Hybrid K-Nearest Neighbors Models with Metaheuristic Optimization for Predicting Undrained Shear Strength

Ning Cao[1,2,*], Xin-E Yan [1,2], Lijuan Zhang[1], Genqi Xu[1] and Jing Ma[3,4]
[1]Xi'an Key Laboratory Of Monitoring And Prevention Of Railway Roadbed Damage, Xi'an, 710300, China
[2]Xi'an Traffic Engineering Institute, Xi'an, 710300, China
[3]China Gezhouba Group No.3 Engineering Co.,Ltd, Xi'an, 710077, China
[4]China Gezhouba Groupco.,Ltd., Yichang,443002, China
E-mail: Caoning0231@163.Com
[*]Corresponding author

Keywords: undrained shear strength, k-nearest neighbours, machine learning algorithms, mountain gazelle optimizer, coronavirus herd immunity optimizer

Received: February 12, 2024

Around the world, soft soils can be found in many areas close to seas and rivers. These areas play an indispensable and crucial role in the development of government plans, especially in the population growth sector. Due to maintaining a weak shear power and vast settlement under the buildings, soft soils are considered problematic soil. The significant risks associated with building structures and infrastructures in soft soil are high, requiring engineers' extreme attention. It depends on undrained shear strength (USS) that the foundation of structures can bear in soft soil, and this factor vigorously controls the selection of soil improvement techniques. In recent years, there have been enhancements and extensions in the methodologies employed for estimating soil characteristics, including USS. These methods are divided into three main sections: Laboratory Testing, Field Testing, and Correlation with Other Soil Parameters. In recent research, data science techniques have created more reliable and accurate models for predicting USS. This study aims to apply the K-Nearest Neighbors (KNN) classifying method for predicting USS. Mountain Gazelle Optimizer (MGO) and Coronavirus Herd Immunity Optimizer (CHIO) appeal for developing hybrid models with KNN and facilitating accuracy enhancement. The dataset which utilized in this study contains four input variables including liquid limit (LL), plastic limit (PL), and sleeve friction (SF), overburden weight (OBW). Comparative analysis across all data phases reveals that the KNCH model, optimized using the CHIO, achieved superior predictive performance with the highest coefficient of determination (R^2 = 0.993), and the lowest values in root mean square error (RMSE = 85.19), mean squared error (MSE = 7256.15), normalized RMSE (NRMSE = 0.470), and scatter index (SI = 0.065). In contrast, the KNN model without optimization reported R^2 = 0.971, RMSE = 168.17, and SI = 0.132, while the KNMG model—optimized using the MGO—resulted in R^2 = 0.983, RMSE = 128.15, and SI = 0.101.

Povzetek: Raziskava uvaja hibridne modele KNN, optimizirane z metahevristikama CHIO in MGO, za napredno napovedovanje nedrenirane strižne trdnosti tal v geotehničnem inženirstvu.

1 Introduction

Construction of geotechnical structures like dam embankments, dikes, levees, as well as road embankments is a common challenge in most countries worldwide, particularly when such structures are situated on organic subsoils [1]. Due to the significant compressibility and low initial Undrained Shear Strength (USS) of organic soils, constructing even moderately-sized embankments become problematic in engineering practice. When designing embankments on such subsoils, determining the stability of the embankment and anticipating subsurface deformations are the main challenges. The embankment stability analysis during construction predominantly relies on the USS of organic soils [2,3].

Common field investigations include the Cone Penetration Test with Pore Pressure Measurement (CPTu) and the Vane Shear Test (VST) to assess the USS of soils. The VST utilizes cruciform blades rotated at a defined speed as per ASTM D2573-08 standards to determine USS values in saturated clay deposits, while the CPTu involves a 60° cone penetrometer with a friction sleeve, adhering to ASTM D-5778 standards, providing measurements of cone tip resistance (CR), sleeve friction (SF), and porewater pressure (PP). Dissipation tests are often conducted to evaluate equilibrium porewater pressure (pp_0) by halting cone penetration and measuring porewater pressure dissipation. Even if human errors are assumed to be reduced to zero in these experiments, they are time- and cost-consuming [4].

To overcome all issues, utilizing novel machine learning (ML) algorithms in various aspects of engineering geotechnical studies has experienced remarkable growth lately [5–8] . Among the vast number of research articles on applying machine learning models in geotechnical engineering published over the past years, approximately 70% have emerged within the last decade (data spanning from 1984 to 2019 [9]). A recent

comprehensive review article [10] provides an overview of the use of artificial neural networks (ANN) for estimating soil mechanical substances, particularly in terms of compressive and shear strength. [11–16]. support vector machines (SVM) and ANN have found application in forecasting bearing capacity for driven piles and settlement of shallow foundations [17,18]. These techniques have also been widely adopted for soil slope steadiness evaluation and estimation, including decision trees and logistic regression [19–22]. In geotechnical earthquake engineering, ANN and other deep learning frameworks were effectively employed to evaluate soil liquefaction potential [23–25]. Additionally, within the field of dynamic soil-foundation-structure interaction investigations, researchers have utilized multivariate linear regression and distance-weighted K-Nearest Neighbors (KNN) regression techniques to construct preliminary predictive models. These models aim to estimate seismic energy dissipation, permanent settlement, and the acceleration amplification ratio (which signifies the reduction in maximum acceleration transferred to the structure) for shallow foundations exhibiting rocking behaviour [26–28].

In geotechnical engineering, designers rely on crucial soil properties to accurately determine USS. These fundamental properties are also integral in developing predictive models through ML techniques. For example, the concept of soil plasticity, initially proposed by Atterberg in 1911, was introduced into soil mechanics by Casagrande in 1932 and further elaborated by Terzaghi and Peck in 1948. Soil behaviour is categorized based on water content (w), defined as the ratio of moisture mass to the mass of dry soil particles, resulting in four states: solid, semi-solid, plastic, and liquid. Two primary consistency limits are employed to characterize soil behaviour. The plastic limit (PL) is the minimum water content at which soil exhibits plastic deformation, which is irrecoverable deformation without failure. The liquid limit (LL) is defined as the minimum water content at which soil behaves as a viscous fluid, essentially marking the boundary where the shear strength of the soil is nearly zero [29]. The liquidity index (LI) normalizes soil's water content concerning its plastic and liquid limits. When the LI is less than zero, the soil exhibits brittle behaviour, falling into the categories of solid or semi-solid. Conversely, when the LI exceeds 1, the soil behaves like a liquid. In the literature, it has been documented that the USS of remoulded soil at the liquid limit can range from 0.5 to 1.7 kPa, with variations depending on soil types and testing apparatus or measurement methods [30].

This investigation's goal is to develop data-driven predictive models for soil USS utilizing supervised learning and non-linear ML techniques. Data from a 200 rock samples database consisting of sleeve friction (SF), LL, PL, and overburden weight (OBW) have been used to train, validate, and test machine learning models developed in this investigation. The non-linear, nonparametric ML algorithm considered is KNN. The performances of this single model are compared with the performance of optimized versions, which were developed using CHIO and MGO. The following sections explain the intake attributes and the theories of the ML methods in brief and their process of implementation in this research, provide performance evaluation parameters and present the major conclusions and outcomes of the research.

- **Justification for algorithms' selection**

In this study, the MGO and CHIO were selected to enhance the predictive performance of the KNN model due to their demonstrated efficiency in handling complex, nonlinear optimization problems. These recently developed metaheuristic algorithms offer strong exploration-exploitation balance, fast convergence, and robustness against local minima, making them suitable for hyperparameter tuning in machine learning models. Their unique mechanisms—MGO's predator-prey-inspired movement strategy and CHIO's immunity-based adaptation—provide diverse search capabilities and improved accuracy, which are critical in developing reliable predictive models for USS.

- **Novelties and Contributions**

The novelty and distinctiveness of this study, particularly in the context of USS prediction using ML, can be articulated through the following key contributions:

1. Pioneering Application of KNN-Based Models in USS Prediction: To the best of the author's knowledge, this study represents only the second documented academic attempt to employ KNN-based models for the prediction of USS of soils. This rare application underscores the innovative nature of the study, especially given the limited exploration of KNN in geotechnical engineering contexts, where more conventional regression-based or deep learning approaches are typically favored.
2. Development of Two Novel Hybrid Models with State-of-the-Art Optimization Algorithms: This research introduces two cutting-edge hybrid models—KNCH and KNMG—by integrating the KNN algorithm with two recently developed metaheuristic optimization techniques: the CHIO and the MGO, respectively. These algorithms have shown substantial promise in diverse domains for hyperparameter tuning and global search efficiency, and their application here represents a significant methodological
3. advancement for enhancing the predictive capability and convergence behavior of KNN models in civil engineering problems [31–33].
4. Robust Performance Evaluation Using Multiple Statistical Metrics: A comprehensive model evaluation framework is established through the implementation of five widely recognized statistical performance indicators: R^2, RMSE, MSE, NRMSE, and the SI. This multi-metric approach facilitates an in-depth and multi-dimensional performance assessment of the models, enabling rigorous benchmarking of their prediction quality, generalization capacity, and error distribution.

Table 1 reports the summary of comparing between previous publications.

Table 1: The summary of the literature review

Study	Dataset Size	Methodologies	Key Results	Limitations
[30]	~100+ laboratory-tested remoulded clay soil samples	Experimental investigation and empirical modeling based on Atterberg limits	- Reported that USS at liquid limit (LL) ranges between 0.5–1.7 kPa depending on soil type - Strong empirical correlation observed between USS and the Liquidity Index (LI)	- Not ML-based - Focused only on remoulded soils in specific consistency states - Limited generalizability across soil types
[12]	~150 samples	Multilayer ANN	- Showed ANN could outperform conventional empirical equations - Strong correlation between input features (e.g., PL, LL) and USS	- Overfitting risk due to small dataset - No feature importance analysis
[13]	~200 samples	Support Vector Machines (SVM)	- SVM accurately predicted soil cohesion - $R^2 \approx 0.9$ with optimized kernel parameters	- Highly sensitive to kernel selection - Requires tuning for different soil types
[14]	~180–220 data points	ANN (Feedforward)	- Developed regression-based ANN models - Achieved better accuracy than linear models	- Lacked uncertainty quantification - No real-time or field validation
[34]	300+ samples	ANN with cross-validation	- ANN demonstrated good generalization after k-fold validation - Applied in pile foundation design	- Focused on bearing capacity - Input features not aligned with USS prediction
[15]	~200 soil specimens	Hybrid SVM-GA (Genetic Algorithm)	- Combined SVM with GA for parameter tuning - RMSE reduced by ~15% compared to standalone SVM	- Computationally intensive - No interpretability analysis (e.g., SHAP or PCC)
[19]	~300 slope case studies	Decision Tree (DT), Random Forest (RF)	- Achieved classification accuracy > 90% - Identified top influencing features (e.g., angle, cohesion)	- Binary classification (stable/unstable) - Not suitable for continuous USS regression
[20]	GIS-based dataset (~500+ sites)	Logistic Regression (LR)	- Used for landslide hazard zonation - Provided probabilistic maps for decision-making	- Not designed for strength prediction - Model outputs are probabilities, not magnitudes
[21]	~400 data points	Decision Tree	- High interpretability - Good for categorizing soils by failure type	- Not regression-capable - May oversimplify nonlinear soil behavior
[22]	150–250 slope instances	SVM, DT, ANN	- Compared multiple ML algorithms - SVM performed best (AUC > 0.93)	- Application limited to classification - Soil shear strength not directly modeled

2 Materials and methodology

This section provides a comprehensive overview of the data preparation process, a detailed description of the K-Nearest Neighbors (KNN) model used as the baseline, and the implementation of the MGO and CHIO algorithms for optimization. Additionally, it outlines the evaluation metrics employed for performance assessment and presents the cross-validation results. These components are organized and discussed in detail across subsections 2.1 to 2.6.

