Determine the Value of Maximum Dry Density by Gaussian Process Regression in Individual and Hybrid Models

Jianmei Feng[1] and Xiao Chen[2,*]
[1]School of Urban Construction Engineering, Chongqing Technology and Business Institute, Chongqing, 400052 China
[2]Computer Engineering Technical College Guangdong Institute of Science and Technology, Zhuhai 519000, China
E-mail: 13277095763@163.com
[*]Corresponding author

Keywords: maximum dry density, Gaussian process regression, COOT optimization algorithm, Dingo optimization algorithm

Received: July 20, 2025

This investigation presents a new method for applying GPR technology to predict the MDD in soil stability mixes. The method involves creating detailed schemes that connect various natural soil properties, such as linear shrinkage, plasticity, particle size variation, and the type and amount of stabilization additives, to the MDD of stabilized soil. These schemes are developed and evaluated using different soil types from previously published destabilization test results. The study includes two meta-heuristic algorithms: the Dingo Optimization Algorithm (DOA) and the COOT Optimization Algorithm (COA). As a result, two hybrid schemes, GPDO and GPCO, were developed. The GPCO model achieved a high R2 = 0.9905 and a favorable RMSE = 26.13 during training, indicating superior predictive and generalization capabilities compared to other schemes in this study. Overall, this approach provides a practical solution for accurately predicting the MDD of soil-strengthening mixes using GPR combined with meta-heuristic algorithms, which could be useful in various engineering applications.

Povzetek: Raziskava predstavlja novo metodo z uporabo GPR in metahevrističnih algoritmov za zelo natančno napoved največje suhe gostote (MDD) stabiliziranih tal.

1 Introduction

Soil is a time-honored and widely utilized building material that presents many options in its natural state. Nonetheless, it is crucial to recognize that certain kinds of soil may not be appropriate for every construction project. Each soil has its scientific method for handling force components, similar to other building materials. In civil engineering, soil is crucial in designing structures with high safety factors, especially those in direct contact with the ground, like foundations, embankments, and soil-based structures. These constructions need greater safety factors since soil [1], [2] and structural appraisal are unpredictable [3], [4], [5], [6]

Nevertheless, in their natural and modified states, including with treatments like compaction, reinforcement, and consolidation, soil and rock will continue to play important parts in construction [7], [8], [9]. As one of the most common methods used to enhance soil engineering performance, mechanical soil compaction methods are widely adopted for engineering purposes. The quality of soil compaction is usually related to the dry density using the MDD. In these construction projects, like earth dams, road and railway embankments, landfill liners, and backfilling of retaining structures, understanding the traits of soil compaction becomes highly essential, especially MDD. MDD provides the improvement measurement and enables the evaluation of the soil's state following compaction but before remediation [10], [11], [12], [13]. It is helpful to create a theoretical scheme for calculating MDD values to circumvent the laborious laboratory assessment of MDD in each new building project and to comprehend the complex interactions involving the soil qualities and the influencing elements. Such a model should include the following soil traits before stabilization: texture, ductility, linear shrinkage, type, and amount of the stabilizing additives [14]. However, the test procedures are time-consuming and expensive, relying significantly on the expertise and experience of operators in both sampling and result verification.

When manual analysis is difficult, machine learning (ML) [15] is a fantastic technology that allows machines to examine data and extract information from it. The availability of vast volumes of data has increased the demand for machine learning across many businesses. Generally speaking, machine learning's primary goal is to create schemes that can learn from data without clear-cut coding; as a result, many advancements are occurring in this field. To overcome this difficulty and develop machine learning schemes that can control big and complicated databases, physicists and developers have used diverse tactics. It works well for seeing trends and generating precise forecasts. Numerous tactics, including deep learning (DL), reinforcement learning, and supervised and unsupervised learning, have been created and published. Industries including production, logistics,

finance, and civil engineering have embraced machine learning (ML) extensively, and its application has shown great promise in enhancing worker efficiency and productivity [16], [17], [18], [19], [20], [21], [22], [23], [24]. The created machine learning model can accurately forecast the MDD of soil drawing on data. The scheme's property to control irregular interactions between variables of input and output is among its most significant advantages [15].

Table 1 summarizes previous research on MDD prediction. Earlier studies successfully employed machine learning models like ANN, SVM, and Random Forest. However, most of these were based on small datasets and did not include explicit hyperparameter tuning, which could cause overfitting or limit their ability to generalize. Many also lacked proper cross-validation or clear data partitioning strategies. This study addresses these gaps by integrating meta-heuristic optimizers such as COA and DOA into the GPR framework for more effective hyperparameter tuning. This combined approach enhances both the accuracy and stability of MDD forecasts, leading to improved generalization across different soil and additive scenarios.

Table 1: Comparative summary of previous studies on MDD prediction

Study	Method / Model	Dataset Description	Performance (R^2)	RMSE	Remarks
Alavi et al.	RBF Neural Network	120 stabilized soil samples (cement–lime)	0.952	41.2	Early AI-based approach for MDD prediction
Das et al.	Artificial Intelligence (ANN, GP)	Experimental soil stabilization data	0.965	36.8	Improved prediction via non-linear mapping
Hossein Alavi et al.	ANN and Regression Models	Mixed soil dataset from lab tests	0.971	33.5	Highlighted ANN flexibility
Suman et al.	AI Techniques (ANN, SVM)	Cement-stabilized soil samples	0.976	29.8	Robust nonlinear performance
Taffese & Abegaz	ML Algorithms (RF, SVR, ANN)	350 soil samples, multiple regions	0.982	27.6	Comparative machine learning study
Present Study	Hybrid GPR + COA (GPCO)	200+ samples; varied soil, cement, and lime proportions	0.9905	26.13	Proposed state-of-the-art hybrid model

Additionally, the scheme only optimizes a small count of parameters, which lowers the possibility of overfitting. Consequently, it is a trustworthy and useful model for estimating MDD for civil engineering projects. Its extensive use in exploration and real-world utilizations attests to its efficacy and legitimacy.

This investigation introduces a new machine learning tactic to accurately predict crucial soil properties, in particular, MDD outputs that are vital in the design of a civil engineering project. In this respect, due to the difficulties associated with collecting empirical data, this investigation focuses on leveraging GPR. Enhancement of the parameters is highly essential to ensure the GPR model operates at its best. This can be addressed with the incorporation of two schemes, namely the COOT Optimization Algorithm and the Dingo Optimization Algorithm. It will result in a much-improved version of GPR regarding accuracy and efficiency. By streamlining the creation and fabrication of MDD structures, the schemes' integration has a major positive effect on the transportation industry. To appraise the efficacy of the recommended scheme, a sizable MDD database is collected, and comparative studies are also conducted. The outcomes of this investigation offer helpful perspectives for forecasting MDD in projects for civil engineering. This research proposes an efficient method of MDD anticipation by incorporating the GPR algorithm as part of the ML tactic. The GPR model whose parameters are optimized using the COA and DOA schemes can effectively solve the complexity of gathering empirical data related to MDD. In general, this research offers practical solutions and important knowledge for tackling the anticipation of MDD, which is a critical aspect of soil behavior in civil engineering projects.

