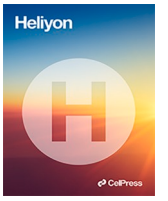
Heliyon 9 (2023) e18673

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Neural networks based linear (PCA) and nonlinear (ISOMAP) feature extraction for soil swelling pressure prediction (North East Algeria)

Bahloul Ouassila [a,*], Tebbi Fatima Zohra [b], Lekouara Laid [c,d], Bekhouche Hizia [a]

[a] LGCROI. Civil Laboratory Risks and Structures Interactions, Faculty of Technology, University of Mostepha Ben Batna 2, Fesdis, Algeria
[b] LRNAT Laboratory, Earth Sciences and Universe Institute, University of Mostepha Ben Boulaid Batna 2, Fesdis, Algeria
[c] Faculty of Science and Technology, University of Abbés Laghrour, Khenchela, Algeria
[d] Material and Hydrology Laboratory, University of Sidi Bel Abbes, Faculty of Technology, Sidi Bel Abbes, Algeria

ARTICLE INFO

Keywords:
Swelling pressure
Feature extraction
ANN
PCA
ISOMAP
Nonlinear

ABSTRACT

The swelling pressure (SP) of expansive soils is crucial for both geotechnical studies as well as practitioners. Multiple attempts have been made to correlate the SP with the properties of soil due to the difficulty of determining it in the laboratory. However, the large number of environmental and physical governing parameters makes accurate SP predictions difficult. In this paper, Artificial Neural Networks (ANNs) are used to assess accurate prediction of SP of soil. Dimension reduction techniques are intensely required for ANNs inputs. Feature extraction (FE) based dimension reduction (DR) methods map original multidimensional space into a space of reduced dimensionality. This paper presents a comparative study of linear FE using Principal Component Analysis (PCA) and nonlinear FE using ISOmetric MAPping (ISOMAP) for feed forward neural models to predict SP. Results showed that FE technique improves ANNs models compared to multiple linear regression (MLR) and ANNs model without DR. Moreover, nonlinear ISOMAP based DR technique has proven its effectiveness regarding performance metrics for five dimensions inputs (Dims), Determination coefficient ($R^2 = 0.923$), Mean absolute percentage error (MAPE = 0.072), and Root mean square error (RMSE = 54.937) and Root relative squared error (RRSE = 0.383). Therefore, ISOMAP-ANN models can be adopted to solve geotechnical problems specially those of expansive soils which have a very complex and nonlinear structure.

1. Introduction

Expansive soils are a worldwide issue [1–3]. They can be found especially in arid and semi-arid climate zones among many countries from different continents [4–7]. These regions have recorded infrastructure damage resulting from expansive soil movements which can cause significant strain and, subsequently, serious damage to overlying structures. If the damage is substantial, this might result in expensive repair expenses and service interruptions which may increase construction's cost [8,9]. Millions of dollars are lost annually due to infrastructure damage caused by expansive soil, which is higher than catastrophic natural occurrences like earthquakes and tornadoes. The swelling mechanism of expansive soils is highly complex and governed by a multitude of factors. These

* Corresponding author. LGCROI. Civil Laboratory Risks and Structures Interactions, Faculty of Technology, University of Mostepha Ben Boulaid Batna 2, Fesdis, Algeria.
E-mail address: o.bahloul@univ-batna2.dz (B. Ouassila).

https://doi.org/10.1016/j.heliyon.2023.e18673
Received 27 March 2023; Received in revised form 22 July 2023; Accepted 25 July 2023
Available online 26 July 2023

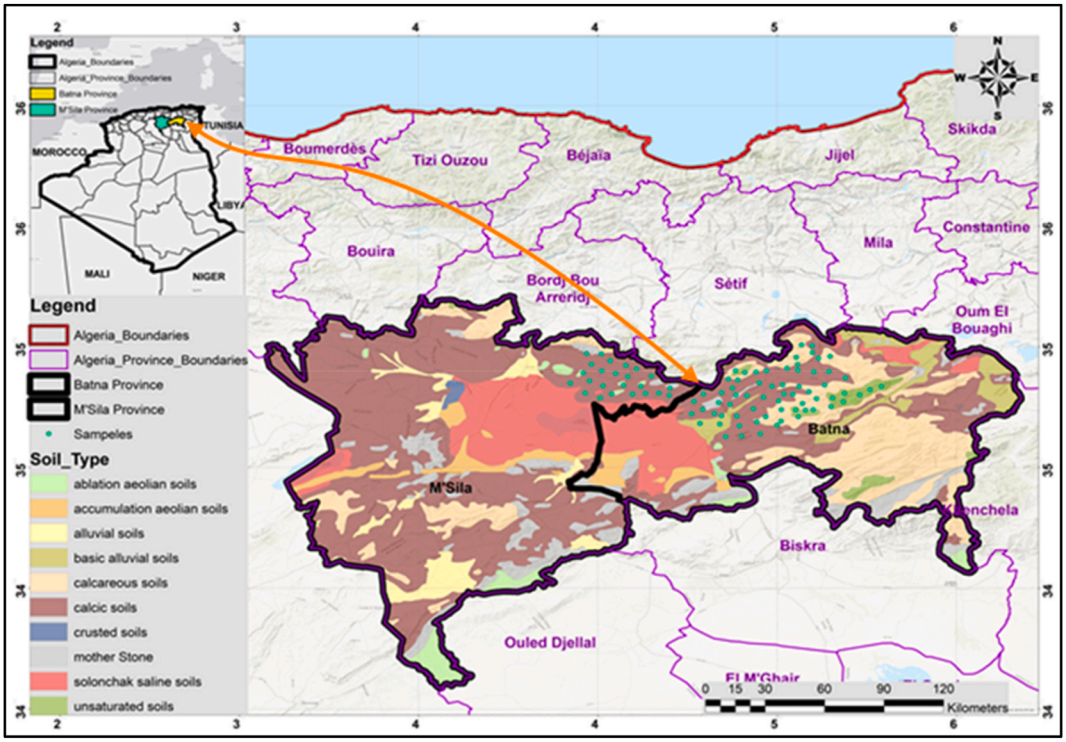


Fig. 1. The study area sampling sites (Batna-MSila East of Algeria).

factors can be categorised into three major categories: soil characteristics, environmental factors and soil stress [1,2,10,11].

At the stage of foundation design preparation, the heave information and swelling pressure are essential for accommodating the predicted volume change and foundation movement [12,13]. Commonly, oedometer tests are performed to assess swelling pressure. Swelling pressure measurement is affected by testing methodologies employed. Moreover, sample disturbance can also led to a significant impact on the measurement of swelling pressure [14–16]. Overall, oedometer tests are inconvenient and time-consuming for application in geotechnical engineering [17–19]. In addition, the conventional methods based on the elasticity and plasticity theories cannot accurately represent the true behavior of the unsaturated soils. This is because of the complexity and variability of soil structure [20]. Furthermore, the use of empirical correlations for swelling pressure can not be generalized for all sites and all soils. For this reason, geotechnical scientists have attempted to develop new models based on intelligent systems such as Artificial Intelligence (AI) or Machine Learning (ML) [18]. Artificial Neural Networks (ANNs) mimic human brains, learning relationships and interdependencies, and extracting information from data. They have potential for predicting soil properties [1,2,11,21–27].

Reducing the number of variables of data not only reduces complexity but it also makes machine learning algorithms computationally efficient and provides better prediction performance [9,28,29]. To deal with high-dimensional data one significant and commonly used technique is Dimensionality Reduction (DR). The latter has been accomplished using variety of methods. Principal component analysis (PCA) maximizes data variance in lower-dimensional representations through linear mapping. Considerable studies undertaken the use of PCA in both feature extraction (FE) and feature selection (FS) to neural network models for soils properties prediction [18,30,31].

Taylor [32] investigated a combination of PCA and a sigmoid-feedforward Neural Network to predict soil nitrate concentration results and showed that the PCA based neural network approach reduces errors than the integration and cross correlation methods. Amin Benbouras and Petrisor [18] elaborated through several advanced machine learning methods after using feature selection techniques namely PCAs, Gamma test, and forward selection to predict soil swelling index parameter. The feature selection random forest (FS-RF) provides the most accurate prediction values. However, PCA is less efficient for data with complex and nonlinear structures, than some non-linear DR algorithms, such as kernel PCA, locally linear embedding, Laplacian eigen-maps and ISOMAPs [28,33]. ISOMAP is a nonlinear DR technique based on spectral theory; it attempts to maintain the geodesic distances in the lower dimension. This method can effectively learn nonlinearly correlated variables [34–36]. This paper presents a comparative analysis of algorithms belonging to manifold learning and nonlinear dimensionality reduction feature extraction (ISOMAP) to a traditional linear method (PCA) to predict soil pressure prediction using a neural model.

Table 1
Literature proposed models for swelling pressure SP prediction.

