Improved prediction of clay soil expansion using machine learning algorithms and meta-heuristic dichotomous ensemble classifiers

E.U. Eyo, S. J. Abbey, T. T. Lawrence, and F. K. Tetteh,

DOI 10.1016/j.gsf.2021.101296
ISSN 1674-9871

Publisher: Elsevier

Geoscience Frontiers 13 (2022) 101296

HOSTED BY

Contents lists available at ScienceDirect

Geoscience Frontiers

journal homepage: www.elsevier.com/locate/gsf

Research Paper

Improved prediction of clay soil expansion using machine learning algorithms and meta-heuristic dichotomous ensemble classifiers

E.U. Eyo [a,*], S.J. Abbey [a], T.T. Lawrence [b], F.K. Tetteh [c]

[a] Department of Geography and Environmental Management, University of the West of England, Bristol, United Kingdom
[b] Research Centre for Fluid and Complex Systems, Coventry University, Coventry, United Kingdom
[c] Department of Civil Engineering, University of Birmingham, Birmingham, United Kingdom

ARTICLE INFO

Article history:
Received 24 May 2021
Revised 19 August 2021
Accepted 10 September 2021
Available online 13 September 2021
Handling Editor: M. Santosh

Keywords:
Artificial neural networks
Machine learning
Clays
Algorithm
Soil swelling
Soil plasticity

ABSTRACT

Soil swelling-related disaster is considered as one of the most devastating geo-hazards in modern history. Hence, proper determination of a soil's ability to expand is very vital for achieving a secure and safe ground for infrastructures. Accordingly, this study has provided a novel and intelligent approach that enables an improved estimation of swelling by using kernelised machines (Bayesian linear regression (BLR) & bayes point machine (BPM) support vector machine (SVM) and deep-support vector machine (D-SVM)); (multiple linear regressor (REG), logistic regressor (LR) and artificial neural network (ANN)), tree-based algorithms such as decision forest (RDF) & boosted trees (BDT). Also, and for the first time, meta-heuristic classifiers incorporating the techniques of voting (VE) and stacking (SE) were utilised. Different independent scenarios of explanatory features' combination that influence soil behaviour in swelling were investigated. Preliminary results indicated BLR as possessing the highest amount of deviation from the predictor variable (the actual swell-strain). REG and BLR performed slightly better than ANN while the meta-heuristic learners (VE and SE) produced the best overall performance (greatest R^2 value of 0.94 and RMSE of 0.06% exhibited by VE). CEC, plasticity index and moisture content were the features considered to have the highest level of importance. Kernelized binary classifiers (SVM, D-SVM and BPM) gave better accuracy (average accuracy and recall rate of 0.93 and 0.60) compared to ANN, LR and RDF. Sensitivity-driven diagnostic test indicated that the meta-heuristic models' best performance occurred when ML training was conducted using *k*-fold validation technique. Finally, it is recommended that the concepts developed herein be deployed during the preliminary phases of a geotechnical or geological site characterisation by using the best performing meta-heuristic models via their background coding resource.

1. Introduction

Practicing ground engineering experts and stakeholders in geology and related disciplines have been seriously occupied with the challenges posed by swelling soils on foundations of structures and similar land development ventures. One of the reasons for this continuing interest is because in the past 50 decades, the reported cost of damage to infrastructure due to undesirable expansion has amounted to several billions of dollars (Charlie et al., 1985; Du et al., 1999; Puppala and Cerato, 2009; Zumrawi, 2015). Moreover, studies have revealed that soil swelling-related disaster is one of the most devastating geo-hazards at least in modern history (Jones and Jefferson, 2015). Therefore, the proper identification and characterisation of an expansive soil's ability to swell is very vital for achieving a secure and safe construction ground for civil structures. Accordingly, most published researche and field studies have in the recent past, examined the nature and characteristics of potential swelling or expansive soils at micro and micro levels (Nelson et al., 2015).

Several methods of soil's volume change estimation and prediction in the recent past some of which have either referenced or incorporated the techniques established over the past 50 or more decades, have been reported in literature (Erguler and Ulusay, 2003; Yilmaz, 2006; Erzin and Erol, 2007; Vanapalli and Lu, 2012; Elbadry, 2016). Both indirect and/or direct empirical correlations that rely on soil index properties such as Atterberg limits, moisture content and unit weight have been used by most of these studies. Some of the studies have attempted evaluating the behaviour of soils during the course of swelling by considering

* Corresponding author.
E-mail address: eyo.eyo@uwe.ac.uk (E.U. Eyo).

https://doi.org/10.1016/j.gsf.2021.101296
1674-9871/

microstructure, texture, chemical, and the effects of the external environment on the soil (Likos and Wayllace, 2010; Chittoori and Puppala, 2011; Rani, 2013; Puppala et al., 2014; Eyo et al., 2019). Meanwhile other investigations have implemented techniques that are regarded as an extension of the theories and concepts from related engineering fields to estimate the swelling of soils (Buzzi, 2010; Buzzi et al., 2011; Adem and Vanapalli, 2014; Berrah et al., 2018). In general, some of the limiting factors of the foregoing methods are the costly instrumentation and cumbersome experimental setup and procedures involved in the determination of soil expansion.

An artificial intelligence using the machine learning (ML) paradigm can serve as a means of tackling some of the problems of soil behaviour not least the setbacks already mentioned above relating to swelling soils under inundation (Kayadelen et al., 2009; Das et al., 2010; Ikizler et al., 2010; Yilmaz and Kaynar, 2011; Bekhor and Livneh, 2014; Toksoz and Yilmaz, 2019; Ermias and Vishal, 2020; Alizamir et al., 2021; Eyo and Abbey, 2021; Zhang et al., 2021a,b). The techniques relying on the principles of data-driven decision making, which is the bedrock of ML, to examine and predict the performance of swelling clays are few in literature. Moreover, the application of ML in this regard by most authors have stopped short of validating the used algorithms through a robust sensitivity analysis to ensure that the learning that was done involved sufficient discrimination of the types of clays used as defined according to the Casagrande-Unified Soil Classification System (USCS) for soil plasticity.

Consequently, this research aims to utilise both ML algorithms and binary classifiers namely, stand-alone models such as kernelized machines, linear & logistic regressors and neural networks; tree-ensembles such as random and boosted decision forests and meta-heuristic ensembles that incorporate the voting and stacking methods, to study and predict the swelling of soils. Since most studies have reported ML performances using traditional statistical metrics such as standard mean error scores and coefficient of determination, this research has been set up to extend the boundaries of swell estimation by including sensitivity-driven diagnostic tests to promote the confirmation and validation of the used statistical measures. Finally, it is important to emphasise that for the first time, dichotomous classifiers are herein being applied across different multiple cross-validation techniques, to analyse various soils of broad ranging plasticity properties.

Finally, an implementation of the intelligent prediction concepts in this study can be performed at the preliminary phases of civil construction and related land developments most especially those involving geotechnical or geological site characterisation.

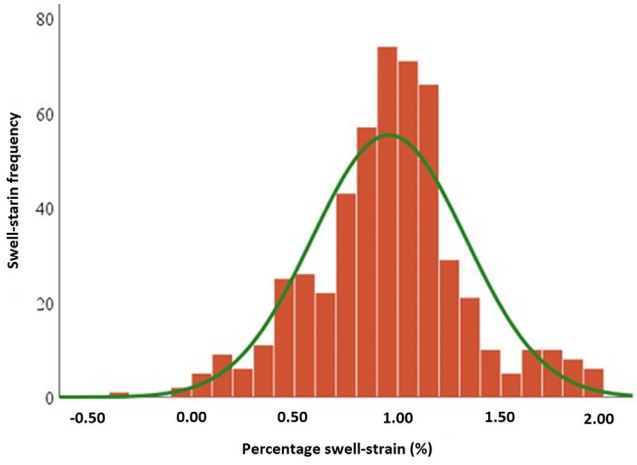

Fig. 1. Normal distribution for swell strain dataset.

2. Methodology

2.1. Database construction, integration & pre-processing

Very high-quality database of soils of low-to-high plasticity subjected to expansion under water were diligently and carefully mined from previous research (Ashayeri and Yasrebi, 2009; Çimen et al., 2012; Erzin and Gunes, 2013; Yilmaz and Kaynar, 2011; Zumrawi, 2012) These studies utilised extremely sound standard techniques and followed recommended procedures during testing to achieve the goal of vertical one-dimensional movement of the soil masses under inundation. A total number of 517 data records of soil swelling that includes the properties that can potentially have some form of influence on such behaviour were adjudged to be comprehensively ample for this research. Soil properties such as, moisture content, void ratio, unit weight, liquid limit, plasticity index, clay content, maximum dry unit weight, coarse content, cation exchange capacity and activity all obtained from the database are used as explanatory variables or features in this research. It should be borne in mind that some differences exist in the procedures followed by the authors in the testing and measurement of one-dimensional swelling of the soils hence as it is expected, noisy observations are inevitable. In order to deal with this issue, data of the dependent variable (which in this case is the soil swelling) were subjected to a rigorous transformation to obtain an approximate normal distribution (Fig. 1) with some of the vital descriptive statistical metrics given in Table 1. From Table 1 it is observed that measures of the skewness and kurtosis which are very low (0.45 and 0.709, respectively), do lend further credence to the reliability of the swell-strain dataset that will be subsequently used in the training, testing and validation of the ML algorithms adopted in this study. Important statistical metrics were also derived from the raw dataset of independent features and presented in Table 2. The values and ranges of liquid limit (minimum value of 4.145% and maximum value of 112.093%) and plasticity indices of the clay soils (minimum value of 24.504% and maximum value of 213.000%) both suggest a very wide coverage of the soils' plasticity properties. The frequency distribution of the independent soil features shown in Fig. 2 indicates a uniform distribution of the consistency limits properties of the soils compared to the other variables. Soil texture and cation exchange capacity (CEC) seem to possess equal distribution. The frequency distribution of soil Activity shows a very good representation of the soils ranging from 'Normal' soils (activity between 0.75 and 1.25) to 'very active' soils (activity greater than 1.25). Thus, further validating the broadness of soil plasticity used in this study.

Table 1
Descriptive statistics of swell strain dataset.

Statistical parameters											
Missing	Mean	Median	Mode	Std. dev.	Kurtosis	Skewness	Range	Min.	Max.	Sum	Count
0	0.957	0.970	0.780	0.373	0.709	0.045	2.270	0.100	1.970	495.010	517

Table 2
Relevant statistics of independent features/variables dataset.

Statistical parameter	Moisture content	Void ratio	LL (%)	PI (%)	Unit weight kN/m^3	Fine content (%)	Coarse content (%)	Max. dry unit wt. kN/m^3	CEC Meq./100 g	Activity
Mean	0.204	0.912	56.996	24.504	15.533	49.635	46.922	4.665	48.851	0.898
Std. dev.	0.072	0.313	27.436	23.950	2.296	14.832	16.089	5.939	24.048	0.329
Range	0.321	1.351	107.948	205.900	10.357	66.500	66.500	17.569	89.402	1.353
Min.	0.079	0.430	4.145	7.100	11.500	13.000	20.500	1.431	5.693	0.484
Max.	0.400	1.781	112.093	213.000	21.857	79.500	87.000	19.000	95.096	1.838

To enable an analysis, assessment, and subsequent comparison of the degree of expansion of the swelling soils obtained from different sources, dataset integration was needful. Hence, the data with records of very similar features and attributes were appended and combined. An aggregation of the dataset in this manner resulted in three different scenarios of dataset integration an evaluation of which is discussed in later sections of this study.