2.1 Dataset description

To predict the undrained shear strength of soil, outstanding 200 sample records from publication [35] were categorized randomly into validation (15%), testing (15%), and training (70%) phases. Here, SF, LL, PL, and

OBW—five variables that can impact the USS's value—were taken into account as inputs. As the preprocessing tasks, the dataset is evaluated for miss values and no miss value is observed. In addition, the samples are randomized utilizing randperm method.

SF refers to the resistance a pile or shaft encounters as it is driven or inserted into the subsurface soil or rock. It arises due to the shear resistance between the surface of the pile or shaft and the surrounding soil or rock material. LL represents the moisture content at which soil shifts from a plastic to a semi-solid state, determined by the Casagrande cup test, while PL signifies the moisture content at which soil transitions from a plastic to a brittle state, determined through the plasticity index test. Finally, OBW is the total vertical load the soil or rock exerts above a specific point underground. Table 2 presents the statistical characteristics of the dependent input variables' maximum, minimum, average, and standard deviation values as well as USS's target variable.

The line-symbol plot in Figure 1 demonstrates the values of each input parameter and the output for 200 samples. As it is obvious, among input parameters, the broadest and narrowest range of values are related to the LL and SF, respectively. OBW illustrates more fluctuation, and USS as an output parameter fluctuates less with more values near 1000 (MPa). In addition, the correlation between input and output parameters is illustrated in Figure 2.

Table 2: The statistic properties of the input variable of USS.

Indicator	Variables				
	SF	LL	PL	OBW	USS
Max	1.400	133	50	3.640	5670
Min	0.020	24	12	0.030	100
Avg	0.425	63.49	27.48	1.569	1271.9
St. Dev.	0.289	19.886	6.940	0.824	968.61

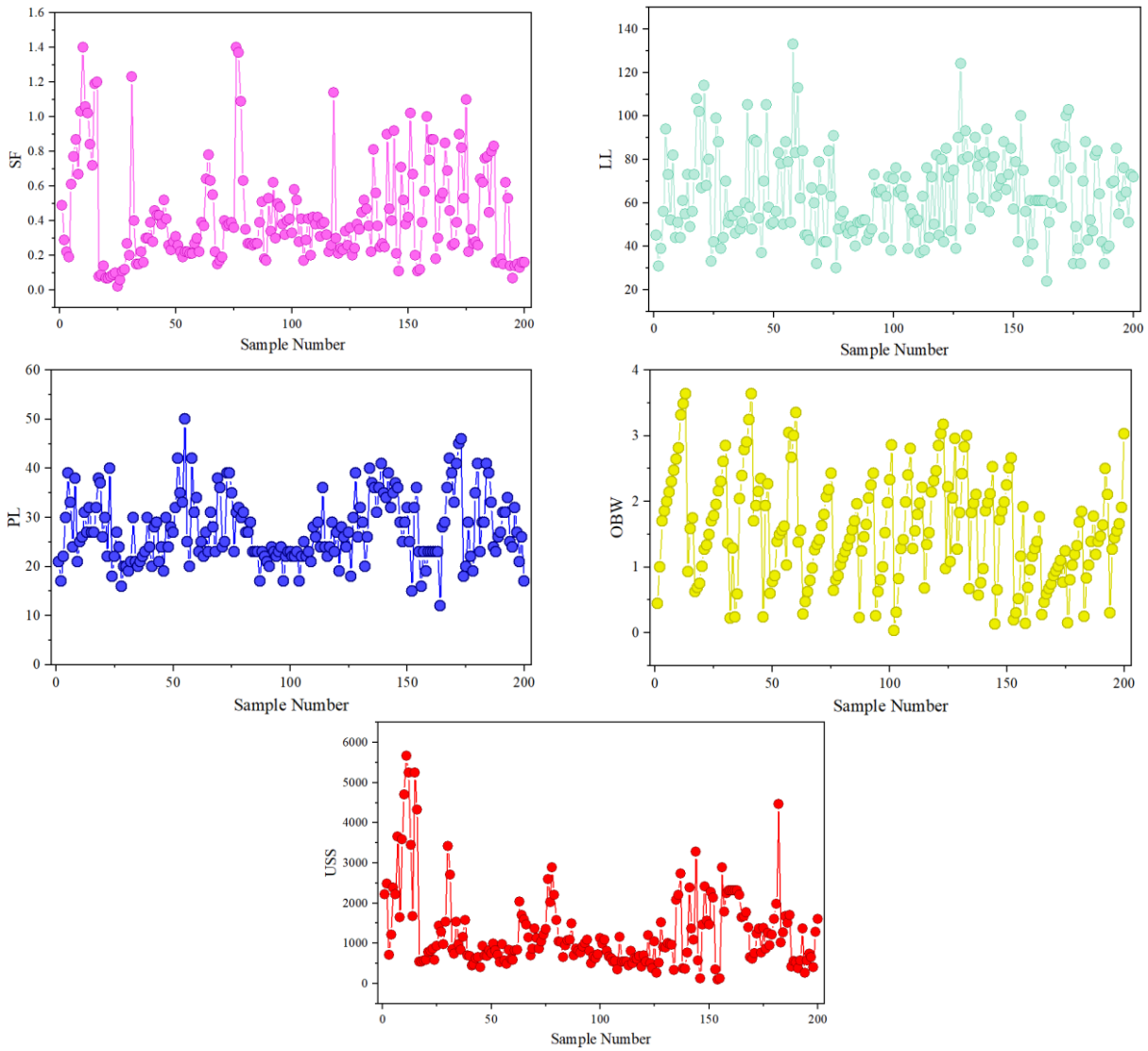


Figure 1: The Line-Symbol plot for input and output

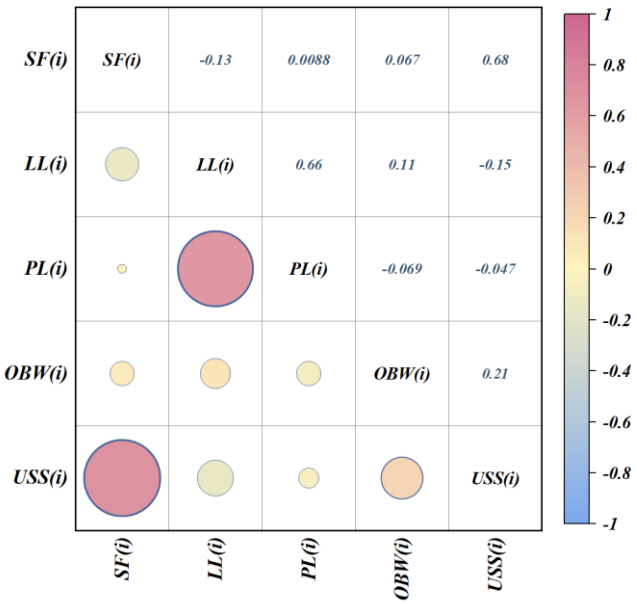


Figure 2: The Correlation plot for relationship between parameters.

The computational framework utilized in this study is optimized to meet both hardware and software requirements efficiently. At its core, the system operates on an Intel® Core™ i7-3770K CPU clocked at 3.50 GHz, paired with 16 GB of RAM to support seamless multitasking and high-performance operations. The system architecture is 64-bit, featuring an x64-based processor, and it runs on the Windows 11 Pro operating system, ensuring modern compatibility and robust functionality. Graphical processing is managed by an NVIDIA GeForce GT 640 GPU, which provides reliable support for rendering and visualization tasks. Storage needs are met with a 1-terabyte hard drive, offering ample capacity for large-scale data processing and storage. On the software front, Python serves as the primary programming environment. Machine learning implementations are facilitated through the scikit-learn library, while Pandas and NumPy are employed for data manipulation and analysis. For visualization purposes, Matplotlib is utilized to effectively present analytical results, forming a comprehensive and capable toolkit for data-driven applications.

2.2 Machine learning techniques

2.2.1 K-Nearest Neighbors (KNN)

The KNN algorithm is well known for its ease of use, simplicity, as well as effectiveness [36]. KNN is comparable to random forests (RF) and ANN in that it may be utilized in both regression and classification applications. The utilization of this technique offers several advantages:

1. It exhibits a clear and intelligible essence, which qualifies it for practical application.
2. When used for classification as well as regression, it can learn non-linear decision boundaries and becomes more versatile by allowing adjustment of the K value to define these boundaries.
3. In contrast to other algorithms, KNN doesn't require a specific training phase.
4. The technique utilizes a single hyperparameter, represented as K, simplifying the adjustment of other hyperparameters.

Finding a collection of K samples—often identified by a distance function—that show proximity to unknown samples within the calibration dataset is the basic idea behind KNN. Reaching this goal involves identifying groups of matching samples. To identify the classes of unidentified samples, KNN then computes the average of the response variables and compares the outcome to that of a set of K samples [37]. In this way, the KNN algorithm's selection of value for K is crucial to its effectiveness [38]. Equations (1) through (3) are the three distance functions that are used for this purpose in regression tasks to calculate the distances between neighboring data points:

$$F(e) = \sqrt{\sum_{i=0}^{f}(x_i - y_i)^2} \tag{1}$$

$$F(ma) = \sum_{i=0}^{f}|x_i - y_i| \tag{2}$$

$$F(mi) = \left(\sum_{i=0}^{f}(|x_i - y_i|)^q\right)^{\frac{1}{q}} \tag{3}$$

In this context, $F(e)$ denotes the Euclidean distance function, $F(ma)$ represents the Manhattan distance function, and $F(mi)$ stands for the Minkowski distance function. x_i and y_i That is, it concerns the ith dimension

of the data points x and y, where q is an order parameter utilized to calculate the distances between them.

