Williams, CG and Ojuri, OO

 Predictive modelling of soils' hydraulic conductivity using artificial neural network and multiple linear regression

http://researchonline.ljmu.ac.uk/id/eprint/15309/

Article

Research Article

Predictive modelling of soils' hydraulic conductivity using artificial neural network and multiple linear regression

Charles Gbenga Williams[1] · **Oluwapelumi O. Ojuri**[1]

Abstract

As a result of heterogeneity nature of soils and variation in its hydraulic conductivity over several orders of magnitude for various soil types from fine-grained to coarse-grained soils, predictive methods to estimate hydraulic conductivity of soils from properties considered more easily obtainable have now been given an appropriate consideration. This study evaluates the performance of artificial neural network (ANN) being one of the popular computational intelligence techniques in predicting hydraulic conductivity of wide range of soil types and compared with the traditional multiple linear regression (MLR). ANN and MLR models were developed using six input variables. Results revealed that only three input variables were statistically significant in MLR model development. Performance evaluations of the developed models using determination coefficient and mean square error show that the prediction capability of ANN is far better than MLR. In addition, comparative study with available existing models shows that the developed ANN and MLR in this study performed relatively better.

Keywords Hydraulic conductivity · Artificial neural network · Multiple linear regression · Predictive model

1 Introduction

Solutions to many geotechnical and geo-environmental engineering issues require extensive understanding of soil hydraulic conductivity. According to Murthy [1], hydraulic conductivity is a measure of the ease with which water flows through permeable materials. It is a measured indicator of the soil's ability to convey water when exposed to a hydraulic gradient. This soil parameter plays a key role in solving problems relating to leachate transportation in landfill design, earth dam design as it dictates, among other important parameters, the selection of suitable material for the liner system and the core material for earth dam, respectively [2–4]. Like many geotechnical parameters, hydraulic conductivity is simple in concept, but has some very complex aspects in practice, especially when trying to obtain realistic measurements. Approaches taken to estimate hydraulic conductivity of soils include laboratory and field methods of measurement and calculation from empirical formulae. Laboratory and field measuring methods include constant head test, falling head test, flexible wall permeameter test, rigid wall permeameter test, ring infiltrometer, instant profile method, test basins [5, 6], etc. Meanwhile, the different empirical equations currently in use correlate the hydraulic conductivity of fine-grained soils with the index properties which in accordance with Freeze and Cherry [7], hydraulic conductivity is established to associate with the grain-size distribution of porous granular media. The major advantage of the available empirical methods [8, 9] is that the hydraulic conductivity value is rapidly estimated than the direct measurement. The application of these relationships, however, may be incorrect and may lead to random errors [5]. Field test methods have the advantages that the soil profile is

often undisturbed but cannot control the soil environment unlike the laboratory test method [10]. Although laboratory methods are relatively easy, the common inconvenience is that they take time. Consequently, getting a quick result to a problem in the field within the given time frame can be challenging. As a result, appropriate consideration has now been given to determining the hydraulic conductivity of soils using predictive techniques [3].

Over the last few decades, computational intelligence (CI) methods otherwise known as soft computing have been applied to various fields of science and engineering. Complex tasks such as learning, modelling or getting a pattern from experimental approach can be handled with CI methods with accurate precision [11]. CI technique can either be single (such as artificial neural networks (ANNs), fuzzy, support vector machine (SMV), particle swarm optimisation (PSO), genetic algorithm (GA)) or hybrid (e.g. adaptive neuro-fuzzy inference system (ANFIS), GA-ANN, etc.). Single CI technique is employed mainly for predicting, modelling or exploring data [11, 12]. ANN as a subset of CI has been the most popular and significant tool in various engineering fields [13–15]. Its high ability to predict nonlinear behaviours has unfolded its uniqueness to many researchers [16]. ANN is inspired by how the human brain works. The human brain consists of a large number of highly networked neurons working together to solve a specific problem. Like human brain that possessed tremendous ability to process huge amount of information using data sent by human senses, ANNs too learn by training [17–19]. Neural network's basic processing elements are called nodes, and the weighted connections perform the same work as synapses in biological systems. Nodes are simple elements of information processing, while the connection weights modulate the effect of the related input signals and a transfer function represents the nonlinear characteristic displayed by the neurons. The output of a neuron is then computed as the weighted sum of the inputs plus the bias activated by the transfer function [17, 20]. By contrast, traditional linear regression is one of the oldest statistical methods that still maintains its relevance in the academic world, especially as a benchmark for measuring the performance of currently developed predictive tools. Multiple linear regression (MLR) examines the relationship between a response variable and the collection of independent variables. It is a generalisation of the linear regression model [21]. The assumption in regression modelling is that the output can be explained by a linear combination of input values.

In recent years, the use of ANNs has increased in several areas of civil engineering profession. Its application to many geotechnical and geo-environmental engineering problems has shown a commendable degree of success. ANN can be trained with experimental data; as a result, it is

esteemed superior among popular modelling tools [12]. A series of studies show that ANNs have been used successfully in the prediction of pile capacity, soil behaviour modelling, soil retention structures, settlement of structures, stability of slope, tunnel design and underground openings, liquefaction, soil compaction, soil swelling, soil chemical properties such as cation exchange capacity (CEC) and classification of soils [22–29], in addition, ANN and MLR as tools for prediction of geotechnical properties [24–26], prediction of cation exchange capacity by ANN and MLR [30, 31], prediction of tropical soil's hydraulic conductivity using eight different algorithms [10], and prediction of hydraulic conductivity of clays [3, 18, 32]. Minasny et al. [33] used the neural network tool to predict unsaturated hydraulic conductivity of alluvial soils.