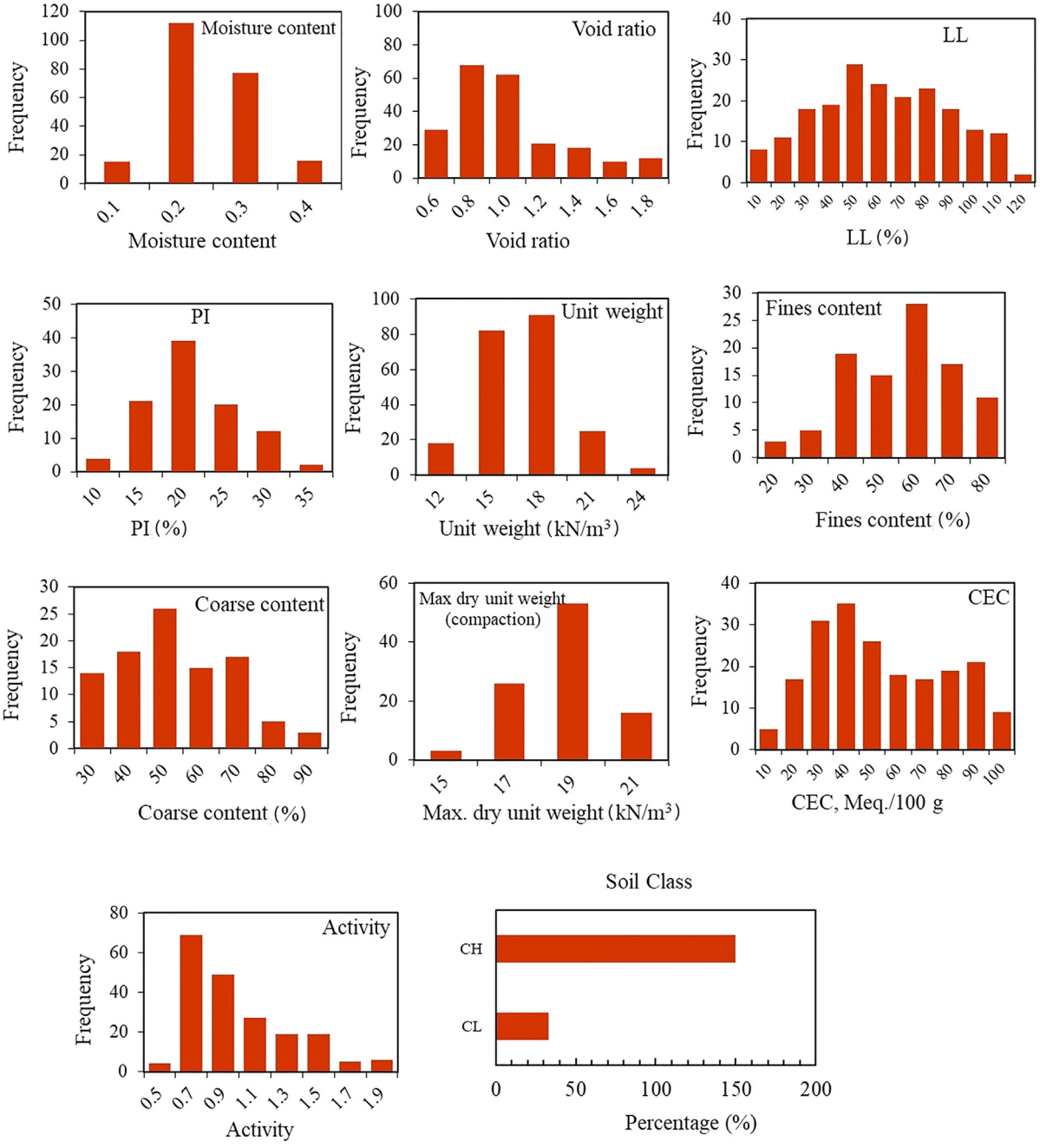

Fig. 2. Frequency distribution of soil features.

2.2. Dataset wrangling and optimisation

It is very pertinent to state that the data records of soil swelling used in this research for ML prediction did not contain missing values. Nevertheless, it was needful to perform some further forms of processing and feature engineering to ensure that as much as possible, unwanted redundancy was reduced thus, presenting a clean dataset for training, testing, and validation.

2.2.1. Normalisation of data

Without distorting the existing differences in the values of both the explanatory and predictor variables, these were subjected to a common scale provided by the z-score normalisation technique that ensures the avoidance of outliers. Z-score transformation can be mathematically expressed as Eq. (1):

$$K_n = \frac{k_0 - m_n}{\sigma_n} \tag{1}$$

where K_n and k_0 are the normalized and influence factors of soil swelling; m_n and σ_n denote the respective values of mean and corresponding deviation from the mean of the distribution.

2.2.2. Cross-validation

Cross-validation (CV) techniques were applied to minimise the possibility of coincidental features having more importance most especially during dichotomous classification (Joshi, 2020). Hence, to optimise model hyperparameters, train-validation split (TVS), *k*-fold cross-validation (*k*FCV) and the Monte Carlo cross-validation (MCCV) were used.

ML dataset training using the TVS involved random splitting of the dataset each of which was used for either training or testing of the ML models. 75% of the entire data records were utilised to rigorously train the proposed models while the remainder (i.e., 25% of randomly selected dataset) were employed to test and evaluate ML models' performance. It is suggested that for the testing of models, about 10%–30% of the parent data should be used (Han et al., 2020; Joshi, 2020).

The *k*-fold cross-validation (*k*FCV) method involves an initial dataset (N) division into equal *k*- subsets where one of the subsets is used for validating and the remainder utilised to train the ML prediction. This is an iterative process that involves *k*-cycles with either of the *k*-subsets excluded from each cycle. This study adopts 10-fold CV subsets in the ML prediction.

The Monte Carlo cross-validation (MCCV) technique tends to combine both the TVS and *k*-FCV in dataset training and validation. Just like *k*-FCV, the dataset is first broken up into two sets but without *replacement* of either of them. The non-replaced dataset is then used in the training while validation is carried out on the remainder. Given the existence of unique training dataset, MCCV may not go through similar iterative cycles like the *k*-FCV technique. In this research, a combination of 25% of the dataset were used in testing coupled with 10-fold cross validations.

3. Machine learning algorithms and binary classifiers

3.1. Multiple linear regression (REG)

Multilinear regression is applied in most technical analyses to provide estimates of unknown variables, coefficients, or parameters by indicating how a change in one or more of some given independent variables can affect the predictor variable. The general form of a simple REG is expressed thus as Eq. (2):

$$Y_n = \mu + \sum_{n=1}^{m} \alpha_n \cdot x_n \tag{2}$$

where the dependent variable is assumed to be influenced by independent variables – X_1, X_2, X_3,. . ., X_m plus an error term which does account for various other factors.

3.2. Logistic regression (LR)

Logistic regression is a non-linear model which is usually preferred for class prediction over simple regression. For instance, if it is desired that a binary classification problem is to be carried out then, the predictor variable *Y* will take up the value of 0 or 1 and the general relationship of this variable with the independent variables will become Eq. (3):

$$y(x) = \frac{exp^{x'\beta}}{1 - exp^{x'\beta}} = \frac{1}{exp^{x'\beta}} \tag{3}$$

where $y(x) = P(Y|X = x)$ is a probabilistic outcome; β = regression parameters' vector. Hence, in contrast to a simple multiple regression, LR determines the probability that the dependent variable would belong to a certain category or class. Nevertheless, the logistic transformation of the $y(x)$ is carried out to ultimately determine the linear form of the model given as Eq. (4):

$$log\left[\frac{y(x)}{1 - y(x)}\right] = x'\beta \tag{4}$$

3.3. Artificial neural networks (ANN)

The human nervous system which is simply a network of interconnected neurons between synapses (Fig. 3a), does provide the inspiration and architecture for a family of data-processing models regarded as artificial neural networks (ANN). Hence, ANN is basically a network of input – processes (decisions) – output system. Just like the simple neuron, ANN can be set up to process data while being supported by an enhanced filtering function in order to ensure that the inputs at certain nodes do not affect the entire network as much. In technical parlance, the input – processes (decisions) – output system of an ANN is represented as a network of input, hidden, and an output layers as shown in (Fig. 3b).

This interconnectedness of an ANN can be mathematically given as Eq. (5):

$$\alpha_n = \sigma\left(\sum_j \omega_{ij} \cdot y_j\right), \sigma(x) = \frac{1}{1 - e^{-x}} \tag{5}$$

where α_n = activities of the network; w_{ij} = neurons' weight; $x = n^{th}$ neuron activation; y_n = output signal; $\sigma(x)$ = activation function that allows input transformation into the output by inputs' (processing neuron) multiplication by the corresponding weights.

3.4. Kernel machines

Kernelized machines are commonly referred to as systems of transforms generally used to map out or provide predictors from one variable space to a greater or higher-dimensional feature space. Predictions carried out in this manner are more complex compared to a simple polynomial approach given previously. The transformations of kernels are typically represented mathematically as Eq. (6):

$$k(x, y) = [(\phi(x), \phi(y)] \tag{6}$$

where k = kernel function, x and y = N-dimensional input vectors, N = no. of predictor variables, φ = mapping from some m-dimensions to an m-dimensional space.

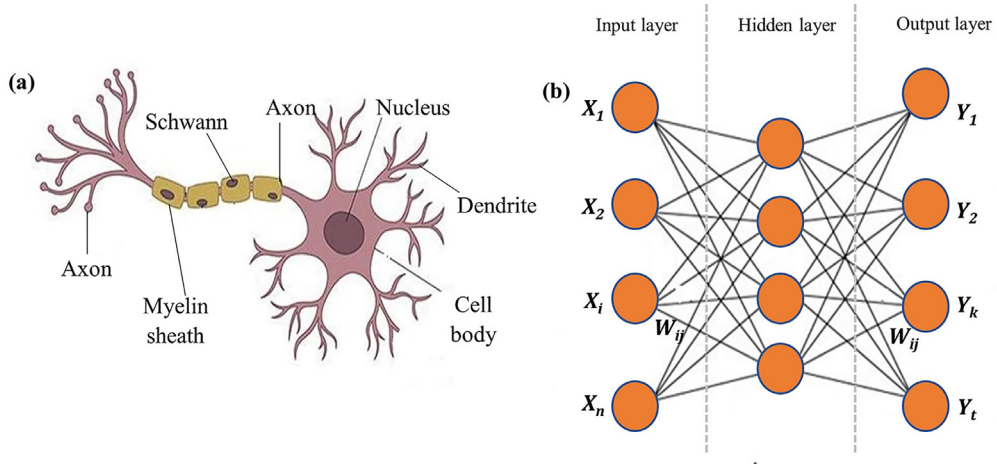

Fig. 3. Neural network structure (a) Neurone (b) Artificial neural network (ANN).

3.4.1. Bayesian approach

Bayesian regressor could be regarded as a typical kernel machine and a very special case of REG. Thus, when adopting this kernelized approach in regression problems, the analysis is mostly performed by inferring the probability of an event occurring based on a *prior* knowledge of some conditions bearing a relationship to such events which in turn is a description of what is commonly referred to as the 'Bayes' theorem.

Bayesian model could be expressed simply as Eq. (7):

$$y_i = \alpha + \beta x_i + \epsilon_i \tag{7}$$

where i = 1, 2, 3,. . .,n.

The above assumes that the error terms, ϵ_i, are independent while also being normally distributed identically as random variables with zero mean and a constant variance, σ^2. The goal will then be to update the distributions of unknown parameters α, β and σ, given a set of data x_1, y_1, . . ., x_n, y_n.

Now, given the observed dataset x_1 and parameters α, β and σ^2 a random variable of each response say Y_i will be normally distributed:

$$Y_i|x_i, \alpha, \beta, \sigma^2 \ Normal(\alpha + \beta x_i, \sigma^2),$$

This therefore follows that the likelihood of each Y_i given the unknown parameters is expressed as Eq. (8):

$$p(y_i|x_i, \alpha, \beta, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e\left\{\frac{[y_i - (\alpha + \beta x_i)]^2}{2\sigma^2}\right\} \tag{8}$$

Going forward, we then consider a reference prior (non-informative prior) which in turn will form the basis for posterior distributions of the previously stated unknown parameters.