2.2.2 Mountain gazelle optimizer (MGO)

In order to replicate the behavior of mountain gazelles, the MGO (Mountain Gazelle Optimization) algorithm divides them into three groups: territorial males, maternity herds, and bachelor males. It balances exploitation and exploration to approach optimal solutions, focusing on cost-effective individuals inspired by the least competitive gazelles among bachelor herds [39].

- **Territorial solitary males**

Upon maturing and gaining physical strength, male mountain gazelles establish solitary territories they vigorously defend. These territories are typically spaced apart, leading to confrontations and dominance battles over territory and access to females. Younger males seek to establish themselves and mate with females, while adult males staunchly protect their territories. Equation (4) serves as a model for these territories maintained by adult male gazelles.

$$TSM = male_{gzl} - \left|\left(ri_1 \times YH - ri_2 \times X(t)\right) \times F\right| \times Cof_r \tag{4}$$

Equation (4) describes $male_{gzl}$ as the situation vector shows the greatest universal resolution, which is the adult male. The variables ri_2 and ri_1 show arbitrary integers that can take on a value of either 1 or 2. YH denoted the coefficient vector of the young male herd and is obtained by Equation (5). Similarly, F is obtained by Equation (6). During every iteration, the randomly chosen coefficient vector Cof_r is updated and utilized to boost the search power. Equation (7) can be employed to define this coefficient vector.

$$YH = X_{ra} \times \lfloor r_1 \rfloor + M_{pr} \times \lceil r_2 \rceil, \quad ra = \left\{\left\lceil \frac{N}{3} \right\rceil \ldots N\right\} \tag{5}$$

Here, X_{ra} denotes an arbitrary resolution (young male) in ra's range. M_{pr} refers to the search experts' mean amount, which is equal to $\left\lceil \frac{N}{3} \right\rceil$ And N shows the whole gazelles' amount when r_1 and r_2 show arbitrary amounts in [0,1].

$$F = N_1(D) \times \exp\left(2 - Iter \times \left(\frac{2}{MaxIter}\right)\right) \tag{6}$$

Equation (6) incorporates various problem-specific variables, incorporating the use of the exponential function (exp) and N_1, a number chosen at random from a standard distribution. It also considers the current iteration ($Iter$) and the total number of iterations ($MaxIter$) in the process.

$$Cof_i = \begin{cases} (x+1) + r_3, \\ x \times N_2(D), \\ r_4(D), \\ N_3(D) \times N_4(D)^2 \times \cos\left((r_4 \times 2) \times N_3(D)\right), \end{cases} \tag{7}$$

$$x = -1 + Iter \times \left(\frac{-1}{MaxIter}\right) \tag{8}$$

Additionally, r_3, r_4, and $rand$ are random numbers from 0 to 1. The random numbersN_2, N_3, and N_4 are related to the dimensions of the problem and are drawn from a standard distribution. $Iter$ represents the number of iterations that are being performed, and $MaxIter$ is the number of iterations that are to be executed.

- **Maternity herds**

Maternity herds are crucial for producing robust male gazelles. Equation (9) mathematically describes the interaction between male gazelles assisting in childbirth and juvenile guys trying to mate with females.

$$MH = \left(YH + Cof_{1,r}\right) + (ri_3 \times male_{gzl} - ri_4 \times X_{rand}) \times Cof_{1,r} \tag{9}$$

Here, YH signifies the influence factor vector of adolescent males. Coefficient vectors $Cof_{2,r}$ and $Cof_{3,r}$ are independently established through random selection. Two random integers, ri_3 and ri_4, can have values of 1 or 2. The maximum universal resolution (adult male) in the current iteration was indicated by the symbol $male_{gzl}$. In the end, the position vector of a gazelle selected at random from the entire population is represented by X_{rand}.

- **Bachelor male herds**

As male gazelles mature, they establish territories and compete intensely with younger and with adult males in the mating chase for access to females and territory dominance, as mathematically depicted in Equation (10).

$$YMH = (X(t) - D) + (ri_5 \times male_{gazelle} - ri_6 \times YH) \times Cof_r \tag{10}$$

$$D = (|X(t)| + |male_{gzl}|) \times (2 \times r_6 - 1) \tag{11}$$

Where, $X(t)$ indicates the position vector of the Gazelle in the current iteration. The variables ri_5 and ri_6 are random integers that can take a value of either 1 or 2. The position vector of the male Gazelle (the best solution) is denoted by $male_{gzl}$. Another random number between 0 and 1 is r_6.

- **Migration to search for food**

Mountain gazelles have a strong tendency for food foraging, which sometimes involves extensive journeys to locate suitable food sources or engage in migration. Their exceptional running speed and jumping abilities are

crucial in this behaviour, as mathematically represented in Equation (12).

$$MSF = (ul - ll) \times r_7 + ll \quad (12)$$

Where ul and ll define the upper and lower problem limits in order. Moreover, r_7 is a randomly selected integer within the range [0,1]. The Algorithm 1 is the pseudocode for the MGO. In addition, the flowchart of MGO illustrated in Figure 3.

Algorithm 1. Pseudocode of MGO.

```
MGO setting
Inputs: The population size N and maximum number of iterations I
Outputs: Gazelle's location and fitness potential

Initialization
Create a random population using Xi (i = 1, 2, ..., N)
Calculate Gazelle's fitness level.

While (stopping condition is not met) do
    for (each Gazelle Xi) do
        % Single male
        Calculate TSM
        Mother and child herd
        Calculate MH
        Young male herd
        Calculate YMH
        Migration to search for food
        Calculate MSF
        Calculate the fitness values of TSM, MH, YMH, and MSF, then add them to the habitat
    end for
    Sort the entire population in ascending order
    Update bestGazelle
    Save the N Best Gazelles in the Max number of population
end while

Return XBestGazelle, best Fitness
```

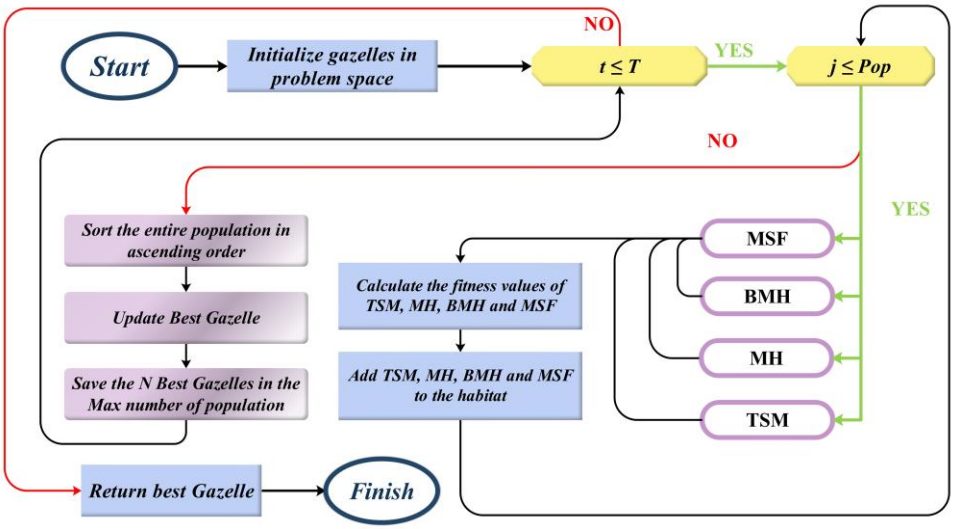


Figure 3: The flowchart of the MGO algorithm.

2.2.3 Coronavirus herd immunity optimizer (CHIO)

Six main stages of the CHIO algorithm are described below [40]:

Step (1): At the start, the CHIO parameters are configured along with the initiation of the optimization issue. This includes constructing the optimization issue with the utilization of an objective function, which will be further detailed below.

$$Min\ f(x)\ \ x \in [LB, UB] \tag{13}$$

Here, the immunity rate, represented by the objective function $f(x)$, is evaluated for a particular instance (or person) represented by $x = (x_1, x_2, \dots, x_n)$, where each x_i relates to the gene (or decision variable) linked to the index i. Each individual contains a total of n genes and the values for each gene, x_i, are confined within the range [LB, UB], determined by their respective lower (LB) and upper bounds (UB).

The CHIO system consists of two control parameters and four algorithmic parameters.

The algorithm parameters are as follows:

- I_0: Stands for the initial count of infection cases, initially set to one.
- $Max - Itr$: Specifies the maximum number of iterations.
- Ps: Governs the population size.
- n: Represents the dimensionality of the problem.

At this stage, two crucial control parameters for CHIO are set to their initial values:

- The Basic Reproduction Rate BR_r, which quantifies viral pandemic transmission and significantly influences the decisions made by CHIO operators.
- The outcome of infected individuals is significantly influenced by the maximum age for infection (Max_{Age}). Individuals who reach this age are documented as having recovered or having passed away due to the illness.

Step (2): This stage aims to establish a population possessing herd immunity. It commences by generating a set of cases equivalent to the capacity of the host immune system (Ps). These instances are subsequently kept in the population classified as persons with herd immunity (P_{HI}) in a matrix-sized $n \times Ps$, as elaborated below:

$$P_{HI} = \begin{bmatrix} x_1^1 & x_2^1 & \cdots & x_n^1 \\ x_1^2 & x_2^2 & \cdots & x_n^2 \\ \vdots & \vdots & \vdots & \vdots \\ x_1^{Ps} & x_2^{Ps} & \cdots & x_n^{Ps} \end{bmatrix} \tag{14}$$

In this step, each casex_j is generated using the formula $x_i^j = LB_i + (UB_i - LB_i) \times U(0,1)$, for all $i = 1,2, \dots, n$. Additionally, a status vector (S) of length Ps is initialized for all cases in P_{HI}, with values of 0 indicating susceptibility and 1 indicating infection. With a count equal to I_0, the initial number of units in set S with a value of 1 is created at random.

Step (3): The genes (x_i^j) in each case x_j can experience stability or be influenced by social distancing measures, depending on three rules defined by the percentage of the Basic Reproduction Rate (BR_r).

$$x_i^j(t+1) \leftarrow \begin{cases} x_i^j(t) & r \geq B \\ C\left(x_i^j(t)\right) & r < \frac{1}{3} \times B.\ //infected\ case \\ N\left(x_i^j(t)\right) & r < \frac{2}{3} \times B.\ //susceptible\ case \\ R\left(x_i^j(t)\right) & r < B.\ //immuned\ case \end{cases} \tag{15}$$

A variable "r" generates a random number from zero to one, and B stands for BR_r. The three regulations set out the following terms and conditions.