Authors	Methods and Equations
[37]	**MLR:** $S_P = 13720.02 + 149.24E_c - 49401G_s - 25.65C + 20.33PI + \varepsilon\ R^2 = 0.9665$
[38]	**MLR:** $P_s = -7.14 + 4.345\gamma_d + 0.947 \log s\ R^2 = 0.81$
[39]	**ANN:** $R^2 = 0.997$
[40]	**DENN:** $R^2 = 0.76$, **BRNN:** $R^2 = 0.81$, **LMNN:** $R^2 = 0.77$, **SVM:** $R^2 = 0.88$
[1]	**ANN:** $0.9486 \leq R^2 \leq 0.9653$ **MRA:** $0.6687 \leq R^2 \leq 0.9511$
[10]	**MLR:** $Log\ P_s = 0.0276PI - 365.2118{\gamma_d}^{-2.4616} - 0.0320w_0 + 2.2292\ R^2 = 0.974$
[41]	**ANN:** $R^2 = 0.999$, **ANFIS:** $R^2 = 0.998$
[42]	**SLR:** $P_s = 428 - 4.52C\ R^2 = 0.36$
	MLR: $P_s = 323 - 4.87C + 3.02I_P\ R^2 = 0.18$
[6]	**DA:** $P_s = w^{-0.82}{W_L}^{0.37}{I_P}^{1.04}{F_f}^{-0.38}{\rho_d}^{-1.86}{P_c}\ R^2 = 0.94$
[25]	**MLR:** $P_s = -5.599 - 0.153Z + 0.034C + 0.145W + 2.374\gamma_d\ R^2 = 0.42$
	ANN: $R^2 = 0.88$
[43]	**MLR:** $P_s = 339.43 - 25.97\gamma_d - 9.41w - 0.68F_f + 2.06W_L + 6.1I_P + 1.23P_c\ R^2 = 0.933$
[44]	**DOE:** $P_s = 2604.6 - 1041.3A - 11440B + 4034C - 2568.7D + 1420.3E + 133.6F$
[45]	**MLR:** $R^2 = 0.933$
[46]	**MLR:** $S_P = 21.466 + 3.021F + 0.336LL - 0.307PI - 8.071\gamma_d - 2.846W_L + 6.444\alpha - 82.963D_r\ R^2 = 0.951$

Table 2
Descriptive Statistics of geotechnical parameters.

Characteristics	LL (%)	PI (%)	WS (%)	Ff2mm (%)	Ff80 μm (%)	Ff2μm (%)	VBS (g/100 g)	SP (kPa)
Min	63.84	41.66	5.800	92.56	70.80	30.10	8.00	389.0
1st Qu	72.70	47.34	7.600	95.89	88.84	40.78	10.66	516.5
Median	79.66	50.94	8.800	97.23	91.60	48.58	12.50	580.0
Mean	79.53	51.47	9.089	96.96	90.07	47.39	12.26	610.8
3rd Qu	86.48	55.94	10.700	98.24	93.85	54.08	13.66	693.5
Max	95.30	66.97	12.600	99.37	97.28	62.00	16.00	936.0
SD	8.35	5.49	1.82	1.62	5.65	8.02	2.06	147.97

2. Materials and methods

2.1. Study area and data description

This study was carried out in two regions in the North East of Algeria (Batna-M'Sila), which are characterised by an arid climate with very common clay and marl geological formations that are generally subject to the shrinkage-swell phenomenon, the main cause of damage observed on various structures due to the high pressures generated (Fig. 1).

In order to recognise and evaluate the swelling pressure of soils, an experimental study of 64 samples from the two regions mentioned above was carried out. The soil parameters used in this research were measured in accordance with international or European standards in the laboratory. Based on many studies (Table 1) selected variables in the present investigation are Liquid Limit (LL %), Shrinkage Limit (WS %), Plasticity Index (PI %), Methylene blue Value (VBS g/100 g), Swelling Pressure (SP KPa). Principal descriptive statistics are indicated below (Table 2). They have also been plotted using a boxplot. This facilitates a greater understanding of data distribution, as depicted in Fig. 2. The median of the data is represented by the middle line in the graph, allowing for easy differentiation between the quartiles.

2.2. Proposed methodology

Principal steps of the adopted methodology Fig. 3 are divided in two parts after preprocessing of raw data.

2.2.1. Artificial Neural Networks (ANN)

An artificial neural network is a highly simplified model of a brain based on the main computational element, the artificial neuron. It is a simplification of a biological neuron. Several neurons with activation functions and interconnected by weights form a network. Most neural networks are structured in a multi layer perceptron (MLP) [47]. In each formal neuron, a summation and bias $\sum (w_i\ x_i) +$ bias are performed with inputs x_i associated with weights w_i and bias b_i. After that, activation function is used to convert the sum of the weighted inputs from the node into an output value for the subsequent hidden layer or as output (Fig. 4).

A neural network works in two consecutive phases: a learning design phase considering 70% of the data and a testing phase with the remaining 30%. The first phase consists of choosing the architecture of the network and its parameters. Namely the number of hidden layers and the number of neurons in each of these layers. The activation functions of the network are adjusted to produce a minimum error on the output variable [48]. The network is then tested on the test sample. The best architecture is the one that will present better performance criteria.

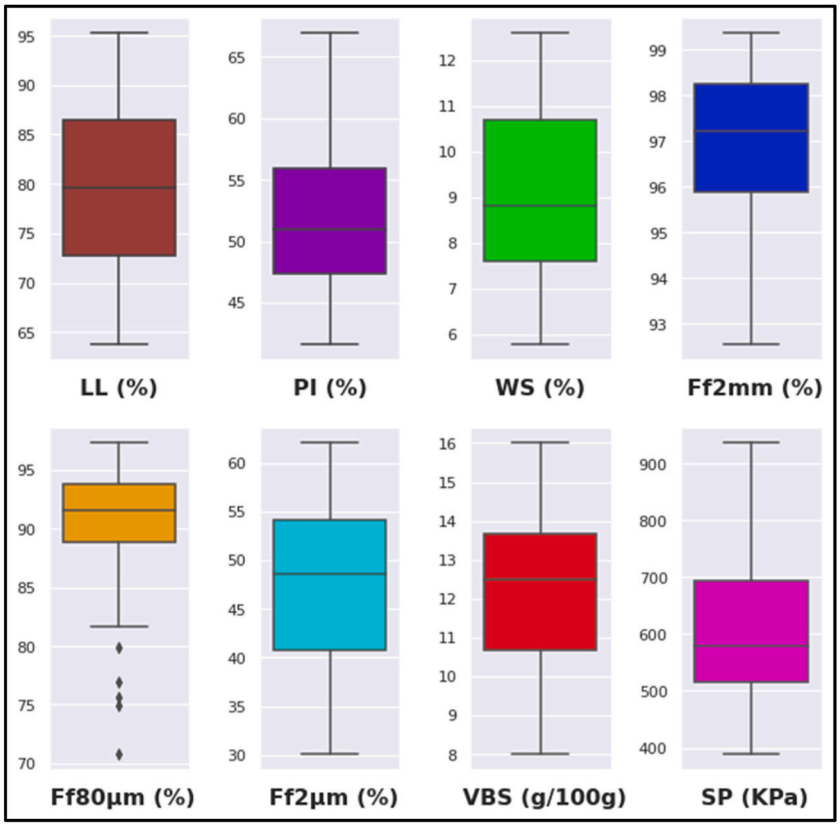


Fig. 2. Box plot of geotechnical parameters.

2.2.2. Multiple regression model (MLR)

Multiple regression analyzes the relationship between a single dependent variable and multiple independent variables, using known predictor values and weighing their relative contribution to the overall prediction. The MLR model is expressed in Eq. (1)

$$Y = a + b_1X_1 + b_2X_2 \ldots\ldots\ldots b_nX_n \tag{1}$$

Here Y is the dependent variable, and $X_1, \ldots, X_n$ are the *n* independent variables. The weights a, $b_1, \ldots, b_n$, are calculated usually by least squares estimation for optimal prediction of dependent variable.

To build a regression model, stepwise method [49] is used by adding or removing predictor variables, typically through a series of F-tests or T-tests. Based on the test statistics of the estimated coefficients, variables to be added or omitted are selected.

Stepwise regression minimizes the independent variables, by using forward and backward algorithms (Table 3).

2.2.3. Principal component analysis (PCA)

PCA [51] is a classic form of dimensional reduction. PCA uses orthogonal basis vectors to maximize variance, constructs an m × k projection matrix Uk, and obtains feature vectors by projecting X onto Uk ($\mathbf{Y} = \mathbf{XUk}$) [52] (Table 4).

2.2.4. ISOmetric feature MAPping (ISOMAP)

ISOMAP [54] is a dimension reduction method capable of estimating the degree of non-linear freedom underlying a dataset. The algorithm (Table 5) constructs a graph linking each point to its nearest *k* neighbours by assuming the approximation of the geodetic distance to the Euclidean distance. As follows.

- All the neighbourhoods are represented as edges of a graph. Each edge is weighted by the distance delta_{i, j}.
- From the neighborhood graph. The shortest path between each pair of vertices is determined. The shortest path is estimated by the Djikstra or Floyd Warshall algorithm.
- The matrix of the shortest paths. Thus constituting a matrix of geodetic distances. Is then used with the MDS (MultiDimensional Scaling) technique [55] which is none other than an PCA of the geodetic distance table to construct a Euclidean subspace.

2.2.5. Merits and demerits of technique

Table 6 summarises the main characteristics of the two methods used for linear PCA and nonlinear ISOMAP dimension reduction.