The prediction or estimation of the hydraulic conductivity of soils using ANN and MLR by many researchers was based on a specific soil. To the best of our knowledge, there is no or limited recommendations in the literature with regard to ANN and MLR application for predicting hydraulic conductivity of all soil types. Therefore, the specific aim of this study is to develop models for the prediction of saturated hydraulic conductivity of soils (fine grain and coarse grain) through a comparative study using artificial neural network and multiple linear regression analysis. These models were developed by increasing the spectrum of test soils used by Sinha and Wang [20] by adding more results from various reliable experimental studies on hydraulic conductivity of naturally occurring soil types published in the literature using different input variables and training parameters for an optimised result. The input data used are: percentages of sand (S), fines (Fi), clay (C) of the soil samples, plasticity index (PI), and the compaction characteristics. Comparative studies were done with the selected existing MLR models and networks to evaluate the reliability of the developed models.

2 Methodology

2.1 Data collation and analysis

The reliability of the data set used is the most significant phase that can influence the ANN modelling, particularly in geotechnical and geo-environmental engineering. In addition, the efficiency of ANN relies on the data width selected. For more complicated issues, more examples are needed that show all the distinct features of the problem [17, 27]. In this study, data were collated from several experimental studies on hydraulic conductivity of different soil types published worldwide in the literature. Factors affecting hydraulic conductivity of soils include soil density, moulding water content, degree of saturation, void

ratio, soil composition, soil structure, permeant properties and others. Most of these factors are not really independent but interrelated complexly with each other, for example, grain size and void ratio, etc**.** [20, 34]. The smaller the grain size, the smaller the voids which leads to the reduced size of flow channels [1]. Hence, low hydraulic conductivity is likely to be achieved when the soil is well graded and the clay fraction governs the hydraulic behaviour of the soil matrix. As stated by Lambe [35], five factors had the greatest influence on hydraulic conductivity: soil composition; soil structure; permeant characteristics; void ratio and degree of saturation. Considering the aforementioned factors, the experimental studies selected provide data on the particle size distribution, namely percentages of sand (*S*), fines (*Fi*), clay (*C*) of the soil sample, plasticity index (*PI*), and the compaction properties (optimum moisture content, OMC and the maximum dry density, MDD) and the corresponding hydraulic conductivity, *k* (i.e. seven variables). These input variables are factors considered easily obtainable that influence the hydraulic conductivity of most soils. The data set was divided into two parts: 75% for training and 25% for testing. To avoid overfitting, the training set was chosen so that each soil class was properly represented and samples in each class contain a wide range of variations. This data set was analysed using R Software to develop MLR model and network for hydraulic conductivity prediction.

2.2 R software

R is a software language for carrying out simple and complicated statistical analyses. R is free software and comes with totally no guarantee. R was originally written by Robert Gentleman and Ross Ihaka from the Statistics Department of the University of Auckland in New Zealand. It is a collaborative effort with many contributors, since mid-1997 there has been a key team with written access to the R source [36]. R has a number of benefits for scholars; it is open source. Additionally, using R implies having access to a global group of individuals who are continuously creating new R packages and fresh teaching resources. Diverse packages for all machine learning techniques, especially neural network (e.g. nnet, NeuralNet, etc.) are available on R. Its language is easier to learn compared to other proprietary software and offers rich and better options for statistics. Different user-friendly interfaces to execute R commands are available (e.g. RStudio) that are free and simple to install. Although MATLAB provides some good options to create, train, validate and test neural networks, it seems that there are not too many options for Windows and also requires license. For this research, RStudio version 1.1.463 was used along with R version 3.5.2. R software is available for Mac, Windows and Linux operating systems

and can be obtained via www.r-project.org, and RStudio is accessible at www.rstudio.com.

2.3 Multiple linear regression model (MLR)

Regression modelling aimed at using numbers of independent measurements to determine a mathematical function that describes the relationship between the input parameters and the output. In engineering and science, many problems revolve round the relationship between two or more variables. MLR is a linear regression technique that is very beneficial for predicting the best relationship between a response variable and several independent variables unlike the simple linear regression analysis [31, 37]. One of the assumptions in multiple linear regression is non-existence of collinear relation between independent variables. Variance inflation factor (VIF) is an index that is used for collinear determination. If there is no linear relationship between independent variables, VIF value will be one and the deviation of this factor from 1 reveals the tendency to collinearity. Having VIF values more than 10 for each variable show the multiple collinearity and it may result in estimation problems [38].

Multiple linear regression was developed using 75% of the training data set and the remaining 25% to evaluate the efficiency of the developed model. The MLR model for hydraulic conductivity prediction was executed using the 'lm' function in R Software. Equation (1) shows the general form of the MLR equation:

$$Y = a + b_1X_1 + b_2X_2 + \dots + b_iX_i \tag{1}$$

where *Y* is the response variable representing hydraulic conductivity *k*,

a is the intercept,

$b_1 \dots b_i$ are regression coefficients, and

$X_1 \dots X_i$ are independent variables referring to basic soil properties (i.e. the input data).

At first, all the input variables were utilised in developing the MLR model and subsequently the variables that were significantly less effective on the hydraulic conductivity (output parameter) were eliminated.