$infected\ case$: When $r \in [0, \frac{1}{3}B)$, the impact of social distance on the gene expression $x_i^j(t+1)$ relies on the difference between the gene expression of an infected case x^m and the present gene expression, which is defined as follows:

$$x_i^j(t+1) = C\left(x_i^j(t)\right) \tag{16}$$

Where:

$$C\left(x_i^j(t)\right) = x_i^j(t) + r \times (x_i^j(t) - x_i^c(t)) \tag{17}$$

The status vector (S), where c denotes the set of i values for which S_i equals 1, determines the value of $x_i^c(t)$, which is selected at random from an infected case x^c.

$susceptible\ case$: The gene value $x_i^j(t+1)$ falls within the range of $r \in [\frac{1}{3}B, \frac{2}{3}B)$, and it is influenced by social distancing measures. These measures are determined by the difference between the current gene and a gene selected from a susceptible case x^m, as elaborated below:

$$x_i^j(t+1) = N\left(x_i^j(t)\right) \tag{18}$$

Where:

$$N\left(x_i^j(t)\right) = x_i^j(t) + r \times (x_i^j(t) - x_i^m(t)) \tag{19}$$

From a vulnerable situation x^m specified by the status vector (S), where m is the set of i values where S_i equals 0, $x_i^m(t)$is a randomly selected value.

$Immuned\ case$: The gene value $x_i^j(t+1)$ within the range of $r \in [\frac{2}{3}B, B)$, is modified to social distancing measures, which are influenced by the disparity between the current gene and a gene selected from a susceptible case x^ω as described:

$$x_i^j(t+1) = R\left(x_i^j(t)\right) \tag{20}$$

Where:

$$R\left(x_i^j(t)\right) = x_i^j(t) + r \times (x_i^j(t) - x_i^\omega(t)) \tag{21}$$

In a susceptible situation x^ω, which is defined by the status vector (S), $x_i^\omega(t)$ is a randomly selected value such that $f(x^\omega) = \min_{j \sim \{k | S_k = 2\}} f(x^j)$.

Step (4): The immunity rate, denoted as $f(x^j(t+1))$, is computed for each new case, $x^j(t+1)$. If the new rate is superior, meaning when $f(x^j(t+1)) < f(x^j(t))$, the current case $x^j(t)$ is replaced with a new one. In cases where S_j equals 1, the age vector A_j increases by one. The status vector (S_j) is updated for each individual (x^j) based

on the herd immunity threshold, as determined by the subsequent equation:

$$S_j \leftarrow \begin{cases} 1\ f\left(x^j(t+1)\right) < \dfrac{f(x)^j(t+1)}{\Delta f(x)} \wedge S_j = 0 \wedge is\_corona(x^j(t+1)) \\ 2\ f\left(x^j(t+1)\right) > \dfrac{f(x)^j(t+1)}{\Delta f(x)} \wedge S_j = 1 \end{cases} \quad (22)$$

Here's the term: $is\_corona(x^j(t+1))$ is given a value of 1 if the condition is inherited by the newly diagnosed case $x^j(t+1)$ from any previously infected case. The population's mean immunity rates are represented by the symbol$\Delta f(x)$, which is defined as$\frac{\sum_{i=1}^{Ps} f(x_i)}{Ps}$.

Step (5): When a person who has developed a certain illness and is classified as a case (with $S_j = 1$), and their immunity rate $f\left(x^j(t+1)\right)$ stays constant for a predetermined number of iterations as given by the parameter Max_Age $(i.e., A_j \geq Max_{Age})$, then the case is classed as a fatality.

Then, it is entirely recreated with $x^j(t+1) = LB_i + (UB_i - LB_i) \times U(0,1)$ for all $i = 1,2,\ldots,n$. Additionally, both A_j and S_j values are set to zero.

Step (6): The CHIO framework advances from $Step$ 3 to $Step$ 6 until it fulfils a pre-established termination condition, often set by a maximum iteration threshold. The population comprises a considerably larger number of susceptible and immune individuals than the population itself, and no further infections are occurring. Figure 4 displays the flowchart of the CHIO.

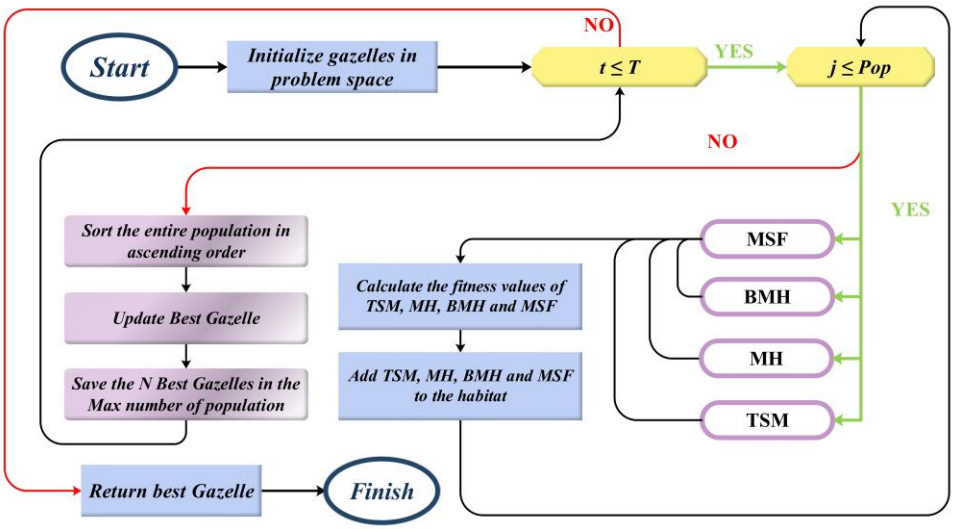


Figure 4: The flowchart of the CHIO algorithm.

2.3 K-Fold cross validation

To ensure robust evaluation of the prediction model, this study adopts the k-fold cross-validation (KCV) technique—a widely accepted and systematic approach to model validation. In k-fold cross-validation, the dataset is divided into *k* equally sized folds, where each subset is used once as the testing set while the remaining *k-1* subsets serve as the training set. This cycle is repeated *k* times, ensuring that every data point is used for both training and testing exactly once.

In this work, 5-fold cross-validation (k = *5*) is employed. The dataset is partitioned into five equal parts, and the algorithm is executed over five iterations. In each iteration, one-fold is designated for testing and the remaining four for training. The performance metrics across all five iterations are then averaged to provide a comprehensive assessment of the model's predictive capabilities.

As illustrated in Figure 5, the evaluation of the KNN model reveals that the fifth fold outperformed the others, achieving the highest coefficient of determination (R^2 = 0.971) and the lowest root mean squared error (RMSE = 168.23), marking it as the most optimal split.

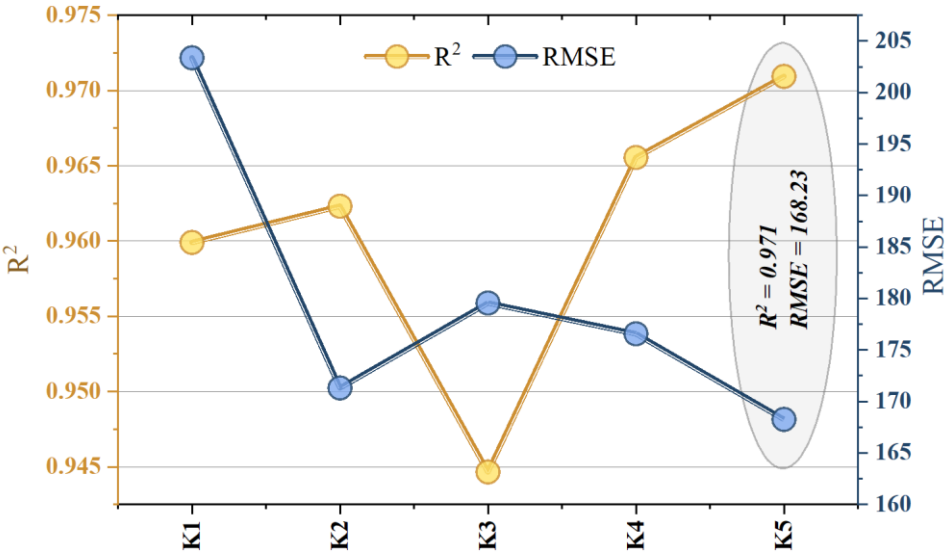


Figure 5: The results of 5-Fold Cross validation.

2.4 Metrics for performance assessment

To evaluate the performance and predictive accuracy of the developed models, a set of widely accepted statistical metrics are employed, each offering unique insight into different aspects of model behavior in regression tasks. The selection of these metrics—R^2, RMSE, NRMSE, MSE, and SI—is based on their proven effectiveness and interpretability in capturing model accuracy, error magnitude, relative performance, and data dispersion. These indicators are particularly suitable for this study, which involves continuous prediction of geotechnical properties, where both absolute and relative errors are of interest. While alternative metrics (e.g., MAE, MAPE) were considered, the chosen set offers a well-rounded, interpretable, and comprehensive assessment framework.

Coefficient of Determination (R^2): This metric quantifies the proportion of variance in the observed data that is predictable from the model. It captures the strength and direction of the linear relationship between observed and predicted values. A higher R^2 indicates better model fit. It is defined as:

$$R^2 = \left(\frac{\sum_{i=1}^{n}(t_i - \bar{w})(v_i - \bar{v})}{\sqrt{[\sum_{i=1}^{n}(v_i - w)^2][\sum_{i=1}^{n}(v_i - \bar{v})^2]}} \right)^2 \quad (23)$$

Root Mean Square Error (RMSE): RMSE measures the average magnitude of prediction errors, offering an intuitive understanding of error in the same units as the target variable. It is particularly sensitive to large deviations, making it effective for detecting significant outliers:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(v_i - w_i)^2} \quad (24)$$

Normalized Root Mean Square Error (NRMSE): This is a scale-independent version of RMSE, enabling comparison across datasets or models with differing output ranges. It provides a normalized perspective on model performance:

$$\text{NRMSE} = \sqrt{\frac{\frac{1}{n}\sum_{i=1}^{n}(v_i - w_i)^2}{\sum_{i=1}^{n}(w_i)^2}} \quad (25)$$

Mean Squared Error (MSE): As the square of the standard deviation of prediction errors, MSE penalizes larger errors more severely, thus highlighting models that consistently produce significant deviations:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(v_i - w_i)^2 \quad (26)$$

Scatter Index (SI): SI measures the dispersion of the prediction errors relative to the mean observed value, providing a normalized indicator of how tightly predictions cluster around the true values. It is particularly useful for comparing performance across datasets with different scales:

$$SI = \frac{RMSE}{w_i} \quad (27)$$

Together, these metrics offer a comprehensive evaluation of both the absolute accuracy and relative performance of the predictive models, ensuring robust, multi-faceted performance validation.