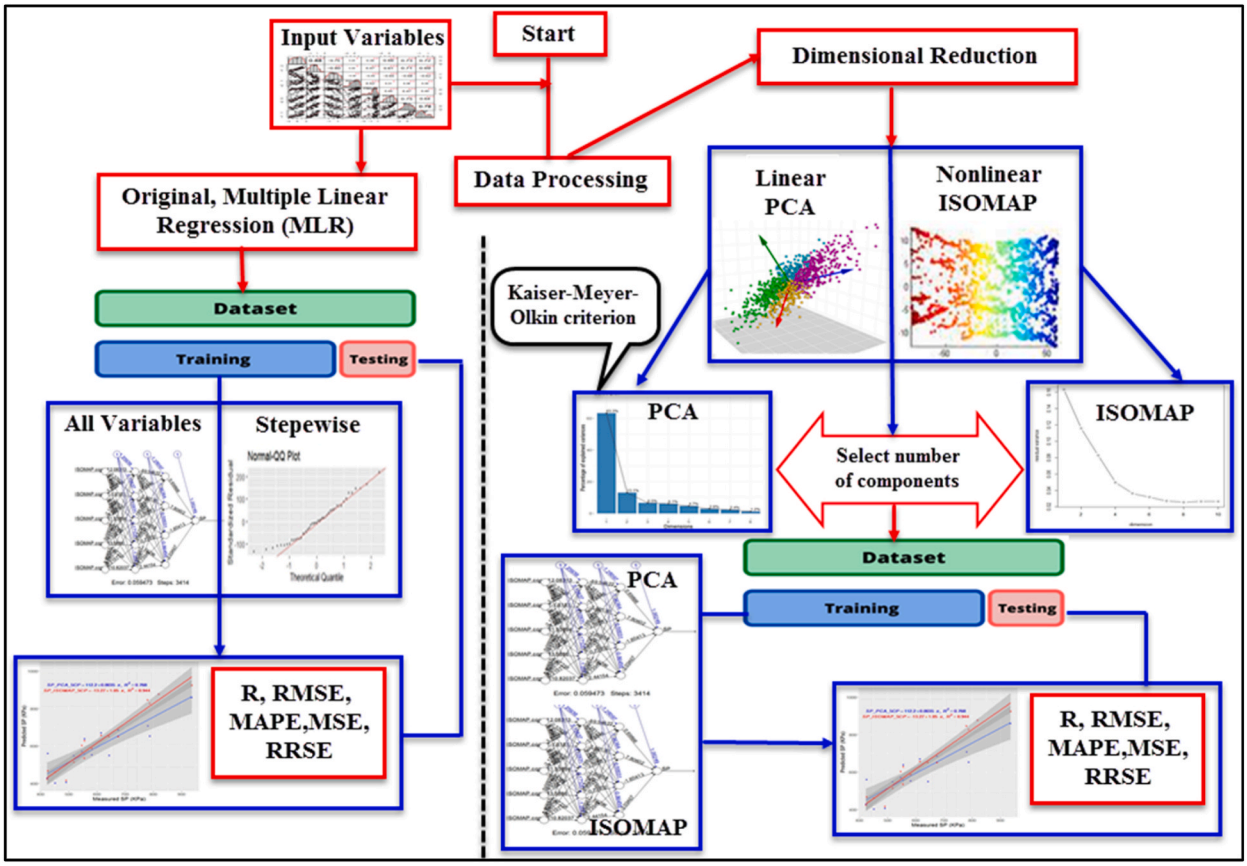


Fig. 3. Methodology flow chart of the present study.

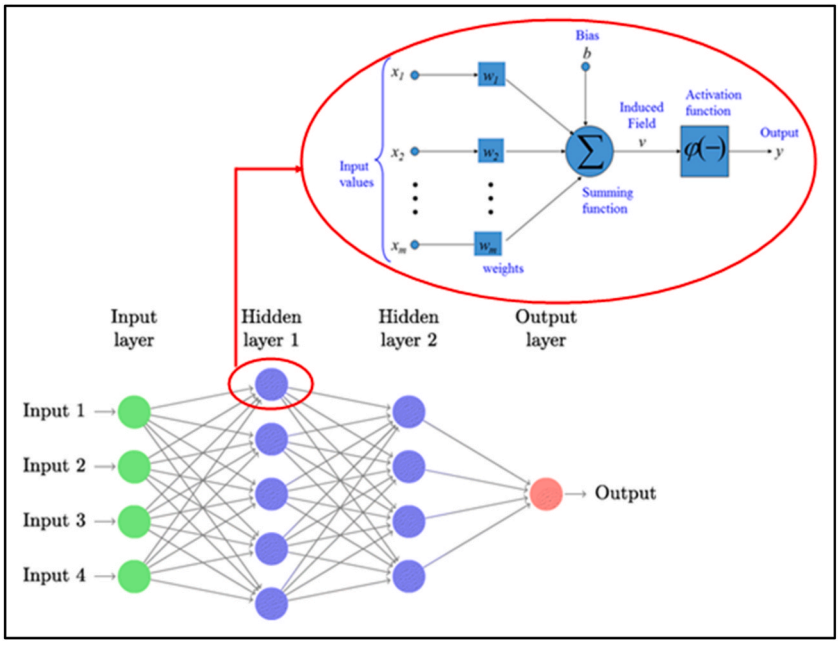


Fig. 4. Multi-Layers Perceptron MLP and Formal neuron.

Table 3
Multiple regression model using stepwise Algorithm [50].

Algorithm Forward Stepwise Selection
1. Let M_0 represent the null model, which lacks predictors.
2. For k = 0, …, p − 1: where: k is kth iteration, p: number of predictors (a) Consider all p − k models that supplement the predictors in $M_k$with one additional predictor.
(b) Select the best model among these p − k models, and label it M_k+1. It is defined as having the minimum RSS or highest R^2.
3. Best model selection using cross-validated prediction error, C_p (AIC)**1**, BIC**2**, or adjusted R^2.
1: Akaike information criterion
2: Bayesian information criterion
Algorithm Backward Stepwise Selection
1. Let M_p denote the comprehensive model that includes all p predictors.
2. For k = p, p − 1, …, 1: (a) Consider all k models containing all but one of the predictors in M_k, for a total of k − 1 predictors. (b) Choose the most advantageous of these k models, and call it M_{k-1}. It is defined as having the lowest RSS or the greatest R^2.
3. Best model selection using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

Table 4
PCA algorithm [53].

Algorithm PCA
1. Compute the mean feature vector
$\mu = \frac{1}{p}\sum_{k=1}^{p} x_k$, where x_k is a pattern ($k = 1$ *to* p), p:number of patterns, x is the feature matrix
2. Covariance matrix
$C = \frac{1}{p}\sum_{k=1}^{p}\{x_k - \mu\}\{x_k - \mu\}^T$ where, T represents matrix transposition
3. Compute Eigen values λ_i and eigen vectors v_i of covariance matrix,
$Cv_i = \lambda_i v_i$ (i = 1,2,3, ….q), where q is the number of features
4. The estimation of high-valued eigen vectors i. All the eigen values (λ_i) in descending order
ii. Threshold value θ
iii. Number of high-valued λ_i can be selected so as to satisfy the relationship
$(\sum_{i=1}^{s}\lambda_i)(\sum_{i=1}^{a}\lambda_i)^{-1} \geq \theta$, where, s = number of high valued λ_i selected
iv. Eigen vectors corresponding to chosen high valued λ_i
5. Extract low dimensional feature vectors. $p = V^T x$, where, V is the matrix of principal component (PCs) and x is the feature matrix

Table 5
ISOmetric feature MAPping (ISOMAP) algorithm [36].

Algorithm ISOMAP
1. Squared pairwise similarities matrix $D_i D_j = \left\|y_i - y_j\right\|^2$, also known as the distance matrix measured on the temporal dimension. The indices I and j are temporal.
2. Weighted graph based on the D according to the number of neighbouring point for each node
3. Geodesic distances estimation D_G by locating the shortest path on the weighted graph (Dijkstras algorithm)
4. Define $B = -{}^1\!/_2 J D_G J^T$, where $J = I - {}^1\!/_N$, where, I is the identity matrix, and N is the total number of data points
5. Solve eigen problem $BP = P\sum$
6. Compute prinicipal vectors by using the formula $X = P\sum^{1/2}$

Table 6
Merits and demerits of PCA and ISOMAP.

Technique	Merits	Demerits
PCA (Linear)	• Linear • Dimension reduction • Popular technique and widely used • Removes correlated features • Improves up other machine learning algorithms • Helps visualization of hight dimensional data	• Requires scaling data • Lose of remaining informations • PCs may not be interpretable • May not handle non linearity in data structure
ISOMAP (Nonlinear)	• Non linear • Dimension reduction • Widely used • Non iterative • Technique is able to compute a globally optimal solution	• Graph discreteness overstimates the geodesic distance • May not handle more complex domains • Computationally expensive • Known to have difficulties with holes