2.4 Artificial neural network architecture

Multilayer perceptron (MLP) and radial basis function (RBF) are two of the most widely used neural network architecture in the literature for classification or regression problems. General difference between MLP and RBF is that RBF is a localist type of learning which is responsive only to a limited section of input space [39]. However, MLP, being the most predominant network architecture and due to the simplicity of its design, was utilised for this study [18,

31, 40]. Figure 1 demonstrates a typical two-layer perceptron to simulate an input–output reaction. The network which consists of nodes is structured into input, hidden and output nodal layers. The input layer is not regarded a neuron layer as it does not process any signal. Every node is linked to all nodes in the adjacent layers. The training set which is made up of 108 observations (i.e. 75% of the data set) was used to developed ANN model for predicting hydraulic conductivity using the R Neuralnet library package. The six input variables that were employed for the neural network training are; *S, Fi, C, PI*, OMC, and MDD of the soil. The input layer therefore has six neurons. The only output is the hydraulic conductivity, so the output layer has only one neuron. Since there are no overall guidelines for defining the number of neurons in each hidden layer, Bahmed et al. [27] suggested the use of a simple architecture of one hidden layer with a limited number of neurons to earn time in the training stage. The rule of thumb in deciding the number of hidden layers is normally to start the training process with one hidden layer, and if one hidden layer does not train well, then the number of hidden neurons can be increased before considering adding more hidden layers [41]. The choice of the number of hidden neurons depends on the complexity of the problem. In this study, the number of hidden neurons equal half of the input variables was used to start the network training. This was further increased as the training error remains above the training error tolerance until the training error drops. After several network trainings with different number of neurons in the hidden layer, one hidden layer with the number of neurons that produced the least error was selected**.**

The NeuralNet presents the training set to the ANN and modifies the weights to minimise the error generated between the actual and desired output. In other words, a neuron's output is the weighted sum of inputs plus the bias activated by the transfer function [20]. Lim and

Kolay [10] observed that backpropagation (BP) training algorithm yields the best prediction model for hydraulic conductivity of tropical soils compared to other learning algorithms such as Levenberg–Marquardt algorithm, scale conjugate gradient, BFGS quasi-Newton, conjugate gradient with Powell/Beale Restarts, Fletcher–Powell conjugate gradient, and one-step secant**.** As a result, feed-forward neural network, with backpropagation training algorithm was used to develop the ANN model for this study. The goal of BP training is to iteratively change the connections weights between the neurons in a direction that minimises the error. Connection weights in the network are adjusted by the algorithm using a sample-by-sample updating rule. In one algorithm iteration, a training sample is presented to the network. The signal is then fed in a forward manner through the network until the network output is obtained. The error between the actual and desired network outputs is calculated and used to adjust the connection weights [41]. After the completion of the training process, a new set of data was presented to the network, the testing data, to validate and evaluate the integrity of the trained network.

2.5 Performance evaluation

The following statistical indices, which were deemed significant, were used to assess the predictability of the developed ANN and MLR: mean squared error (MSE), root of the mean squared error (RMSE), multiple coefficient of determination (R^2), and mean absolute error (MAE). During the ANN training, a minimum network error is repeatedly tried by altering the weights as earlier mentioned and the number of the hidden layer neurons.

The mean square error (MSE) indicates the error obtained while training, and it measures the average square gap between the anticipated response value and its prediction. MSE is calculated using Eq. (2):

$$MSE = \frac{1}{N} \sum_{n=1}^{N} (actual - predicted)^2 \qquad (2)$$

where N is the overall number of data.

The root mean squared error (RMSE) is calculated between the measured values and the predicted values using Eq. (3)

$$RMSE = \sqrt{\frac{1}{N} \sum_{n=1}^{N} (actual - predicted)^2} \qquad (3)$$

Coefficient of determination, R^2 defined by Eq. (4) expressed the proportion of the total variation in response variable (predicted value) that is explained by different independent variables. The lower the difference
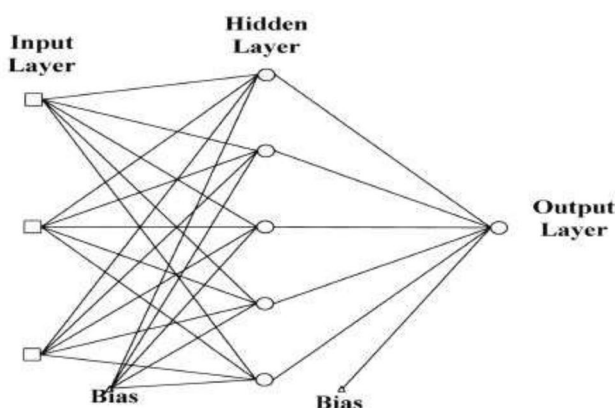

Fig. 1 Two-layer perceptron network architecture [42]

between actual and forecast values the higher the value of the determination coefficient. The value of R^2 is between 1 and 0. R^2 is near 1, for a good fit model, and R^2 near 0 indicates a poor fit model

$$R^2 = 1 - \frac{SSE}{SSa} \qquad (4)$$

where $SSE = \sum (a - \hat{a})^2$; $SSa = \sum \left(a - \overline{a}\right)^2$,
  a is the true value; $\hat{a}$ is the forecast value of a, and the mean value of the a values is $\overline{a}$.
  Mean absolute error (MAE) measures difference between two continuous variables.
  The correlation coefficient (r) described the strength of the linear connection between the predicted and actual response variable, ranging from -1 to $+1$.

3 Results and discussion

3.1 Data collated

The data set collated constitutes 144 observations, from different regions. Table 1 shows the summary of the data set collated of which 108 observations (75% of the data set), representing the characteristics of all the soil class present,
  Were used as training set for both the ANN and MLR models for hydraulic conductivity prediction. The descriptive statistics for each variable as contained in the collated data set is shown in Table 2. The PI ranged between 0.00 and 480 with a mean value of 83.64 showing that the soils collated ranged from pure sands to extreme swelling clay (sodium bentonite).