In all the above equations:

n: number of samples,
v_i: predicted value,
$\bar{v}$: average predicted value,
w_i: experimentally measured value,
$\bar{w}$: average experimentally measured value.

3 Results

The primary objective of this study was to investigate the reliability of a single KNN model and optimized versions based on MGO and CHIO in predicting the Undrained Shear strength (USS) of soils. To achieve this goal, the models were divided into three categories: a training set, a

validation set, and a testing set, with 70%, 15%, and 15% of the models, respectively. Subsequently, the precision of each model in forecasting the USS was evaluated based on five evaluation metrics, and the results were presented through comparative charts and graphs, as presented in the following sections.

3.1 Results of hyperparameters and convergence curves

Effective hyperparameter tuning plays a critical role in enhancing the performance and generalization capability of machine learning models. While several strategies exist—including manual tuning, grid search, random search, and Bayesian optimization—the choice of method often depends on model complexity, the number of tunable parameters, computational resources, and time constraints.

In this study, a random search strategy was adopted for hyperparameter optimization due to its efficiency in exploring high-dimensional spaces without exhaustive computation. This method randomly samples combinations from predefined hyperparameter ranges, increasing the likelihood of identifying optimal configurations with reduced evaluation overhead.

As summarized in Table 3, the hyperparameter search focused on tuning three key parameters for the KNN-based hybrid models (KNCH and KNMG): n_neighbors, leaf_size, and P. These settings were varied systematically to maximize model performance while maintaining computational efficiency.

Table 3: The result of developed models for KNN

Hyperparameters	Hybrid Models	
	KNCH	KNMG
n_neighbors	14	20
leaf_size	958	736
p	813	905

Tracking RMSE over iterations is a common strategy to monitor how well a machine learning model is learning. Convergence plots help visualize this process by revealing trends in error reduction and indicating the model's ability to reach an optimal state.

As shown in Figure 6, the convergence performance of two hybrid models—KNCH and KNMG—was analyzed across 200 iterations. The KNCH model displayed a consistent downward trajectory in RMSE, ultimately stabilizing around 60, which points to efficient learning and robust optimization. On the other hand, KNMG started with a higher RMSE and leveled off closer to 90, reflecting a comparatively slower convergence.

These findings clearly demonstrate the superior convergence efficiency of KNCH over KNMG, making it a more promising candidate for modeling under the given setup.

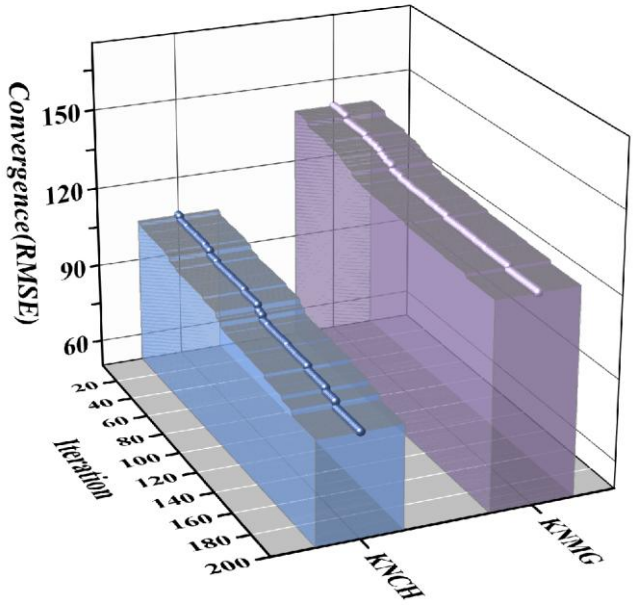


Figure 6: The Convergence curves for hybrid model

3.2 Results of evaluation metrics for models

This section compares the employed models' productivity by utilizing five statistical evaluators, R^2, RMSE, MSE, NRMSE, and SI, as depicted in Table 4. In terms of R^2, it is evident that the best result is for the KNCH in the training, validation, and testing parts, in which R^2 is 0.993, 0.987, and 0.977, respectively. Among the three models, KNCH consistently achieved the best results across the training, validation, and testing phases, with R^2 values of 0.993, 0.987, and 0.977, respectively. It is important to note that R^2 values in the testing phase are naturally lower than in the training phase, as the testing process is performed solely on unseen input data, without access to the corresponding output values. In contrast, during training, the model learns from both inputs and known outputs, allowing it to optimize its parameters. Therefore, this drop does not necessarily indicate improper training; rather, it reflects the realistic challenges of generalizing to new data. The relatively small decline in KNCH's R^2 values indicates strong generalization capability, especially when compared to KNMG (R^2: 0.985 → 0.976 → 0.967) and KNN (R^2: 0.975 → 0.958 → 0.950), where larger performance gaps are evident.

Comparing values of the SI in which lower values mean the highest model accuracy, it is evident that the optimal value of 0.065 was obtained in the validation phase of the KNCH, while this value for KNMG and KNN was approximately 50% and 30% higher. RMSE, NRMSE, and MSE outcomes demonstrate that the KNCH with the least error values ($RMSE = 85.192$, $NRMSE = 0.470$, and $MSE = 7259.15$) exhibited the highest level of performance, while KNMG and KNN placed second and third in rank, respectively.

Table 4: The result of developed models for KNN.

Phase	Model	Index values							
		RMSE	R^2	MSE	NRMSE	SI	Confidence Interval-lower	Confidence Interval-upper	St. dev of Errors
Train	KNN	163.81	0.975	26866	1.171	0.134	22163	34447	21.909
	KNCH	85.19	0.993	7259	0.609	0.070	6374	9907	10.991
	KNMG	123.05	0.985	15131	0.879	0.101	12756	19827	16.296
Validation	KNN	173.38	0.958	30061	5.779	0.106	20698	76344	19.327
	KNCH	106.90	0.987	11428	3.563	0.065	7028	25922	11.461
	KNMG	137.24	0.976	18835	4.575	0.084	9137	33701	14.011
Test	KNN	182.35	0.950	33252	6.078	0.159	19319	71260	26.407
	KNCH	116.62	0.977	13600	3.887	0.102	8554	31552	14.600
	KNMG	141.40	0.967	19994	4.713	0.123	12056	44469	19.458
All	KNN	168.17	0.971	28303	0.841	0.132	23571	34951	22.325
	KNCH	93.99	0.991	8836	0.470	0.074	7358	10911	11.701
	KNMG	128.15	0.983	16416	0.641	0.101	13623	20201	16.620

Figure 7 presents scatter plots that depict the relationship between the predicted USS values and their corresponding experimentally measured values for all three models across the training, validation, and testing phases. These plots serve as a visual diagnostic tool to evaluate both the accuracy and consistency of model predictions.

The x-axis represents the actual (measured) USS values, while the y-axis corresponds to the predicted values, adhering to standard regression visualization conventions. Each subplot also incorporates three key graphical elements:

1. A diagonal reference line (Y = X), representing perfect predictions where predicted values exactly match measured values.
2. A best-fit linear regression line, indicating the actual trend followed by the predictions.
3. Two boundary lines (Y = 0.9X and Y = 1.1X), denoting a ±10% margin of error—used to assess the practical range within which most predictions fall.

The inclusion of Root Mean Square Error (RMSE) and Coefficient of Determination (R^2) values in each panel quantitatively supports the scatter distributions. RMSE measures the average magnitude of prediction errors, with lower values indicating less dispersion and higher data point density around the ideal fit. R^2 reflects the proportion of variance explained by the model; values closer to 1 imply a stronger correlation and tighter clustering of data along the ideal line.

Among the three models, KNCH demonstrates superior performance, with the highest R^2 values of 0.993 (training), 0.987 (validation), and 0.977 (testing). These results show a strong linear agreement between predicted and measured values. Concurrently, it exhibits the lowest RMSE values of 85.19, 106.90, and 116.62, respectively, indicating minimal prediction error and more compact clustering of data points.

While some overlap of data points is observed—due to the relatively narrow range of USS values in the dataset—this visualization remains valuable in confirming both the accuracy and consistency of the KNCH model across different data splits. Future work may incorporate density plots or zoomed views to improve interpretability when data point congestion is high.