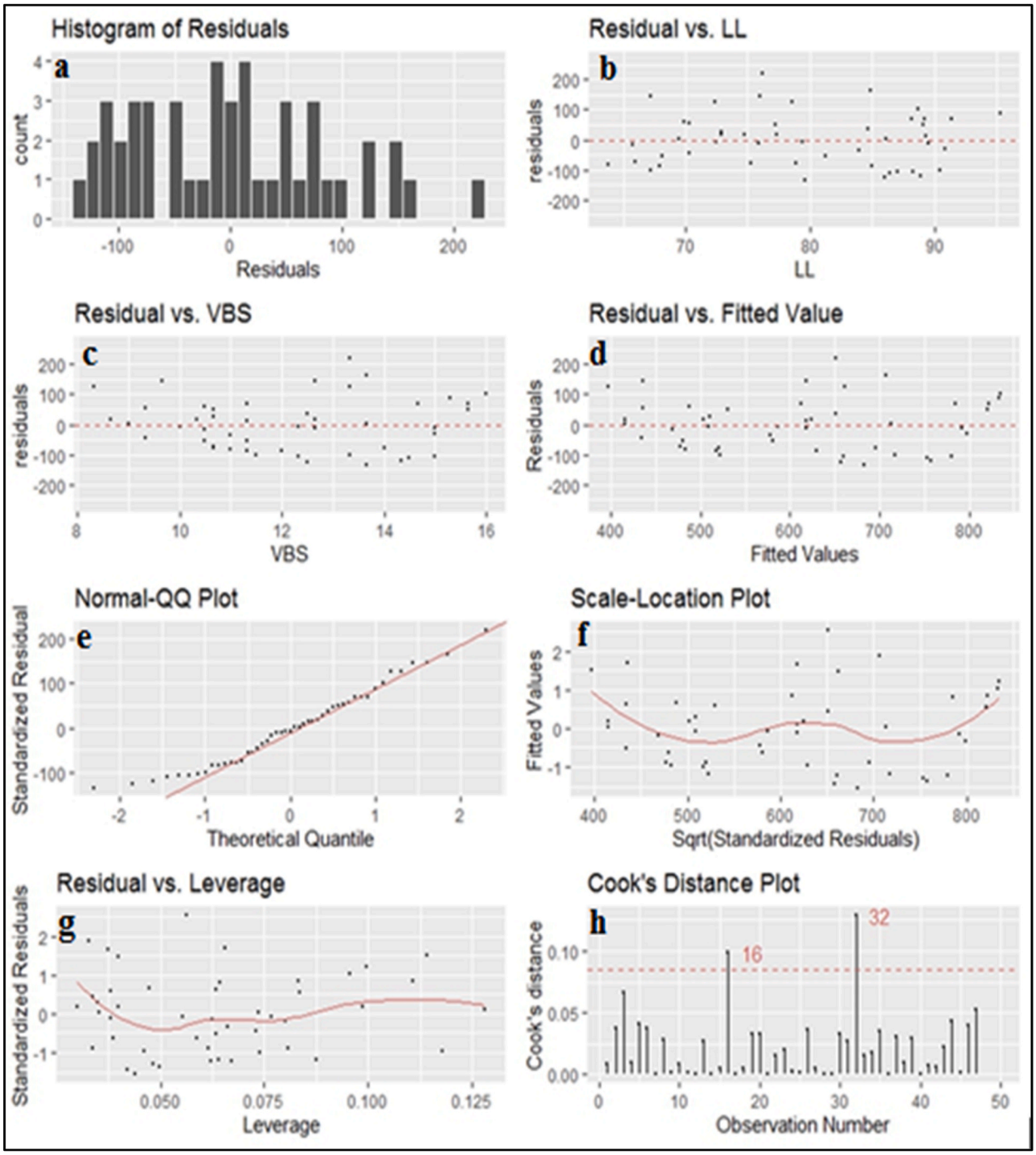


Fig. 5. Residuals diagnostics of stepwise regression, (a) histogram, (b) residual vs. LL, residual vs. VBS, residual vs. fitted value, normal QQ plot, scale-location plot, residual vs. leverage, cooks distance plot.

2.2.6. *Performance metrics*

In order to evaluate the performance of the developed models, well-known performance indices [32] were calculated. These included the mean absolute percentage error (MAPE), root mean square error (RMSE), root relative squared error (RRSE) and determination coefficient (R^2). These indices' mathematical expressions are given respectively by Eq. (2)-(5):

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{(y_i - \hat{y})}{y_i}\right| 100 \quad (2)$$

Table 7
Stepwise regression analysis results.

	Estimate	Std.Error	VIF	t-value	p-value	$SP = -325.85 + 3.91LL + 50.89VBS$
Intercept	−325.853	116.153	/	−2.805	0.00746	
LL	3.90	2.105	2.106	1.855	0.07029	
VBS	50.89	9.591	2.106	5.306	3.49e-06	
Model performance results						
	RMSE	**MAPE**	RRSE	**p-value**	**R^2**	**F-satistic**
Train	86,166	0,117	0,555	6.04e-12	0.687	49.18
Test	97,287	0,125	0,695	/	0.594	/

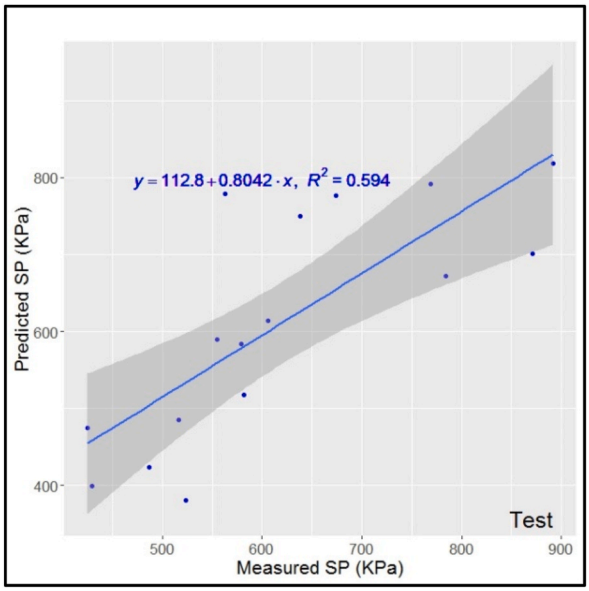


Fig. 6. Measured-Predicted SP (test) stepwise regression.

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \widehat{y})^2}{\mathrm{n}}} \tag{3}$$

$$RRSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \widehat{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}} \tag{4}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \tag{5}$$

Where $\widehat{y}$ is the mean value, y is the actual value, $\widehat{y}$ is the predicted value, and n is the number of data samples.

3. Results and discussion

This study made use of R programing language accessed through RStudio [56], and some popular R packages namely caret package [57,58], RDRToolbox [59], NeuralNetTools [60] neuralnet [61], factoextra and FactoMineR [62,63], performanceanalytics [64] and ggplot2 [65] to predict soil swelling pressure using different neural networks models with both linear and nonlinear dimension reduction techniques. According to the methodology described above different models are presented as follows.

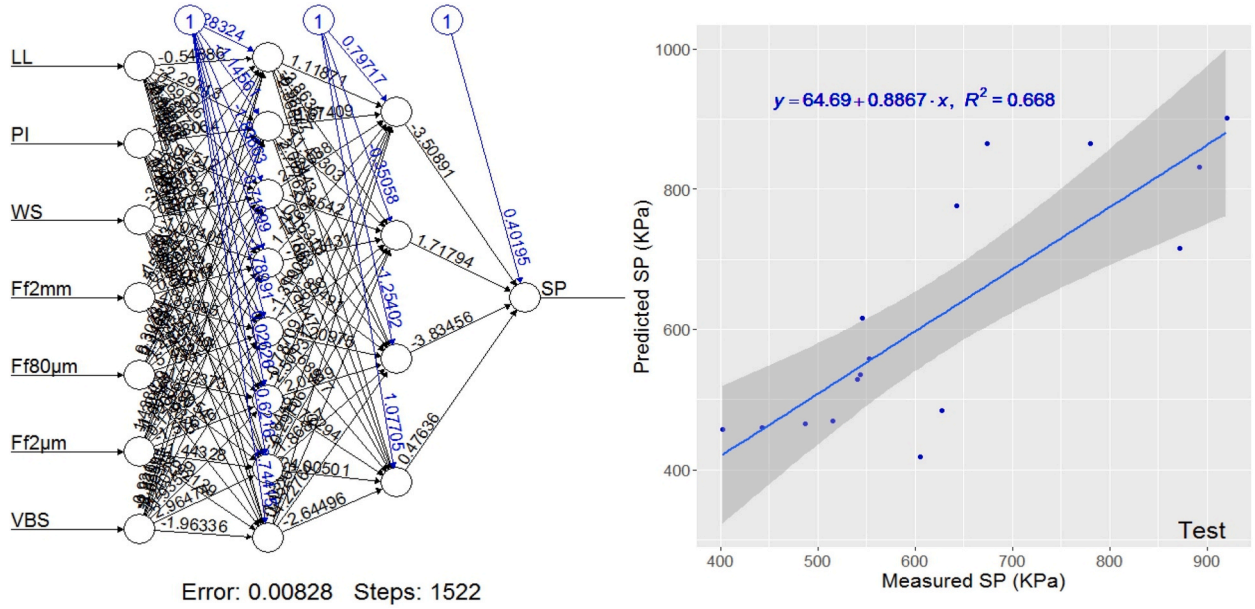


Fig. 7. Architecture of ANN model and Measured-Predicted SP (test) ANN all variables.

Table 8
Performance of all variables ANN model.

	RMSE	MAPE	RRSE	R^2
Train	2778	0,002	0,018	0.998
Test	122,902	0,143	0,834	0.668

3.1. Multiple linear regresion throught stepwise

StepAIC function [66] performs a regression model in a stepwise algorithm using both directions, regression model obtained from this algorithm is set according to Table 2 all the variables were used.

Further, Fig. 5 presents diagnostics of the residuals of the model wich shows normality and no evidence of curvilinearity or heteroscedasticity, so the linear model is accepted. Dataset is randomly splited into 70% for training and 30% for testing. Table 7 lists selected performance measures for the model. Stepwise regression yielding SP as linear model with LL and VBS is statiscaly significant, given the f-satistics and the associated p-values. However, this linear model explains only 59% of the total variation of swelling pressure for testing phase (Fig. 6).

3.2. Neural model without dimension reduction

The "neuralnet package [61] is used to build the neural model. 2/3 of the data are retained for training and the remaining 1/3 for the testing phase [57].