The present research shall utilise the Bayes model to formulate and provide predictions on both regression and classification type problems. Hence, in order to properly differentiate the ML problems, the model for regression is given as BLR (Bayesian Linear Regressor) while BPM (Bayes Point machine) is used for binary classification.

3.4.2. Support vector machines (SVM)

Support vector machines are supervised kernelised models with the capacity to map input vectors unto higher-dimensional spaces much like the bayes machines. For a binary classification problem, SVM would utilise a linear hyperplane by assuming that the observed dataset belongs to either of two classes. Given a set of N dataset samples, the training vectors, D can be expressed as Eq. (9):

$$D = [(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots, (x_n, y_n)], x_i \in \Re^n, y_i \in \Re \tag{9}$$

When applying a penalty with the kernelized trick, SVM can be used as non-linear classifiers. In this case a function in the trick can be represented as Eq. (10):

$$f(x) = w^T \vartheta(x) + b \tag{10}$$

where w^T = transposed form of the vector of the output layer; b = a bias; x = input variable matrix of $N \times n$ (N = observed dataset points and n = No. of input variables); $\varphi(x)$ = kernel function. The solution for w and b can be formulated further through a minimisation problem (Vapnik et al., 1997).

3.4.3. Deep-support vector machines (D-SVM)

A deep-support vector machine hybridises both SVM and several deep learning systems to improve its predictive capacity. Hence, the architecture and characteristics of a D-SVM may tend to resemble those of a neural or belief network with multiple hidden layers (Abdullah et al., 2009). In order to formulate a D-SVM, the basic or standard SVM needs to be first trained. This is then followed by inputting the activation kernels of the SVM which in turn would serve as inputs for the subsequent layers. Each hidden layer's net input can then be represented as following Eqs. (11)–(13):

$$net_{m1} = k_{1_1}(x) \cdot X_1 + k_{1_2}(x) \cdot X_2 + \cdots + k_{1_n}(x) \cdot X_n + h_1 \tag{11}$$

$$net_{m2} = k_{2_1}(x) \cdot X_1 + k_{2_2}(x) \cdot X_2 + \cdots + k_{2_n}(x) \cdot X_n + h_1 \tag{12}$$

$$net_{mn} = k_{n_1}(x) \cdot X_1 + k_{n_2}(x) \cdot X_2 + \cdots + k_{n_n}(x) \cdot X_n + b_1 \tag{13}$$

where X_1, X_2, . . ., X_n are the input layer data points.

3.5. Tree-ensembles

An ensemble of trees is a machine learning paradigm that enables decisions to be made and final predictions provided based on a set of rules gathered from continuous pattern detection in each dataset. Fig. 4 depicts a basic decision tree system whereby an initially obtained result undergoes some forms of averaging before the final predicted output (DeRousseau et al., 2018). Depending on the domain of application or purpose of prediction, the structure of tree-based algorithms is built and utilised differently. Hence, in order to achieve the goal of this study, both random decision forest (RDF) and boosted decision trees (BDT) are adopted.

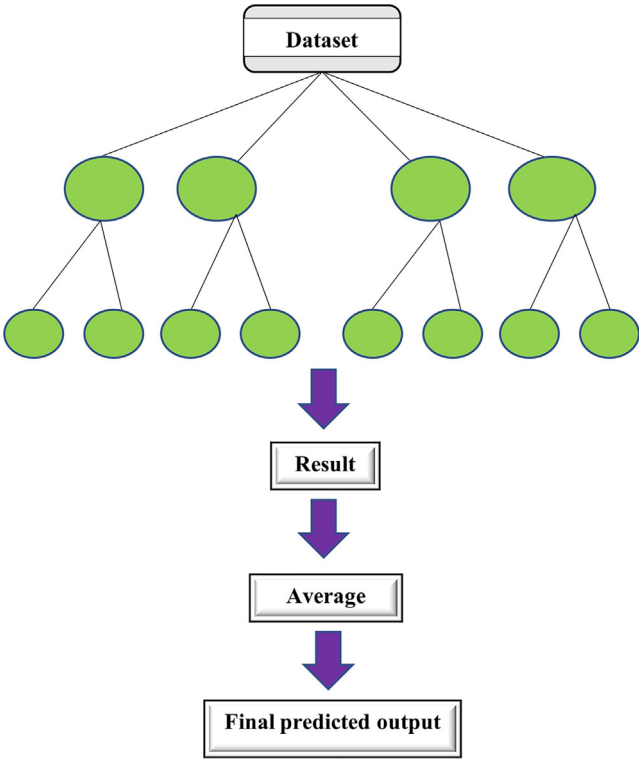

Fig. 4. Structure of a single regression (decision) tree.

3.5.1. Random decision forest (RDF)

RDF are often formulated to mitigate some of the instabilities frequently observed when using just a single or relatively less-deep regression trees for ML prediction. Hence, in order to improve decision-making, a method called bootstrapping or bagging is used to generate predictions from almost similar subsets of data derived from the parent source. The bagging technique results from an aggregation or assemblage of multiple tree models in the training (Kang et al., 2021). If not approached with extreme caution, RDF may sometimes become prone to overfitting given its minute biases and enormous variances.

3.5.2. Boosted decision trees (BDT)

Much like RDF, the idea of boosting used in BDT is to enhance the functionality of a hitherto less deep or weaker trees in an iterative manner. The structure of BDT is considered hierarchical and each participatory tree layer is generated in a recursive manner unlike the RDF where the trees are made to be of similar or equal importance (DeRousseau et al., 2019). Hence, it is expected that BDT would function efficiently especially when applied to nonlinear problems. The interpretability capacity of a BDT can sometimes be very low which may hinder an enormous gain in its learning rate.

3.6. Meta-ensembles

Meta-ensembles or the so-called model or models could be constructed by aggregating some of the best performing hyperparameters of the stand-alone and tree ensemble algorithms to improve the accuracy and outcome of ML predictions. Through a system of averaging and/or boosting, the combined models could be tiered, bagged, stacked or voting applied to produce predictions all depending on the purpose of the ML exercise. In this study, the voting and stacking techniques are adopted.

3.6.1. Voting (VE)

Voting is applied to basically compute and evaluate the average predictive outcome of aggregated models. For a dichotomous classification, the predictions offered by each of the categories or class labels are summed and the result that has a majority vote is ultimately considered as shown in Fig. 5a (Chou et al., 2016)

3.6.2. Stacking (SE)

Stacking is mostly an adjunct of the voting strategy where the models are made to learn and decide on when to rely on themselves to allow for a general multistage prediction. In this case, results obtained from previously integrated models, ($X_{m=1\sim j}$), will then become inputs, Y for subsequent ones, X_{pred}, as predictions are being made Fig. 5b (Wolpert, 1992; Chou et al., 2016).

3.7. ML model implementation

Model implementation and execution was conducted on a ML cloud computing platform that promotes Python programme developments and its associated libraries. Vital features of the variables and optimised parameter settings for the ML models utilised for the ML prediction are given in Tables 3 and 4 respectively.

3.8. Performance evaluation

Three major statistical indicators of ML performance namely coefficient of determination (R^2), Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) are utilised in the regression problem to assess the predictive capacity of the models. Detailed descriptions of these statistical metrics are presented in lierature (Chatterjee and Hadi, 2012; Galwey, 2014). However, for the classification problem, metrics such as accuracy, recall and F1 score are further described below as Eq. (14):

Accuracy assesses the correctness or precision of a ML model and can be mathematically expressed as Eq. (14):

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \tag{14}$$

where:

TN = True Negative and refers to the total collated number of instances from a negative class or category such that the true class label is equal to the predicted class.
FN = False Negative, refers to the total collated number of instances from a positive class category given that the model is seen as having misclassified these classes by incorrectly predicting them as a negative class.
TP = True Positive, refers to the total collated number of instances from a positive class prediction where the true class label is said to be equal to the predicted class
FP = False Positive, refers to the total collated number of instances from a negative class category given that the model is seen as having misclassified these classes by incorrectly predicting them as positive.

Recall – also called **sensitivity,** measures a model's capacity to ascertain the the proportion of vital data points. Recall is expressed mathematically as Eq. (15):

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{15}$$

F1 Score is normally used to enable a much balanced optimisation of predictive precision given by the rate of recall. Hence, F1 score may also be taken to represent the harmonic mean of precision which produces an optimised algorthm or model. The F1 score is given as Eq. (16):

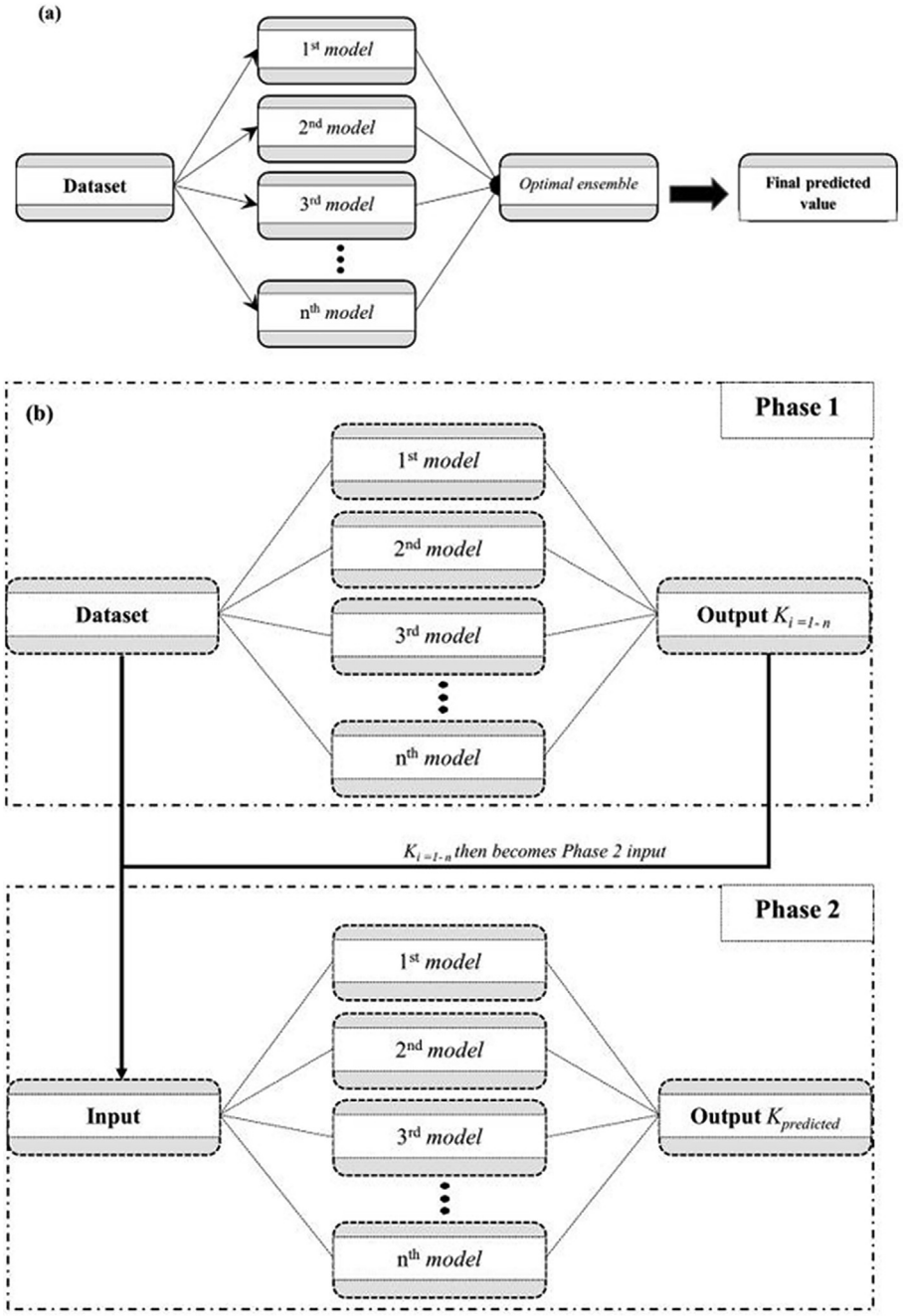

Fig. 5. Meta-heuristic ensembles (a) voting structure and (b) stacking structure.