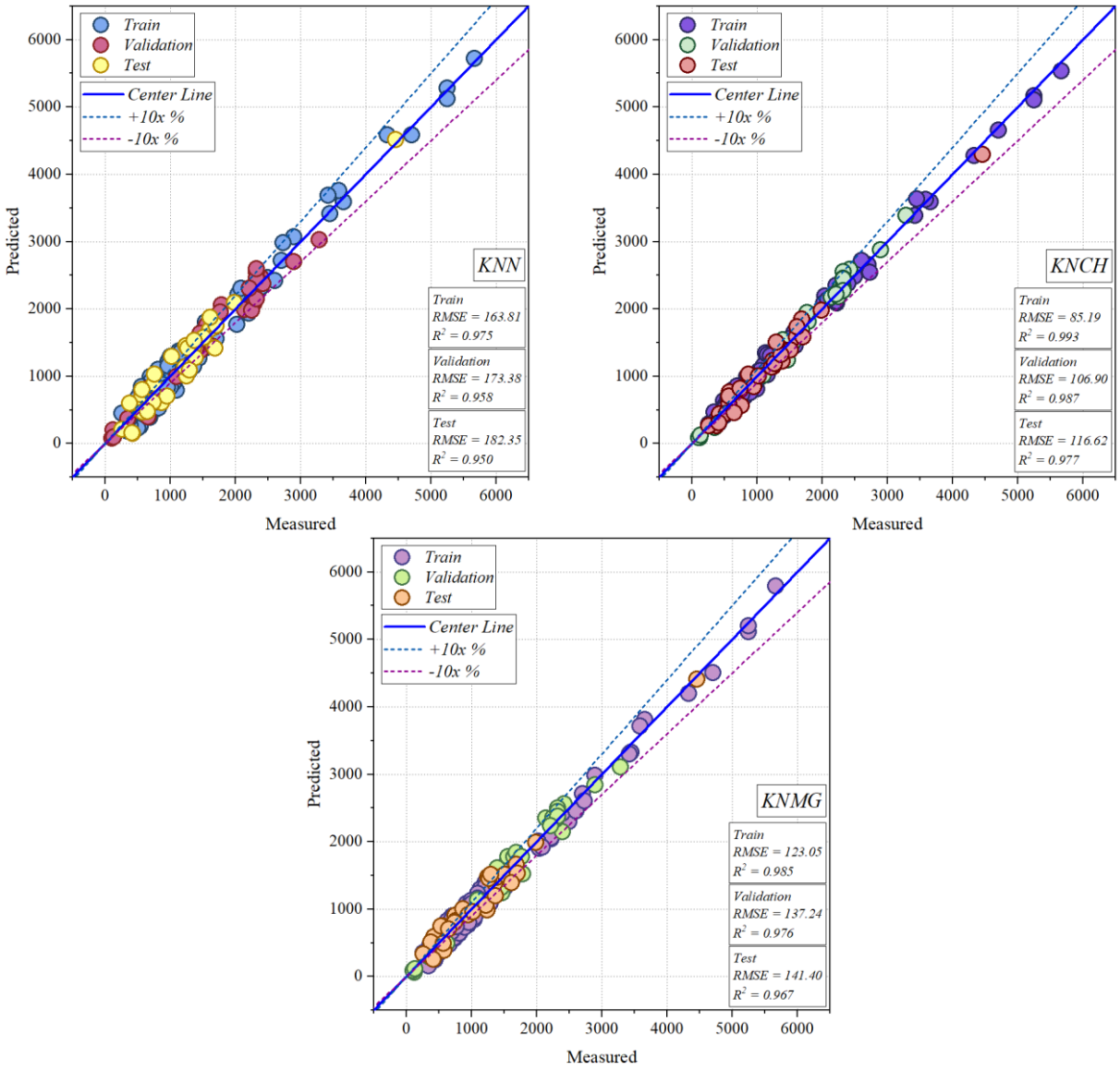


Figure 7: The scatter plot for developed hybrid models

The bar chart in Figure 8 presents a visual comparison of R^2, RMSE, and NRMSE (vertical axes) for three developed models in divided phases of training, validation, and testing (horizontal axes). Differences between R^2 values for the two hybrid models are marginal, with KNCH being better than KNMG, but KNN has a noticeable 3% difference with them. The discrepancy of models in the case of error values is more remarkable, especially in RMSE, in which the maximum value of 163.814 related to the KNN is almost twice the minimum value of 85.192 related to the KNCH model. Turning to the NRMSE charts, again, the same pattern as mentioned for RMSE is clear there. Also, corresponding values of NRMSE in validation and testing parts are 5 to 6-fold higher than that for the training part.

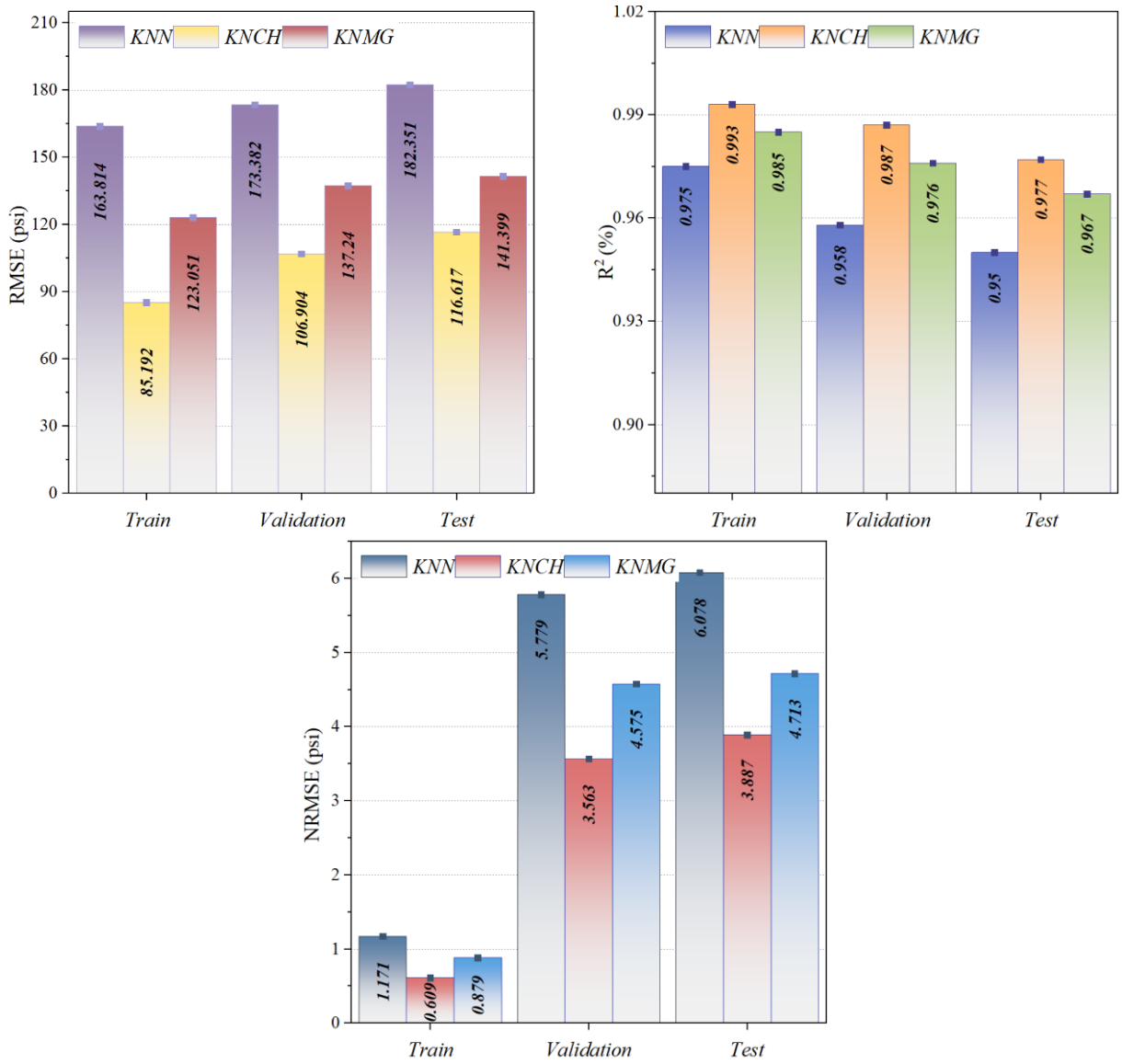
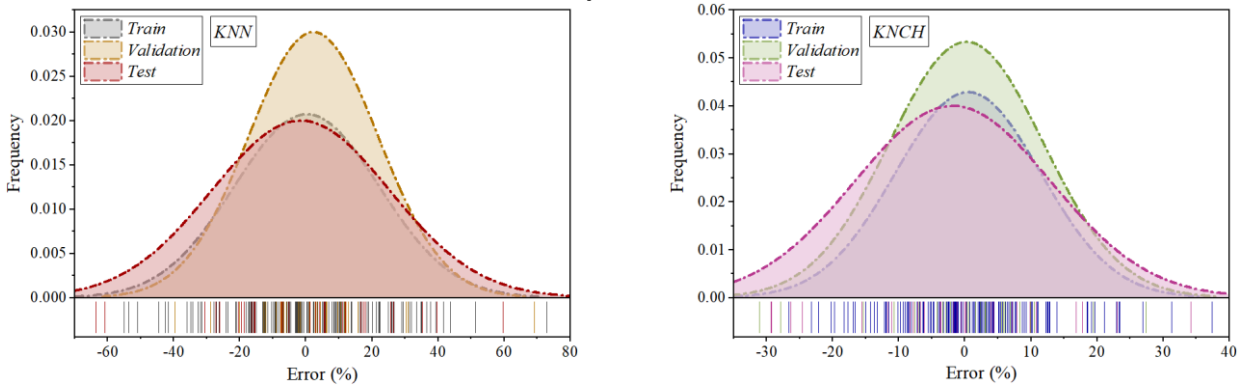


Figure 8: Comparison between models based on RMSE, R2, NRMSE

The normal distribution plot in Figure 9 exemplifies a graphical analysis of errors for the KNN, KNCH, and KNMG models. As the error range of KNN is more widely spread and diagrams related to the KNCH are more narrow bell-shaped ones, KNN and KNCH respective are less and more accurate models.



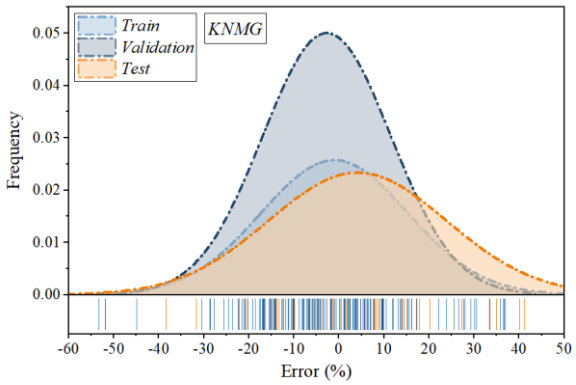


Figure 9: The error percentage for the models based on the normal distribution plot.

Figure 10 presents a comprehensive visualization of percentage prediction errors across all 200 data samples for the three evaluated models—KNN (baseline), KNCH (CHIO-optimized KNN), and KNMG (MGO-optimized KNN)—during the training, validation, and testing phases. This line plot was intentionally chosen for its ability to visually capture the behavior and variability of each model's prediction errors across individual samples, offering deeper insights beyond aggregate statistical metrics such as RMSE or R^2. The x-axis represents the sample number (1–200), while the y-axis indicates the prediction error (%). This enables a sequential view of how the models performed on each data point. The three vertical dashed lines divide the dataset into its respective training, validation, and testing partitions, which were used consistently across all models.

The baseline KNN model shows significant fluctuations and a tendency toward larger error spikes, especially in the training and validation phases. The maximum error value of 72.8%, highlighted in red, indicates susceptibility to outliers or poor generalization on certain samples. The KNCH model exhibits the most stable error profile, with error values closely clustered around zero and minimal large deviations, especially in the test phase. This suggests a higher predictive reliability and generalization capability. The KNMG model demonstrates performance improvements over the original KNN but displays slightly more variability than KNCH, indicating moderate robustness.

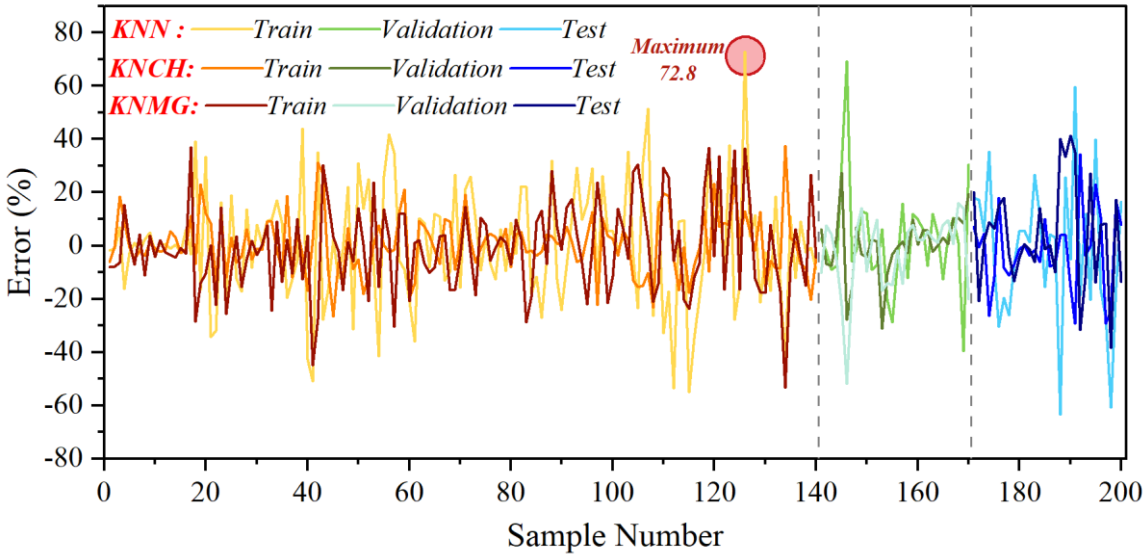


Figure 10: The Line plot for errors of the developed models.