In order to balance between underfitting and over fitting (bias-variance tradeoff) with trial and error approach, the adopted model in our study is a back propagation neural network composed of two hidden layers with eight nodes in the firt layer and four nodes in the second wih a logistic activation function for the two first layers and a linear function for the output layer (SP). The present model has showed a slight improvement in the explained proportion of variation of the predicted swelling pressure compared to the stepwise regression model (Fig. 7), performance metrics of the model are given in Table 8.

3.3. Feature extraction using linear and nonlinear dimension reduction technique

In this section both PCA and ISOMAP algorithms are applied to the dimension reduction process of FE for neural model. Dataset is normalized before applying the PCA and ISOMAP feature extraction approachs.

3.3.1. feature extraction

Kaiser Meyer Olkin (KMO) and Bartlett's test were performed using EFAtools [67]. KMO test assesses sampling adequacy and data suitability for factor analysis, with acceptable values greater than 0.50. Bartlett's sphericity test tests correlations among factors. Based

Table 9
Kaiser-meyer-olkin, Bartlett's test.

Kaiser-Meyer-Olkin (KMO)							
Variables	LL	PI	WS	Ff2mm	Ff80 μm	Ff2μm	VBS
KMO	0.853	0.808	0.911	0.779	0.823	0.893	0.893
Bartlett's test of sphericity							
Approx chi Square: 363			df: 28			Significant $p < 0.001$	

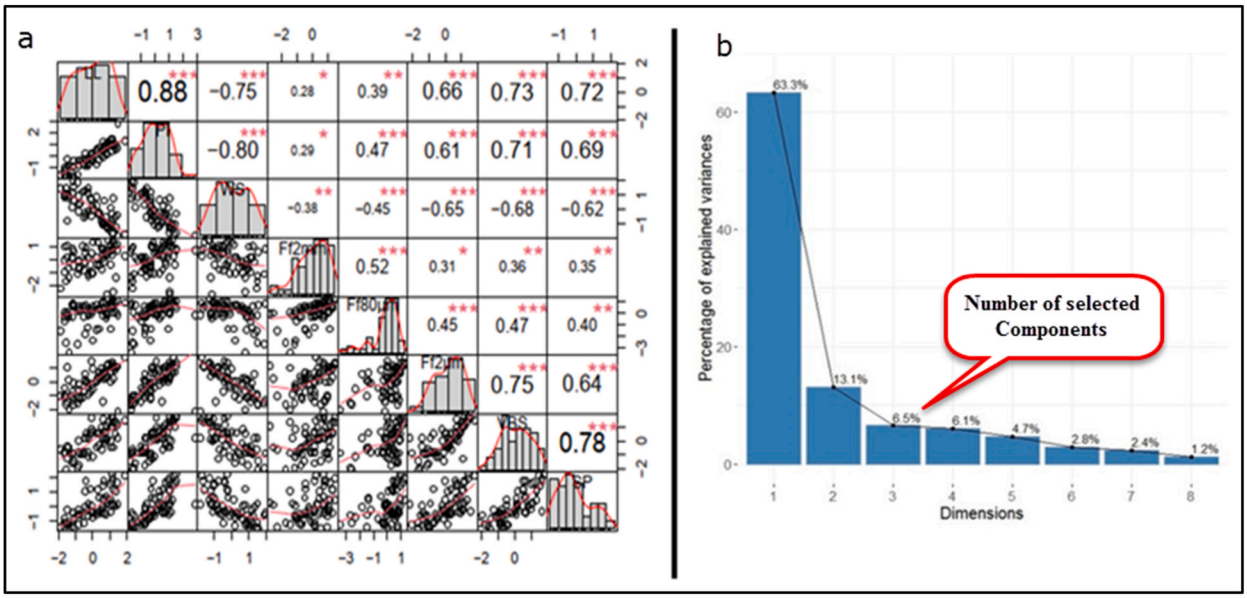


Fig. 8. Correlation matrix(a) and decomposition of the total inertia (b).

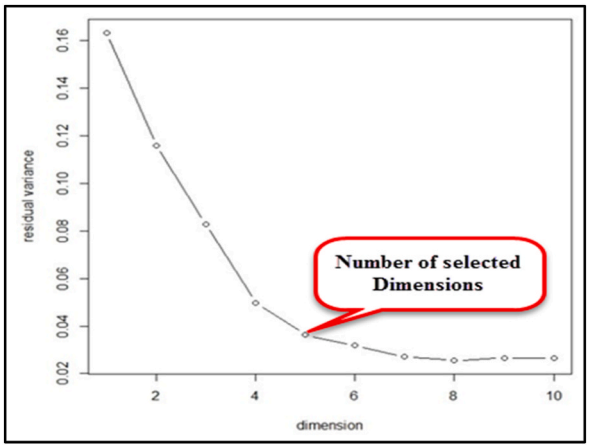


Fig. 9. Residual variance for the 10 first ISOMAP dimension.

on Table 9, factor analysis can be processed and there is significant correlation between all variables.

Correlation matrix (Fig. 8(a)) of geotechnical parameters provides a summary of Spearman R and its significance and correspandant scatter plots. Results indicate a strong positive relationship between SP and VBS, LL, Ff2μm and IP. Moderate positive correlation with Ff2mm and Ff80 μm. Also, SP has a strong negative correlation with WS. In addition, regression coefficients between all the geotechnical parameters is statistically significant. The first three dimensions explain 82.90% of the total dataset inertia then number of components to keep regarding elbow method is three (Fig. 8(b)).

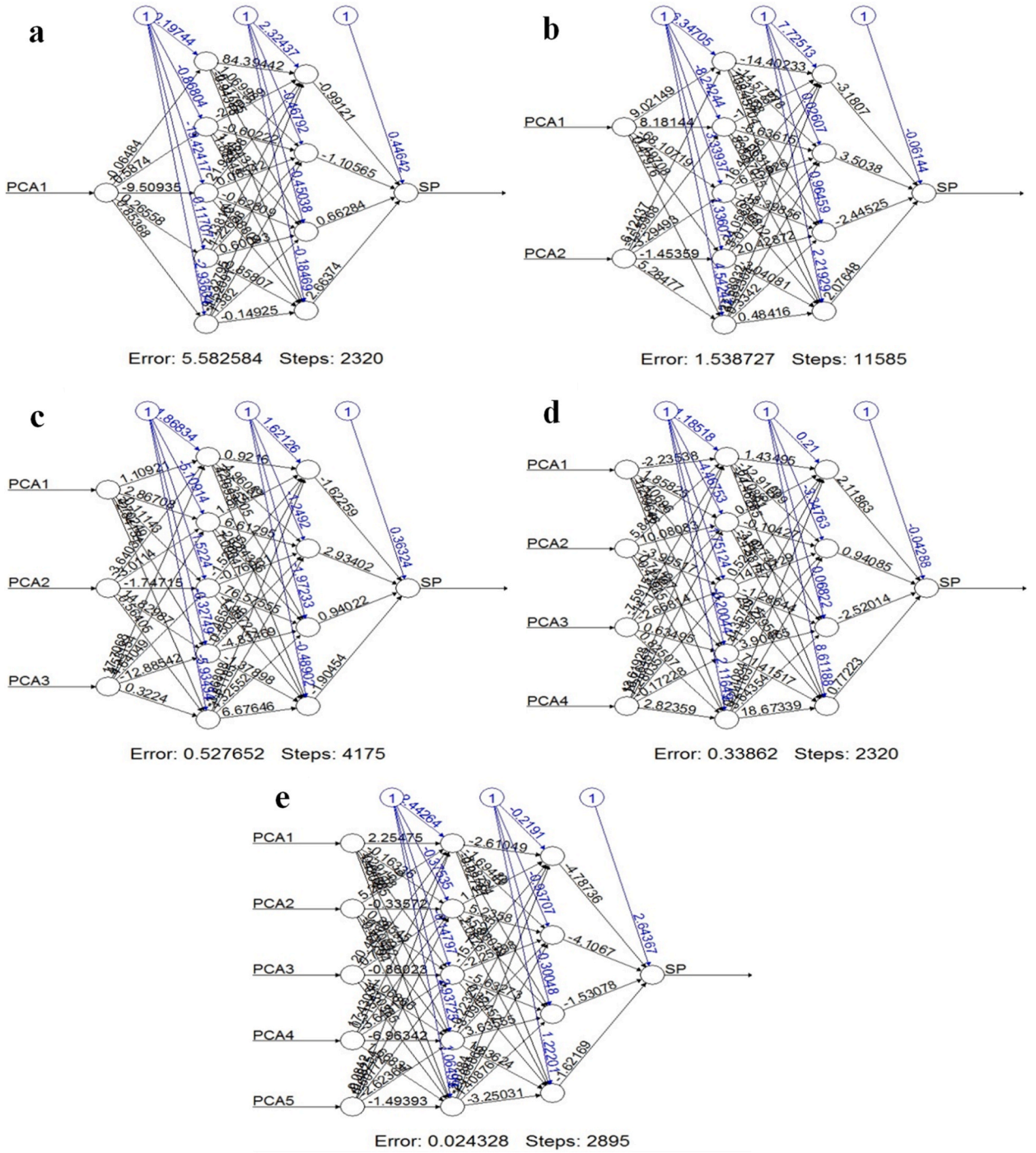


Fig. 10. Architecture of PCA-ANN models inputs: (a) first PC, (b) first two PCs, (c) first three PCs, (d) first four PCs, (e) first five PCs.