Table 3
Data features and attributes or independent variables.

Feature	Abbreviated attrib.	Instances	Type of data
Moisture content	Mct.	221	Integer
Void ratio	Vr.	221	Integer
LL	Ll	198	Integer
PI	Pi	122	Integer
Unit weight	Uwt.	221	Integer
Fine content	Fct.	122	Integer
Coarse content	Cct.	122	Integer
Max. dry unit weight	Mdd.	122	Integer
CEC	Cec	198	Integer
Activity	Ac.	198	Integer
High plasticity clay	Ch	181	String
Low plasticity clay	Cl	40	String

$$\text{F1 score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{16}$$

where: Precision is defined as the TP divided by the sum of the positively predicted outcomes. In order words, precision expresses the model's unit proportion that are positive as being positive. Precision is expressed thus as Eq. (17):

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{17}$$

4. Results and discussion

Analyses of the performance of ML regression carried out using REG, ANN, BLR, BDT RDT and the meta-ensemble models (VE and

Table 4
Models' parameter settings.

Model	Parameter	Option/value
Linear regressor (REG)	Regularisation wt. (L_2)	0.001
	Solution method	Ordinary least squares (OLS)
Logistic regressor (LR)	Optimisation tolerance	1 × 10E-7
	Regularisation wt. (L_1)	1.00
	Regularisation wt. (L_2)	1.00
	Mode of training	Single parameter
Bayesian regressor (BLR)	Regularisation wt. (L_2)	1.00
Boosted decision tree (BDT)	No. of constructed trees	100.00
	Learner rate	0.20
	Tree-forming training instances	10.00
	Max. no. of leaves/tree	20.00
	Mode of training	Single parameter
Random decision forest (RDF)	Ma. tree depth	32.00
	No. of trees	8.00
	Method of resampling	Bagging
	Min. no. of samples/leaf node	1.00
	No. of randomised splits/node	128.00
	Mode of training	Single parameter
Artificial neural networks (ANN)	Normaliser	Min.-Max
	Learner rate	0.005
	No. of hidden nodes	100.00
	Learning wt. diameter (initial)	0.10
	No. of iterative learning	100.00
	Specification for hidden layer	The fully connected case
Bayes point machine (BPM)	Categorical feature values (unknown)	True
	Included biases	True
	No. of training iterations	30.00
Support vector machine (SVM)	Lambda	0.001
	No. of iterations	1.00
	Categorical feature values (unknown)	True
	Normalised features	True
	Mode of training	Single parameter
Deep support vector machine (D-SVM)	Normaliser	Min.-Max.
	No. of iterations	15,000.00
	Lambda W	0.10
	Lambda theta prime	0.01
	Lambda theta	0.01
	Sigmoid sharpness	1.00
	Tree depth	3.00
	Mode of training	Single parameter

SE) are provided first. Following this, an assessment and evaluation including sensitivity analyses shall be considered on binary classification of the clay soil swelling by using the dichotomous elements of ML models namely ANN, LR, RDF, SVM, D-SVM, BPM, and the meta-ensembles (VE and SE).

4.1. ML regression

Evaluation of test set indices

Fig. 6 indicates the standard deviations of ML predictions from the mean of the dependent variable (one-dimensional swelling) of the soils. A comparison of the degree of variations between three scenarios each of which represents a combination of ML explanatory features based on the nature of data obtained from different sources are given in Fig. 6. Scenario 1 (or S1) incorporates moisture content, void ratio and unit weight as independent variables in the prediction, scenario 2 (or S2) includes clay content, plasticity index, coarse content, and maximum dry unit weight as independent variables meanwhile scenario 3 (or S3) combines activity, liquid limit, and cation exchange capacity (CEC) as independent features.

Generally, it is observed from Fig. 6 that higher degrees of deviations from the mean of soil expansion are demonstrated by all ML regression models when S1 is considered. The highest variation given by a peak of about 0.90 is exhibited by BLR which is then closely followed by ANN with a standard deviation of approximately 0.75. A closer examination of Fig. 6 indicates that the tree-based models all seem to have the lowest values of standard deviations across all three scenarios. Nevertheless, as it is also observed, variations under S3 are the lowest in general for all the ML algorithms. Further investigations and assessments of the ML models' performance provided by statistical metrics are presented in the sections following.

4.1.1. Performance forecast of machine learning regression models

Indicators of the actual performance regression models utilised for ML predictions of soil expansion are presented in Table. 5. Attention and discussions are provided herein for the RMSE and R^2 metrics (highlighted in bold-face fonts). Lower values of RMSE and a corresponding higher R^2 scores would indicate better performance of a ML algorithm. Across the three scenarios of independent variable combinations, it is observed that the R^2 does range between 0.33 and 0.94 while values for RMSE are between 0.05% and 0.78%. For the stand-alone ML models, BLR and REG both appear to produce more accurate predictions compared to ANN across all the independent variables' combination. Meanwhile, relatively greater performance accuracy is exhibited by the tree-based ML ensembles when compared to the stand-alone algorithms. The lower accuracy of prediction demonstrated by the stand-alone models (BLR, REG and ANN) is traceable to the inadequacy of dataset fitting to these models' underlying assumptions unlike the tree-ensembles. This behaviour may stem from both

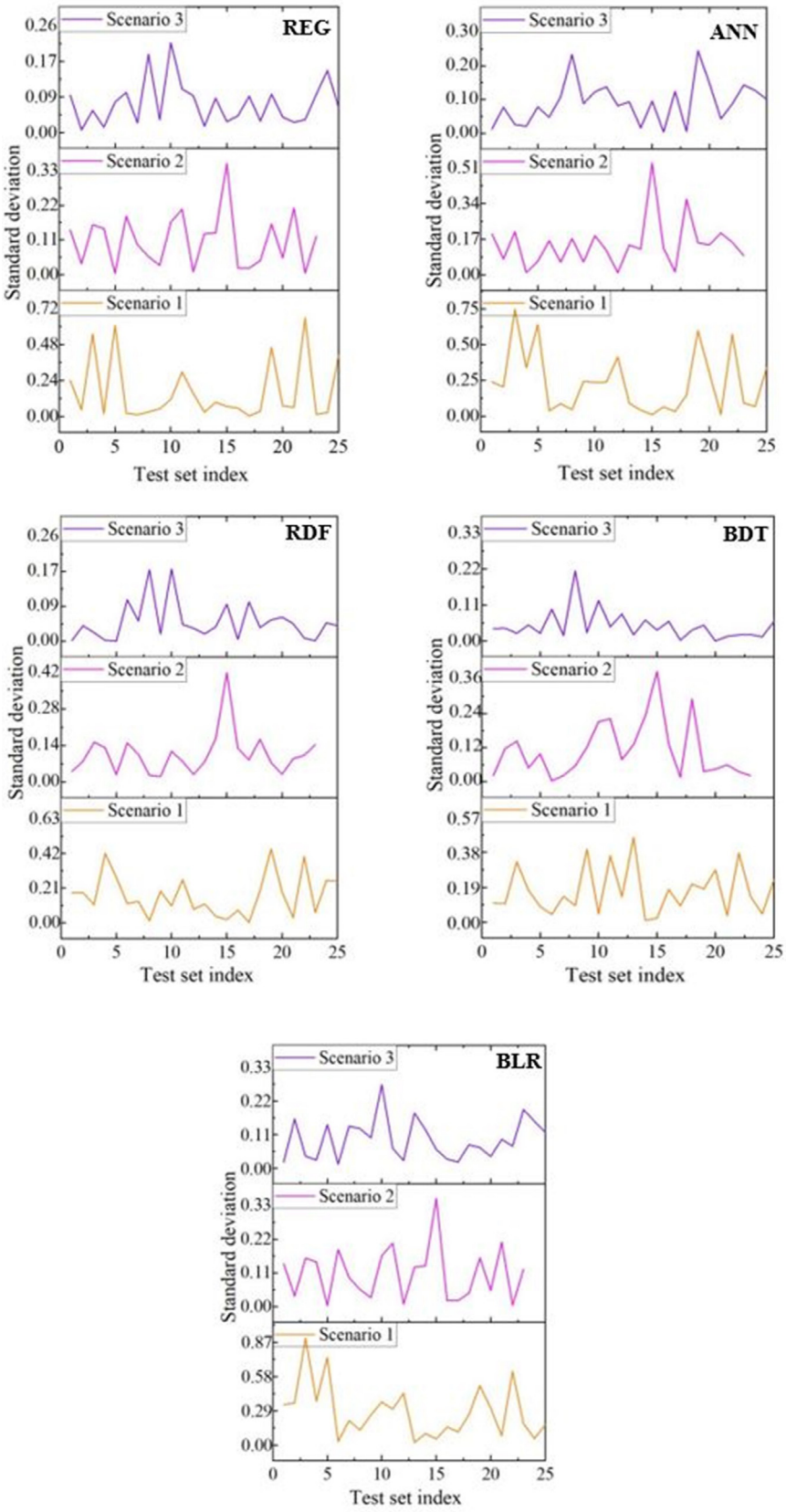

Fig. 6. Standard deviation of predictor variable – REG; ANN; BLR; BDT; RDF.

the nonlinearity as well as the non-normality of the dataset residuals hence, indicating that these models especially, BLR and REG are not able to approximate some of the unobserved phenomena arising from independent variable combinations. It may be surprising to note that the worst prediction of swell expansion within the context of the independent features used is demonstrated by ANN

Table 5
ML models' swell prediction performance metrics.

Model	Feature combinations	R^2	RMSE	MAE
			(%)	(%)
REG	S (1)	0.364	0.762	0.532
	S (2)	0.565	0.627	0.523
	S (3)	**0.919**	**0.070**	0.057
ANN	S (1)	0.333	0.781	0.550
	S (2)	0.395	0.739	0.608
	S (3)	**0.912**	**0.073**	0.057
BLR	S (1)	0.366	0.761	0.533
	S (2)	0.566	0.627	0.521
	S (3)	**0.919**	**0.070**	0.056
BDT	S (1)	0.680	0.540	0.425
	S (2)	0.582	0.615	0.520
	S (3)	**0.923**	**0.068**	0.052
RDF	S (1)	0.578	0.621	0.489
	S (2)	0.655	0.559	0.457
	S (3)	**0.930**	**0.065**	0.050
SE	S (1)	**0.719**	**0.231**	0.186
	S (2)	0.443	0.139	0.115
	S (3)	**0.923**	0.068	0.053
VE	S (1)	**0.738**	**0.223**	0.178
	S (2)	0.637	0.112	0.092
	S (3)	**0.940**	**0.061**	0.049

(average R^2 of 0.55 and average RMSE of 0.53) even though this model has been recognised in some past studies as being capable of providing reasonable degrees of accuracy in the prediction of swelling albeit, within a different set of circumstances (Bekhor and Livneh, 2014; Ermias and Vishal, 2020). However, it should be borne in mind that irrespective of the initial parameter settings adopted in this study, ANN does have an inherent shortcomings given that the process of training followed by this algorithm does rely on backpropagation during its intuitive searching and optimisation in order to continuously upgrade its weights and biases over certain error spaces that includes or that tend to converge to a local minima rather than a more globalised one (Ben Chaabene et al., 2020). Further modification of the initial parameters and specification of the ANN in this study did not result in any improvement in the accuracy of its prediction. Hence, one suggestion that could be advanced for better performance would be to use a deeper reinforcement learning that incorporates multiple hidden-layer and feed-forward mechanism such as that provided by most extreme ML algorithms. However, this technique should be pursued with extra caution because excessive introduction of hidden layers or black-boxes may result in overfitting and over-estimation of the prediction (Behnood and Golafshani, 2018).