2 Materials and methodology

2.1 Data collection

A new advanced strategy has been developed for accurately estimating the maximum dry density (MDD) of soil by using six key variables. To ensure that the proposed method produces reliable and accurate results, the dataset was systematically split into training, validation, and testing subsets. This approach not only improves prediction accuracy but also enhances understanding of

how MDD behaves under different soil compositions and physical conditions. The input variables include the proportions of cement and lime, along with critical geotechnical indicators—Liquid Limit (LL), Plastic Limit (PL), and Plasticity Index (PI). The prediction process utilizes Gaussian Process Regression (GPR), offering a probabilistic framework capable of modeling non-linear and uncertain relationships between inputs and the target variable (MDD). Details about data sources, sample characteristics, variable ranges, and statistical summaries are provided in Table 1. Standard testing procedures determined soil composition, cement content, and quicklime proportions according to relevant geotechnical standards, while Atterberg limit tests measured LL, PL, and PI. The Liquid Limit (LL) indicates the moisture content where soil shifts from plastic to liquid, whereas the Plastic Limit (PL) marks the moisture content where soil transitions from plastic to semi-solid. The Plasticity Index (PI)—the difference between LL and PL—reflects soil deformability and cohesiveness, providing a measure of its plasticity and engineering characteristics [25], [26], [27]. To predict MDD accurately, custom empirical correlations and equations were integrated into the GPR framework. These models incorporate the physical and chemical properties of soil—especially the effects of lime and cement additives and the soil's plastic traits—to improve estimation over traditional regression methods. The model's strength lies in its ability to capture complex, non-linear relationships within the data while maintaining high generalization through probabilistic inference. Ultimately, this hybrid approach offers civil engineers an advanced tool for estimating MDD without relying solely on extensive laboratory tests. By combining the clarity of traditional soil mechanics with the predictive capabilities of GPR-based modeling, this method significantly reduces experimental effort and time while delivering superior prediction accuracy and a broader understanding of soil behavior under various stabilization conditions.

A new advanced strategy has been developed to estimate the maximum dry density of soil using six key variables. To ensure the reliability and accuracy of the model, the dataset was divided into training, validation, and testing subsets. So, this more advanced way means greater precision and an expansive understanding of the MDD of soil. Herein, the proportions of cement and lime, along with Liquid Limit (LL), Plasticity Index (PI), and Plastic Limit (PL), represent the variables employed in these estimations. MDD anticipation is based on the GPR model. Data on different variable sources have been presented in Table 2, providing information concerning samples, sources, numbers required, and variables. So, by following standard procedure, the determination of soil content, cement content, and quicklime will be determined as per specification. While others like LL, PL, and PI are determined through Atterberg limit tests. The water content at which soil transitions from plastic to liquid is denoted by LL, and the water content at which soil transitions from plastic to semi-solid is denoted by PL. The soil's plasticity is determined by the plasticity index, which is the difference between LL and PL [25], [26], [27]. To project the MDD of soil, custom-developed equations and correlations are being used, considering the soil properties and the percentages of cement, lime, LL, PL, and PI obtained from the data collected. These formulas have been developed to offer estimations of MDD in a very accurate manner according to the available information.

Table 2: MDD and the statistical characteristics of inputs.

Indicators	Variables						
	Input						Targets
	Soil (%)	Cement (%)	Lime (%)	LL (%)	PL (%)	PI (%)	MDD (kN/m^3)
Max	100	30	30	102	58.24	70	2210
Min	70	0	0	18	12	0	1200
Avg	93.604	3.807	2.588	39.428	22.673	16.755	1780.61
St. Dev.	4.6366	4.316	4.086	16.763	9.412	12.694	227.508

2.2 Gaussian process regression (GPR)

GPR employs a non-parametric regression, and its involvement is empirical data with stochastic integer output values (yn) $D = \{(y_m, x_m), m = 1,2,3, \dots, M\}$ of M groups of vector input x_m IR^L. GPR then constructs a scheme that can efficiently extrapolate to the output transportation at new input situations. The output noise, which is believed to be combined, zero-mean, fixed, and uniformly distributed, is unknown due to outside influences like shortening or measurement mistakes. Eq. 1 shows the formula for y:

$$y = f(x) + \delta, \quad \delta \text{ "} M(0, s_{noise}^2) \tag{1}$$

GPR displays the latent variables of f as a Gaussian process, using x as an index into these variables. This is done by constraining the analysis to such functions whose values for a set of $\{f(x_1), \dots, f(x_k)\}$ with different indices are jointly Gaussian distributed; this is achieved by drawing any finite set of variables from a consistent Gaussian. That's putting a GP prior over functions in a Bayesian way. Since it's specified how to define mean function $w(x)$ and covariance function $k(x, x')$, the functions can be conveniently defined using this tactic. This means having the values of new input functional easy by spending a few data. The variance of s_{noise}^2, is taken for modeling noise.

As illustrated in Eq. 2, the mean function $w(x)$represents the expected value of the Gaussian process for each input x, defining the central tendency around which the function values are distributed.

$$w(x) = E[f(x)], \quad k(x,x') = E[(f(x) - w(x))(f(x') - w(x'))] \tag{2}$$

To indicate anticipation, use the notation $E[.]$. Only the invisible part of the input space is affected by the $w(x)$ selection, normally set to 0. The coefficient of variation function, which is by nature positive and symmetrically semi-definite when appraised for any two sets of input location points, is the only factor affecting the conduct of the procedure [28]. Many hyperparameters are often included in the covariance function, which establishes the prior distribution of $f(x)$. It is usual practice to employ the squared exponential covariance function.

Eq. (3) explicitly presents the mathematical formulation of $k(x,x')$, which governs how information propagates through the model and determines the overall flexibility and generalization capacity of the GPR framework.

$$k(x,x') = q_1 \exp\left(\frac{\|x - x'\|}{2q_2}\right) \tag{3}$$

In this case, k displays a norm defined on the input space. It is noteworthy that small associations across $f(x)$ and $f(x')$ are shown by the covariance function swiftly decaying as the separation among input pairs x and x' grows. Three hyperparameters are used: q_1 establishes the greatest permitted covariance, q_2 is an entirely optimistic hyperparameter that establishes the rate at which correlation decays as points get farther apart, and q_3 is an extra hyperparameter that, while not stated explicitly in Eq. (2), embodies the undetermined variability s^2_{noise} in Eq. (1). These hyperparameters are grouped to form an agent (q), which is then regarded as the emergence of a vector of randomness (Q). Utilizing the training data, the understanding that offers the best fit for the database is selected to provide predictions. If it is assumed that the hyperparameters have been identified previously, the inference procedure is straightforward. Establishing the resulting vector used to train latent features as f and the vector of test concealed components as f^* may result in the joint distribution of the Gaussian shown below.

$$p(f, f^*) = M\left(0, \begin{bmatrix} k_{f,f} & k_{*,f} \\ k_{f,*} & k_{*,*} \end{bmatrix}\right) \tag{4}$$

Using the variation perform $k(.,.)$ in Eq. (4) and the associated hyperparameters [29], the correlation associated with the i_th factor in the set referred to as the first underlining and the j_th attribute in the group portrayed by the second underlining ($*$ is employed in place of f^* for short) is calculated to create the asymmetrical covariance matrix K. Figure 1 depicts the scenario scheme.

