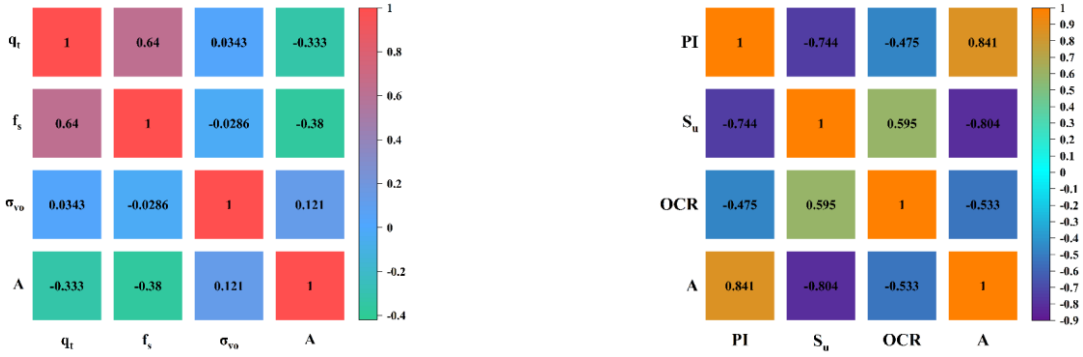
Fig. 5 Circular distribution of data type 1 for training and testing data set; (a) $q_t$, (b) $f_s$, (c) $\sigma_{vo}$ and (d) $A$




(a) (b)

(c) (d)

Fig. 6 Circular distribution of data type 1 for training and testing data set; (a) $PI$, (b) $S_u$, (c) $OCR$ and (d) $A$




(a) (b)

Fig. 7 Pearson correlation coefficient heat-map for: (a) data category I and (b) data category II

positive or negative correlation coefficient between the parameters may indicate that the procedures have a low yield and that it is difficult to discern how the explanatory factors affect the result. As it can be obtained from Fig. 7(a) for data type I, there are moderate positive correlations between $f_s$ and $q_t$. $A$ and $f_s$ also have a moderately negative correlation. Additionally, there is no remarkable linear link among $\sigma_{vo}$ or any of the parameters. Turning to data type 2 (Fig. 7(b)), there are moderate correlations between $OCR$ and $PI$, as well as between $OCR$ and $S_u$. Moreover, an almost high correlation is depicted between $A$ with $PI$ and $S_u$.

### 2.2 Extreme gradient boosting ($XGB$)

In XGB, gradient boosting trees are used as a hybrid algorithm (Chen *et al.* 2015). Collection algorithms are reinforced using gradient reinforcement of agent algorithms. A very strong efficiency of the XGB algorithm is its gradient reinforcing. A dense graph with excessive GB is similar to a gradient-reinforced tree according to the retrogression and categorization method (Le *et al.* 2019, Yang *et al*. 2019, Zhou *et al*. 2019, Zhang *et al.* 2021). In order to analyze the data collection in multiple ways, it could summarize the modeling results from weak assessments. While simultaneously addressing categorization and regression errors, the XGB technique effectively produced a large output in place of multiple separate outputs (Zhou *et al.* 2019). The new algorithm could be described using the gradient reinforcement tree approach using a regular computational library.

In the hybrid XGB technique, the fit conditions are diminished with multiple parts to generate the desired function (Chen and Guestrin 2016). The first part is used to obtain the difference between the predicted and measured values, and the second part is used for adjustment. The prediction authenticity of the algorithm could be measured by the deviation and the type of algorithm used. D = {(Xi, Yi)} is a series of data including m properties, n samples, and the predictor is an excess system created of k basic systems. Samples predicting outcomes could be shown as

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in \varphi \tag{4}$$

$$\varphi = \{f(x) = w_s(x)\}(s: R^m \to T, w_s \in R^T) \tag{5}$$

$x_i$ : The specimen
$f_k(x_i)$ : Estimation concession for a particular sample
$\varphi$ : The set of regression trees
$s$ : principle parameter
$w$ : weighted leaf
$T$ : number of leaf
$K$ : Number of trees
$\hat{y}_i$ : The forecasted label

According to XGB , the algorithm's performance technique needs to be calculated with the greatest degree of difficulty in order to obtain the least loss in a collection tree, in other words optimized performance. According to this statement, the efficiency obtained includes the efficiency of loss and the hardness of the model.

$$Obj^{(t)} = \sum_{i=1}^{m} l\left(y_i^t, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \sum_{k=1}^{t} \Omega(f_k) \tag{6}$$

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \|w\|^2 \tag{7}$$

In this equation, i represents the number of samples in the database and m represents all the data entered in the kth tree. The first part of Eq. (7) gives the difference between the predicted and measured values and also shows the normal loss function. The second part of Eq. (7) shows the degree of difficulty of the mentioned algorithm. In addition, γ and λ are variables which could diminish the degree of difficulty of the tree, also, by adjusting the final weights, the arrangement point prevents over-compatibility.

And finally, to simplify the goal term, Taylor's law is used

$$Obj^{(t)} = \sum_{i=1}^{m}\left[f_t(x_i)g_i + \frac{1}{2}\left(f_t(x_i)\right)^2 h_i\right] + \gamma T + 0.5\lambda\sum_{j=1}^{T} w_j^2 \quad (8)$$

That $gi$ and $hi$ state the first and second derivatives gained on the loss function, respectively.

### *2.3 Sine–cosine algorithm* ($SCA$)

After a set of solutions is randomly collected, the SCA approach is a population technique to achieve the best result. In performing equations, defining sufficient solutions increases the likelihood of achieving outstanding results, and finally selecting the most appropriate result. As the exploration and exploitation phases of the algorithm's computing technique progress, the location of the solution is updated according to the given rules (Mirjalili 2016).

$$X_i^{t+1} = \begin{cases} X_i^t + r_1 \times \sin r_2 \times \left|r_3 P_i^t - X_i^t\right| & r_4 < 0.5 \\ X_i^t + r_1 \times \cos r_2 \times \left|r_3 P_i^t - X_i^t\right| & r_4 \geq 0.5 \end{cases} \quad (9)$$

Where $X^{t+1}$ expresses the solution's location in the $i_{th}$ dimension and $t_{th}$ repetition, and $P_t$ indicates the target point in the $i_{th}$ dimension. Also, $r1$ to $r4$ shows the random values.

This optimization strategy accounts for the spread of sine and cosine by using the formula below to balance exploration and exploitation (Mirjalili 2016).

$$r_1 = a - t\frac{a}{T} \quad (10)$$

That t indicates the repetitions' maximum number, T expresses the present repetition and an Abbreviation for a constant.

### *2.4 Particle swarm optimization* ($PSO$)

A particle-based method for optimizing a specific subject is actually a computational method (Kennedy and Eberhart 1995). Every position vector of a particle is expressed by $X_i^k$, speed vector by $V_i^k$, and fit value, that i and k are the existing output and the $i_{th}$ particle. Based on their respective premiere locations and local speeds, particles in the exploration region are directed to global premier locations. Hence, the location of the optimal value of fit depends on the smallest value of fit (Xue 2018). Eqs. (11) and (12) update the speed of every particle.

$$V_i^{k+1} = \omega V_i^k + c_1 r_1 P_i^k - X_i^k + c_2 r_2 P_g^k - X_i^k \quad (11)$$

$$X_i^{k+1} = X_i^k + V_i^{k+1} \quad (12)$$

$C_1$ & $C_2$ = Coefficients of Acceleration
ω = Inertia weight =1
$r_1$ & $r_2$ = Accidental members (0,1)
$P_i$ = The best present location of $i_{th}$ particles
$P_j$ = The best global

The lowest and highest values of $V_i^k$ are -1 and 1, respectively. In this technique, parameters of ω, c1 and c2 must be corrected, because their values affect the isotopy speed. This increased optimization performance can be achieved by choosing different PSO parameters in the literature, so that the performance of this optimization could be improved by using variables (Shi and Eberhart 1998, Trelea 2003, Khoshaim *et al*. 2021).