The relatively higher accuracy of the tree-ensembles is attributable mostly to their unique architecture. As could be observed in Table 5, RDF and BDT seem to have an average R^2 of 0.72 and RMSE of 0.41. The strategy of bagging and boosting have certainly aided the higher performance of these models compared to the above-mentioned stand-alone models. Unlike a stand-alone model such as the ANN, a sequence of training and testing are required to be conducted on the dataset to logically split or partition them so that previous sets of data would serve as inputs to successive partitioning. This technique should be carefully carried out to avoid the inconsistent learning of some features. On the other hand, overfitting can be minimised, and generalisation errors mitigated if deep trees are left unpruned or without smoothening (Han et al., 2020).

Table 5 also indicates that the performance of the tree-based ensembles is lower when compared to the meta-heuristic ensemble models (VE and SE). Another major setback in using tree-ensembles for ML prediction is the seeming "greedy" contrivance inherent in each step whereby an aggregation of the best performing entity is preferred and selected without taking into consideration another successful forward-looking entity which may provide even better prediction. This process is closely associated with another shortcoming which involves information loss during the training and testing of the algorithms caused by the sustained discretisation of features of the dataset in the splitting or partitioning (Dreiseitl and Ohno-Machado, 2002).

A meta-heuristic aggregation of some of the above-mentioned models' hyperparameters to enable much improved predictions compared to the tree-based ensembles are done by using the technique of voting and stacking. Even though the RMSE scores provided by using the meta-heuristic ensembles or model of models (VE and SE) seem to fluctuate between about 0.2 and 0.05%, these values are still very desirable because they are approximately 2–10 folds better than the stand-alone and tree-ensembles. The VE model has the highest accuracy (average R^2 of 0.77 and average RMSE of 0.13%). The much-enhanced predictive performance demonstrated by using the meta-ensembles is attributed to their capacity to combine and meta-heuristically utilise the individual strengths of stand-alone and tree-based models' hyperparameters.

With regards the effect of independent features in the swell prediction, it is observed in Table 5 that except for the meta-ensembles, ML prediction carried out under S3 are the most accurate followed by those conducted under S2. For BDT, the ML prediction carried out under S1is also more accurate than S2. However, since the meta-ensembles are the highest performing models, they do in turn provide the basis for comparison of the scenarios of independent variable combinations. For this reason, the prediction offered under S1 are better than those of S2. This could be because of the lesser number of variables involved in the ML regression when considering S1 compared to S2 even though lower deviations from the mean of the predictor variable observed earlier was exhibited by the later. However, compared to S1, the independent variables (LL, CEC) and activity of S3, when taken together, may bear more direct influence on soil expansion. Further discussions on the importance of the independent features are given in a later section of this research.

4.1.2. Verification of the quality of ML regression models

By using the lag plots of residual data points, an evaluation and verification of the ML regression model's capacity to forecast the performance of the swelling clay soils considering the different scenarios are carried out. Lag plots of residuals tend to allow a critical assessment of some of the statistical presuppositions that undergirds the independent variables including the normality of the frequency distribution of these features in a data-driven decision or prediction (Galwey, 2014). Residual plots of the predictor variable against its actual or observed values are presented in Fig. 7. One of the requirements for the validity of any statistical assumption made in the ML prediction, is the scatter or random distribution of the data about the zero axis as shown (Chatterjee and Hadi, 2012). For a ML algorithm to be regarded as having the capacity to predict soil expansion under inundation, the residual lag plot should be seen as capturing an uncontrolled disorderly arrangement of the independent variables' data points. Hence, a rather conspicuous or perceived trends in the pattern of the data points would depict or suggest the residual's error term non-independence. It is necessary to reiterate that before dataset training, these datasets were firstly normalised, and some randomised testing carried out to avoid any form of imbalance and/or overfitting. At first glance, the ML models' residuals all appear to fall on both sides of the zero line and less concentration of the same on either side of the axes. A much closer examination of the plot indicates an aggregation of the data points when considering the stand-alone ML models with tendencies to exhibit some form of trending or less disorderliness when compared with the ensemble learners. The sheer trending in the data points pattern demonstrated by ANN under S1 and S2 further serves to stress its least

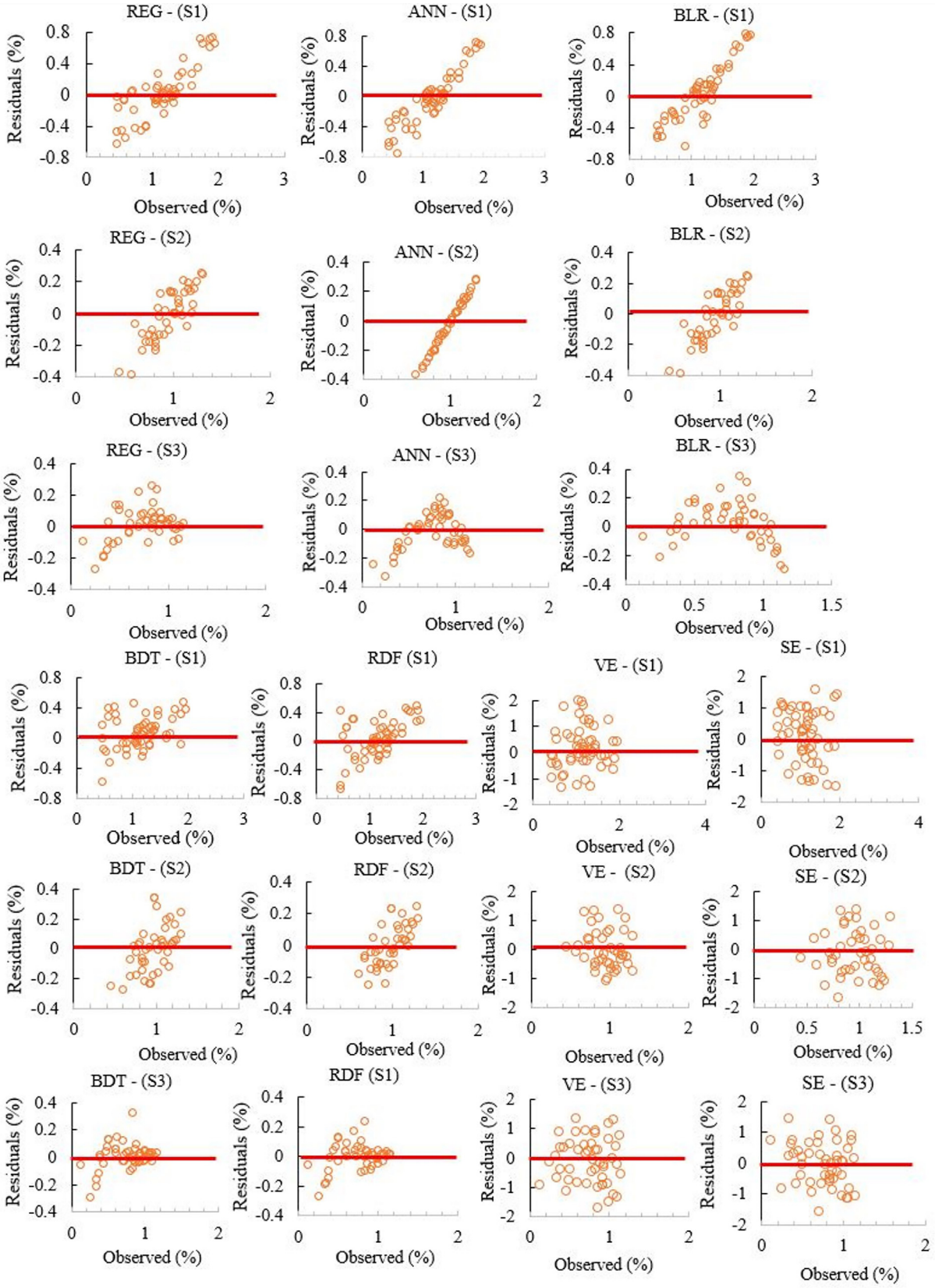

Fig. 7. Residuals – REG, ANN, BLR, BDT, RDF, SE, VE.

performing ability as already discussed in the previous sections. In general, as could be observed from Fig. 7, the meta-ensembles do show the best independence of error terms given their forms of data positioning in the lag plot.

When comparing the lag plots across the different scenarios of independent feature combination it is observed that much more randomness is displayed under S3 and followed by S1.

4.1.3. Distribution of residuals of the best ensembles

Since the best performing ML models are those of the meta-ensembles, it was pertinent to consider the distribution of their residuals and normality hence, providing a further basis through which their effectiveness and authenticity can be assessed. Unlike residual plots, the assumptions of distribution of random errors in the prediction is evaluated by observing their degree of symmetry with respect to the origin (Galwey, 2014). Thus, good ML models will tend to have their residuals peaking at zero but with few of the stochastic errors at its extremes while a low performing model on the other hand will have its distribution spreading out but with fewer errors around zero. Going by this description, Fig. 8 confirms the VE model with predictions carried out using the scenario 3 (S3) variables as being the most accurate compared with S1 and S2. On the other hand, the most normally distributed residuals as well as one with the most bin counts is shown by S3 suggesting that its variable combination may have had the highest influence on soil swelling when compared with S1 and S2. Further comparison of the accuracy of prediction of the VE model across the different scenarios is given in Fig. 9. Again, predictions using the variables of S3 appear to follow the ideal line.

4.1.4. Feature importance

ML predictions with a consideration of the degree of importance of the individual explanatory variables is very essential. An analysis of feature importance can give insights into the nature of the dataset used in the prediction while also aiding the efficiency of ML predictions. An examination of feature importance also allows the assignment of scores to input variables so that relative significance of the variables can be ranked in ML modelling. In this study, 'permutation feature importance' was applied after the training and testing of the dataset during ML prediction. The statistical parameters R^2, MAE and RMSE were used to assign scores to each of variables to allow an evaluation of their importance. As depicted in Fig. 10, although the soils' plasticity index (for S2) appears to have the highest R^2 value of about 0.99, its MAE of 0.56% and RMSE of 0.59% all seem to be the highest compared to the other independent variables. For S3, CEC seems to exhibit the greatest significance or bearing on the soil expansion due to its reasonably lower values of MAE (0.05%) and RMSE (0.18%) with a corresponding high R^2 value of approximately 0.85.

For S1, the feature with the greatest importance is the moisture content (R^2, RMSE and MAE values approximately 0.8, 0.4% and 0.3% respectively). The effect of the moisture content when compared to the other variables in S1 is in no doubt given its more direct influence on swelling behaviour of soils (Ikizler et al., 2010; Bekhor and Livneh, 2014). Both unit weight and void ratio are observed as wielding equal level of significance on the capacity of the soils to swell. When considered on their own, the feature which have to do with soil texture seem to possess relatively less scores and hence, the least importance within the context of the soils used in this study. Interestingly, features such as liquid limit and activity though appearing to be apparently less important including the coarse and fines content of the soils, can affect soil swelling most especially when combined with the other features in the ML regression analysis as indicated by this study.