3.3.2. ISOMAP feature extraction

ISOMAP calculates residual variance by comparing geodesic distances, computed in G represented by matrix D_G (G is a neighborhood graph), to the pairwise distances of the mapped data Y, represented by matrix D_Y [68], $R = 1 - \rho(D_G, D_Y)^2$, Here, ρ is standard linear correlation coefficient, taken over all entries of D_G and D_Y. Fig. 9 shows the residual variance of ISOMAP with dimensions of 1–10 from which the dimension of ISOMAP space can be selected as five. It can be seen that the residual variance for ISOMAP with one dimension is much lower than for PCA.

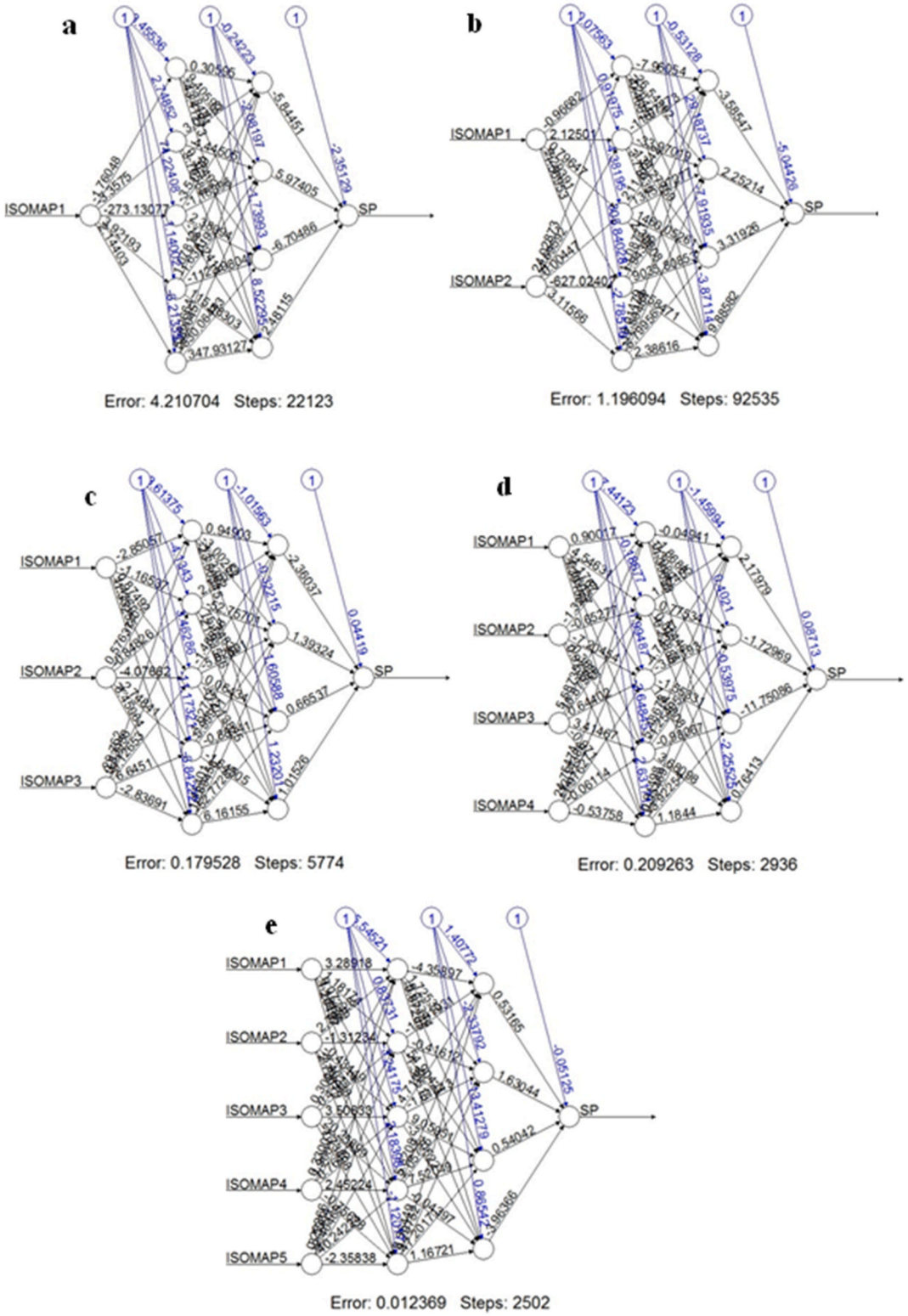


Fig. 11. Architecture of ISOMAP-ANN models inputs: (a) first Dim, (b) first two Dims, (c) first three Dims, (d) first four Dims, (e) first five Dims.

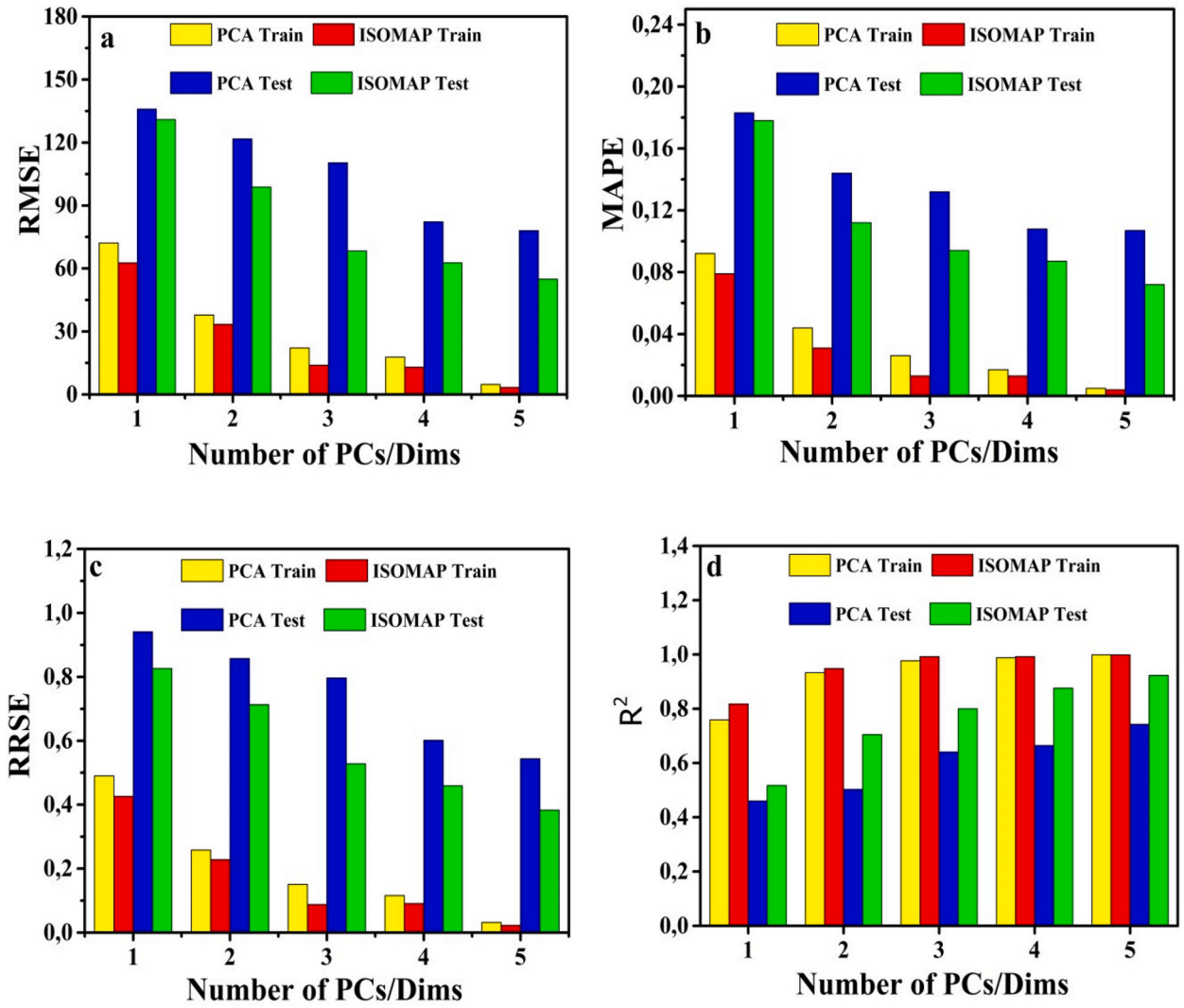


Fig. 12. Performance of PCA-ANN and ISOMAP-ANN models, (a) RMSE, (b) MAPE, (c) RRSE, (d) R^2.

3.4. Neural model with feature extraction (PCA/ISOMAP)

In the implementation of PCA and ISOMAP into the neural model, the search for the optimal number of PCs (number of inputs) is a significant challenge. The most common method is to set the threshold of explained variance such as 80%. Fig. 8(b) shows that three components are retained. The following approach is chosen in the present study. At each stage, the PCs are increased by one. Similarely, ISOMAP dimensions are increased by one reaching five dimensions (Fig. 9), to make results comparable for the cases being considered in the present study, the same back propagation neural network composed of two hidden layers with five nodes in the first layer and four nodes in the second wih a logistic activation function for the two first layers and a linear function for the output layer (SP) (Figs. 10 and 11).

Fig. 12 shows the performance of each model for specified number of PCs and Dims respectively inputs to the neural model during training and testing phases. Using four statistical performance metrics: MAPE, RRSE, RMSE and R^2.