### *2.5 Social spider optimization* ($SSO$)

A collaborative approach to treating social spiders has led to $SSO$. Both genders female and male spiders are shown by the optimization technique (Cuevas *et al*. 2013).

The social spider community consists of two parts: its community network and its members. According to the different spiders' genders, all members are divided into two variable groups and every factor is managed by a group of different operators to emulate the collaborative treatment in the group. They should be divided into categories of effective and ineffective people of the male gender. Spiders that are in the effective category show better adaptation than ineffective spiders. In the general network, these spiders are captured by the nearest female spider. Also, ineffective male spiders concentrate on using resources found by effective male spiders. Depending on the fit value of the solution represented by a social spider, every spider will carry a certain amount of weight

$$w_t = \frac{fitness_t - Worst}{Best - Worst} \quad (13)$$

The proportional value of fit(t) was obtained by appraising the location of the $t^{th}$ spider, t=1,2,…T There are two values in the population: best and worst.

Social spiders interact with each other in a general web based on SSO assumptions. Every solution in the exploration region represents the spider's location in the general network.

### *2.6 Arithmetic optimization algorithm* ($AOA$)

In 2021, Abulaigah *et al*. Published the $AOA$, a math operator in science that was motivated by (means Division, Multiplication, Addition, and Subtraction) (Abualigah *et al*. 2021). In the continuation of the discussion, you can see the $AOA$ search terms. The Pseudo code of the original $AOA$ is shown.

The matrix in Eq. (14) includes the primitive answers to the $AOA$.

$$\mathrm{X_i} = \begin{bmatrix} \mathrm{x_1^1} & \mathrm{x_1^1} & \cdots & \mathrm{x_{1,D}^1} \\ \mathrm{x_1^2} & \mathrm{x_2^2} & \cdots & \mathrm{x_D^2} \\ \vdots & \ddots & \ddots & \vdots \\ \mathrm{x_1^N} & \mathrm{x_D^N} & \cdots & \mathrm{x_D^N} \end{bmatrix} \quad (14)$$

The Math Optimizer Accelerated (MOA) is used in Eq. (15) to select the searching operator for all repetition.

$$\mathrm{MOA(C_{Iter})} = \mathrm{Min} + \mathrm{C_{Iter}} \times \left(\frac{\mathrm{Max} - \mathrm{Min}}{\mathrm{M_{Iter}}}\right) \quad (15)$$

Using these two values as a comparison, the Min and Max values of the accelerated function can be determined. That $M_{Iter}$ is all of the repetitions and MOA($C_{Iter}$) is the MOA defined for every repetition. The present repetition is $C_{Iter}$.

**Exploration level:**

In the exploration phase of AOA, different locations are investigated using two search methods ($D$ and $M$) which are mentioned in Eq. (16). This phase (M) is allowed while the accidental number r1 > MOA; otherwise, the other operator (M) is allowed.

$$x_{i,j}(C_{Iter}+1) = \begin{cases} bset(x_j) \div (MOP+\varepsilon) \times \left((UB_j - LB_j) \times \mu + LB_j\right), & r_2 < 0.5 \\ best(x_j) \times MOP \times \left((UB_j - LB_j) \times \mu + LB_j\right), & otherwise \end{cases} \quad (16)$$

In present equation, best ($x_j$) defined the $j_{th}$ location in the best gained location in the $i_{th}$ answer, μ shows a control value equal to 0.5, and $x_{i,j}$ ($C_{Iter}$ + 1) show the $j_{th}$ location of the $i_{th}$ answer (Premkumar *et al*. 2021).

$$MOP(C_{Iter}) = 1 - \frac{C_{Iter}^{\left(\frac{1}{\alpha}\right)}}{M_{Iter}^{\left(\frac{1}{\alpha}\right)}} \quad (17)$$

According to this equation, α shows a control value equal to 0.5, MOP($C_{Iter}$) indicates the MOP's value at the $t_{th}$ repetition.

**Exploitation level:**

Due to the Eq. (18), the S operator is using while r3 < 0.5 and the A operator using in other cases.

$$x_{i,j}(C_{Iter}+1) = \begin{cases} best(x_j) - MOP \times \left((UB_j - LB_j) \times \mu + LB_j\right), & r_3 < 0.5 \\ best(x_j) + MOP \times \left((UB_j - LB_j) \times mu + LB_j\right), & otherwise \end{cases} \quad (18)$$

1. ***Initialization level***
2. *Initialize the algorithm parameters: N, Min, Max, $M_{iter}$, μ, **α***
3. *Generate the initial population of solution randomly*
4. *Identify the best solution of initial population*
5. $C_{iter} = 1$
6. ***While** ($C_{iter} \leq M_{iter}$) **do***
7. $MOA(C_{iter}) = Min + C_{iter} * ((Max-Min)/ M_{iter})$
8. $MOP(C_{iter}) = 1 - (\frac{C_{iter}}{M_{iter}})^{\frac{1}{\alpha}}$
9. ***for** i = 1 to N **do***
10. ***for** j = 1 to n **do***
11. ***if** $MOA(C_{iter}) < r_1$ **then***
12. ***if** $r_2$>0.5 **then (Exploration level)***
13. $x_{i,j} = x_{best,j} / (MOP* (UB_j - LB_j) * \mu) + LB_j)$
14. ***else***
15. $x_{i,j} = x_{best,j} / (MOP* (UB_j - LB_j) * \mu) + LB_j)$
16. ***end if***
17. ***else***
18. ***if** $r_3$> 0.5 **then (Exploitation level)***
19. $x_{i,j} = x_{best,j} / (MOP* (UB_j - LB_j) * \mu) + LB_j)$
20. ***else***
21. $x_{i,j} = x_{best,j} / (MOP* (UB_j - LB_j) * \mu) + LB_j)$
22. ***end if***
23. ***end if***
24. ***end for***
25. *Enforce design variable constraints*
26. *if $p(\acute{x}_i) < p(x_i)$ **then***
27. $x_i = \acute{x}_i$
28. ***end if***
29. *Update the best solution found so far: $x_{best}$*
30. ***end for***
31. $C_{iter} = C_{iter} + 1$
32. ***end while***

### 2.7 Model evaluator indices

Model evaluation and rationalization are the primary levels of use. After the required model is built, it is necessary to determine whether the results are useful and sufficient for the purposes of prediction and simulation now. To evaluate the accuracy of the approaches used, the performance of training and testing data is examined. For the purpose of evaluating the model's accuracy and efficiency, six statistical criteria have been used. They are the coefficient of determination ($R^2$), root mean square error (RMSE), mean absolute error (MAE) and the variance accounted factor (VAF). There is a greater correlation between low RMSE andMAE and a larger , indicating more reliable statistical analysis. The larger $R^2$ indicates a stronger correlation between observed and simulated data.

$$R^2 = \left( \frac{\sum_{p=1}^{P}(t_P - \bar{t})(y_P - \bar{y})}{\sqrt{\left[\sum_{p=1}^{P}(t_P - \bar{t})^2\right]\left[\sum_{p=1}^{P}(y_P - \bar{y})^2\right]}} \right)^2 \quad (19)$$

$$RMSE = \sqrt{\frac{1}{P}\sum_{p=1}^{P}(y_p - t_p)^2} \quad (20)$$

$$MAE = \frac{1}{P}\sum_{p=1}^{P}|y_p - t_p| \quad (21)$$

$$VAF = \left(1 - \frac{var(t_P - y_P)}{var(t_P)}\right) * 100 \quad (22)$$

$$OBJ = \frac{n_{train}}{N} * \left(\frac{RMSE_{train} + MAE_{train}}{R^2{}_{train} + 1}\right) + \frac{n_{test}}{N} * \left(\frac{RMSE_{test} + MAE_{test}}{R^2{}_{test} + 1}\right) \quad (23)$$

So the $y_P$ & $t_P$ are the simulated and observed values, $\bar{t}$ & $\bar{y}$ are the mean of the observed and simulated values, respectively. Also, M indicates the specimen number, and M10 is the number of records with a ratio of measured to predicted value among 0.9 and 1.1.