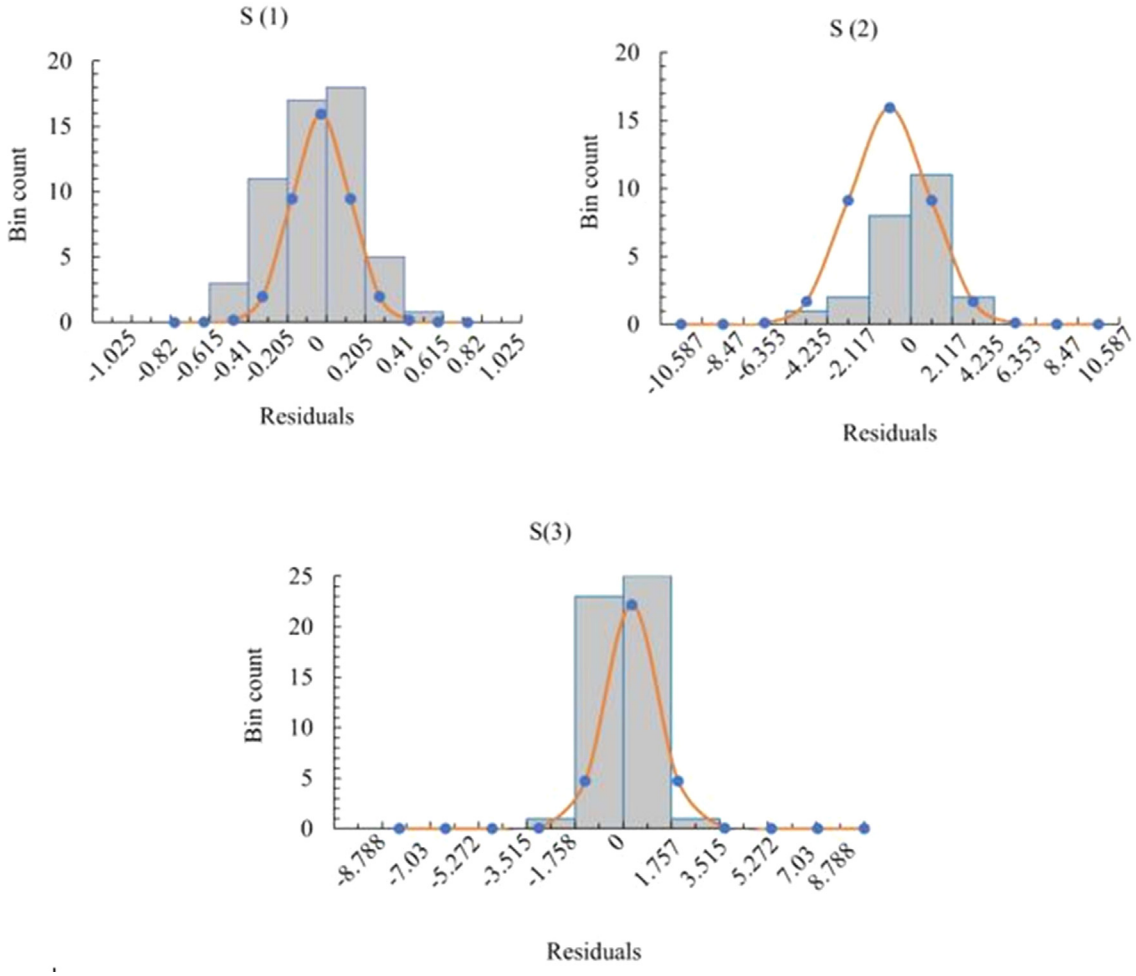

Fig. 8. Residuals and normal distribution of VE model – S(1); S(2); S(3).

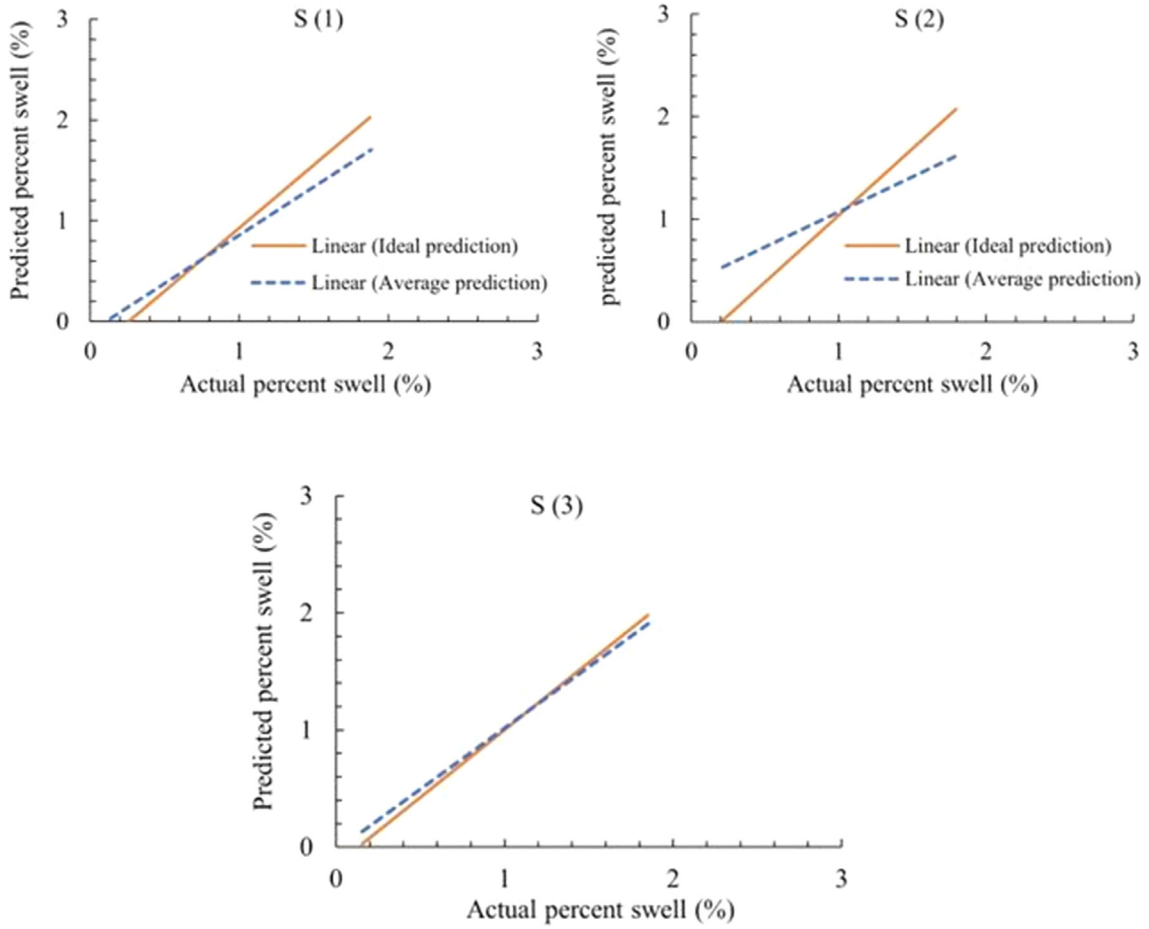

Fig. 9. Predicted vs. actual percent swell (%) – S(1); S(2); S(3).

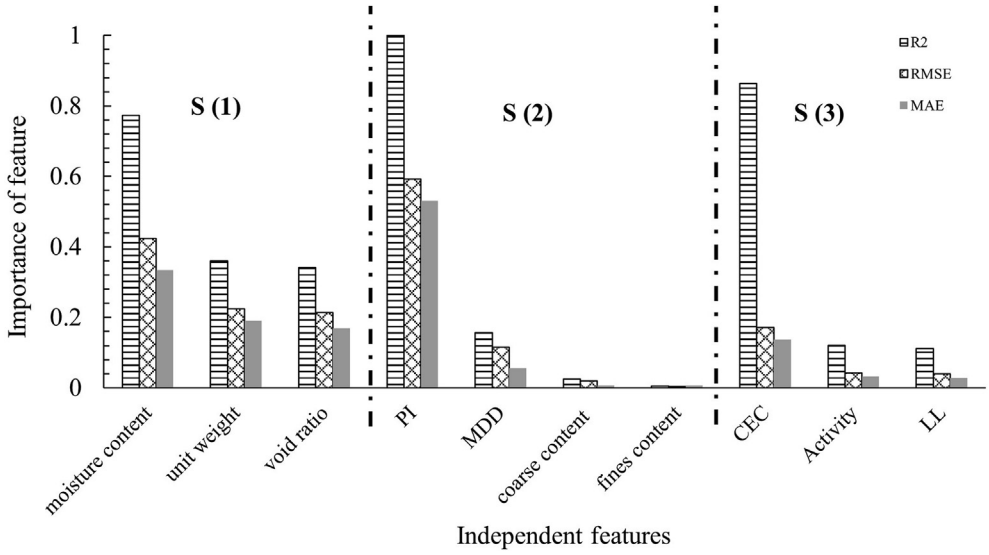

Fig. 10. Importance of features metrics.

4.2. ML classification

The dataset imported and used for ML training, testing and validation in this study are those that represent two categories of clay soils namely – clay of low plasticity (CL) and clay of high plasticity (CH) as defined by the Casagrande – Unified Soil Classification System (USCS). The statistical distribution of the categories was given previously in Fig. 2. Hence, an application of ML dichotomous classification was necessary to confirm that the best performing meta-ensembles were able to learn the soil patterns and provide swell predictions accordingly.

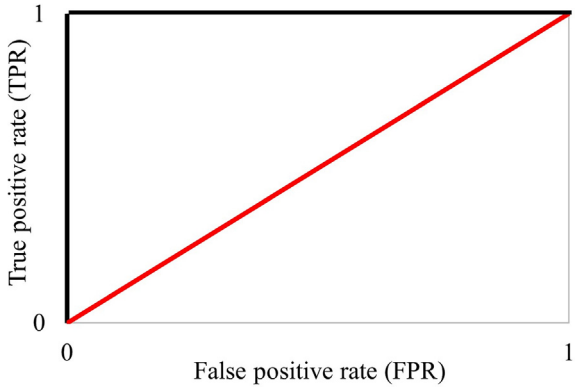

Fig. 11. Receiver operating characteristic curve, ROC.

4.2.1. Sensitivity analysis of best ML models

It is typically required that for most dichotomous classification problems a boundary for classifiers between different classes be set by a threshold point. Therefore, an analysis of sensitivity carried out using ROC (receiver operating characteristics curve) can serve as an important tool for the measurement of the capacity of a model to recognise or identify what the true positives are between the class labels. The ROC as shown in Fig. 11, tends to evaluate the best performing ML classifier from the relationship that exist between the true and false positive rates during the changes that occur in the decision threshold. Within the ROC curve, a region exists on either side of the red diagonal line that is called area under curve (AUC) that measures the degree of separation between classes. A perfect classification is achieved when the curve is tangent or in alignment with the upper left corner of the plot. This behaviour does also suggest 100% of sensitivity with no false negatives either being assumed or existing in the prediction. On the other hand, an act of random prediction or guessing may mean the curve being in alignment with the diagonal line of no-discrimination (AUC of 0.5) and therefore suggesting a bad or the worst prediction.

The ROCs of Fig. 12 show the meta-heuristic ensembles (VE and SE) as having the ability to discriminate or distinguish between the positive and negative classes. This can be further confirmed in Table 6 which indicates very high values of the classification metrics resulting from the prediction. The AUC and the accuracy of prediction for instance are observed to be generally above 0.95. It can then be inferred that the behaviour of the meta-ensembles in the binary class prediction has ensured a reduction of both type 1 and type 2 prediction errors.

Further examination of the models with dataset training performed using the cross-validation techniques namely *k*FCV and MCCV are also indicated in Fig. 12. As mentioned previously, using various cross-validation methods ensures a non-biased estimation and evaluation of the model's calibration during the prediction. Hence, both *k*FCV and MCCV are used to improve the prediction (prevent overfitting) of ML regression analyses while also serving as fine-tuning mechanisms to the TVS technique. It is observed that predictions resulting from the 10-fold validation technique provide the best outcome between the two meta-heuristic models (VE and SE). Further confirmation of this claim is given by their classification scores in Table 6. On the other hand, dataset training and testing performed by using the TVS technique seems to produce the least outcome in terms of the models' predictive capability.

4.2.2. Lift curve

Another means of sensitivity analysis of a classifier is by using the lift curve (LC). Through the LC, further evidence can be provided to indicate how effective a model is compared to one that involves an act of randomly guessing between the different classes. This is because for a dichotomous classification problem, a random model may give an inaccurate prediction when compared to a much better model with greater proportions of the sampled dataset. LC represents the cumulative gains ratio between a model plus a baseline lifting portion of the curve coinciding with the horizontal percentile axis. Fig. 13 is the LC for the meta-heuristic ensembles comparing between the dataset validation methods used. There appears to be a slight difference in the LC between the cross-validation techniques with the entire area of the curves ris-

Table 6
Meta-ensemble models' classification metrics.