Using PCA feature extraction dimensions reduction for the neural model and in the training phase and by increasing the number of CPs from one to five, one at a time, we can observe that RMSE, MAPE and RRSE decrease: (72.121–4.760), (0.092–0.005), (0.490–0.032) respectively and R^2 increases (0.759–0.999). Similarly for ISOMAP and by increasing the number of Dims one at a time, RMSE, MAPE and RRSE decrease from (62.636–3.395), (0.079–0.004), (0.426–0.023) respectively and R^2 increases (0.818–0.999). Following, the same procedure in the testing phase, we note for PCA observe that RMSE, MAPE and RRSE decrease from (135.936–78.055), (0.183–0.107), (0.941–0.544) respectively and R^2 increases (0.460–0.743). Also for ISOMAP, RMSE, MAPE and RRSE decrease from (130.94–54.937), (0.178–0.072), (0.826–0.383) respectively and R^2 increases (0.517–0.923).

It is clearly showed that by increasing number of PCs/Dims model performance increases. It is also proved that feature extraction

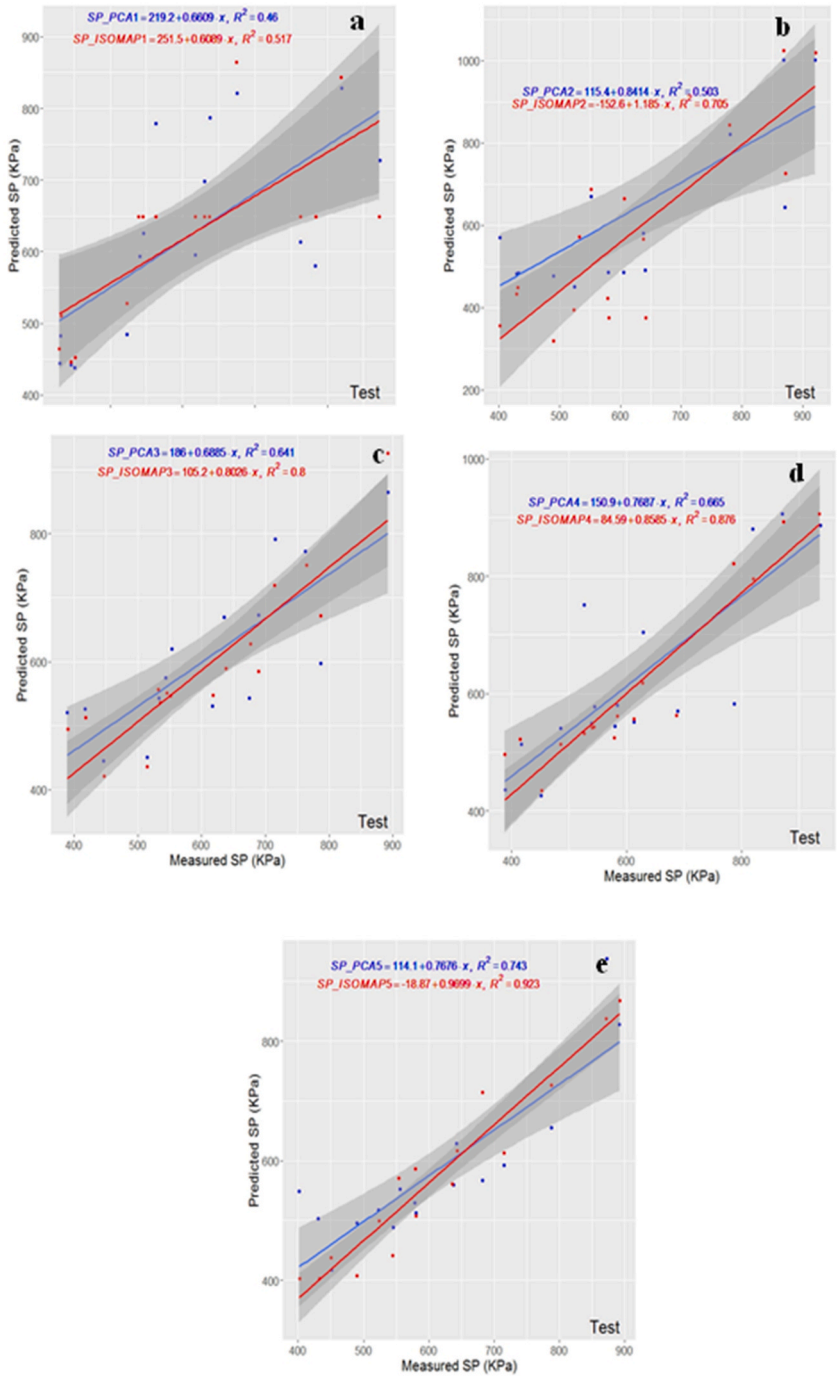


Fig. 13. Measured-Predicted SP (test) PCA-ANN and ISOMAP-ANN models: (a) first PC/Dim, (b) first two PCs/Dims, (c) first three PCs/Dims, (d) first four PCs/Dims, (e) first five PCs/Dims.

either by linear or nonlinear techniques improves efficiency and accuracy of the neural model compared to all variables as inputs in the neural model and also to the multiple linear regression model (Stepwise regression). Furthermore, best performance is obtained by nonlinear (ISOMAP) feature extraction dimension reduction. At three selected CPs/Dims to the neural models highest determination coefficient is that of ISOMAP based neural model $R^2 = 0.80$. Although the number of CPs/Dims has been increased to 5 the coefficient by ISOMAP model performance still higher than that of the PCA one (Fig. 13).

4. Conclusion

In this paper, PCA and ISOMAP methods are applied to extract low-dimensional features. These extracted features are then used as inputs to neural models to predict soil swelling pressure. The results show that the ANN model with feature extraction using PCA and ISOMAP can achieve better performance than that without feature extraction (all variables as inputs and multiple regression models). This can be explained by the fact that dimension reduction using feature extraction allows non-linear problems to be addressed while improving model performance. Determination coefficient increases by 28.3 and 40.6% from one to five extracted features for PCA and ISOMAP-based neural model respectively. In general, it is shown that ISOMAP-based feature extraction algorithm provides more informative feature extraction than the linear PCA algorithm. Indeed, for the five CPs/Dims inputs selected in this study, the ISOMAP-based neural model remains the best performing model in terms of metrics compared to the PCA-based model. Determination coefficient increases by 18% for five CPs/Dims selected. Given the complex, heterogeneous structure of unsaturated soils, ISOMAP-based feature extraction method provides a relevant and powerful technique for predicting the various parameters of this type of soil, especially as this method takes into account the often noisy and sparse nature of laboratory data. It is therefore recommended that the present ISOMAP-ANN model be adopted to predict SP. The study opens up new possibilities for future research into unsaturated soils, particularly the swelling behavior of expansive soils, as this is a complex phenomenon. However, the interpretability of black box models remains a challenge in the geotechnical field.

Author contribution statement

Ouassila Bahloul: Conceived and designed the experiments; Analyzed and interpreted the data; Wrote the paper.
Tebbi Fatima Zohra: Contributed reagents, materials, analysis tools or data; Wrote the paper.
Lekouara Laid: Analyzed and interpreted the data.
Bekhouche Hizia: Performed the experiments.

Data availability statement

The data that has been used is confidential.

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

[1] Y. Erzin, N. Güneş, The prediction of swell percent and swell pressure by using neural networks, Math. Comput. Appl. 16 (2) (2011) 425–436.
[2] D. Dam Nguyen, et al., Bagging and multilayer perceptron hybrid intelligence models predicting the swelling potential of soil, Transp. Geotech. 36 (2022), 100797.
[3] S. Azam, Large-scale Odometer for Assessing Swelling and Consolidation Behavior of Al-Qatif Clay, 2006.
[4] Z.A. Erguler, R. Ulusay, A simple test and predictive models for assessing swell potential of Ankara (Turkey) Clay, Eng. Geol. 67 (3) (2003) 331–352.
[5] A.J. Puppala, A. Pedarla, Innovative ground improvement techniques for expansive soils, Innovat. Infrastruct. Solut. 2 (1) (2017) 24.
[6] Y. Berrah, et al., Application of dimensional analysis and regression tools to estimate swell pressure of expansive soil in Tebessa (Algeria), Bull. Eng. Geol. Environ. 77 (2016).
[7] M. Khemissa, A. Mahamedi, Cement and lime mixture stabilization of an expansive overconsolidated clay, Appl. Clay Sci. 95 (2014) 104–110.
[8] J. Du, et al., Prediction of swelling pressure of expansive soil using an improved molecular dynamics approach combining diffuse double layer theory, Appl. Clay Sci. 203 (2021), 105998.
[9] B. Devkota, et al., Accounting for expansive soil movement in geotechnical design—a state-of-the-art review, Sustainability 14 (23) (2022) 15662.
[10] Ö. Çimen, S.N. Keskin, H. Yıldırım, Prediction of swelling potential and pressure in compacted clay, Arabian J. Sci. Eng. 37 (6) (2012) 1535–1546.
[11] F.E. Jalal, et al., Indirect estimation of swelling pressure of expansive soil: GEP versus MEP modelling, Adv. Mater. Sci. Eng. 2023 (2023), 1827117.
[12] S. Vanapalli, L. Lu, A state-of-the art review of 1-D heave prediction methods for expansive soils, Int. J. Geotech. Eng. 6 (2012) 15–41.
[13] W. Diana, A. Muntohar, A. Widianti, Review of oedometer method for predicting heave on the expansive soil, International Journal of Sustainable Construct. Eng. Technol. 11 (2020).
[14] N. Honne, M. Munnas, A. Sridharan, Critical evaluation of determining swelling pressure by swell-load method and constant volume method, J. ASTM Int. (JAI) 32 (2009) 1–10.
[15] T. Berre, Effect of sample disturbance on triaxial and oedometer behaviour of a stiff and heavily overconsolidated clay, Can. Geotech. J. 51 (8) (2014) 896–910.
[16] H. Tu, S. Vanapalli, Prediction of the variation of swelling pressure and 1-D heave of expansive soils with respect to suction using the soil water retention curve as a tool, Can. Geotech. J. 53 (2016).
[17] D. Fredlund, Prediction of Ground Movements in Swelling Clays, 1983.
[18] M. Amin Benbouras, A.-I. Petrisor, Prediction of swelling index using advanced machine learning techniques for cohesive soils, Appl. Sci. 11 (2) (2021) 536.
[19] A. Zhafirah, A.K. Somantri, S. Permana, Analysis of oedometer and rowe cell consolidation compared to experimental testing, J. Phys. Conf. 1402 (2) (2019), 022006.
[20] M. Shahin, Artificial intelligence in geotechnical engineering: applications, modeling aspects, and future directions. Metaheuristics in water, Geotech. Transp. Eng. (2013) 169–204.
[21] A. Albalasmeh, et al., Artificial neural network optimization to predict saturated hydraulic conductivity in arid and semi-arid regions, Catena 217 (2022), 106459.
[22] A.H. Abdulkareem, D.O. Aziz, Multiple regression and ANN (MLP) model for predicting swelling index of ramadi cohesive soil, IOP Conf. Ser. Mater. Sci. Eng. 737 (1) (2020), 012116.
[23] N. Dounane, H. Trouzine, Integrated approach of ANN and Taguchi algorithms for soils swelling estimation, Int. J. Sustain. Build. Technol. Urban Dev. 11 (4) (2020) 244–257.