To develop the mentioned optimized models, Matlab software, R2020B ver. 9.9 has been employed. Convergence curve against R2 is illustrated in Fig. 8.

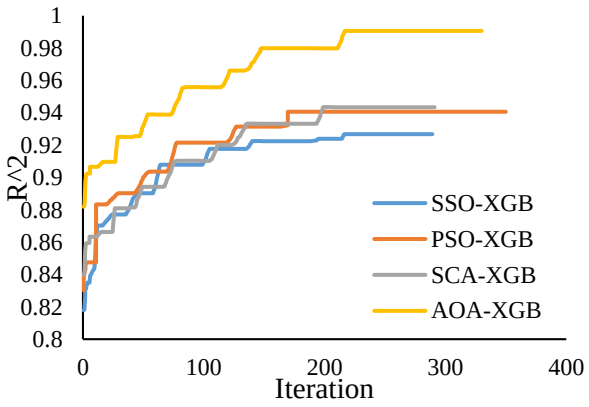



Fig. 8 Adjusting hyper-parameters utilizing hybrid models

Table 2 hybrid $XGB$ based approaches' performance evaluation to estimate the $A$ parameter

| Model | $AOA-XGB$ | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | *Training data* | | | | *Testing data* | | | | |
| | $R^2$ | $RMSE$ | $MAE$ | $VAF$ (%) | $R^2$ | $RMSE$ | $MAE$ | $VAF$ (%) | $OBJ$ |
| Data type I | 0.9962 | 0.0078 | 0.0015 | 99.6189 | 0.9807 | 0.0141 | 0.0112 | 98.0394 | 0.0067 |
| Data type II | 0.9941 | 0.0105 | 0.0074 | 99.2795 | 0.982 | 0.0137 | 0.0126 | 97.8834 | 0.01 |
| Model | $PSO-XGB$ | | | | | | | | |
| | *Training data* | | | | *Testing data* | | | | |
| | $R^2$ | $RMSE$ | $MAE$ | $VAF$ (%) | $R^2$ | $RMSE$ | $MAE$ | $VAF$ (%) | $OBJ$ |
| Data type I | 0.9811 | 0.0195 | 0.0166 | 97.6494 | 0.9745 | 0.0166 | 0.0129 | 97.222 | 0.0174 |
| Data type II | 0.9857 | 0.0196 | 0.0241 | 97.4588 | 0.9712 | 0.0173 | 0.0151 | 96.5321 | 0.0178 |
| Model | $SCA-XGB$ | | | | | | | | |
| | *Training data* | | | | *Testing data* | | | | |
| | $R^2$ | $RMSE$ | $MAE$ | $VAF$ (%) | $R^2$ | $RMSE$ | $MAE$ | $VAF$ (%) | $OBJ$ |
| Data type I | 0.967 | 0.0245 | 0.021 | 96.631 | 0.962 | 0.0231 | 0.0214 | 94.8507 | 0.0473 |
| Data type II | 0.9618 | 0.0241 | 0.017 | 96.1663 | 0.9621 | 0.027 | 0.0198 | 95.5134 | 0.0217 |
| Model | $SSO-XGB$ | | | | | | | | |
| | *Training data* | | | | *Testing data* | | | | |
| | $R^2$ | $RMSE$ | $MAE$ | $VAF$ (%) | $R^2$ | $RMSE$ | $MAE$ | $VAF$ (%) | $OBJ$ |
| Data type I | 0.9365 | 0.0335 | 0.0277 | 93.65 | 0.8967 | 0.0351 | 0.027 | 87.7283 | 0.0319 |
| Data type II | 0.9331 | 0.033 | 0.0243 | 92.9523 | 0.8934 | 0.0309 | 0.0251 | 89.3354 | 0.0296 |

## 3. Results and discussion

### *3.1 Predictive capability of the models*

The results of XGB-based hybrid approaches which have been employed to predict "A" parameter, are presented in this section. As mentioned, four optimization algorithms ($AOA$,$PSO$, $SCA$, and $SSO$) have been utilized for tuning the internal hyper-$parameters$ of the XGB approach. In order to achieve a comprehensive evaluation of the performance and accuracy of the developed models, five evaluation indexes $R^2$, $RMSE$, $MAE$, $VAF$ and $OBJ$ have been used to ensure the accuracy of the predicted values.

For data type I, considering the training dataset, the amounts of $R^2$ for $XGB-AOA$, $XGB-PSO$, $XGB-SCA$ and $XGB-SSO$approaches are 0.9962, 0.9811, 0.967 and 0.9365, respectively. Similarly, these values for the testing dataset are 0.9807, 0.9745, 0.962 and 0.8967, respectively. For data type II, considering the training dataset, the amounts of $R^2$ for $XGB-AOA$, $XGB-PSO$, $XGB-SCA$, $XGB-SSO$ approaches are 0.9941, 0.9857, 0.9618 and 0.9331, respectively. Similarly, these values for the testing dataset are 0.982, 0.9712, 0.9621 and 0.8934, respectively. Consequently, considering both data type I and II, the ranking of models based on the $R^2$ index are $XGB-AOA$, $XGB-PSO$ , $XGB-SCA$ and $XGB-SSO$ models, respectively, for training and testing data.

Considering the fact that R2 is a linear and conventional index, it cannot reliably be employed. Another index that is more popular among data scientists and has a high ability to rank AI-based models is $OBJ$. Considering the data type I, the amounts of $OBJ$ for $XGB-AOA$, $XGB-PSO$, $XGB-SCA$ , $XGB-SSO$ approaches are 0.011003, 0.0178, 0.02167, 0.0472 and 0.02963, respectively. Similarly, these values for data type II are 0.0067, 0.0174, 0.0472 and 0.0319, respectively.

In order to better understand the difference in accuracy and performance of the models developed in this study, the measured values of parameter A against the values predicted by these models are shown in Figs. 9 and 10. Figs. 7 $A1$ to $A4$ show the graph of the measured values of parameter A against the predicted values for $XGB-AOA$, $XGB-PSO$, $XGB-SCA$ and $XGB-SSO$ models considering data type I for both training and testing dataset with different legends, respectively. As can be seen, for all models, the linear fit of training prediction has better agreement prediction with the best fitted line in comparison to the linear fit of testing prediction. According to Figs. 9, the XGB − AOA model provides the highest accuracy for both training and testing steps.

Figs. 10(a) to 10(d) illustrate the graph of the measured amount of parameter A against the predicted amounts for the $XGB-AOA$ , $XGB-PSO$ , $XGB-SCA$ and $XGB-SSO$ approaches considering data type II for both training and testing dataset with different legends, respectively. As can be seen, for all approaches, the linear fit of training prediction has better agreement prediction with the best fitted line in comparison to the linear fit of testing prediction. According to Figs. 10, the XGB − AOA model provides the highest accuracy for both training and testing steps.