Cross validation method	Metrics	Meta-ensemble models	
		VE	SE
TVS	Accuracy	0.975	0.971
	Recall	0.891	0.873
	F1 score	0.891	0.873
	AUC	0.778	0.78
k-FCV	Accuracy	0.977	0.958
	Recall	0.932	0.905
	F1 score	0.932	0.905
	AUC	0.925	0.947
MCCV	Accuracy	0.869	0.958
	Recall	0.914	0.891
	F1 score	0.896	0.891
	AUC	0.869	0.914

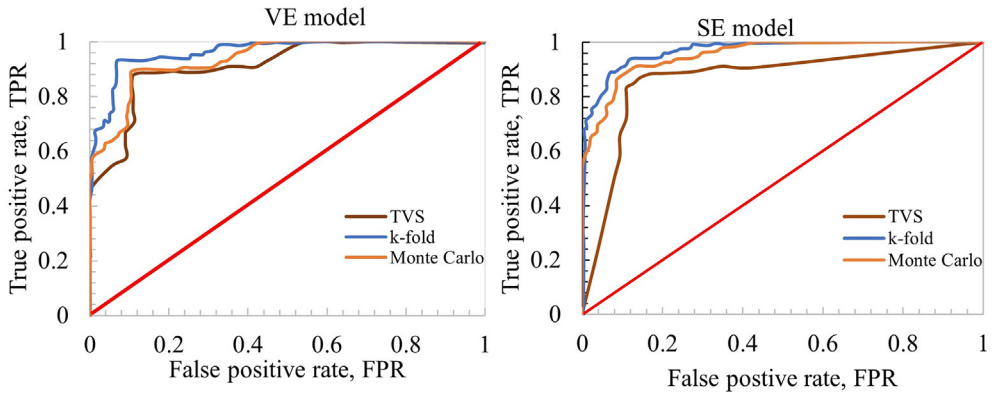

Fig. 12. Receiver operating curve (ROC) of meta-ensemble models.

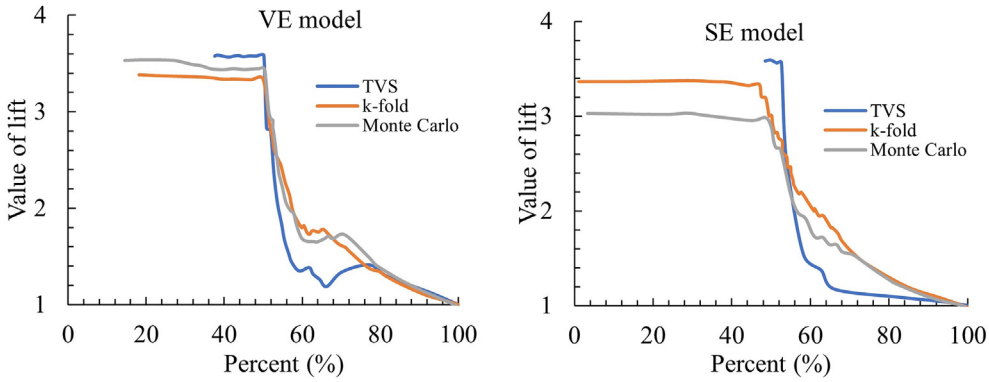

Fig. 13. Meta-heuristic ensemble models' lift curve.

Table 7
Models' classification metrics.

Metrics	Binary classification models							
	LR	BPM	ANN	SVM	D-SVM	RDF	VE	SE
Accuracy	0.855	0.927	0.909	0.927	0.927	0.909	0.975	0.971
Recall	0.200	0.600	0.600	0.600	0.600	0.600	0.891	0.873
F1 score	0.333	0.750	0.706	0.750	0.750	0.706	0.867	0.837
AUC	0.931	0.918	0.924	0.924	0.893	0.882	0.778	0.780

ing from the percentile axis being somewhat indistinguishable although the method of dataset training and testing procedure relying on TVS is the lowest for the VE model. On the other hand, the 10-fold cross validation technique seems to produce the best lifting while the TVS is just slightly lower when considering the SE model.

4.2.3. Comparison between traditional classifiers and meta-heuristic ensembles

In order to further validate the superior performance of the meta-heuristic ensembles as applied to the binary classification of the swelling soils' clay categories, a comparison is now made between them and other frequently used dichotomous classifiers namely logistic regressor (LR), bayes point machine (BPM), artificial neural networks (ANN), support vector machines (SVM), deep-support vector machines and random decision forest (RDF). Table 7 presents the scores of classifiers used in the prediction of the soil clay categories. Just like the meta-ensembles, the other stand-alone and tree-based classifiers do at least possess the ability to distinguish between categories given that their AUC scores are more than 0.5. With the exception of the meta-ensembles, BPM, SVM, D-SVM appear to outperform (higher AUC, accuracy, recall and F1 score) ANN and LR as stand-alone models. LR has the least predictive performance (accuracy of 0.86, recall rate of 0.42 and F1 score of 0.33). One of the reasons for the inability of a LR to learn categorical features is due to its assumption of extreme linearity having been an extension of a linear regressor although there may be some occasions of multiple collinearities among inputs and output labels. Interestingly, ANN seems to do much better as a dichotomous classifier compared to both the LR as a stand-alone model and when used in a regression problem. This does indicate that within the context of this study and by using the same optimised parameter settings given previously in Table 4, ANN as a classifier may have been more suited to handle the complex non-linear response of binary classification. It is also observed in Table 7 that the D-SVM does not appear to demonstrate a massive difference in their performance when compared to SVM hence, suggesting that the integration of a kernel machine and several deep learning networks may not yield as much at least within the nature of classification problem considered in this study.

Nevertheless, a hybridisation of multiple classifier models through the voting and system of stacking does enable a much higher accuracy of prediction as the classification metrics of Table 7 demonstrates. It is very needful to bear in mind that the classification metrics resulting from an aggregation of models as produced by VE and SE are considered as "weighted" scores due to the processes of constant averaging during the prediction. Hence, although the weighted AUC for the meta-heuristic ensemble is lower than those of the traditional classifiers, however when considered under micro-averaging as observed in Table 8 are observed to be much higher. In conclusion, the micro-score as depicted in Table 8 is much preferrable due to the seeming class imbalance of the categorical data of CL and CH as used in this study (Fig. 2).

Table 8
Meta-ensemble class scores by type.

Class metrics	Type	Meta-ensembles	
		VE	SE
Recall	Weighted	0.891	0.873
	Macro	0.667	0.611
	Micro	0.891	0.873
F1 score	Weighted	0.867	0.837
	Macro	0.719	0.646
	Micro	0.891	0.873
AUC	Weighted	0.778	0.780
	Macro	0.778	0.780
	Micro	0.916	0.914

5. Study significance, recommendations, and application

This study has certainly extended the evolving machine learning and artificial intelligence niched within the geotechnical engineering discipline (Chou and Ngo, 2018; Yin et al., 2018; Jin and Yin, 2020; Tinoco et al., 2020; Zhang et al., 2020; Zhang et al., 2021a,b; Zhang and Wang, 2021). Quite a handful of swelling soils'

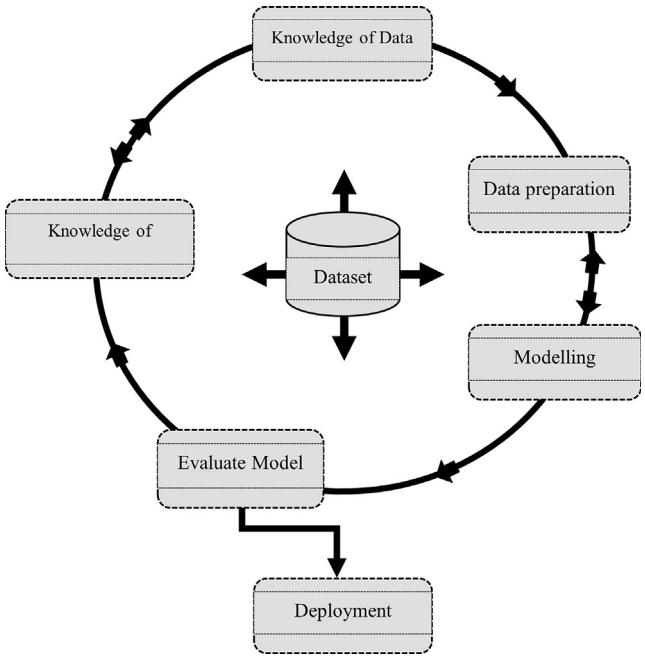

Fig. 14. CRISP-DM models' data mining cycle.

influential factors have been examined in this research. For the sake of further data-driven decision in this area of soil mechanics, it is recommended that additional soil properties such as mineralogy, suction, etc and including environmental factors like vegetation, groundwater movement, drainage, etc be evaluated and analysed using the concept developed from the present study. This research and its adopted machine learning techniques sits succinctly within a typical lifecycle of data mining as depicted by the CRISP-DM (CRoss Industry Standard Process for Data Mining model as shown in Fig. 14. Hence, in order to practically apply its concepts and ideas, the last step in the cycle – the deployment stage; all the resources and background python coding could be successfully deployed in any ground engineering-related organisation's server and the best performing models used on new data to determine and assess the behaviour of swelling soils. The implementation of this machine learning prediction can be performed at the preliminary phases of civil construction and related land developments most especially those involving geotechnical or geological site characterisation.

6. Conclusion

Machine learning regression and binary algorithms have been applied to evaluate and predict the behaviour of soils of wide-ranging plastic properties subjected to expansion under inundation. Machine learning binary classification was performed across multiple cross-validation techniques. Based on the nature of dataset used, three different scenarios of independent variable combinations were investigated. An evaluation of the significance of the features utilised as explanatory variables in the prediction was carried out using the concept of "permutation feature importance". The following are the main conclusions derived from this study:

(i) Preliminary results indicated several levels of deviations from the predictor variable of soil swelling across the stand-alone and tree-based ensemble learners used. BLR ML algorithm had the highest standard deviation (about 0.90) from the scored mean of prediction while the tree-based ensemble learners produced the lowest deviations from the mean.

(ii) The accuracy of predictions produced by both REG and BLR were slightly greater than those that relied on ANN but with the tree-base ensemble learners outperforming the stand-alone algorithms. The meta-ensembles (VE and SE) gave the best overall performance (with the greatest R^2 value of 0.94 and RMSE of 0.06% exhibited by VE).

(iii) CEC, plasticity index and moisture content were the features considered to have the highest level of importance because of their overall tremendous influence on soil expansion whereas the features based on soil texture possessed the lowest influence on soil swelling and thereby regarded as having less importance.

(iv) Kernelised dichotomous classifiers (SVM, D-SVM and BPM) produced the most accurate results with an average accuracy and recall rate of 0.93 and 0.60 respectively when compared to ANN, LR and RDF whose average accuracy and recall rate of were 0.89 and 0.46 respectively.

(v) An examination of the sensitivity of ML classification of the swelling soils by using the receiver operating characteristics (ROC) and the lifting curve (LC) indicated the capacity of all classifiers as having the capacity to discriminate between positive and negative classes. Meta-ensembles assessed across three cross validation techniques showed that ML training and testing performed using *k*-fold cross validation technique, produced the best performance when compared to the train-validation split and Monte Carlo methods.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Abdullah, A., Veltkamp, R.C., Wiering, M.A., 2009. An ensemble of deep support vector machines for image categorization. SoCPaR 2009 - Soft Comp. and Pattn. Recog. 301–306. https://doi.org/10.1109/SoCPaR.2009.67.

Adem, H.H., Vanapalli, S.K., 2014. Elasticity moduli of expansive soils from dimensional analysis. Geotech. Res. 1, 60–72. https://doi.org/10.1680/gr.14.00006.

Alizamir, M., Kim, S., Zounemat-Kermani, M., Heddam, S., Shahrabadi, A.H., Gharabaghi, B., 2021. Modelling daily soil temperature by hydro-meteorological data at different depths using a novel data-intelligence model: deep echo state network model. Artif. Intel. Rev. 54, 2863–2890. https://doi.org/10.1007/s10462-020-09915-5.