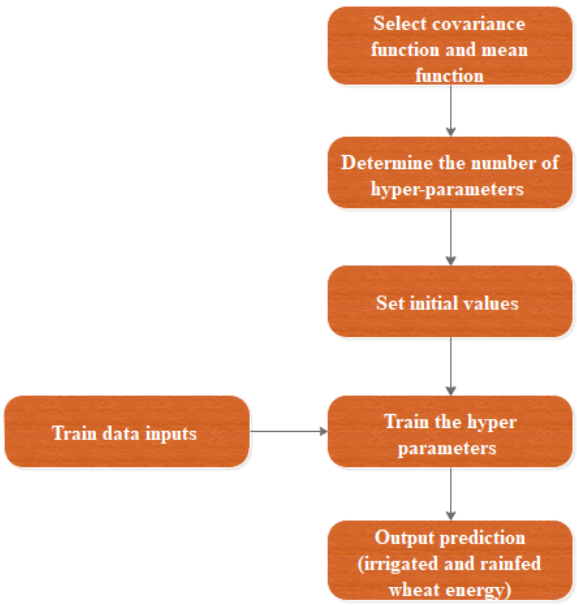

Figure 1: The anticipation scheme drawing on GPR

2.3 COOT Optimization Algorithm (COA)

The unique movement patterns that coot populations display on water surfaces serve as the foundation for the COOT optimization method. Coots are small birds that engage in a variety of group behaviors on water surfaces, mostly to get closer to food supplies or pre-established spots [30]. The coot swarm exhibits four primary behavioral traits on the watery plane: directed movement prompted by the person in charge, location adaption with the dominating individual, chain migration, and stochastic movement [31]. The application process of the COOT algorithm consists of four distinct behavioral movements. The following is the algorithm's procedure set of instructions:

According to Eq. (5), the team of persons is going to start using a randomized tactic.

$$CP(i) = rand\ (1,b) \times (vc - kc) + kc \tag{5}$$

The variable that showcases the quantity or length of the problem of optimization is denoted by d, while the location of the $i-th$ coot is displayed by $CP(i)$. The highest limit vc and the lower boundary kc, which specify the search space, set the highest and lowest values for every variable in this issue space. Specifically, vc and kc specify the size of the sought-after region for the optimization problem.

$$vc = [vc_1, vc_2, \dots, vc_b], \quad kc = [kc_1, kc_2, \dots, kc_b] \tag{6}$$

Each coot's situation is refreshed depending on four diverse locomotion behaviors after the flock has been started.

2.3.1 Random movement

Eq. (7) is employed to establish a point Q that displays the movement's initial action at random.

$$G = rand\ (1, b) \times (vc - kc) + kc \tag{7}$$

To avert getting stuck in a locally optimal, the role has been changed by Eq. (14):

$$CP(i) = CP(i) + E \times S_2 \times (G - CP(i)) \tag{8}$$

Eq. (8) is employed to ascertain *E*'s significance.

$$E = 1 - Z \times (\frac{1}{Iter}) \tag{9}$$

Z showcases the current number of revisions, whereas the variable $Iter$ showcases the peak count of repetitions.

2.3.2 Chain movement

The mean position of two coots can be determined using Eq. (10) to execute the chain movement.

$$CP(i) = \frac{CP(i-1) + CP(i)}{2} \tag{10}$$

where $CP(i-1)$ is the location of the second coot bird.

2.3.3 Adjusting position according to the leader

An individual adjusts its position during the leader movement based on the leader's location within the group. Specifically, it moves toward the leader. Eq (11) is used to select the leader.

$$P = 1 + (i\ MOD\ MZ) \tag{11}$$

Eq. (11) uses the numbers P for the individual in charge, i for the people who follow him, and MZ for the overall number of leaders [32].

Eq. (12) is used to change a coot bird's location throughout its transition motion:

$$CP(i) = LP(P) + 2 \times S_1 \times \cos(2s\pi) \times (LP(P) - CP(i)) \tag{12}$$

The coot bird's present spot is displayed by $CP(i)$ in Eq. (12), the chosen leader's location by $LP(P)$, an arbitrary value in the region [0, 1] by S_1, and a random variable in the period [-1, 1] by R.

2.3.4 Leander movement

To discover the best situation, the person in charge should shift from the present local posture to the worldwide optimum position [33]. Eq. (13) is used to update the leadership position to achieve this:

$$LP(i) = \begin{cases} B \times S_3 \times \cos(2\pi S) \times \left(qBest - LP(i)\right) + qBest & S_4 < 0.5 \\ B \times S_3 \times \cos(2\pi S) \times \left(qBest - LP(i)\right) - qBest & S_4 \geq 0.5 \end{cases} \tag{13}$$

Figure 2 shows the flowchart for COA.