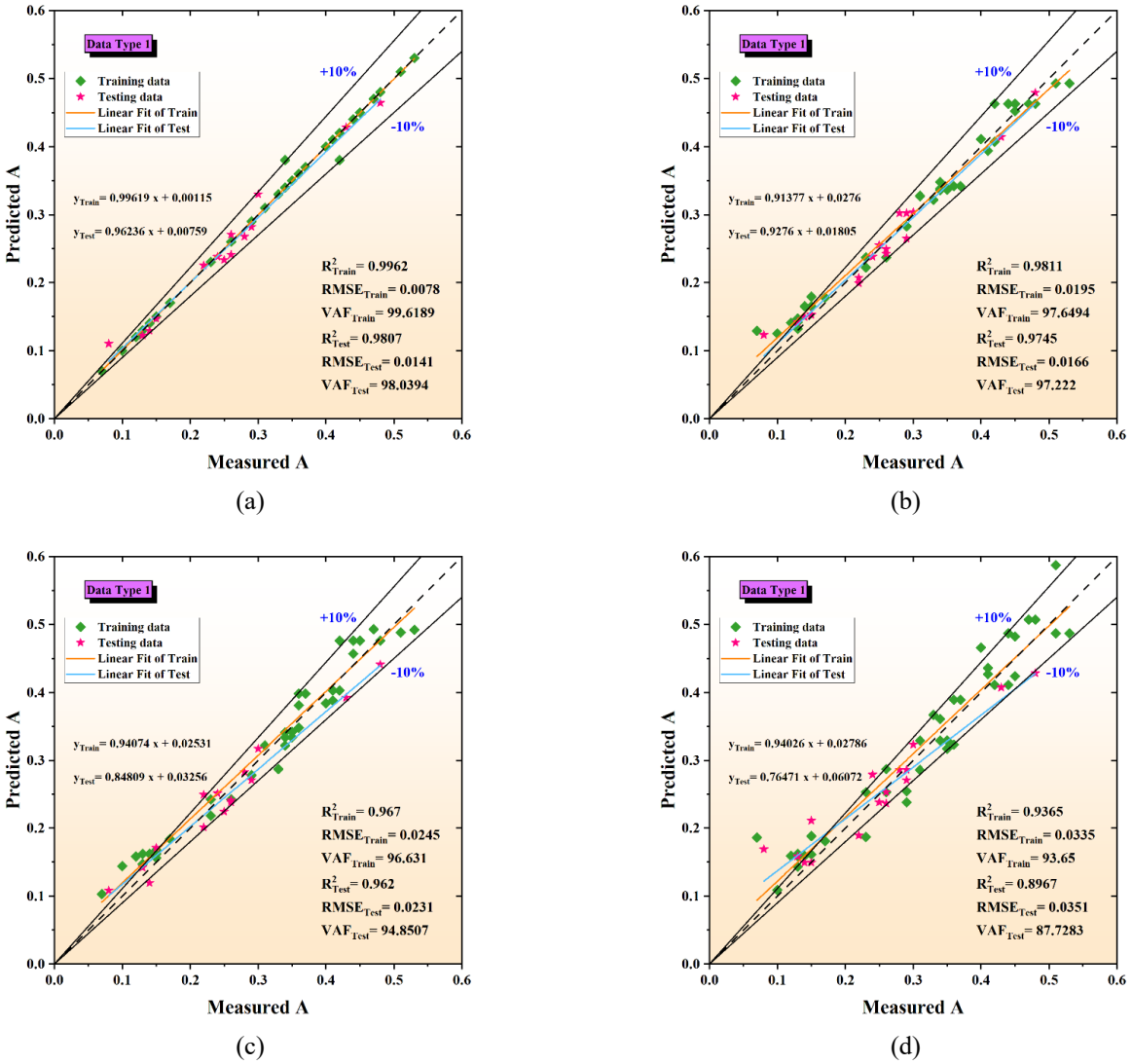



Fig. 9 The results of $XGB$ based approaches to estimate the $A$ using data category I – predicted against measured for: (a) $AOA-XGB$ model, (b) $PSO-XGB$ model, (c) $SCA-XGB$ model; and (d) $SSO-XGB$ model

Each study encounters various limitations and this study, similar to the others, has some limitations. The most important limitation of the present research is the dataset quantity. However two types of dataset have been utilized to extend the models of the present study, but the larger numbers of dataset may lead to more capable models. Also, this research is focused on the XGB optimized model. Other hybrid or optimized models may lead to more accurate results.

### *3.2 Sensitivity analysis*

In order to specify the most influential input parameter, sensitivity analysis of the optimized model was conducted using the best fitted model. Various datasets from the training phase were developed by simultaneously removing one of the input parameters, and the testing dataset, together with the training dataset, prepared the amounts of two efficiency evaluation criteria, $RMSE$, and $VAF$. Considering the last section of the present paper, the dataset is divided to 75% and 25% for training and testing dataset, respectively. The outcomes illustrate that eliminating every variable leads to reducing the efficiency of the approach. The outcomes are in Table 3, which is presented that the $S_u$ is the most effective parameter to estimate the A parameter raised RMSE and VAF from 0.0305 to 0.0396, and 0.0231 to 0.0309 for training dataset, and from 0.0415 to 0.0502, and from 0.0312 to 0.0379 for testing dataset, respectively. The lowermost influence is related to the OCR considering the model with all inputs. It is worth mentioning that removing inputs only leads to a small performance damage for models, but in the present study, because the approach was developed based upon empirical results, removing inputs may decrease the generalizability of the developed approaches. Considering that the being multi-linear problem has no essential effect on approach suitability and often does not influence a lot on estimations, the present study does not suggest removing any inputs.