Table 1 Data set sources

Source	Number of data	Soil description	Region/country
Wang and Huang [44]	55	Ranges from sandy to clayey	-
Benson et al. [45]	55	Majorly clayey	USA
Anderson and Brandon [46]	2	Silt and silty clay	Hawaii
Benson and John [47]	13	Majorly clays and few sands	USA
Othman and Benson [48]	3	Clay	Southern Wisconsin
Alhassan [49]	1	Lateritic	Minna, Nigeria
Cuisinier et al., [50]	1	Silt	Paris, France
Amiralian et al., [51]	1	Quartz sand	Western Australia
Aytekin and Akcanca [52]	4	Sand	Turkey
Elsharief et al. [53]; Mohammed and Elshariel [54]	2	Clayey and silty clay	Eastern Sudan
Umar et al. [55]	1	Clayey sand	Bauchi, Nigeria
Govindasamy and Taha [56]	1	Silty sand	Malaysia
Maurya et al. [57]	2	Silty clay	Sultanpur, India
Umar et al. [58]	1	Sandy clay	Bauchi, Nigeria
Nithis et al. [2]	1	Clayey silt	India
Ojuri and Oluwatuyi [4]	1	Sand	S-W, Nigeria
Total number of data	144		

Table 2 Descriptive statistics of data set collated

	PI	Sand %	Fines %	Clay %	MDD	OMC	k
Min. value	0.00	0.00	0.00	0.00	1250.00	8.00	2.500e-13
First quantiles	14.00	8.50	58.00	20.88	1540.75	14.00	8.975e-11
Median	26.00	18.95	78.10	33.75	1680.00	18.15	2.700e-10
Mean	83.64	25.52	70.0005	34.88	1672.35	19.36	8.987e-07
Third quantiles	60.50	33.58	70.0005	47.05	1802.50	24.13	2.625e-09
Max. value	480.00	100.00	100.00	84.00	2083.00	38.10	7.670e-05

3.2 Data cleaning and multicollinearity analysis

Sanity checks were carried out on the data set to ensure that there are no variable values that fall outside the expected boundaries (Table 2). This also includes checking for missing values in each column for proper handling and to ensure there are sufficient observations to utilise for analysis. R software is gracefully designed to handle missing values with.

annotation **'NA'** to indicate the existence of missing values [43]. Sanity checks revealed the absence of missing data and values of each parameter used is within the expected range.

Correlation assessment was conducted to explore the likelihood and degree of multicollinearity relationship between each independent variable and all other variables. Correlation analysis was computed on R software using the 'cor ()' command. From Table 3, the independent variables satisfied the conditions for non-existence of multicollinearity except for sand–fine pair ($r = 0.969$) and OMC–MDD pair ($r = 0.907$). The correlation values between the independent variable are expected to be less than 0.8 in order to avoid the multicollinearity problem in the predicting model. The strong relationship between the two pairs of variables as revealed by the correlations analysis suggests that one of the two variables for each pair is needed in the regression analysis. It can also be observed that the independent variables chosen have

weak correlation relationship with the response variable, k. Another significant parameter that R also offers as a measure of multicollinearity is the Variance Inflation Factor (VIF). VIF analysis in R software is executed through the library 'car'. VIF value less than 5 and close to one indicates that there is no linear connection between input variables, if more than 10, is an indication that the variable is not needed and can be removed from the model [38].

3.3 Multiple linear regression model

3.3.1 MLR model training

The training of the MLR model for hydraulic conductivity prediction was accomplished using 75% of the data as earlier mentioned with six input variables, namely *PI*, *S*, *Fi*, *C*, OMC and MDD of the soils. This was executed in R using the 'lm ()' function. Table 4 which was generated by R gives the summary of the developed MLR model. The residuals as indicated in Table 4 give the differences between the experimental values and the predicted values. Positive residuals indicate that the model predicted a value that was lower than the observed value, and a value less than zero indicates that the regression model predicted a value higher than the observed value. As shown in Table 4, '**Min**' as used by R indicates minimum value of residual, and '**Max**', maximum value of residual. Residuals' median value is denoted using '**Median**'. The variables 1*Q* and 3*Q*

Table 3 Pearson correlation coefficient

	PI	Sand %	Fines %	Clay %	MDD	OMC	*k*
PI	1						
Sand %	0.016	1					
Fines%	−0.030	−0.969	1				
Clay %	0.395	−0.655	0.697	1			
MDD	−0.446	0.464	−0.505	−0.661	1		
OMC	0.275	−0.622	0.635	0.680	−0.907	1	
k	−0.085	0.343	−0.306	−0.213	0.037	−0.119	1

Table 4 Summary table of parameter estimates, residuals, and standard errors for the linear model of hydraulic conductivity fitted with six predictors

Residuals	Min	1Q	Median	3Q	Max
	−1.235e−05	−1.857e−06	−2.430e−07	1.158e−06	6.412e−05

	Input variables	Coefficients	Std Error	Pr (>|T|)	VIF
	Intercept	3.300e−05	3.633e−05	0.3657	–
	PI	−1.416e−08	7.229e−09	0.0529	1.814
	S	1.413e−07	1.872e−07	0.4522	45.955
	Fi	3.092e−08	1.843e−07	0.8671	46.484
	C	5.189e−10	6.234e−08	0.9934	3.1897
	MDD	−1.909e−08	1.128e−08	0.0935	8.385
	OMC	−2.591e−07	3.436e−07	0.4525	9.739

Significant codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

are the points that label the first and third quartiles of the residuals. The median of the residual values of a good model is expected to be close to zero, whereas the minimum and maximum values of almost the same value and the first and third quartile values should be approximately the same [43]. The residuals, as shown in Table 4, deviate slightly from these conditions for this present model. A section

Table shows the coefficient of each of the input variables. Hence, the developed regression equation is as shown in Eq. (5):

$$k = (3.300E - 05) + (-1.1416E - 08)PI + (1.413E - 07)S$$
$$+ (3.092E - 08)Fi + (5.189E - 10)C$$
$$+ (-1.909E - 08)MDD + (-2.591E - 07)OMC$$

(5)

The column named 'Std. Error' in Table 4 displays the standard statistical error for each coefficient. The standard error is expected, for a reliable model, to be at least 5–10 times less than the corresponding coefficient [43]. The resulting statistical standard errors for this model was almost greater than the corresponding coefficient for each input variable. The column marked $Pr(>|t|)$, gives the p value of the coefficient. The p value indicates the probability that the corresponding coefficient is not needed in the model; it ranges from zero (no chance) to unity (absolute certainty). In other words, subtracting this value from one gives the significance level. In science generally, results yielding a p value less than or equal to 0.05 are considered to be statistically significant and statistically highly significant if p value is less than or equal to 0.001. The p values revealed PI and MDD to be the only variables that are statistically significant with p value equal 0.0529 and 0.0935, respectively (i.e. 94.71% and 90.65% significance, respectively), while other inputs were not statistically strong enough to establish a significant model by MLR.