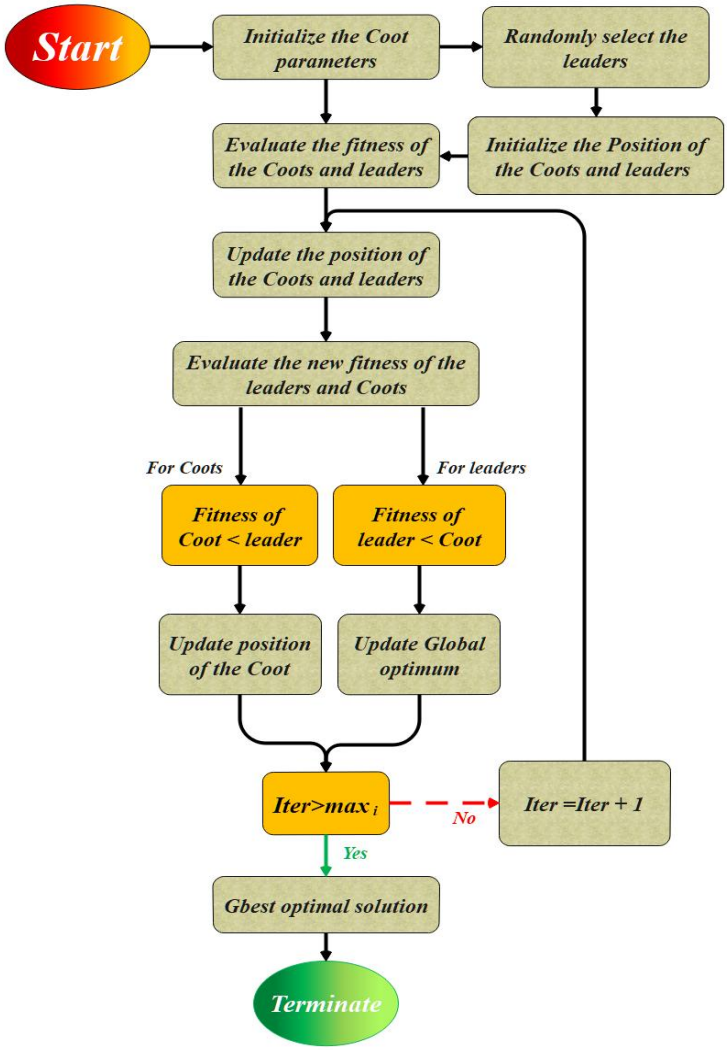

Figure 2: Flowchart for COA

Eq. (13) uses S as an arbitrary amount in the gap [-1, 1], S_3 and S_4 as random integers in the range [0, 1], and $qBest$ to indicate the optimal location. The value of B is determined using Eq. (14) as a reference.

$$B = 2 - Z \times (\frac{1}{Iter}) \tag{14}$$

2.4 DOA

DOA [34], a newly developed bio-inspired scheme for global enhancement, mimics hunting tactics such as persecution, group tactics, and scavenging behavior [35].

DOA considers the likelihood that dingoes will survive sin ce the Australian dingo dog is now in danger of extinction. Fig. 3 shows a graphic representation of the complete procedure.

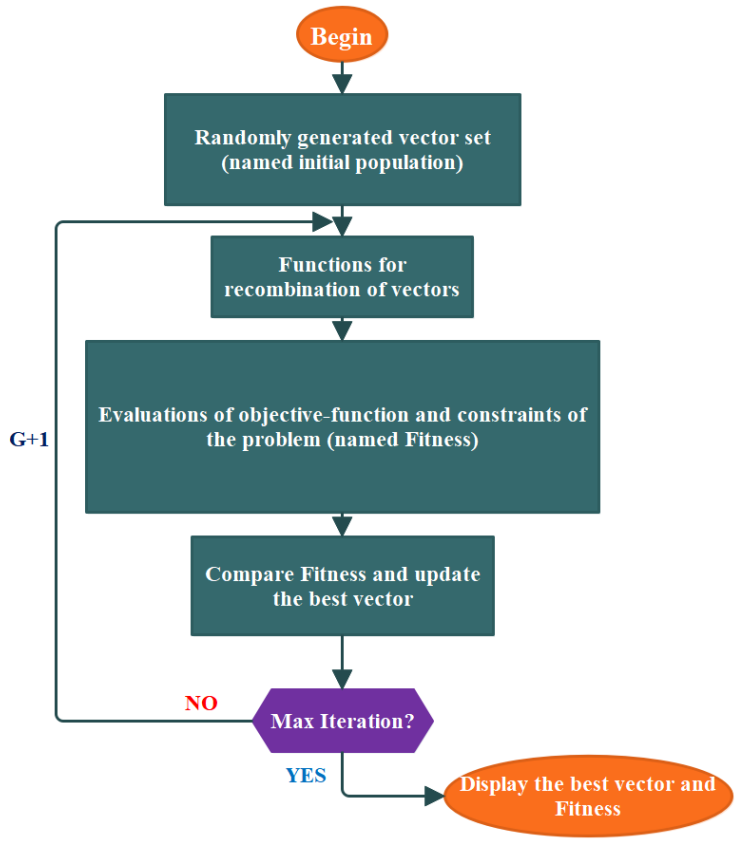

Figure 3: DOA flowchart.

In the computational representations of the DOA algorithm, Strategy 1 is displayed by Eq. (15), which entails group assault, a foraging technique frequently employed by dingoes that hunt in packs, find the prey's situation, and surround it [36].

$$\underset{x_i}{\rightarrow}(h+1) = \alpha_1 \sum_{l=1}^{ze} \frac{\left| \underset{\phi_l(h)}{\longrightarrow} - \underset{x_i}{\rightarrow}(h) \right|}{ze} - \underset{x_*}{\rightarrow}(h) \qquad (15)$$

In mathematical schemes of the DOA algorithm, Eq. (16) shows a fresh situation of a search agent displayed as $\underset{x_i}{\rightarrow}(h+1)$ since the position means a movement for dingoes, where $\underset{(\phi_l(h))}{\longrightarrow}$ displays the subset of attack that will attack $\phi \subset X$, while X is defined as a population of dingoes that haphazardly created. The current search agent is expressed as $\underset{x_i}{\rightarrow}(h)$, while the best search agent found from the previous cycle is expressed as $\underset{x_*}{\rightarrow}(h)$. The scale factor α_1, generating an evenly distributed random number within the interval of $[-2, 2]$, alters the size and direction of the dingoes' routes. Also, a random integer number ze is produced within the interval of $\left[2, \frac{SizePop}{2}\right]$, in which SizePop shows the entire population size of dingoes [37].

Dingoes use the second tactic, intimidation, to hunt tiny prey by pursuing each one separately until it is captured. This behavior is modeled by the subsequent calculation:

$$\underset{x_i}{\rightarrow}(h+1) = \underset{x_*}{\rightarrow}(h) + \alpha_1 * a^{\alpha 2} * \left(\underset{x_{s_1}}{\longrightarrow}(h) - \underset{x_i}{\rightarrow}(h) \right) \qquad (16)$$

while the rotation of dingoes is displayed by $\underset{x_i}{\rightarrow}(h+1)$, the at present searching agent is indicated by $\underset{x_i}{\rightarrow}(h)$, and the top query result from the previous session is displayed by $\underset{x_*}{\rightarrow}(h)$. In contrast to α_2, which is a uniformly random integer generated from inside a specified interval $[-1,1]$, α_1 preserves the identical quantity as in Eq. (16). $\underset{x_{s_1}}{\longrightarrow}(h)$ showcases the $s_1 - th$ search agent selected, where $i \neq s_1$, and s_1 is a random integer produced between 1 and the highest number of queries (dingoes).

Dingoes using the third tactic, Scavenger, wander aimlessly around their environment in search of carrion to consume. Eq. (17) is used to represent this behavior:

$$\underset{x_i}{\rightarrow}(h+1) = \frac{1}{2}\left[a^{\alpha 2} * \underset{x_{s_1}}{\longrightarrow}(h) - (-1)^{\varrho} * \underset{x_i}{\rightarrow}(h) \right] \qquad (17)$$

The process of dingoes is displayed by $\underset{x_i}{\rightarrow}(h+1)$, α_2 stays the same as in Eq. (16), s_1 is a number created at random between 1 and the peak count of search engines (dingoes), $\underset{x_{s_1}}{\longrightarrow}(h)$ showcases the $s_1 - th$ search agent is chosen, $\underset{x_i}{\rightarrow}(h)$ showcases the present searching agent in which $i \neq s_1$, and ϱ is a haphazardly produced binary number.