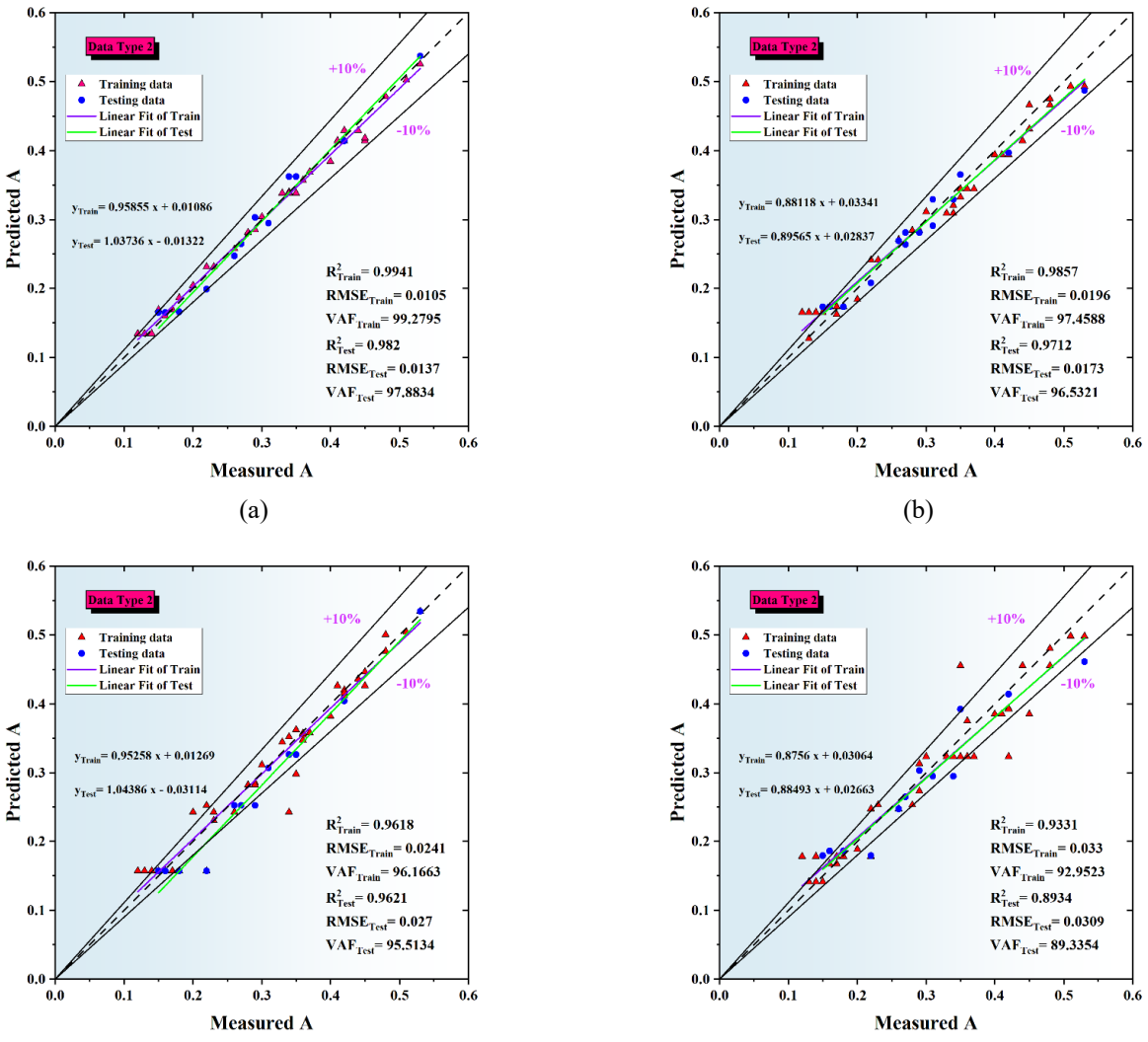



Fig. 10 The results of $XGB$ based approaches to estimate the $A$ using data category II – predicted against measured for: (a) $AOA-XGB$ model, (b) $PSO-XGB$ model, (c) $SCA-XGB$ model; and (d) $SSO-XGB$ model

Table 3 sensitivity analysis using $XGB-AOA$ approach for data category II

| Inputs | Removed parameter | $RMSE$ | | $VAF$ | |
|---|---|---|---|---|---|
| | | $Train$ | $Test$ | $Train$ | $Test$ |
| $PI$ $S_u$ $OCR$ | – | 0.0105 | 0.0137 | 99.2795 | 97.8834 |
| | $PI$ | 0.0195 | 0.0234 | 73.658 | 65.9211 |
| | $S_u$ | **0.0308** | **0.0392** | **41.9756** | **34.1495** |
| | $OCR$ | 0.0249 | 0.0293 | 69.5322 | 61.3843 |

## 4. Conclusions

The principal aim of the present study is to develop a novel AI-based approach called hybrid XGB which hybridizes the Arithmetic Optimization Algorithm (AOA), Particle Swarm Optimization (PSO), Sine-cosine Algorithm (SCA) and Social Spider Optimization (SSO) to superior predict the pile set-up parameter "A" CPT employing two different types of inputs. Data category I involves average developed cone head bearing $q_t$, average wall friction $f_s$, and effective vertical stress $\sigma_{vo}$, and data category II contains the index of plasticity $PI$, shear strength of soil in undrained state $S_u$), and the ratio of over-consolidation $OCR$. For this purpose, CPT data and relevant borehole log have been collected from 70 several positions on all parts of Louisiana. The results show that:

The histogram and distribution curve analysis demonstrate that, there is no discontinuity in the input data for both type I and II.

XGB – AOA has the best accuracy for both data type I and II so that $R^2$ 0.9962 and 0.9941 have been obtained for the training data, respectively. These values considering the $RMSE$ index are 0.0078 and 0.0105, respectively.

XGB – AOA has the best accuracy for both data type I and II so that $R^2$ 0.9807 and 0.982 have been obtained for the testing data, respectively. These values considering the $RMSE$ index are 0.0141 and 0.0137, respectively.

XGB − SSO has the lowest accuracy for both data type I and II so that $R^2$ 0.9365 and 0.9331 have been obtained for the training data, respectively. These values considering the RMSE index are 0.0335 and 0.033, respectively.

XGB − SSO has the lowest accuracy for both data type I and II so that $R^2$ 0.8967 and 0.8934 have been obtained for the testing data, respectively. These values considering the $RMSE$ index are 0.0351 and 0.0309, respectively.

The $S_u$ is the most effective parameter to estimate the A parameter raised $RMSE$ and $VAF$ from 0.0305 to 0.0396, and 0.0231 to 0.0309 for training dataset, and from 0.0415 to 0.0502, and from 0.0312 to 0.0379 for testing dataset,

## Acknowledgments

This work was supported by self-financing project of Cangzhou City Science and Technology Plan, "Study on internal force of precast piles based on quality soil of Huanghuagang in Cangzhou under" (No.204105005), and research project of basic scientific research and operation fee of Hebei University of Water Resources and Electric Engineering, "Study on non-limit passive earth pressure considering soil arching and displacement effect under" (No.SYKJ1901)..

## References

Abdelsalam, M., Diab, H.Y. and El-Bary, A.A. (2021), "A metaheuristic Harris hawk optimization approach for coordinated control of energy management in distributed generation based Microgrids", *Appl. Sci*., **11**(9), 4085. https://doi.org/10.3390/app11094085.

Abellán-García, J. and Guzmán-Guzmán, J.S. (2021), "Random forest-based optimization of UHPFRC under ductility requirements for seismic retrofitting applications", *Constr. Build. Mater*., **285**, 122869. https://doi.org/https://doi.org/10.1016/j.conbuildmat.2021.122869.

Abu-Farsakh, M.Y. and Mojumder, M.A.H. (2020), "Exploring artificial neural network to evaluate the undrained shear strength of soil from cone penetration test data", *Transport. Res. Rec*., **2674**, 11-22. https://doi.org/10.1177/0361198120912.

Abualigah, L., Diabat, A., Mirjalili, S., Abd Elaziz, M. and Gandomi, A.H. (2021), "The arithmetic optimization algorithm", *Comput. Method. Appl. M.*, **376**, 113609. https://doi.org/10.1016/j.cma.2020.113609.

Aghakhani, S., Larijani, A., Sadeghi, F., Martín, A., Shahrakht, A.A. (2023), "A novel hybrid artificial bee colony-based deep convolutional neural network to improve the detection performance of backscatter communication systems", *Electronics*, **12**(10), 2263. https://doi.org/10.3390/electronics12102263.

Ahmadi, S., Motie, M. and Soltanmohammadi, R. (2020), "Proposing a modified mechanism for determination of hydrocarbons dynamic viscosity, using artificial neural network", *Pet. Sci. Technol*., **38**, 699-705. https://doi.org/10.1080/10916466.2020.1780256.

Akbarzadeh, M.R., Ghafourian, H., Anvari, A., Pourhanasa, R. and Nehdi, M.L. (2023), "Estimating compressive strength of concrete using neural electromagnetic field optimization", *Materials*, **16**(11), 4200. https://doi.org/10.3390/ma16114200.

Anbari, M., Arıkan Öztürk, E. and Ateş, H. (2020), "Evaluation of sustainable transport strategies for Tehran with thetheir urbanization rate criterion based on the fuzzy ahp method", *J. Xi'xxan Univ. Archit. Technol*., 12. https://doi.org/10.37896/jxat12.07/2394.

Asadi, S., Roshan, S. and Kattan, M.W. (2021), "Random forest swarm optimization-based for heart diseases diagnosis", *J. Biomed. Inform*., **115**, 103690. https://doi.org/10.1016/j.jbi.2021.103690.

ASTM D2850-03 (2017), Standard Test method for unconsolidated-undrained triaxial compression test on cohesive soils. ASTM International

ASTM D3441-16 (2018), Standard test method for mechanical cone penetration testing of soils. Annu B ASTM Stand West Conshohocken, PA. https://doi.org/10.1520/D3441-16.

ASTM D422-63 (2007), Standard test method for particle-size analysis of soils. ASTM D422-63, West Conshohocken, PA.

ASTM D4318-17e1 (2018), Standard test methods for liquid limit, plastic limit, and plasticity index of soils. ASTM, West Conshohocken. https://doi.org/10.1520/D4318-17E01.

ASTM D7263-21 (2021), Standard test methods for laboratory determination of density and unit weight of soil specimens. ASTM, West Conshohocken. https://doi.org/doi.org/10.1520/D7263-21.

Axelsson, G. (1998), "Long-term set-up of driven piles in non-cohesive soils evaluated from dynamic tests on penetration rods", In: Geotechnical site characterization, 895-900.