Retraining of the MLR was carried out using the backward elimination approach as explained by Lilja [43] to determine the predictor that should be utilised in developing the model and the ones to discard. The 'summary ()' function in R computes the significance level of each input variable used in the model. As earlier stated, the variable with the largest p value is least significant statistically, while threshold of p value equals 0.05 is predetermined below which the input variable has more than 95% chance that it is significant. Having p value higher than the set threshold value, such variable or predictor is removed from the model and re-computed. With the back-elimination process, regression Eq. (6) was developed with three predictors. The model outcomes as shown in Tables 4, 5 indicate that Eq. (6) is more reliable than Eq. (5). The three variables, namely *PI*, percent sand (*S*) and MDD, prove to be statistically significant with p values below 0.05 and VIF close to 1.

3.3.2 MLR model validation

Having obtained the regression Eq. (6), the equations was fitted with the test data to predict hydraulic conductivity of the test data. Figure 2 shows the scatter plots for observed values of *k* against its predicted values for the test data. Performance evaluation of the developed MLR models was carried out using the obtained values of MSE, RMSE, MAE and determination coefficient R^2 between the observed and predicted values as presented in Tables 6, 7. As noted from Table 7 for test samples, coefficient of determination, R^2, and correlation coefficient, r, for model Eq. (6) demonstrate a stronger and more accurate output than model Eq. (5) with six input variables. 40.4% variation in hydraulic conductivity for model Eq. (6) was explained by the three input variables utilised (*PI, S* and *MDD*), whereas 36.9% of hydraulic conductivity variability was explained by the six inputs utilised for MLR model Eq. (5). Correlation coefficient r, for MLR model Eq. (6) indicates a stronger linear relation between the observed and predicted values of *k* compared to *r* value for MLR model Eq. (5). Since R^2 and r values could give a biased estimate of model performance, the MLR models are also compared with respect to their mean square error, MSE and mean absolute error, MAE. As shown in Table 7, MSE and MAE values for model Eq. (6) are lower showing that the MLR model with three input variables (PI, S and MDD) is more accurate. The result of the

Table 5 Summary table of parameter estimates, residuals, and standard errors for the linear model of hydraulic conductivity fitted with three predictors

Residuals	Min	1Q	Median	3Q	Max
	−1.162e−05	−1.842e−06	−2.680e−07	1.000e−06	6.462e−05
	Input variables	Coefficients	Estimate Std Error	Pr (>|T|)	VIF
	Intercept	1.858e−05	8.283e−06	0.0270 *	–
	PI	−1.362e−08	6.339e−09	0.0340 *	1.42402
	S	1.291e−07	3.139e−08	7.87e−05 ***	1.31927
	MDD	−1.198e−08	5.037e−09	0.0192 *	
	1.70804				

Significant codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Fig. 2 The scatter plots of observed versus predicted values of hydraulic conductivity for MLR Eq. (6) model
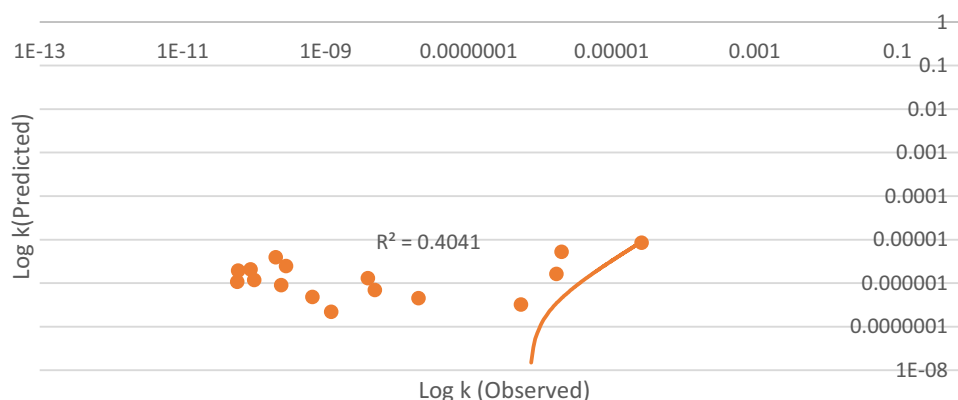

Table 6 Summary of performance indices for MLR models (training)

MLR equations	Training				
	R^2	r	MSE	RMSE	MAE
5	0.154	0.392	4.719e-11	6.870e-06	2.641e-06
6	0.146	0.383	4.760e-11	6.899e-06	2.614e-06

Table 7 Summary of performance indices for MLR models (testing)

MLR equations	Testing				
	R^2	r	MSE	RMSE	MAE
5	0.369	0.607	1.294e-11	3.597e-06	1.993e-06
6	0.404	0.636	1.192e-11	3.452e-06	1.902e-06

regression analysis revealed that *PI*, percent sand (*S*) and MDD of soils have more significant impact with respect to hydraulic conductivity of soils:

$$k = (1.858E - 05) + (-1.362E - 08)\text{PI} + (1.291E - 07)S$$
$$+ (-1.198E - 08)\text{MDD}$$

(6)