The amount to be used for the dingo survivor rate, which serves as a component of the fourth tactic, is given by Eq. (18).

$$S(i) = \frac{fit_{max} - fit\ (i)}{fit_{max} - fit_{min}} \tag{18}$$

The chance of survival of dingoes is determined by Eq. (18), where fit (i) is the present-day quality value of the $i-th$ search agent, and fit_{max} and fit_{min} are the best and worst fitness values of each generation, accordingly. The normalized fitness values in the survival vector in Eq. (18) fall between 0 and 1. Eq. (19) is applied when the overall survival probability is inadequate, equivalent to or underneath 0.3:

$$\underset{x_i}{\rightarrow}(h) = \underset{x_*}{\rightarrow}(h) + \frac{1}{2}\left[\underset{x_{s_1}}{\longrightarrow}(h) - (-1)^{\varrho} * \underset{x_{s_2}}{\longrightarrow}(h)\right] \tag{19}$$

Finding an operator that has poor probability of survival, denoted by $\underset{x_i}{\rightarrow}(h)$, is updated in the DOA algorithm by Eq. (19). Between 1 and the greatest amount of search agents (dingoes), the random integers s_1 and s_2 are created, with $s_1 \neq s_2$. The $s_1 - th$ and $s_2 - th$ search agents chosen for the update are denoted by $\underset{x_{s_1}}{\longrightarrow}(h)$ and $\underset{x_{s_2}}{\longrightarrow}(h)$. The top search engine found in the earlier cycle is displayed by $\underset{x_*}{\rightarrow}(h)$, and ϱ is a binary value that is produced at random [38].

2.5 Performance evaluation schemes

The schemes are appraised in this investigation utilizing a variety of measures, including the degree of doubt (95% U95), symmetrical mean absolute percentage error (SMAPE), MAE, R^2, and RMSE. Below are the related formulas for these measurements. A high R^2 value showcases that a scheme performs well during the training, validation, and testing stages. Conversely, as they show a lower level of model error, lower values of metrics like RMSE, U95, SMAPE, and MAE are preferred. These measures are calculated using Eqs. (20) through (24) accordingly.

$$R^2 = \left(\frac{\sum_{i=1}^{w}(k_i - \bar{k})(q_i - \bar{q})}{\sqrt{[\sum_{i=1}^{w}(k_i - k)^2][\sum_{i=1}^{w}(q_i - \bar{q})^2]}}\right)^2 \tag{20}$$

$$RMSE = \sqrt{\frac{1}{w}\sum_{i=1}^{w}(q_i - k_i)^2} \tag{21}$$

$$MAE = \frac{1}{w}\sum_{i=1}^{w}|q_i - k_i| \tag{22}$$

$$U_{95} = \frac{1.96}{w}\sqrt{\sum_{i=1}^{w}(q_i - k_i)^2 + \sum_{j=1}^{w}(q_i - k_j)^2} \tag{23}$$

$$SMAPE = \frac{100}{w}\sum_{j}^{w}\frac{2 \times |q_i - k_i|}{|q_i| + |k_i|} \tag{24}$$

Here, the anticipated and observed outcomes are denoted by k_i and q_i, correspondingly. $\bar{K}$ and $\bar{q}$ display the standard deviation values of the forecasted and empirical specimens, accordingly. Conversely, w depicts the count of samples under examination.

The figure 4 shows the R^2 values of the GPR model across five folds (K1–K5) in a five-fold cross-validation for predicting Maximum Dry Density (MDD). The R^2 values range from 0.848 to 0.886, indicating how well the model fits the data in each fold. Fold K5 has the highest R^2 value of 0.886, signifying the strongest correlation between predicted and observed MDD values. This suggests that Fold 5 (K5) offers the best model performance and generalization among the folds. As a result, K5 was chosen as the best fold for further evaluation and model validation due to its superior predictive accuracy and consistency in estimating MDD.

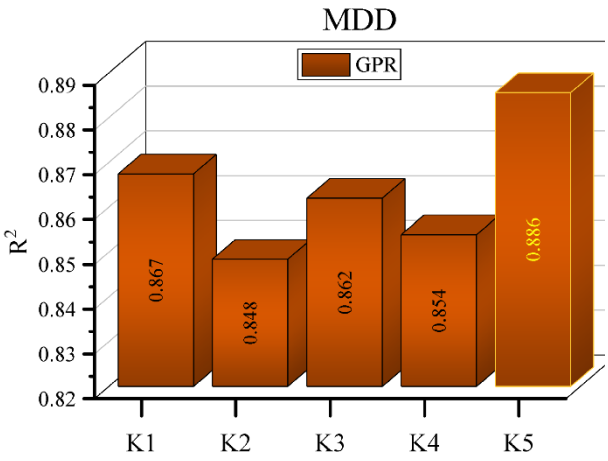

Figure 4: Cross-Validation

3 Outcomes and Discussion

Table 3, titled "Critical Parameters (MDD)," presents three models—GPRT, GPSC, and GPR—and their key hyperparameters: n_restarts, length_scale, and alpha. In the GPRT model, n_restarts is 16, length_scale is 486, and alpha is 0.01. For GPSC, these values are 11, 396, and 0.14, respectively, while the basic GPR model has zero restarts, an undefined length scale, and a very small alpha value (1.00E-10). Overall, the results indicate that increasing the number of restarts and properly tuning the length scale greatly improves model accuracy and stability in predicting MDD. The simple GPR model mainly serves as a baseline with lower generalization capability. Implementing optimization algorithms such as COOT within GPR frameworks directly enhances performance, reduces error, and increases the model's adaptability to complex soil data.

Table 3: Criticalparameter (MDD)

Models	n_restarts	length_scale	alpha
GPRT	16	486	0.01
GPSC	11	396	0.14
GPR	0	---	1.00E-10

This portion of the work assesses the anticipation ability of different schemes with MDD. Table 2 displays the outcomes of their testing, validation, and instruction effectiveness. The GPR model is used in the paper to predict positive MDD treatments. To enhance the Gaussian Process Regression (GPR) model, novel COA and DOA schemes are employed in the GPCO and GPDO hybrid scheme. The scheme is trained on 70% of the samples, validated on 15%, and tested on 15%. The findings in Table 2 indicate that GPCO exhibits the highest R2 of 0.9905 during the training stage, while GPR demonstrates the lowest value of 0.939 during the validation stage.