Azizi, S., Dadarkhah, A. and Masouleh, A.A. (2020), "Multi-objective optimization method for posture prediction of symmetric static lifting using a three-dimensional human model", *Ann. Mil. Heal. Sci. Res.*, **18**(2). https://doi.org/10.5812/amh.104283.

Bullock, P.J. (2008), "The easy button for driven pile setup: dynamic testing", *From Res. Pract.Geotech. Eng*., 471-488. https://doi.org/10.1061/40962(325)17.

Bullock, P.J. (1999), Pile friction freeze: A field and laboratory study, University of Florida.

Camp III, W.M. and Parmar, H.S. (1999), "Characterization of pile capacity with time in the Cooper Marl: study of applicability of a past approach to predict long-term pile capacity", *Transp. Res. Rec*., **1663**(1), 16-24. https://doi.org/10.3141/1663-03.

Chen, T. and Guestrin, C. (2016), "Xgboost: A scalable tree boosting system", *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*.

Chen, T., He, T., Benesty, M. and Khotilovich, V. (2015), "Xgboost: extreme gradient boosting", R Packag version 04-2 1:1–4.

Cuevas, E., Cienfuegos, M., Zaldívar, D. and Pérez-Cisneros, M. (2013), "A swarm optimization algorithm inspired in the behavior of the social-spider", *Exp. Syst. Appl*., **40**(16), 6374-6384. https://doi.org/10.1016/j.eswa.2013.05.041.

Dawei, Y., Bing, Z., Bingbing, G., Xibo, G. and Razzaghzadeh, B. (2023), "Predicting the CPT-based pile set-up parameters using HHO-RF and PSO-RF hybrid models", *Struct. Eng. Mech*., **86**(5), 673-686. https://doi.org/10.12989/sem.2023.86.5.673.

Dehghan, S. Comparison of seismic behavior factors for reinforced concrete (RC) special moment resisting frames (SMRFs) in Iran in low-, mid-, and high-rise buildings based on Iranian seismic standard 2800 and ASCE

Dehghani, F. and Larijani, A.A. Machine Learning-Jaya Algorithm (Ml-Ijaya) Approach for Rapid Optimization Using High Performance Computing. Available SSRN 4423338

Duan, L., Wu, M. and Wang, Q. (2022), "Predicting the CPT-based pile set-up parameters using HHO-RF and WOA-RF hybrid models", *Arab. J. Geosci*., **15**, 1-19. https://doi.org/10.1007/s12517-022-09843-4.

Esmaeili-choobar, N., Esmaeili-falak, M., Roohi-hir, M. and

Keshtzad, S. (2013), "Evaluation of collapsibility potential at Talesh", *Iran. EJGE*, 2561-2573.

Esmaeili-Falak, M. and Benemaran, R.S. (2023), "Ensemble deep learning-based models to predict the resilient modulus of modified base materials subjected to wet-dry cycles", *Geomech. Eng.*, **32**(6), 583-600. https://doi.org/10.12989/gae.2023.32.6.583.

Esmaeili-Falak, M. and Hajialilue-Bonab, M. (2012), "Numerical studying the effects of gradient degree on slope stability analysis using limit equilibrium and finite element methods", *Int. J. Acad. Res.*, **4**, 216-222

Esmaeili-Falak, M., Katebi, H. and Javadi, A. (2018), "Experimental study of the mechanical behavior of frozen soils-A case study of tabriz subway", *Period Polytech. Civ. Eng.*, **62**, 117-125. https://doi.org/10.3311/PPci.10960.

Esmaeili-Falak, M., Katebi, H., Vadiati, M. and Adamowski, J. (2019), "Predicting triaxial compressive strength and Young's modulus of frozen sand using artificial intelligence methods", *J. Cold Reg. Eng.*, **33**, 4019007. https://doi.org/10.1061/(ASCE)CR.19435495.0000188.

Faramarzi, A., Loghmani, N., Moghadam, R., Allahverdy, A. and Mansoory, M.S. (2021), "Semi-automated glioblastoma tumor detection based on different classifiers using magnetic resonance spectroscopy", *Front. Biomed. Technol.*, **8**(3), https://doi.org/10.18502/fbt.v8i3.7113.

Focht, J.A. and Vijayvergiya, V.N. (1972), "A new way to predict capacity of piles in clay", *Proceedings of the Offshore Technology Conference*, OnePetro.

Ge, D.M., Zhao, L.C. and Esmaeili-Falak, M. (2022), "Estimation of rapid chloride permeability of SCC using hyperparameters optimized random forest models", *J. Sustain. Cem. Mater.*, **12**(5), 1-19. https://doi.org/10.1080/21650373.2022.2093291.

Guang-Yu, Z. (1988), "Wave equation applications for piles in soft ground", *Proceedings of the 3rd International Conference on the Application of Stress-Wave Theory to Piles* (Ed., B.H. Fellenius,), Ottawa, Ontario, Canada.

Gwizdała, K. and Więcławski, P. (2013), "Influence of time on the bearing capacity of precast piles", *Stud. Geotech. Mech.*, 65-74. https://doi.org/10.2478/sgem-2013-0037.

Hammerstrom, D. (1993), "Neural networks at work", *IEEE Spectr.*, **30**, 26-32. https://doi.org/10.1109/6.214579.

Haque, M.N. (2015), Field instrumentation and testing to study set-up phenomenon of driven piles and its implementation in LRFD design methodology.

Haque, M.N., Abu-Farsakh, M.Y., Chen, Q. and Zhang, Z. (2014) "Case study on instrumenting and testing full-scale test piles for evaluating setup phenomenon", *Transport. Res. Rec.*, **2462**(1), 37-47. https://doi.org/10.3141/2462-05.

Hashemi, A. (2022), Advanced data science and physics-based modeling for dynamic systems.

Hashemi, A., Jang, J. and Beheshti, J. (2023), "A machine learning-based surrogate finite element model for estimating dynamic response of mechanical systems", *IEEE Access*, **11**, 54509-54525. https://doi.org/10.1109/ACCESS.2023.3282453.

Hoang, N.D., Chen, C.T. and Liao, K.W. (2017), "Prediction of chloride diffusion in cement mortar using multi-gene genetic programming and multivariate adaptive regression splines", *Measurement*, **112**, 141-149. https://doi.org/10.1016/j.measurement.2017.08.031.

Huo, W., Li, W., Zhang, Z., Sun, C., Zhou, F. and Gong, G. (2021), "Performance prediction of proton-exchange membrane fuel cell based on convolutional neural network and random forest feature selection", *Energ. Convers Manage.*, **243**, 114367. https://doi.org/10.1016/j.enconman.2021.114367. https://doi.org/https://doi.org/10.1016/j.enconman.2021.114367

Iqbal, M., Zhang, D. and Jalal, F.E. (2022), "Durability evaluation of GFRP rebars in harsh alkaline environment using optimized tree-based random forest model", *J. Ocean Eng. Sci.*, **7**, 596-606. https://doi.org/10.1016/j.joes.2021.10.012

Iraji, S., Soltanmohammadi, R., Matheus, G.F., Basso, M. and Campane Vidal, A. (2023), "Application of unsupervised learning and deep learning for rock type prediction and petrophysical characterization using multi-scale data", *Geoenergy Sci. Eng.*, **230**, 212241. https://doi.org/10.1016/j.geoen.2023.212241.

Iravani, S.N., Dehghan, S. An Investigation to the Seismic Performance of Base Isolator-Equipped Moment Frame Steel Structures.

Jamil, U., Sulaiman, M., Ghafoor, N., Malmir, M., Nawaz, F. and Shakoor, R.I. (2023), "Power harvesting towards sustainable energy technology through ambient vibrations and capacitive transducers", *Proceedings of the 2023 International Conference on Emerging Power Technologies (ICEPT)*.