[24] J.C. Egbueri, et al., Development of MLR and variedly optimized ANN models for forecasting the detachability and liquefaction potential index of erodible soils, Geosyst. Geoenviron. 2 (1) (2023), 100104.
[25] F.Z. Merouane, S.M.A. Mamoune, Prediction of swelling parameters of two clayey soils from Algeria using artificial neural networks, Model. Civil Environ. Eng. 14 (3) (2018) 11–26.
[26] Y. Mohamedzein, R. Ibrahim, A. Alsanosi, Prediction of Swelling Pressure of Expansive Soils Using Neural Networks, 2006.
[27] W.A. Bodour, et al., Development of evaluation framework for the unconfined compressive strength of soils based on the fundamental soil parameters using gene expression programming and deep learning methods, J. Mater. Civ. Eng. 34 (2) (2022), 04021452.
[28] W. Jia, et al., Feature dimensionality reduction: a review, Compl. Intelli. Syst. 8 (3) (2022) 2663–2693.
[29] L. Wang, S. Jiang, S. Jiang, A feature selection method via analysis of relevance, redundancy, and interaction, Expert Syst. Appl. 183 (2021), 115365.
[30] M.D. Nguyen, et al., Soft-computing techniques for prediction of soils consolidation coefficient, Catena 195 (2020), 104802.
[31] J.J. Doris, D.M. Rizzo, M.M. Dewoolkar, Forecasting vertical ground surface movement from shrinking/swelling soils with artificial neural networks, Int. J. Numer. Anal. Methods GeoMech. 32 (10) (2008) 1229–1245.
[32] K.E. Taylor, Summarizing multiple aspects of model performance in a single diagram, J. Geophys. Res. 106 (2001) 7183–7192.
[33] L. Wu, L. Noels, Recurrent Neural Networks (RNNs) with dimensionality reduction and break down in computational mechanics; application to multi-scale localization step, Comput. Methods Appl. Mech. Eng. 390 (2022), 114476.
[34] Y.M. Chen, et al., Combination of the manifold dimensionality reduction methods with least squares support vector machines for classifying the species of sorghum seeds, Sci. Rep. 6 (1) (2016), 19917.
[35] M. Kamran, et al., Application of KNN-based isometric mapping and fuzzy c-means algorithm to predict short-term rockburst risk in deep underground projects, Front. Public Health 10 (2022).
[36] J.C.-H. Tseng, An ISOMAP analysis of sea surface temperature for the classification and detection of el niño & La niña events, Atmosphere 13 (6) (2022) 919.
[37] C.M.O. Nwaiwu, I. Nuhu, Evaluation and prediction of the swelling characteristics of Nigerian black clays, Geotech. Geol. Eng. 24 (1) (2006) 45–56.
[38] Y. Erzin, O. Erol, Swell pressure prediction by suction methods, Eng. Geol. 92 (3) (2007) 133–145.
[39] S.B. Ikizler, et al., Prediction of swelling pressures of expansive soils using artificial neural networks, Adv. Eng. Software 41 (4) (2010) 647–655.
[40] S.K. Das, et al., Prediction of swelling pressure of soil using artificial intelligence techniques, Environ. Earth Sci. 61 (2) (2010) 393–403.
[41] S.B. Ikizler, et al., Prediction of swelling pressures of expansive soils using soft computing methods, Neural Comput. Appl. 24 (2) (2014) 473–485.
[42] N. Dounane, H. Trouzine, M. Hamhami, Application de la méthode de Taguchi pour l'étude des sols gonflants, in: Rencontres Universitaires de Génie Civil, Bayonne, France, 2015.
[43] Y. Berrah, A. Boumezbeur, N. Charef, Regression tools to quantify the swelling pressure of expansive soil in tebessa region (Algeria), in: Recent Advances in Geo-Environmental Engineering, Geomechanics and Geotechnics, and Geohazards, Springer International Publishing, Cham, 2019.
[44] Y. Berrah, et al., Design of experiments (DOE) techniques to predict swelling pressure of expansive soils in tebessa (Algeria), in: CIGOS 2019, Innovation for Sustainable Infrastructure, Springer Singapore, Singapore, 2020.
[45] A. Djellali, B. Saghafi, M.S. Laouar, Experimental correlations for the swelling pressure of expansive clays in the city of tebessa, Algeria, in: CIGOS 2019, Innovation for Sustainable Infrastructure, Springer Singapore, Singapore, 2020.
[46] K. Meshram, P.K. Jain, Estimation of swelling pressure of expansive soils reinforced with granular pile, Stavební obzor - Civil Eng. J. 29 (2020) 315–323.
[47] O.A. Montesinos López, A. Montesinos López, J. Crossa, Fundamentals of artificial neural networks and deep learning, in: O.A. Montesinos López, A. Montesinos López, J. Crossa (Eds.), Multivariate Statistical Machine Learning Methods for Genomic Prediction, Springer International Publishing, Cham, 2022, pp. 379–425.
[48] A. Apicella, F. Isgrò, R. Prevete, A simple and efficient architecture for trainable activation functions, Neurocomputing 370 (2019) 1–15.
[49] M. Efroymson, Multiple regression analysis, in: A. w Ralston, H.S. Wilf (Eds.), Mathematical Models for Digital Computers, J. Wiley, New York, 1960.
[50] G. James, et al., An Introduction to Statistical Learning: with Applications in R, Springer, New York, 2013.
[51] I.T. Jolliffe, Principal components used with other multivariate techniques, in: I.T. Jolliffe (Ed.), Principal Component Analysis, Springer New York, New York, NY, 1986, pp. 156–172.
[52] N. Thiamchoo, P. Phukpattaranont, Evaluation of feature projection techniques in object grasp classification using electromyogram signals from different limb positions, PeerJ Comput. Sci. 8 (2022) e949.
[53] V. Gs, et al., Radial basis function neural network based comparison of dimensionality reduction techniques for effective bearing diagnostics, Proc. IME J. J. Eng. Tribol. 227 (6) (2013) 640–653.
[54] J.B. Tenenbaum, V.d. Silva, J.C. Langford, A global geometric framework for nonlinear dimensionality reduction, Science 290 (5500) (2000) 2319–2323.
[55] W.S. Torgerson, Multidimensional scaling: I. Theory and method, Psychometrika 17 (4) (1952) 401–419.
[56] R. Team, RStudio: Integrated, vol. 700, development for R. RStudio. Inc., Boston, MA, 2015, p. 879.
[57] K. Max, Building predictive models in R using the caret package, J. Stat. Software 28 (2008).
[58] M. Kuhn, Building predictive models in R using the caret package, J. Stat. Software 28 (5) (2008) 1–26.
[59] C. Bartenhagen, Package 'RDRToolbox', 2014.
[60] M.W. Beck, NeuralNetTools: visualization and analysis tools for neural networks, J. Stat. Software 85 (11) (2018) 1–20.
[61] F. Günther, S. Fritsch, Neuralnet: training of neural networks, R Journal 2 (2010).
[62] I. Irnawati, et al., The use of software packages of R factoextra and FactoMineR and their application in principal component analysis for authentication of oils, Indones. J. Chemom. Pharmaceut. Anal. (2021) 1–10.
[63] S. Lê, J. Josse, F. Husson, FactoMineR: an R package for multivariate analysis, J. Stat. Software 25 (1) (2008) 1–18.
[64] B.G. Peterson, et al., Package 'performanceanalytics, R Team Cooperation 3 (2018) 13–14.
[65] H. Wickham, ggplot2, 2016.
[66] B. Ripley, et al., Package 'mass', vol. 538, Cran r, 2013, pp. 113–120.
[67] M. Steiner, S. Grieder, EFAtools: an R package with fast and flexible implementations of exploratory factor analysis tools, J. Open Source Softw. 5 (2020) 2521.
[68] F. Schoeneman, et al., Learning manifolds from dynamic process data, Algorithms 13 (2) (2020) 30.