3.3.3 MLR comparative study

The MLR models of Salarashayeri and Siosemarde [29], Merdun, et al. [59] and Arshad, et al.[39] show a higher RMSE and MAE with relatively higher coefficient of determination as shown in Table 8. This indicates that the developed MLR model for this study with three input variables performed relatively better than the above-mentioned models with respect to error generated. The MLR model of Salarashayeri and Siosemarde [29] was developed using three input variables from soil particle diameters, namely D_{10}, D_{50} and D_{60}, where D_{10}, D_{50} and D_{60} are the soil particle diameter (mm) corresponding to 10%, 50% and 60% finer by weight and saturated hydraulic conductivity, *k* is expressed in m/day. The obtained RMSE and MAE were 4.06 and 3.32, respectively, when *k* is expressed in m/day, and the equivalent values when express in m/s are as shown in Table 8. Merdun, et al. [59] utilised seven variables to developed MLR model, namely *S, Si, C, BD, P*1, *P*2, *and P*3, corresponding to percent sand, percent silt, percent clay, bulk density and pore sizes, respectively. The obtained RMSE is 0.938 when hydraulic conductivity is expressed in cm/hr. Arshad, et al. [39] utilised percentages of clay (*C*), silt (*Si*), sand (*S*), and bulk density (ρb) as independent variables

Table 8 MLR models comparison

Model author	Input variables	R^2	r	RMSE	MAE
Salarashayeri and Siosemarde [29]	D_{10}, D_{50} and D_{60},	0.512	0.712	4.697 e-05	3.84 e-05
Merdun, et al. [59]	*S, Si, C, BD, P*1, *P*2, *and P3,*	0.637	0.798	2.606 e-05	NA
Arshad, et al. [39]	*C, Si, S, ρb*	0.5	0.707	1.402 e-04	NA
Present study	PI, S, MDD	0.404	0.636	3.452e-06	1.902e-06

to develop MLR model for hydraulic conductivity prediction. The RMSE obtained was also higher than the one obtained for this present study (Table 8). This shows that for this current study, the MLR model developed with three independent variables for hydraulic conductivity prediction of soils is relatively better.

3.4 Artificial neural network

3.4.1 ANN training

The training of the ANN for hydraulic conductivity prediction was achieved using six input parameters as used for the MLR model. It is considered a good practice to normalise the data before training a neural network in order to avoid unnecessary results or very difficult training processes resulting in algorithm convergence problems [43]. Among the simple methods of data normalisation available, minimum–maximum normalisation approach was utilised to bring the data values between 1 and 0 using Eq. (7):

$$a_i = \frac{a - a_{\min}}{a_{\max} - a_{\min}} \tag{7}$$

where a_i is the normalised value, a is the actual value, $a_{\max}$ is the maximum value and $a_{\min}$ is the minimum value.

The architecture that produced the best result after several network trainings contains one hidden layer with 10 neurons (Fig. 3). This was chosen based on the obtained R^2 and MSE (Fig. 4). The summary of the major training parameters is shown in Table 9. Figure 5 demonstrates the comparative significance of predictors in neural networks to the output prediction as obtained from R using Garson algorithm, a function in the NeuralNetTools library. Based on Garson algorithm, input relevance is calculated using Eq. (8):

$$R_{ik} = \sum_{j=1}^{nh} \left(\frac{|w_{ij}||w_{jk}|}{\sum_{i=1}^{ni} (|w_{ij}||w_{jk}|)} \right) \tag{8}$$

where ni and nh are the number of inputs and hidden units, respectively, w_{ij} is the weight between input i and hidden unit j and w_{jk} is the weight between hidden unit j and output k.

Unlike the MLR, ANN utilised all the input variables to develop the network for hydraulic conductivity prediction with percent fines having the greatest influence, while the least is the percent clay. This indicates that ANN can interpret very complex relationships between the input
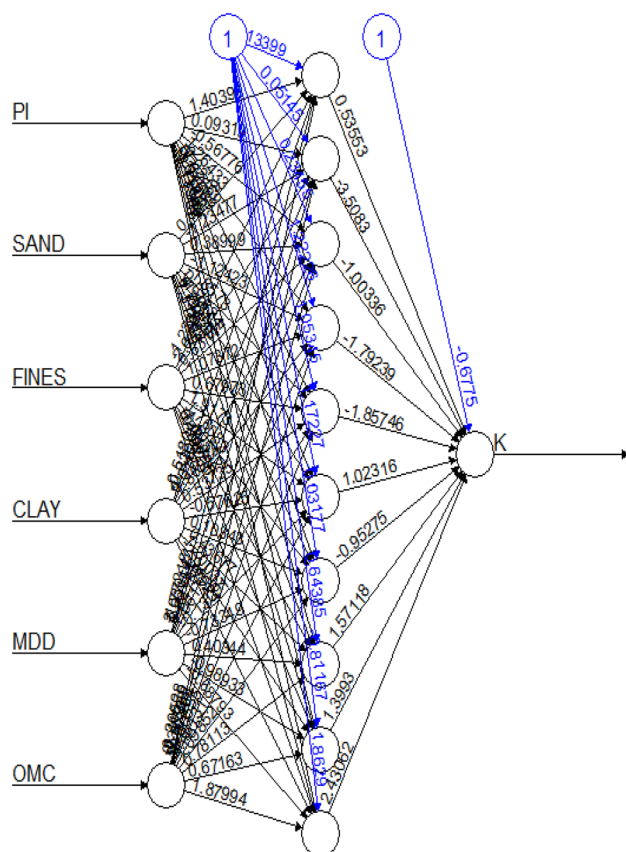