Johari, A., Habibagahi, G. and Ghahramani, A. (2011a), "Prediction of SWCC using artificial intelligent systems: A comparative study", *Sci. Iran*, **18**(5), 1002-1008. https://doi.org/10.1016/j.scient.2011.09.002.

Johari, A., Javadi, A.A. and Habibagahi, G. (2011b), "Modelling the mechanical behaviour of unsaturated soils using a genetic algorithm-based neural network", *Comput. Geotech.*, **38**(1), 2-13. https://doi.org/10.1016/j.compgeo.2010.08.011.

Johari, A., Javadi, A.A. and Najafi, H. (2016), "A genetic-based model to predict maximum lateral displacement of retaining wall in granular soil", *Sci. Iran*, **23**(1), 54-65. https://doi.org/10.24200/SCI.2016.2097.

Joushideh, N., Majidi, A., Tabrizi, H. and Zadeh, S.S. (2023a), "Characterization of scour-Induced subsurface deformations in port structures with contiguous pile walls using Ground-Penetrating Radar (GPR)", *Comput. Res. Prog. Appl. Sci. Eng. CRPASE Trans. Civ. Environ. Eng.*, **9**, 52547.

Joushideh, N., Zadeh, S.S., Bahrami, B. and Mahmoudabadi, N.S. (2023b), "Pseudo-static slope stability analysis and numerical settlement assessment of rubble mound breakwater under hydrodynamic conditions", *World J. Adv. Res. Rev.*, **19**(2). https://doi.org/10.30574/wjarr.2023.19.2.1542.

Kennedy, J. and Eberhart, R. (1995), "Particle swarm optimization", *Proceedings of the ICNN'95-international conference on neural networks*, IEEE.

Khorshidi, M., Ameri, M. and Goli, A. (2023), "Cracking performance evaluation and modelling of RAP mixtures containing different recycled materials using deep neural network model", *Road Mater. Pavement Des.*, 1-20. https://doi.org/10.1080/14680629.2023.2222835.

Khoshaim, A.B., Elsheikh, A.H., Moustafa, E.B., Basha, M. and Mosleh, A.O. (2021), "Prediction of residual stresses in turning of pure iron using artificial intelligence-based methods", *J. Mater. Res. Technol.*, **11**, 2181-2194. https://doi.org/10.1016/j.jmrt.2021.02.042.

Kiannejad Amiri, M., Ghorbanzade Zaferani, S.P., Sarmasti Emami, M.R., Zahmatkesh, S. Pourhanasa, R., Namaghi, S.S., Klemeš, J.J. Bokhari, A. and Hajiaghaei-Keshteli, M. (2023), "Multi-objective optimization of thermophysical properties GO powders-DW/EG Nf by RSM NSGA-II, ANN, MLP and ML", *Energy*, **280**, 128176. https://doi.org/10.1016/j.energy.2023.128176.

Komurka, V.E., Wagner, A.B. and Edil, T.B. (2003), Estimating soil/pile set-up. Citeseer.

Larijani, A. and Dehghani, F. (2023), "Stock price prediction using the combination of firefly (FA) and genetic algorithms", Available SSRN 4448024.

Le, L.T., Nguyen, H., Zhou, J., Dou, J. and Moayedi, H. (2019), "Estimating the heating load of buildings for smart city planning using a novel artificial intelligence technique PSO-XGBoost", *Appl. Sci.*, **9**(13), 2714.

https://doi.org/10.3390/app9132714.

Looney, C.G. (1996), "Advances in feedforward neural networks: demystifying knowledge acquiring black boxes", *IEEE Trans. Knowl. Data Eng.*, **8**, 211-226. https://doi.org/10.1109/69.494162.

Mahmoudabadi, N.S. (2020), "The study of cable behavior with two spring-dampers and one viscous damper", *World*, **9**, 128-136.

Malmir, M., Momeni, H. and Ramezani, A. (2019), "Controlling megawatt class WECS by ANFIS network trained with modified genetic algorithm", *Proceedings of the 2019 27th Iranian Conference on Electrical Engineering (ICEE)*.

McVay, M.C., Schmertmann, J., Townsend, F. and Bullock, P. (1999), "Pile friction freeze: a field investigation study", Res Rep No WPI 510632.

Mirjalili, S. (2016), "SCA: A sine cosine algorithm for solving optimization problems", *Knowledge-Based Syst.*, **96**, 120-133. https://doi.org/10.1016/j.knosys.2015.12.022.

Mojumder, M.A.H. (2020), Evaluation of undrained shear strength of soil, ultimate pile capacity and pile set-up parameter from cone penetration test (CPT) using artificial neural network (ANN), Louisiana State University and Agricultural & Mechanical College.

Moqadam, R., Loghmani, N., Mansoori, M.S. and Allahverdy, A. (2022), "Combination of classifiers to detect grade of glioblastoma using MRS", P*roceedings of the 2022 30th International Conference on Electrical Engineering (ICEE)*.

Moshtaghi Largani, S. and Lee, S. (2023), "Efficient sampling for big provenance", *Proceedings of the ACM Web Conference 2023*.

Motie, M., Bemani, A. and Soltanmohammadi, R. (2018), "On the estimation of phase behavior of CO2-based binary systems using ANFIS optimized by GA algorithm", *Proceedings of the 5th CO2 Geological Storage Workshop. European Association of Geoscientists & Engineers*.

Murdoch, L.C., DeWolf, S., Germanovich, L.N., Moysey, S., Hanna, A., Roudini, S. and Moak, R. (2023a), "Using the shallow strain tensor to characterize deep geologic reservoirs", *Water Resour. Res.*, **59**(2), e2022WR032920. https://doi.org/10.1029/2022WR032920.

Murdoch, L.C., Germanovich, L.N., Roudini, S., DeWolf, S.J., Hua, L. and Moak, R.W. (2021), "A type-curve approach for evaluating aquifer properties by interpreting shallow strain measured during well tests", *Water Resour. Res.*, **57**(9), e2021WR029613. https://doi.org/10.1029/2021WR029613.

Murdoch, L.C., Germanovich, L.N., Slack, W.W., Carbajales-Dale, M., Knight, D., Moak, R., Laffaille, C., DeWolf, D. and Roudini, S. (2023b), "Shallow geologic storage of carbon to remove atmospheric $CO_2$ and reduce flood risk", *Environ. Sci. Technol.*, **57**(23), 8536-8547. https://doi.org/10.1021/acs.est.3c00600.

Nazoktabar, M., Zahedinejad, M, Heydari, P. and Asgharpour, A.R. (2014), "Fabrication and optical characterization of silicon nanostructure arrays by laser interference lithography and metal-assisted chemical etching.

Ng, K.W., Suleiman, M.T. and Sritharan, S. (2013), "Pile setup in cohesive soil. II: Analytical quantifications and design recommendations", *J. Geotech. Geoenviron. Eng.*, **139**(2), 210-222. https://doi.org/10.1061/(ASCE)GT.1943-5606.000075.

Paikowsky, S.G. and Chernauskas, L.R. (1992), "Energy approach for capacity evaluation of driven piles", *Proceedings of the 4th International Conference on the Application of Stress-Wave Theory to Piles*, The Hague, Netherlands.

Premkumar, M., Jangir, P., Kumar, B.S., Sowmya, R., Alhelou, H.H., Abualigah, L. and Yild, A.R. (2021), "A new arithmetic optimization algorithm for solving real-world multiobjective CEC-2021 constrained optimization problems: diversity analysis and validations", *IEEE Access*, **9**, 84263-84295. https://doi.org/10.1109/ACCESS.2021.3085529.

Rashidi Nasab, A. and Elzarka, H. (2023), "Optimizing machine learning algorithms for improving prediction of bridge deck deterioration: A case study of Ohio Bridges", *Buildings*, **13**(6), 1517. https://doi.org/10.3390/buildings13061517.