Regarding the RMSE value, GPR records the highest value of 51.250, while the GPCO model achieves the lowest value of 26.136 during the training stage. When considering the MAE result, the GPCO model performs best with a value of 16.112 during the training stage, whereas the GPR model shows the poorest result at 46.354 during the testing stage. Regarding SMAPE, GPCO delivers the highest performance with a value of 0.000069 during the training stage, while GPR exhibits the weakest performance at 0.00092 during the testing stage. Additionally, GPCO demonstrates desirable performance in U_{95} with a value of 72.401 during the training stage, whereas GPR exhibits the worst performance at 142.88 during the testing stage.

Table 2: Developed appraisal outcomes of schemes by evaluators

Schemes	**Sections**	**Evaluators**				
		RMSE	R^2	MAE	U_{95}	SMAPE
GPR	Train	45.966	0.9618	39.747	127.50	0.00017
	Validation	44.114	0.9390	37.488	121.73	0.00075
	Test	51.250	0.9588	46.354	142.88	0.00092
GPCO	Train	26.136	0.9905	16.112	72.401	0.000069
	Validation	30.536	0.9800	23.495	84.172	0.000467
	Test	42.473	0.9838	33.189	118.61	0.000664
GPDO	Train	34.224	0.9801	29.112	95.045	0.000124
	Validation	33.239	0.9611	28.044	92.179	0.000558
	Test	38.115	0.9805	32.199	105.01	0.00064

The scatter plot of the projected and gauged values of MDD in Fig. 5 provides a correlation and the distribution and density of the sample points, which are controlled by the metrics R2 and RMSE, accordingly. As the values of R2 increase, the points disperse, and the points tend to concentrate more with low values of RMSE. As shown by Fig. 3, GPCO has less dispersion in all three stages of the experiment and therefore is highly precise in its anticipations. On the other hand, GPR has more dispersion, causing overestimation and underestimation, hence not being as accurate as GPCO. GPCO has better accuracy and execution compared to other hybrid schemes during the training, validation, and testing stages.

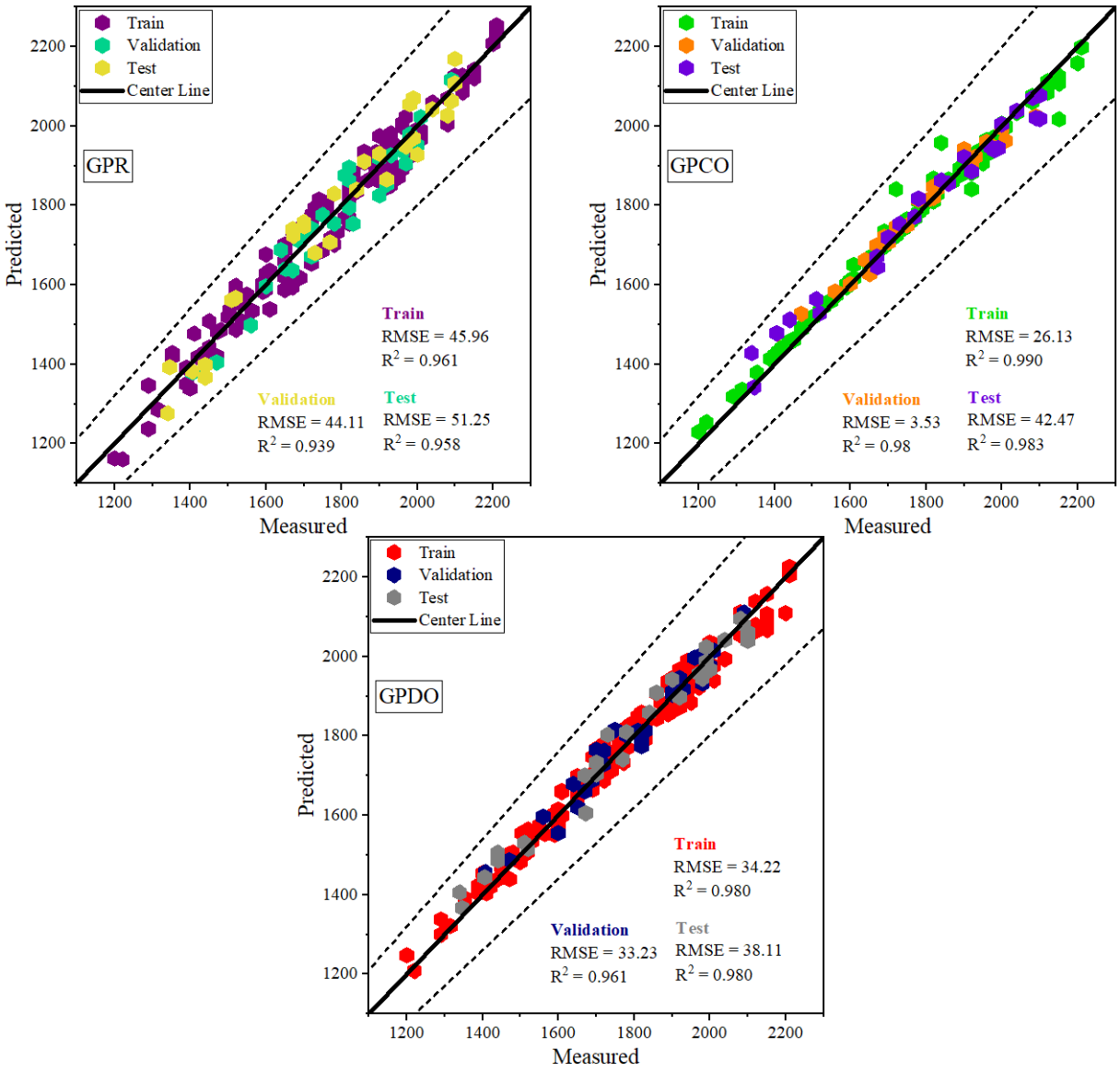

Figure 5: The scatter plot for the projected and gauged MDD.

The error analysis of GPCO, GPR, and GPDO schemes was done. The outcomes are displayed as symbol-line and violin diagrams in Fig (6). For the training stage, GPCO gave the highest percent error of 7%; the distribution is concentrated near zero percent. However, when GPCO is displayed in the validation stage, there is a significant improvement—the percent error value has gone down to 5%. On the other hand, GPR initially had higher errors during the training stage but showed steady improvement in subsequent stages. Overall, the analysis showcases that GPCO surpasses the other schemes regarding lower error rates and consistent execution across all stages.