Ashayeri, I., Yasrebi, S., 2009. Free-swell and swelling pressure of unsaturated compacted clays; experiments and neural networks modeling. Geotech. Geol. Eng. 27, 137–153. https://doi.org/10.1007/s10706-008-9219-y.

Behnood, A., Golafshani, E.M., 2018. Predicting the compressive strength of silica fume concrete using hybrid artificial neural network with multi-objective grey wolves. J. Clean. Prod. 202, 54–64. https://doi.org/10.1016/j.jclepro.2018.08.065.

Bekhor, S., Livneh, M., 2014. Using the artificial neural networks methodology to predict the vertical swelling percentage of expansive clays. J. Mater. Civil Eng. 26, 06014007. https://doi.org/10.1061/(ASCE)MT.1943-5533.0000931.

Ben Chaabene, W., Flah, M., Nehdi, M.L., 2020. Machine learning prediction of mechanical properties of concrete: critical review. Constr. Build. Mater. 260, 119889. https://doi.org/10.1016/j.conbuildmat.2020.119889.

Berrah, Y., Boumezbeur, A., Kherici, N., Charef, N., 2018. Application of dimensional analysis and regression tools to estimate swell pressure of expansive soil in Tebessa (Algeria). B. Eng. Geol. Environ. 77, 1155–1165. https://doi.org/10.1007/s10064-016-0973-4.

Buzzi, O., 2010. On the use of dimensional analysis to predict swelling strain. Eng. Geol. 116, 149–156. https://doi.org/10.1016/j.enggeo.2010.08.005.

Buzzi, O., Giacomini, A., Fityus, S., 2011. Towards a dimensionless description of soil swelling behaviour. Geotechnique 61, 271–277. https://doi.org/10.1680/geot.7.00194.

Charlie, W.A., Asce, M., Osman, M.A., Ali, E.M., 1985. Construction on expansive soils in Sudan. J. Constr. Eng. M-ASCE 110, 359–374. https://doi.org/10.1061/(ASCE)0733-9364(1984)110:3(359).
Chatterjee, S., Hadi, A., 2012. Regression Analysis by Example. Wiley, Hoboken, New Jersey, p. 408.
Chittoori, B., Puppala, A.J., 2011. Quantitative estimation of clay mineralogy in fine-grained soils. J. Geotech. Geoenviron. Eng. 137, 997–1008. https://doi.org/10.1061/(ASCE)GT.1943-5606.0000521.
Chou, J.S., Yang, K.H., Lin, J.Y., 2016. Peak shear strength of discrete fiber-reinforced soils computed by machine learning and metaensemble methods. J. Comput. Civil. Eng. 30, 04016036. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000595.
Chou, J.S., Ngo, N.T., 2018. Engineering strength of fiber-reinforced soil estimated by swarm intelligence optimized regression system. Neur. Comp. Appl. 30, 2129–2144. https://doi.org/10.1007/s00521-016-2739-0.
Çimen, Ö., Keskin, S.N., Yıldırım, H., 2012. Prediction of swelling potential and pressure in compacted clay. Arab. J. Sci. Eng. 37, 1535–1546. https://doi.org/10.1007/s13369-012-0268-4.
Das, S.K., Samui, P., Sabat, A.K., Sitharam, T.G., 2010. Prediction of swelling pressure of soil using artificial intelligence techniques. Envir. Earth Sci. 61, 393–403. https://doi.org/10.1007/s12665-009-0352-6.
DeRousseau, M.A., Kasprzyk, J.R., Srubar, W.V., 2018. Computational design optimization of concrete mixtures: A review. Cem. Concr. Res. 109, 42–53. https://doi.org/10.1016/j.cemconres.2018.04.007.
DeRousseau, M.A., Laftchiev, E., Kasprzyk, J.R., Rajagopalan, B., Srubar, W.V., 2019. A comparison of machine learning methods for predicting the compressive strength of field-placed concrete. Constr. Build. Mater. 228, 116661. https://doi.org/10.1016/j.conbuildmat.2019.08.042.
Dreiseitl, S., Ohno-Machado, L., 2002. Logistic regression and artificial neural network classification models: A methodology review. J. Biomed. Inform. 35, 352–359. https://doi.org/10.1016/S1532-0464(03)00034-0.
Du, Y., Li, S., Hayashi, S., 1999. Swelling-shrinkage properties and soil improvement of compacted expansive soil, Ning-Liang Highway, China. Eng. Geol. 53, 351–358. https://doi.org/10.1016/S0013-7952(98)00086-6.
Elbadry, H., 2016. Simplified reliable prediction method for determining the volume change of expansive soils based on simply physical tests. HBRC J. 13, 353–360. https://doi.org/10.1016/j.hbrcj.2015.10.001.
Erguler, Z.A., Ulusay, R., 2003. A simple test and predictive models for assessing swell potential of Ankara (Turkey) Clay. Eng. Geol. 67, 331–352.
Ermias, B., Vishal, V., 2020. Application of artificial intelligence for prediction of swelling potential of clay-rich soils. Geotech. Geol. Eng. 38, 6189–6205. https://doi.org/10.1007/s10706-020-01427-x.
Erzin, Y., Erol, O., 2007. Swell pressure prediction by suction methods. Eng. Geol. 92, 133–145. https://doi.org/10.1016/j.enggeo.2007.04.002.
Erzin, Y., Gunes, N., 2013. The unique relationship between swell percent and swell pressure of compacted clays. B. Eng. Geol. Environ. 72, 71–80. https://doi.org/10.1007/s10064-013-0461-z.
Eyo, E.U., Abbey, S.J., 2021. Machine learning regression and classification algorithms utilised for strength prediction of OPC / by-product materials improved soils. Constr. Build. Mater. 284, 122817. https://doi.org/10.1016/j.conbuildmat.2021.122817.
Eyo, E.U., Ng'ambi, S., Abbey, S.J., 2019. Effect of intrinsic microscopic properties and suction on swell characteristics of compacted expansive clays. Transport. Geotech. 18, 124–131. https://doi.org/10.1016/j.trgeo.2018.11.007.
Galwey, N., 2014. Introduction to Mixed Modelling: Beyond Regression and Analysis of Variance. Wiley, Chichester, England, p. 366.
Han, T., Siddique, A., Khayat, K., Huang, J., Kumar, A., 2020. An ensemble machine learning approach for prediction and optimization of modulus of elasticity of recycled aggregate concrete. Constr. Build. Mater. 244, 118271. https://doi.org/10.1016/j.conbuildmat.2020.118271.
Ikizler, S.B., Aytekin, M., Vekli, M., Kocabaş, F., 2010. Prediction of swelling pressures of expansive soils using artificial neural networks. Adv. Eng. Softwr. 41, 647–655. https://doi.org/10.1016/j.advengsoft.2009.12.005.
Jin, Y.-F., Yin, Z.-Y., 2020. Enhancement of backtracking search algorithm for identifying soil parameters. Int. J. Num. Anal. Met. Geomech. 44, 1239–1261. https://doi.org/10.1002/nag.3059.
Jones, L.D., Jefferson, I.F., 2015. Expansive Soils, ICE Manual of Geotechnical Engineering. Institution of Civil Engineers.
Joshi, A., 2020. Machine Learning and Artificial Intelligence. Springer International Publishing, p. 261.
Kang, M.C., Yoo, D.Y., Gupta, R., 2021. Machine learning-based prediction for compressive and flexural strengths of steel fiber-reinforced concrete. Constr. Build. Mater. 266, 121117. https://doi.org/10.1016/j.conbuildmat.2020.121117.
Kayadelen, C., Taşkıran, T., Günaydın, O., Fener, M., 2009. Adaptive neuro-fuzzy modeling for the swelling potential of compacted soils. Envir. Earth Sci. 59, 109–115. https://doi.org/10.1007/s12665-009-0009-5.
Likos, W.J., Wayllace, A., 2010. Porosity evolution of free and confined bentonites during interlayer hydration. Clays Clay Min. 58, 399–414. https://doi.org/10.1346/CCMN.2010.0580310.
Nelson, J.D., Chao, K.C.G., Overton, D.D., Nelson, E.J., 2015. Foundation engineering for expansive soils, Wiley, 416 pp.
Puppala, A.J., Cerato, A., 2009. Heave distress problems in chemically-treated sulfate-laden materials. Geo-Strata 10, 28.
Puppala, A.J., Manosuthikij, T., Chittoori, B.C.S., 2014. Swell and shrinkage strain prediction models for expansive clays. Eng. Geol. 168, 1–8.
Rani, C.S., 2013. Prediction of swelling pressure of expansive soils uisng compositional and environmental factors. Inter. J. Civil Eng. Techn. 4, 134–142.
Tinoco, J., Alberto, A., da Venda, P., Gomes Correia, A., Lemos, L., 2020. A novel approach based on soft computing techniques for unconfined compression strength prediction of soil cement mixtures. Neur. Comp. Appl. 32, 8985–8991. https://doi.org/10.1007/s00521-019-04399-z.
Toksoz, D., Yilmaz, I., 2019. A fuzzy prediction approach for swell potential of soils. Arab. J. Geosci. 12, 728. https://doi.org/10.1007/s12517-019-4938-3.
Vanapalli, S.K., Lu, L., 2012. A state-of-the art review of 1-D prediction methods for expansive soils. Inter. J. Geotech. Eng. 6, 15–41.
Vapnik, V., Golowich, S.E., Smola, A., 1997. Support vector method for function approximation, regression estimation, and signal processing. Adv. Neur. Info. Proc. Sys., 281–287
Wolpert, D.H., 1992. Stacked generalization. Neural Networks 5, 241–259. https://doi.org/10.1016/S0893-6080(05)80023-1.
Yilmaz, I., 2006. Indirect estimation of the swelling percent and a new classification of soils depending on liquid limit and cation exchange capacity. Eng. Geol. 85, 295–301. https://doi.org/10.1016/j.enggeo.2006.02.005.
Yilmaz, I., Kaynar, O., 2011. Multiple regression, ANN (RBF, MLP) and ANFIS models for prediction of swell potential of clayey soils. Exp. Sys. Appl. 38, 5958–5966. https://doi.org/10.1016/j.eswa.2010.11.027.
Yin, Z.Y., Jin, Y.F., Shen, J.S., Hicher, P.Y., 2018. Optimization techniques for identifying soil parameters in geotechnical engineering: Comparative study and enhancement. Int. J. Numer. Anal. Meth. Geomech. 42, 70–94. https://doi.org/10.1002/nag.2714.
Zhang, J., Wang, Y., 2021. An ensemble method to improve prediction of earthquake-induced soil liquefaction: a multi-dataset study. Neur. Comp. Appl. 33, 1533–1546. https://doi.org/10.1007/s00521-020-05084-2.
Zhang, W.G., Li, H.R., Li, Y.Q., Liu, H.L., Chen, Y.M., Ding, X.M., 2021. Application of deep learning algorithms in geotechnical engineering: a short critical review. Artif. Intel. Rev, in press. https://doi.org/10.1007/s10462-021-09967-1.
Zhang, W., Wu, C., Zhong, H., Li, Y., Wang, L., 2021b. Prediction of undrained shear strength using extreme gradient boosting and random forest based on Bayesian optimization. Geosci. Front. 12, 469–477. https://doi.org/10.1016/j.gsf.2020.03.007.
Zhang, W., Zhang, R., Wu, C., Goh, A.T.C., Lacasse, S., Liu, Z., Liu, H., 2020. State-of-the-art review of soft computing applications in underground excavations. Geosci. Front. 11, 1095–1106. https://doi.org/10.1016/j.gsf.2019.12.003.
Zumrawi, M.M.E., 2015. Construction problems of light structures founded on expansive soils in Sudan. Intern. J. Sci. Res. 4, 896–902.
Zumrawi, M.M.E., 2012. Prediction of swelling characteristics of expansive soils. Sudan Eng. Soc. J. 58, 55–62.