Reihanisaransari, R., Samadifam, F., Salameh, A.A., Mohammadiazar, M., Amiri, N. and hannumsin, S. (2022), "Reliability characterization of solder joints in electronic systems through a neural network aided approach", *IEEE Access*, **10**, 123757-123768. https://doi.org/10.1109/ACCESS.2022.3224008.

Richardson, B.D. (2011), A case study on pile relaxation in dilative silts, University of Rhode Island.

Rosti, F. and Abu-Farsakh, M. (2015), "Numerical simulation of pile installation and setup for bayou lacassine site", IFCEE 2015. 1152-1161.

Roudini, S., Murdoch, L.C., DeWolf, S. and Germanovich, L.N. (2020), "Interpretation of borehole strain measurements using surrogate modeling-based optimization", AGU Fall Meeting Abstracts. H036-0008.

Sadeghi, F., Larijani, A., Rostami, O., Martín, D. and Hajirahimi, P. (2023), "A novel multi-objective binary chimp optimization algorithm for optimal feature selection: Application of deep-learning-based approaches for SAR image classification", *Sensors*, **23**(3), 1180. https://doi.org/10.3390/s23031180.

Sadeghi, S., Marjani, T., Hassani, A. and Moreno, J. (2022), "Development of optimal stock portfolio selection model in the tehran stock exchange by employing markowitz mean-semivariance model", *J, Financ. Issues*, **20**, 47-71.

Sarkhani Benemaran, R., Esmaeili-Falak, M. and Javadi, A. (2022), "Predicting resilient modulus of flexible pavement foundation using extreme gradient boosting based optimised models", *Int. J. Pavement Eng.*, 1-20. https://doi.org/10.1080/10298436.2022.2095385.

Schmertmann, J.H. (1991), "The mechanical aging of soils", *J. Geotech. Eng.*, **117**(9),1288-1330. https://doi.org/10.1061/(ASCE)0733-9410(1991)117:9(1288).

Shakouri Mahmoudabadi, N. (2021), "The use of viscous dampers for retrofitting of reinforced concrete frames", *Turkish J. Comput. Math. Educ.*, **12**, 7739-7744.

Shi, X., Yu, X. and Esmaeili-Falak, M. (2023), "Improved arithmetic optimization algorithm and its application to carbon fiber reinforced polymer-steel bond strength estimation", *Compos. Struct.*, **306**, 116599. https://doi.org/10.1016/j.compstruct.2022.116599.

Shi, Y. and Eberhart, R.C. (1998), "Parameter selection in particle swarm optimization BT - Evolutionary Programming VII. (Eds., Porto, V.W., Saravanan, N., Waagen, D., Eiben, A.E.), Springer Berlin Heidelberg, Berlin, Heidelberg, 591-600.

Singh, G., Singh, B. and Kaur, M. (2019), "Grasshopper optimization algorithm–based approach for the optimization of ensemble classifier and feature selection to classify epileptic EEG signals", *Med. Biol. Eng. Comput.*, **57**, 1323-1339. https://doi.org/10.1007/s11517-019-01951-w.

Skov, R. and Denver, H. (1988), "Time-dependence of bearing capacity of piles", *Proceedings of the 3rd International Conference on the Application of Stress-Wave Theory to Piles*, Ottawa.

Stone, M. (1974), "Cross-validatory choice and assessment of statistical predictions", *J. R. Stat. Soc. Ser B*, **36**, 111-147.

Sun, D., Shi, S., Wen, H., Xu, J., Zhou, X. and Wu, J. (2021a), "A hybrid optimization method of factor screening predicated on GeoDetector and Random Forest for Landslide Susceptibility Mapping", *Geomorphology*, **379**, 107623. https://doi.org/10.1016/j.geomorph.2021.107623

Sun, D., Wen, H., Wang, D. and Xu, J. (2020), "A random forest

model of landslide susceptibility mapping based on hyperparameter optimization using Bayes algorithm", Geomorphology, **362**, 107201. https://doi.org/10.1016/j.geomorph.2020.107201.

Sun, D., Xu, J., Wen, H. and Wang, D. (2021b), "Assessment of landslide susceptibility mapping based on Bayesian hyperparameter optimization: A comparison between logistic regression and random forest", *Eng. Geol*., **281**, 105972. https://doi.org/10.1016/j.enggeo.2020.105972.

Svinkin, M.R., Morgano, C.M. and Morvant, M. (1994), "Pile capacity as a function of time in clayey and sandy soils", *Proceedings of the Deep Foundations Institute Fifth International Conference and Exhibition on Piling and Deep Foundations*.

Trelea, I.C. (2003), "The particle swarm optimization algorithm: convergence analysis and parameter selection", *Inf. Process Lett.*, **85**, 317-325. https://doi.org/10.1016/S0020-0190(02)00447-7.

Wang, S.T. and Reese, L.C. (1989), "Predictions of response of piles to axial loading", *Predicted and Observed Axial Behavior of Piles: Results of a Pile Prediction SymposiumASCE*, 173-187.

Xue, X. (2018), "Evaluation of concrete compressive strength based on an improved PSO-LSSVM model", *Comput. Concrete*, **21**(5), 505-511. https://doi.org/10.12989/cac.2018.21.5.505.

Yang, M., Tao, B., Chen C., Jia, W. and Sun, S. (2019), "Machine learning models based on molecular fingerprints and an extreme gradient boosting method lead to the discovery of JAK2 inhibitors", *J. Chem. Inf. Model*., **59**, 5002-5012.

Zakariazadeh, A. (2022), "Smart meter data classification using optimized random forest algorithm", *ISA Trans*., **126**, 361-369. https://doi.org/10.1016/j.isatra.2021.07.051.

Zhang, P., Yin, Z.Y. Jin, Y.F. and Chan, T.H.T. (2020), "A novel hybrid surrogate intelligent model for creep index prediction based on particle swarm optimization and random forest", *Eng. Geol*., **265**, 105328.

Zhang, W., Wu, C., Zhong, H., Li, Y. and Wang, L. (2021), "Prediction of undrained shear strength using extreme gradient boosting and random forest based on Bayesian optimization", *Geosci. Front*., **12**, 469-477. https://doi.org/10.1016/j.gsf.2020.03.007.

Zhao, Z., Chen, S., Zhang, D., Peng, B., Li, X. and Zheng, Q. (2022), "Utilizing the GOA-RF hybrid model, predicting the CPT-based pile set-up parameters", G*eomech. Eng*., **31**(1), 113-127. https://doi.org/10.12989/gae.2022.31.1.113

Zhou, J., Li, E., Yang, S., Wang, M., Shi, X., Yao, S. and Mitri, H.S. (2019), "Slope stability prediction for circular mode failure using gradient boosting machine approach based on an updated database of case histories", *Saf. Sci*., **118**, 505-518. https://doi.org/10.1016/j.ssci.2019.05.046.

*GC*

## Nomenclature

| | |
|---|---|
| $A$ | Pile set-up parameter |
| $CPT$ | Cone penetration test |
| $XGB$ | Extreme gradient boosting |
| $PSO$ | Particle swarm optimization |
| $SSO$ | Social spider optimization |
| $SCA$ | Sine–cosine algorithm |
| $AOA$ | Arithmetic optimization algorithm |
| $q_t$ | Mean corrected cone tip resistance |
| $f_s$ | Mean skin friction |
| $\sigma_{vo}$ | Effective overburden pressure |
| $PI$ | Plasticity index |
| $S_u$ | Undrained shear strength of soil |
| $OCR$ | Over consolidation ratio |
| $R^2$ | Determination coefficient |
| $RMSE$ | Root mean square error |
| $MAE$ | Mean absolute error |
| $VAF$ | The variance accounted factor |
| $WOA$ | Whale optimizer algorithm |
| $HHO$ | Harris hawks optimization |
| $GOA$ | Grasshopper optimization algorithm |
| $RF$ | Random forest |
| $ANN$ | Artificial neural network |