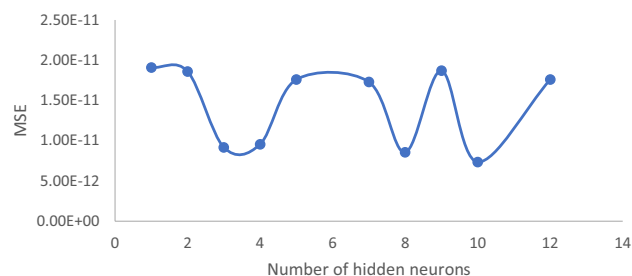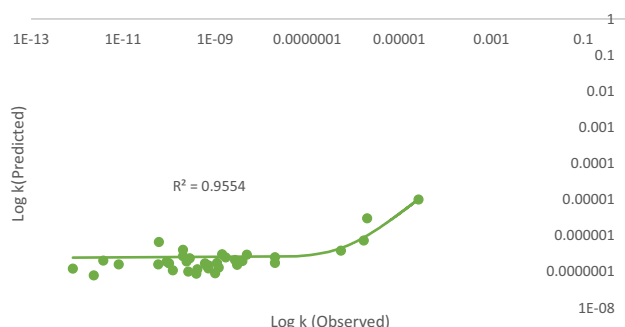**Fig. 3** ANN architecture for hydraulic conductivity prediction (R generated)
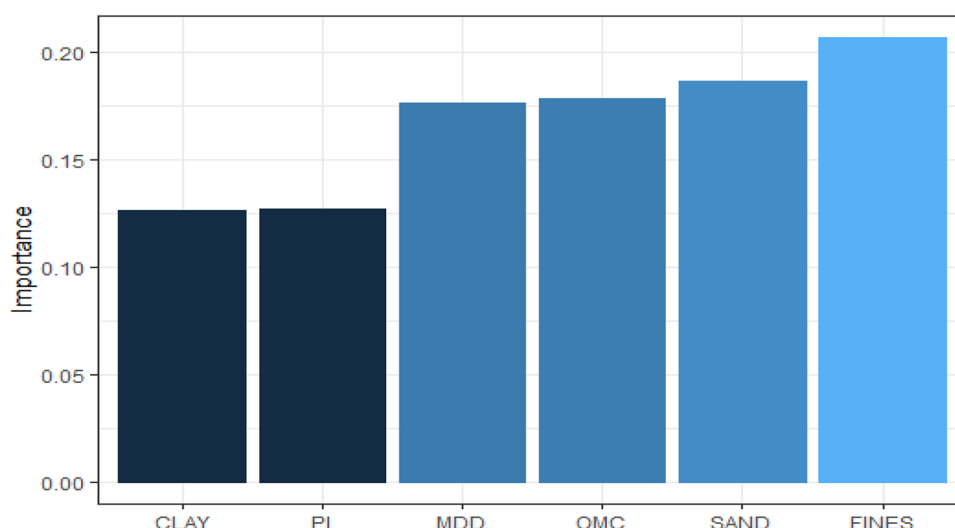

Fig. 4 Number of hidden nodes versus MSE

Table 9 Major training parameter used for ANN model in R

Arguments as used in R	Values and nomination in R
Hidden	10
Stepmax	1e + 05
learningrate	0.0078
algorithm	backprop (refers to backpropagation)
error function	sse (i.e. squared errors)
activation function	logistic
linear.output	F (false)

Fig. 5 Relative importance of input parameters

Fig. 6 Scatter plot of observed versus predicted values of hydraulic conductivity for ANN model (testing)

Table 10 Summary of performance indices for ANN model

	R^2	r	MSE	RMSE	MAE
Training	0.410	0.641	3.309e-11	5.753e-06	1.397e-06
Testing	0.955	0.977	7.366e-12	2.714e-06	6.817e-07

variables and the response variable beyond what MLR can comprehend.

3.4.2 ANN model validation

The ANN model was provided with new data set, the test data, to assess its efficiency and capacity to generalise forecast beyond the learning data. Figure 6 shows the scatter plot for observed values of *k* against its predicted values for the test data, while Table 10 gives the performance estimate of the ANN model. For test data, the obtained R^2 and MSE values of 0.955 and 7.366e-12 indicate an acceptable accuracy showing that 95.5% variation in hydraulic conductivity was explained by the

Table 11 Performance evaluation of MLR and ANN

Model	Testing				
	r	R^2	MSE	RMSE	MAE
MLR (Eq. 6)	0.636	0.404	1.192e-11	3.452e-06	1.902e-06
ANN	0.977	0.955	7.366e-12	2.714e-06	6.817e-07

six input variables considered with a minimised error thereby validating the model.

3.4.3 ANN and MLR model comparison

The MLR analysis of the data collated revealed that half of the input variables (percent fines, percent clay and OMC) did not contribute to the performance of the MLR. This could be attributed to nonlinear relationship or very low correlation between these variables and hydraulic conductivity. However, ANN shows a high level of understanding hidden relationships between these variables and the corresponding hydraulic conductivity. Hence, ANN high ability to predict nonlinear behaviour is worth commending. The results of the performance indices of the developed MLR and ANN models as presented in Table 11 show that ANN produces more reliable estimate of soils hydraulic conductivity than MLR. The higher R^2 of 0.995 and lower error estimates of ANN than those obtained by the MLR models was in support with other previous studies carried out by Sinha and Wang [20], Arshad, et al. [39] and other researchers showing that ANN is a better predictive tool to solving geotechnical/geo-environmental problems than the traditional linear regression.