- **Statistical significance test based on wilcoxon**

Table 5 summarizes the outcomes of the Wilcoxon signed-rank test, which was employed to statistically evaluate the performance differences among the developed models. The test was conducted to determine whether there were significant differences in the distributions of prediction errors.

According to the results, none of the models—KNN_MGO, KNN_CHIO, and KNN—show statistically significant differences, as all p-values are well above the 0.05 threshold. Specifically, the KNN_CHIO model yielded the highest p-value (0.882), followed by KNN (0.668) and KNN_MGO (0.347). The corresponding test statistics (9280, 9929, and 9699, respectively) further support the consistency in performance across these models.

These findings indicate that, under the given conditions, the predictive outputs of the three models are not significantly different from each other in a statistical sense.

Table 5: The result of Wilcoxon test.

Model	P-value	Statistics
KNN_MGO	0.347	9280
KNN_CHIO	0.882	9929
KNN	0.668	9699

- **Comparison with previous studies**

Table 6 presents a comparative analysis between the results of the present study and several recent state-of-the-art articles focused on predicting USS using machine learning techniques. The proposed KNN model in this study, trained on 200 samples using a limited yet effective set of inputs—LL, PL, SF, and OBW—achieved an excellent coefficient of determination ($R^2 = 0.993$). While the RMSE value (85.19) appears higher compared to other studies, it is important to consider that the scale and units of output values may differ, influencing RMSE interpretation.

In comparison, Zarei et al. [41] attained a slightly higher R^2 of 0.9975 using a deep neural network (DNN) with a more complex input set, albeit on a smaller dataset (72 samples). Zhang et al. [42] utilized the XGBoost model on a larger dataset (304 samples) and reported a lower R^2 (0.92), indicating that while their RMSE (2.38) was lower, the model's overall explanatory power was not as strong. Pham et al. [43] applied a hybrid RF-PSO model and obtained an R^2 of 0.89 with an impressively low RMSE (0.453), reflecting a trade-off between error minimization and data generalization. Elsawy et al. [44] implemented a Fine Gaussian Support Vector Regression (SVR), yielding strong performance ($R^2 = 0.96$; RMSE = 1.65) based on 111 data points.

Overall, the present study demonstrates competitive predictive accuracy using fewer input variables and a simpler model, underscoring its practicality and efficiency for real-world geotechnical applications. The high R^2 value confirms the model's robustness, despite RMSE differences which arising from dataset scaling or units of output.

Table 6: The Comparison between the results of present study and available articles.

Article	Dataset size	Inputs	Model	Metrics	
				R^2	RMSE
Zarei et al. [41]	72	vertical stress, percentage of the crushed tire, percentage of clay, size of clay, specific gravity of tires, Liquid limit, Plastic limit and Specific gravity of clay samples	DNN	0.9975	2.42
Zhang et al. [42]	304	vertical effective stress, pre-consolidation stress, liquid limit, plastic limit, and natural water content	XGBoost	0.92	2.38
Pham et al. [43]	127	Clay content, Water content, Specific gravity, Void ratio, Liquid limit, and Plastic limit	RF-PSO	0.89	0.453
Elsawy et al. [44]	111	natural water content, dry unit weight, liquid limit, plasticity index, consistency index, void ratio, specific gravity, and pocket penetration shear	Fine Gaussian SVR	0.96	1.65
Present study	**200**	**liquid limit (LL), plastic limit (PL), and sleeve friction (SF), overburden weight (OBW)**	**KNN**	**0.993**	**85.19**

4 Discussion

4.1 Practical implications

The findings of this study have direct relevance to geotechnical engineering, particularly in regions where soft soils are prevalent, such as coastal and riverine zones. By accurately predicting undrained shear strength (USS) using easily measurable soil parameters (LL, PL, SF, and OBW), engineers can make more informed decisions without relying solely on time-consuming and expensive laboratory or field tests. The hybrid models, especially KNCH, offer a computationally efficient alternative that retains high predictive accuracy, making them suitable for integration into early-stage geotechnical investigations, feasibility assessments, and large-scale infrastructure planning. Moreover, these models can be embedded into software tools or decision-support systems, enhancing design safety and reducing costs associated with soil characterization.

4.2. Limitations of the study

While the results are promising, several limitations should be acknowledged. First, the dataset used in this study, though diverse, was limited to 200 samples. A larger and more geographically varied dataset could improve the generalizability of the models. Second, the input features were selected based on availability and relevance, but additional parameters such as moisture content, particle size distribution, or in-situ test data might further enhance model robustness. Third, although the optimization algorithms (MGO and CHIO) significantly improved model performance, their convergence behavior and computational cost were not thoroughly analyzed, which

could affect scalability for larger datasets or real-time applications.

4.3. Potential future studies

Future research could expand on this work in several directions. Firstly, enlarging the dataset with samples from different soil types and geographic locations would enhance model generalizability and allow for regional model calibration. Secondly, incorporating more advanced machine learning models such as deep neural networks (DNNs), ensemble methods, or transfer learning approaches may yield even better predictive performance. Additionally, future studies could explore real-time or semi-real-time implementation of the proposed models in geotechnical monitoring systems. It is also recommended to investigate the comparative effectiveness of other nature-inspired optimization algorithms and assess their efficiency in terms of convergence speed, stability, and computational load.

5 Conclusion

This study tackled the complex problem of predicting undrained shear strength (USS) in soft soils near water bodies—an essential task in geotechnical engineering for ensuring structural safety and stability. The introduction contextualized the challenge posed by soft soils and highlighted the importance of USS in engineering design, along with the increasing application of machine learning (ML) techniques to enhance prediction accuracy. The research aimed to construct and evaluate data-driven models using non-linear ML approaches, particularly the K-Nearest Neighbor (KNN) algorithm and its two optimized variants: KNMG (KNN with Mountain Gazelle Optimizer) and KNCH (KNN with Coronavirus Herd Immunity Optimizer). The models were trained and tested on a dataset of 200 soil samples with four key input features—sleeve friction (SF), liquid limit (LL), plastic limit (PL), and overburden weight (OBW).

To rigorously evaluate the models, five performance metrics—R^2, RMSE, NRMSE, MSE, and Scatter Index (SI)—were used across training, validation, and test phases. The results show that the KNCH model consistently outperformed both the baseline KNN and the KNMG variant across all evaluation metrics and phases. Specifically, KNCH achieved the highest R^2 values and the lowest RMSE, MSE, NRMSE, and SI values, indicating superior accuracy and model stability. For instance, in the test phase, KNCH's RMSE was approximately 36% lower than KNMG's and 56% lower than KNN's. Its R^2 score was also 0.8% and 2% higher than those of KNMG and KNN, respectively.

While KNMG demonstrated improved performance over the baseline KNN, particularly in reducing prediction errors and increasing correlation strength, it was consistently outperformed by KNCH in every phase and metric. The baseline KNN model, although less accurate, provided a valuable benchmark for gauging the impact of optimization algorithms.

Moreover, the graphical analyses revealed that the predicted values from all three models, especially KNCH, closely followed the trends in experimentally measured USS values. This alignment underscores the effectiveness and feasibility of the proposed hybrid ML-optimization frameworks, with KNCH emerging as the most reliable model for predicting USS in soft soils.

Declarations

Availability of data and materials

Data can be shared upon request.

Competing interests

The authors declare no competing interests.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Authors' contributions

NC performed Data collection also WY carried out simulation and LZ conducted analysis. GX evaluate the first draft of the manuscript, also JM perfomed editing and writing.

Acknowledgements

1. Scientific Research Program Funded by Education Department of Shaanxi Provincial Government（Program No.23JP087）
2. Natural Science Basic Research Program of Shaanxi (ProgramNo.2023-JC-YB-464)
3. Xi'an Traffic Engineering Institute Young and Middle-aged Fund Project (ProgramNo.2023KY-44)

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.

References

[1] Duncan, J.M. and S.G. Wrigth (2005). *Soil strength and slope stability*, John Willey & Sons. Inc., Hoboken, New Jersey, 297.

[2] Lechowicz, Z. and M.J. Sulewska (2022). Assessment of the Undrained Shear Strength and Settlement of Organic Soils under Embankment Loading Using Artificial Neural Networks. *Materials*, MDPI, 16(1), p. 125. https://doi.org/10.3390/ma16010125.

[3] behnam Sedaghat, G.G. Tejani and S. Kumar (2023). Predict the Maximum Dry Density of soil based on Individual and Hybrid Methods of *Machine* Learning. *Advances in Engineering and*

Intelligence *Systems*, Biliji Publisher, 002(03). https://doi.org/10.22034/aeis.2023.414188.1129.

[4] Brandão, H., G. Pinto and T. Santos, Application of K-Nearest Neighbors (KNN) method to undrained shear strength prediction of bauxite tailing.

[5] Yang, X (2025). Economic Cost Prediction Model for Building Construction Based on CNN-DAE Algorithm. *Informatica*, Slovenian Society Informatika, 49(5). https://doi.org/10.31449/inf.v49i5.7029.

[6] Dash, C.S.K., S.C. Nayak, A.K. Behera and S. Dehuri (2023). A Neuro-Fuzzy Predictor Trained by an Elitism Artificial Electric Field Algorithm for Estimation of Compressive Strength of Concrete Structures. *Informatica*, Slovenian Society Informatika, 47(5). https://doi.org/10.31449/inf.v47i5.3951.

[7] Benkaddour, M.K (2021). CNN based features extraction for age estimation and gender classification. *Informatica*, Slovenian Society Informatika, 45(5). https://doi.org/10.31449/inf.v45i5.3262.

[8] Maktum, T., N. Pulgam, V. Chandgadkar, P. Pathak and A. Solanki (2025). A Machine Learning Based Framework for Bankruptcy Prediction in Corporate Finances Using Explainable AI Techniques. *Informatica*, Slovenian Society Informatika, 49(15). https://doi.org/10.31449/inf.v49i15.6745.