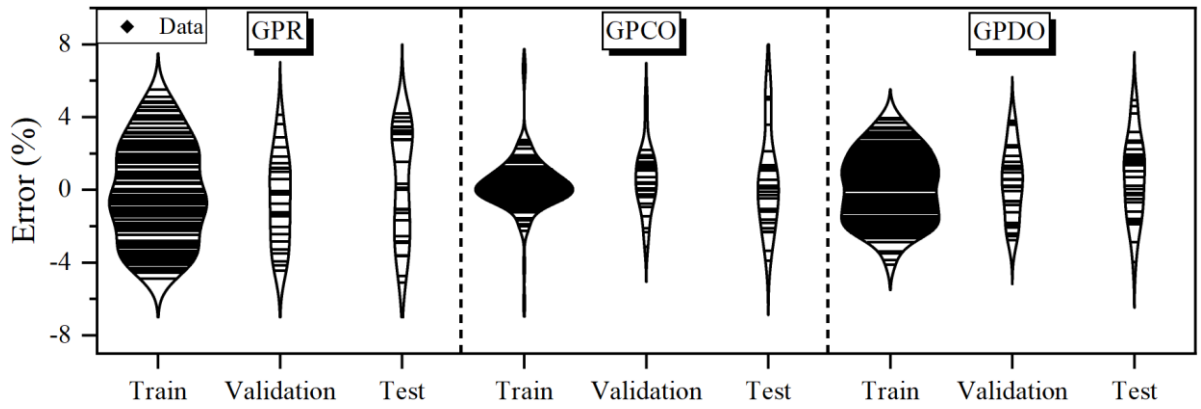

Figure 6: The violin diagram for the error percentage of presented schemes.

The results comparing GPR, GPDO, and GPCO models show that GPCO has the best predictive performance across all evaluation phases. This improvement results from the COA algorithm's effective hyperparameter optimization, which enhances the GPR kernel's convergence and reduces the model's error variance. The consistent R^2 and RMSE values during both validation and testing suggest that the GPCO model has high generalization ability with little sign of overfitting. Although this study successfully develops a hybrid GPR framework, it mainly emphasizes methodological development and validation with existing data. Future research will compare our hybrid models with other leading techniques from the literature, such as Random Forest (RF), Support Vector Regression (SVR), and Deep Learning architectures, to better assess their efficiency, robustness, and suitability for different soil types and stabilization ranges.

While the developed GPR-based hybrid models (GPCO and GPDO) demonstrated high accuracy in predicting maximum dry density (MDD), some limitations are noteworthy. The dataset primarily includes a narrow range of soil types and specific proportions of stabilizers like cement and lime. As a result, the models may not fully capture the variability of natural soils with different textures, mineral content, or regional characteristics. Additionally, laboratory measurements often assume ideal conditions, whereas actual field compaction can fluctuate due to factors such as moisture variations, energy application, and environmental influences. To improve the models' generalizability, future research should use larger, more diverse datasets from various geographic locations and field environments. Incorporating additional input features such as moisture content, compaction energy, grain size, and environmental factors like temperature or humidity could further enhance the models' practical applicability. Despite these limitations, this approach provides a solid foundation for developing intelligent, data-driven tools to predict soil compaction in civil engineering practice.Conclusion

The maximum dry density depends on the type of soil, compaction effort, and moisture content, among other factors, and is a crucial property affecting the stability and strength of soil. In both architecture and engineering, this attribute can frequently be gauged as precisely as possible to guarantee the longevity and safety of buildings. To forecast MDD, a machine learning scheme utilizing GPR has been presented. The COA and DOA meta-heuristic schemes in the study were used to maximize accuracy and reduce errors in model development. Three schemes—GPCO, GPDO, and a single GPR model—are created. Throughout the training, validation, and testing stages, the schemes were verified using testing specimens from several documented publications.

This study presents a hybrid GPCO model combining Gaussian Process Regression (GPR) with the COOT Optimization Algorithm to accurately predict soil's Maximum Dry Density (MDD). The hybrid improves convergence stability and generalization by optimizing kernel parameters. Enhanced data preprocessing, clear workflow, and parameter tuning explanations strengthen the model's interpretability and application. Overall, the refined GPCO framework shows high predictive accuracy and supports future data-driven geotechnical modeling and soil stabilization design.

Schemes' performance was compared utilizing diverse metrics, including R2, RMSE, MAE, U95, and SMAPE. The outcomes of this investigation are discussed below:

1. The R2 values of GPCO schemes were the highest R2, while the lowest was that of GPR, which gave a difference of 3%.
2. GPCO performed the best, with high accuracy, in predicting MDD through all three stages. This was further reflected in its considerably lower error rates, as seen by a 40% lower RMSE and a 20% lower MAE than that of GPR.
3. The research outcomes showed that using the COA optimizer along with GPR resulted in a good combination that yielded accurate anticipations of MDD.

Declarations

Funding

Guangdong Province Continuing Education Quality Improvement Project in 2022, Intelligent Empowerment for Rural Revitalization - Doumen Community Education Demonstration Base" (Project Number: JXJYGC2022GX325).

Authors' contributions

All authors contributed to the study's conception and design. Data collection, simulation and analysis were performed by" Jianmei Feng and Xiao Chen ". Also, the first draft of the manuscript was written by Xiao Chen. Ling Ding commented on previous versions of the manuscript.

Acknowledgements

We would like to take this opportunity to acknowledge that there are no individuals or organizations that require acknowledgment for their contributions to this work.

Ethical approval

The research paper has received ethical approval from the institutional review board, ensuring the protection of participants' rights and compliance with the relevant ethical guidelines.

Research involving human participants and animals

The observational study conducted on medical staff needs no ethical code. Therefore, the above study was not required to acquire an ethical code.

Informed consent

This option is not necessary due to that the data were collected from the references.

Competing of interests

The authors declare no competing interests.

References

[1] Z. S. Janjua and J. Chand, “Correlation of CBR with index properties of soil,” *International Journal of Civil Engineering and Technology*, vol. 7, no. 5, pp. 57–62, 2016.

[2] J. Duque, W. Fuentes, S. Rey, and E. Molina, “Effect of grain size distribution on california bearing ratio (CBR) and modified proctor parameters for granular materials,” *Arab J Sci Eng*, Springer, vol. 45, pp. 8231–8239, 2020. https://doi.org/10.1007/s13369-020-04673-6.

[5] E. G. Akpokodje, “The stabilization of some arid zone soils with cement and lime,” *Quarterly journal of engineering geology*, Geo Science World, vol. 18, no. 2, pp. 173–180, 1985. https://doi.org/10.1144/GSL.QJEG.1985.018.02.06

[6] F. G. Bell, “Lime stabilization of clay minerals and soils,” *Eng Geol*, Elsevier, vol. 42, no. 4, pp. 223–237, 1996. https://doi.org/10.1016/0013-7952(96)00028-2

[7] A. H. Alavi, A. H. Gandomi, M. Gandomi, and S. S. Sadat Hosseini, “Prediction of maximum dry density and optimum moisture content of stabilised soil using RBF neural networks,” *The IES Journal Part A: Civil & Structural Engineering*, Taylor & Francis, vol. 2, no. 2, pp. 98–106, 2009. https://doi.org/10.1080/19373260802659226.

[8] A. B. Ngowi, “Improving the traditional earth construction: a case study of Botswana,” *Constr Build Mater*, Elsevier, vol. 11, no. 1, pp. 1–7, 1997. https://doi.org/10.1016/S0950-0618(97)00006-8

[10] A. Bharath, M. Manjunatha, T. V Reshma, and S. Preethi, “Influence and correlation of maximum dry density on soaked & unsoaked CBR of soil,” *Mater Today Proc*, Elsevier, vol. 47, pp. 3998–4002, 2021. https://doi.org/10.1016/j.matpr.2021.04.232.