[9] Ebid, A.M (2021). 35 Years of (AI) in geotechnical engineering: state of the art. *Geotechnical and Geological Engineering*, Springer, 39(2), pp. 637–690. https://doi.org/10.1007/s10706-020-01536-7.

[10] Jeremiah, J.J., S.J. Abbey, C.A. Booth and A. Kashyap (2021). Results of application of artificial neural networks in predicting geo-mechanical properties of stabilised clays—a review. *Geotechnics*, MDPI, 1(1), pp. 147–171. https://doi.org/10.3390/geotechnics1010008.

[11] Mozumder, R.A. and A.I. Laskar (2015). Prediction of unconfined compressive strength of geopolymer stabilized clayey soil using artificial neural network. *Computers and Geotechnics*, Elsevier, 69, pp. 291–300. https://doi.org/10.1016/j.compgeo.2015.05.021.

[12] Ayeldeen, M., Y. Hara, M. Kitazume and A. Negm (2016). Unconfined compressive strength of compacted disturbed cement-stabilized soft clay. *International Journal of Geosynthetics and Ground Engineering*, Springer, 2, pp. 1–10.

[13] Bunyamin, S.A., T.S. Ijimdiya, A.O. Eberemu and K.J. Osinubi (2018). Artificial neural networks prediction of compaction characteristics of black cotton soil stabilized with cement kiln dust. *Journal of Soft Computing in Civil Engineering*, 2(3), pp. 50–71. https://doi.org/10.22115/scce.2018.128634.1059.

[14] Priyadarshee, A., S. Chandra, D. Gupta and V. Kumar (2020). Neural Models for Unconfined Compressive Strength of Kaolin clay mixed with pond ash, rice husk ash and cement. *Journal of Soft Computing in Civil Engineering*, 4(2), pp. 85–102. https://doi.org/10.22115/scce.2020.223774.1189.

[15] Das, S., P. Samui, S. Khan and N. Sivakugan (2011). Machine learning techniques applied to prediction of residual strength of clay. *Open Geosciences*, Degruyter, 3(4), pp. 449–461. https://doi.org/10.2478/s13533-011-0043-1.

[16] Naeim, B., M.R. Akbarzadeh and V. Jahangiri (2024). Machine learning-based prediction of seismic response of elevated steel tanks. *Structures*, Elsevier, 70, p. 107649. https://doi.org/10.1016/j.istruc.2024.107649.

[17] Samui, P (2012). Application of relevance vector machine for prediction of ultimate capacity of driven piles in cohesionless soils. *Geotechnical and Geological Engineering*, Springer, 30, pp. 1261–1270. https://doi.org/10.1007/s10706-012-9539-9.

[18] Mohanty, R. and S.K. Das (2018). Settlement of shallow foundations on cohesionless soils based on SPT value using multi-objective feature selection. *Geotechnical and Geological Engineering*, Springer, 36, pp. 3499–3509. https://doi.org/10.1007/s10706-018-0549-0.

[19] Samui, P., T. Lansivaara and M.R. Bhatt (2013). Least square support vector machine applied to slope reliability analysis. *Geotechnical and Geological Engineering*, Springer, 31, pp. 1329–1334. https://doi.org/10.1007/s10706-013-9654-2.

[20] Pham, B.T., D. Tien Bui and I. Prakash (2017). Landslide susceptibility assessment using bagging ensemble based alternating decision trees, logistic regression and J48 decision trees methods: a comparative study. *Geotechnical and Geological Engineering*, Springer, 35, pp. 2597–2611. https://doi.org/10.1007/s10706-017-0264-2.

[21] Qi, C. and X. Tang (2018). Slope stability prediction using integrated metaheuristic and machine learning approaches: A comparative study. *Computers & Industrial Engineering*, Elsevier, 118, pp. 112–122. https://doi.org/10.1016/j.cie.2018.02.028.

[22] Sakellariou, M.G. and M.D. Ferentinou (2005). A study of slope stability prediction using neural networks. *Geotechnical & Geological Engineering*, Springer, 23, pp. 419–445. https://doi.org/10.1007/s10706-004-8680-5.

[23] Hanna, A.M., D. Ural and G. Saygili (2007). Neural network model for liquefaction potential in soil deposits using Turkey and Taiwan earthquake data. *Soil Dynamics and Earthquake Engineering*, Elsevier, 27(6), pp. 521–540. https://doi.org/10.1016/j.soildyn.2006.11.001.

[24] Njock, P.G.A., S.-L. Shen, A. Zhou and H.-M. Lyu (2020). Evaluation of soil liquefaction using AI technology incorporating a coupled ENN/t-SNE model. *Soil Dynamics and Earthquake*

Engineering, Elsevier, 130, p. 105988. https://doi.org/10.1016/j.soildyn.2019.105988.

[25] Goh, A.T.C. and S.H. Goh (2007). Support vector machines: their use in geotechnical engineering as illustrated using seismic liquefaction data. *Computers and Geotechnics*, Elsevier, 34(5), pp. 410–421. https://doi.org/10.1016/j.compgeo.2007.06.001.

[26] Gajan, S (2021). Application of machine learning algorithms to performance prediction of rocking shallow foundations during earthquake loading. *Soil Dynamics and Earthquake Engineering*, Elsevier, 151, p. 106965. https://doi.org/10.1016/j.soildyn.2021.106965.

[27] Naeim, B., A.J. Khiavi, P. Dolatimehr and B. Sadaghat (2024). Novel Optimized Support Vector Regression Networks for Estimating Fresh and Hardened Characteristics of SCC. *Advances in Engineering and Intelligence System*, Biliji Publisher. https://doi.org/10.22034/aeis.2024.483317.1239.

[28] Behnam Sedaghat, G.G. Tejani and S. Kumar (2023). Predict the Maximum Dry Density of soil based on Individual and Hybrid Methods of Machine Learning. *Advances in Engineering and Intelligence Systems*, Biliji Publisher, 002(03). https://doi.org/10.22034/aeis.2023.414188.1129.

[29] Wroth, C.P. and D.M. Wood (1978). The correlation of index properties with some basic engineering properties of soils. *Canadian Geotechnical Journal*, Canadian Science Publishing, 15(2), pp. 137–145. https://doi.org/10.1139/t78-014.

[30] Vinod, P., A. Sridharan and K.A. Deepa (2013). Remoulded shear strength at plastic and semi-solid states. *Proceedings of the Institution of Civil Engineers-Geotechnical Engineering*, ICE Virtual Library, 166(4), pp. 415–424. https://doi.org/10.1680/geng.11.00071.

[31] Zafar, M.H., M. Mansoor, M. Abou Houran, N.M. Khan, K. Khan, S.K.R. Moosavi and F. Sanfilippo (2023). Hybrid deep learning model for efficient state of charge estimation of Li-ion batteries in electric vehicles. *Energy*, Elsevier, 282, p. 128317. https://doi.org/10.1016/j.energy.2023.128317.

[32] Chandrasekaran, K., A.S.R. Thaveedhu, P. Manoharan and V. Periyasamy (2023). Optimal estimation of parameters of the three-diode commercial solar photovoltaic model using an improved Berndt-Hall-Hall-Hausman method hybridized with an augmented mountain gazelle optimizer. *Environmental Science and Pollution Research*, Springer, 30(20), pp. 57683–57706. https://doi.org/10.1007/s11356-023-26447-x.

[33] Prakaash, A.S., K. Sivakumar, B. Surendiran, S. Jagatheswari and K. Kalaiarasi (2022). Design and development of modified ensemble learning with weighted RBM features for enhanced multi-disease prediction model. *New Generation Computing*, Springer, 40(4), pp. 1241–1279. https://doi.org/10.1007/s00354-022-00190-2.

[34] Yang, Y. and M.S. Rosenbaum (2002). The artificial neural network as a tool for assessing geotechnical properties. *Geotechnical & Geological Engineering*, Springer, 20, pp. 149–168. https://doi.org/10.1023/A:1015066903985.

[35] 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.

[36] Wu, X., V. Kumar, J. Ross Quinlan, J. Ghosh, Q. Yang, H. Motoda, G.J. McLachlan, A. Ng, B. Liu and P.S. Yu (2008). Top 10 algorithms in data mining. *Knowledge and Information Systems*, Springer, 14, pp. 1–37. https://doi.org/10.1007/s10115-007-0114-2.

[37] Akbulut, Y., A. Sengur, Y. Guo and F. Smarandache (2017). NS-k-NN: Neutrosophic set-based k-nearest neighbors' classifier. *Symmetry*, MDPI, 9(9), p. 179. https://doi.org/10.3390/sym9090179.

[38] Qian, Y., W. Zhou, J. Yan, W. Li and L. Han (2014). Comparing machine learning classifiers for object-based land cover classification using very high-resolution imagery. *Remote Sensing*, MDPI, 7(1), pp. 153–168. https://doi.org/10.3390/rs70100153.

[39] Abdollahzadeh, B., F.S. Gharehchopogh, N. Khodadadi and S. Mirjalili (2022). Mountain gazelle optimizer: a new nature-inspired metaheuristic algorithm for global optimization problems. *Advances in Engineering Software*, Elsevier, 174, p. 103282. https://doi.org/10.1016/j.advengsoft.2022.103282.

[40] Al-Betar, M.A., Z.A.A. Alyasseri, M.A. Awadallah and I. Abu Doush (2021). Coronavirus herd immunity optimizer (CHIO). *Neural Computing and Applications*, Springer, 33, p. 5011–5042. https://doi.org/10.1007/s00521-020-05296-6.

[41] Zarei, C. and L. Rahimi (2021). Prediction of undrained shear strength of crushed tire mixture with fine-grained soil by using machine learning approaches. Research Square. https://doi.org/10.21203/rs.3.rs-820067/v1.

[42] Zhang, W., C. Wu, H. Zhong, Y. Li and L. Wang (2021). Prediction of undrained shear strength using extreme gradient boosting and random forest based on Bayesian optimization. *Geoscience Frontiers*, Elsevier, 12(1), pp. 469–477. https://doi.org/10.1016/j.gsf.2020.03.007.

[43] Pham, B.T., C. Qi, L.S. Ho, T. Nguyen-Thoi, N. Al-Ansari, M.D. Nguyen, H.D. Nguyen, H.-B. Ly, H. Van Le and I. Prakash (2020). A novel hybrid soft computing model using random forest and particle swarm optimization for estimation of undrained shear strength of soil. *Sustainability*,

MDPI, 12(6), p. 2218. https://doi.org/10.3390/su12062218.

[44[Elsawy, M.B.D., M.F. Alsharekh and M. Shaban (2022). Modeling Undrained Shear Strength of Sensitive Alluvial Soft Clay Using Machine Learning Approach. *Applied Sciences*, MDPI, 12(19), p. 10177. https://doi.org/10.3390/app121910177.