[11] S. Suman, M. Mahamaya, and S. K. Das, “Prediction of maximum dry density and unconfined compressive strength of cement stabilised soil using artificial intelligence techniques,” *International Journal of Geosynthetics and Ground Engineering*, Springer, vol. 2, pp. 1–11, 2016. https://doi.org/10.1007/s40891-016-0051-9.

[12] A. Hossein Alavi, A. Hossein Gandomi, A. Mollahassani, A. Akbar Heshmati, and A. Rashed, “Modeling of maximum dry density and optimum moisture content of stabilized soil using artificial neural networks,” *Journal of Plant Nutrition and Soil Science*, Wiley Online Library, vol. 173, no. 3, pp. 368–379, 2010. https://doi.org/10.1002/jpln.200800233.

[13] F. Masoumi, S. Najjar-Ghabel, A. Safarzadeh, and B. Sadaghat, “Automatic calibration of the groundwater simulation model with high parameter dimensionality using sequential uncertainty fitting approach,” *Water Supply*, IWA Publishing, vol. 20, no. 8, pp. 3487–3501, Dec. 2020. https://doi.org/10.2166/ws.2020.241.

[14] B. Mahesh, “Machine learning algorithms-a review,” *International Journal of Science and Research (IJSR).[Internet]*, vol. 9, pp. 381–386, 2020.

[15] Z.-H. Zhou, *Machine learning*. Springer Nature, 2021.

[16] H. Wang, Z. Lei, X. Zhang, B. Zhou, and J. Peng, “Machine learning basics,” *Deep learning*, pp. 98–164, 2016.

[17] G. Biau, “Analysis of a random forests model,” *The Journal of Machine Learning Research*, ACM Digital Library, vol. 13, no. 1, pp. 1063–1095, 2012.

https://dl.acm.org/doi/abs/10.5555/2503308.2343682.

[18] K. Kim, "Financial time series forecasting using support vector machines," *Neurocomputing*, Elsevier, vol. 55, no. 1–2, pp. 307–319, 2003. https://doi.org/10.1016/S0925-2312(03)00372-2.

[19] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science (1979)*, science, vol. 349, no. 6245, pp. 255–260, 2015. https://doi.org/10.1126/science.aaa8415.

[20] T. Chen, J. Morris, and E. Martin, "Gaussian process regression for multivariate spectroscopic calibration," *Chemometrics and Intelligent Laboratory Systems*, Elsevier, vol. 87, no. 1, pp. 59–71, 2007. https://doi.org/10.1016/j.chemolab.2006.09.004.

[21] W. Ni, L. Nørgaard, and M. Mørup, "W. Ni, L. Norgaard, M. Morup, Non-linear calibration models for near infrared spectroscopy, Analytica Chimica Acta 813 (2014) 1–14.," *Anal Chim Acta*, Elsevier, vol. 813, pp. 1–14, 2014. https://doi.org/10.1016/j.aca.2013.12.002.

[22] A. H. Alavi, A. H. Gandomi, and A. Mollahasani, "A genetic programming-based approach for the performance characteristics assessment of stabilized soil," *Variants of evolutionary algorithms for real-world applications*, Springer, pp. 343–376, 2012. https://doi.org/10.1007/978-3-642-23424-8_11.

[23] W. Z. Taffese and K. A. Abegaz, "Prediction of compaction and strength properties of amended soil using machine learning," *Buildings*, MDPI, vol. 12, no. 5, p. 613, 2022. https://doi.org/10.3390/buildings12050613.

[24] S. K. Das, P. Samui, and A. K. Sabat, "Application of artificial intelligence to maximum dry density and unconfined compressive strength of cement stabilized soil," *Geotechnical and Geological Engineering*, Springer, vol. 29, pp. 329–342, 2011. https://doi.org/10.1007/s10706-010-9379-4.

[25] C. E. Rasmussen and C. K. I. Williams, "Gaussian processes for machine learning (adaptive computation and machine learning) the mit press," *Cambridge, MA, USA*, pp. 69–106, 2005.

[26] Z. Y. Wan and T. P. Sapsis, "Reduced-space Gaussian Process Regression for data-driven probabilistic forecast of chaotic dynamical systems," *Physica D*, Elsevier, vol. 345, pp. 40–55, 2017. https://doi.org/10.1016/j.physd.2016.12.005.

[27] I. Naruei and F. Keynia, "A new optimization method based on COOT bird natural life model," *Expert Syst Appl*, Elsevier, vol. 183, p. 115352, 2021. https://doi.org/10.1016/j.eswa.2021.115352.

[28] R. R. Mostafa, A. G. Hussien, M. A. Khan, S. Kadry, and F. A. Hashim, "Enhanced coot optimization algorithm for dimensionality reduction," in *2022 Fifth International Conference of Women in Data Science at Prince Sultan University (WiDS PSU)*, Riyadh, Saudi Arabia, IEEE, 2022, pp. 43–48. https://doi.org/10.1109/WiDS-PSU54548.2022.00020.

[29] H.-Y. Wang *et al.*, "Optimal wind energy generation considering climatic variables by Deep Belief network (DBN) model based on modified coot optimization algorithm (MCOA)," *Sustainable Energy Technologies and Assessments*, Elsevier, vol. 53, p. 102744, 2022. https://doi.org/10.1016/j.seta.2022.102744

[30] B. Milenković, Đ. Jovanović, and M. Krstić, "An application of Dingo Optimization Algorithm (DOA) for solving continuous engineering problems," *FME Transactions*, vol. 50, no. 2, pp. 331–338, 2022. DOI: 10.5937/fme2201331M.

[31] A. K. Bairwa, S. Joshi, and D. Singh, "Dingo optimizer: a nature-inspired metaheuristic approach for engineering problems," *Math Probl Eng*, Wiley Online Library, vol. 2021, pp. 1–12, 2021. https://doi.org/10.1155/2021/2571863.

[32] J. H. Almazán-Covarrubias, H. Peraza-Vázquez, A. F. Peña-Delgado, and P. M. García-Vite, "An improved Dingo optimization algorithm applied to SHE-PWM modulation strategy," *Applied Sciences*, MDPI, vol. 12, no. 3, p. 992, 2022. https://doi.org/10.3390/app12030992.

[33] H. Peraza-Vázquez, A. F. Peña-Delgado, G. Echavarría-Castillo, A. B. Morales-Cepeda, J. Velasco-Álvarez, and F. Ruiz-Perez, "A bio-inspired method for engineering design optimization inspired by dingoes hunting strategies," *Math Probl Eng*, Wiley Online Library, vol. 2021, pp. 1–19, 2021. https://doi.org/10.1155/2021/9107547.