3.4.4 ANN comparative study

Before this present study there are many studies on the ANN applications in geotechnical engineering and few in predicting hydraulic conductivity of a particular soil with different input variables utilised. Therefore, the developed network is compared with other existing models. Table 12 gives the summary of this comparison. The input variables used in all the networks listed include plasticity index (*PI*), percentages of gravel (*Gr*), percent sand (*S*), percent fines (*F*), percent silt (*Si*), percent clay (*C*), bulk density (ρb), dry density (ρd), liquid limit (*LL*), plastic limit (*PL*), maximum dry density (*MDD*), moisture content (*w*), optimum water content (*OMC*), void ratio (*e*), soil particle diameters (D_{10}, D_{50}) and degree of compaction express in percentage (*P%*). Boroumand and Baziar [18] developed network for predicting hydraulic conductivity of clay using 55 data set: 45 for training and 10 for testing. The soil physical properties and compaction properties used as input variables are MDD, PI, Gr, S, C with architecture of 5-4-1 corresponding to one hidden layer with four neurons. The obtained R^2 was 0.54 for test data. The ANN model of Sinha and Wang [20] utilised test soils prepared by mixing different amounts of four main components; gravel, sand, limestone dust and sodium bentonite. These four components were mixed to achieve 55 distinct combinations in distinct ratios. Five variables that were utilised as input include $P (\%)$, D_{10}, D_{50}, PL and LL. The network architecture contains one hidden layer with three neurons, the obtained R^2 was 0.901. Lim and Kolay [10] utilised 10 input variables with 10 neurons in the hidden layer to developed hydraulic conductivity prediction network for tropical soils (Table 12). The data set used contained 144 observations, of which 100 were used as training set and the rest as test data. The obtained R^2 was 0.92. For easier comparison, the R^2 value obtained for Lim and Kolay [10] was calculated from the obtained correlation coefficient and the MSE for Merdun et al. [59] was calculated from the obtained RMSE, the unit of k used was properly considered. It can be observed from Table 12 that R^2 value obtained for this study was the highest (0.955) with the lowest error estimate (MSE = 7.366e-12) showing that the developed ANN for hydraulic conductivity

prediction with six input variables, namely PI, *S, Fi, C*, MDD, and OMC, was well generalised to the validation data set (test data).

4 Conclusions

This research assessed the performance of artificial neural network (ANN) and the multiple linear regression (MLR) in predicting hydraulic conductivity of a wide range of soil types to obtain the appropriate value of soil hydraulic conductivity within the shortest time frame. Based on the analysis and the results obtained, the following conclusions are drawn:

- The correlation and the p value results revealed that only three inputs variables (plasticity index, percent sand and MDD) are statistically significant to the development of the MLR model and others are reductants.
- Relative Importance analysis revealed that the six input variables utilised for ANN development are all significant with percentage of fines being the most influential.
- The results of the statistical indices (R^2, MSE) show that ANN is the most reliable predictive tool and has strong ability to predict nonlinear behaviour when compared with MLR.
- Comparative study analysis shows the developed MLR and ANN to perform better that the corresponding available models considered.
- ANN model developed in this study can be efficiently utilised to predict the hydraulic conductivity of most soil types since the input variables are easily obtainable parameters thereby making soil investigation with respect to hydraulic conductivity faster.
- This study utilised the earliest and the most used activation function (sigmoid) for the developed network. It is suggested that the performance of rectified linear unit (ReLU) for ANN development to predict hydraulic conductivity of soils should be investigated. Furthermore, it will be important to develop a network that can predict hydraulic conductivity of different soil types

Table 12 ANN models comparison

Model author	Input variables	Architecture	Training data	Testing data	R^2	MSE
Boroumand and Baziar [18]	MDD, *PI, Gr, S, C,*	5–4–1	45	10	0.54	NA
Merdun et al. [59]	*S, Si, C, ρb, P1, P2,* and *P3*	NA	–	–	0.52	9.512e-11
Sinha and Wang [20]	*P(%), d10, d50, PL, LL*	5–3–1	–	–	0.901	NA
Lim and Kolay [10]	*w, ρb, ρd, e, LL,PL, Gr, S, Si, C,*	10–10–1	100	44	0.92	NA
Arshad et al. [39]	*C, Si, S, ρb*	4–7–1	130	45	0.66	2.314e-09
Present study	*PI, S, Fi, C,* MDD, OMC	6–10–1	108	36	0.955	7.366e-12

stabilised with the same additive (e.g. lime, cement, fly ash, etc.).

Resolving to collate data from previous studies on hydraulic conductivity of soils was a result of unavailability of database on basic soil's properties in Nigeria. Since the performance of ANN depends on the reliability of the training data, the data set used for this study was carefully selected to minimise data error. However, the integrity of the collated data set cannot be fully ascertained.

Compliance with ethical standards

Conflict of interest No potential conflict of interest was reported by the authors.

References

1. Murthy VNS (2002) Principle and practices of soil mechanics and foundation engineering. Madison Dekker Inc, New York
2. Nithis ST, Relha B, Uma S (2017) Use of lateritic soil ammended with bentonite as landfill liner. Rasayaan J Chem 10(4):1431–1438
3. Vakili AH, Arab A, Davoodi S (2015) Use of artificial neural network in predicting permeability of dispersive clay treated with lime and pozzolan. Int J Sci Res Environ Sci 10(4):0023–0037
4. Ojuri OO, Oluwatuyi OE (2017) Strength and hydraulic conductivity characteristics of sand-bentonite mixtures designed as a landfill liner. Jordan J Civil Eng 11(4):614–622
5. Stibinger IJ (2014) Examples of determining the hydraulic conductivity of soils: Theory and applications of selected basic methods. Ústí nad Labem: J. E. Purkyně University in Ústí n. Labem, Faculty of the Environment
6. Dimenescus MA, Dumitran GE, Vuţā LI (2019) Experimental Method to determine the hydraulic conductivity. In E3S Web of conferences. 85, p. 06010. EDP Sciences
7. Freeze RA, Cherry JA (1979) Groundwater. Prentice Hall Inc, Englewood Cliffs, New Jersey
8. Qi Shipeng et al. (2015) A new empirical model for estimating the hydraulic conductivity of low permeability media. Proceedings of the international association of hydrological sciences, 368, pp.478.
9. Philip HJ, Ngala AL, Waniyo UU (2014) Comparison of empirical models and laboratory saturated hydraulic conductivity measurements. Ethiop J Environ Stud Manag 7(3):305–309
10. Lim DKH, Kolay PK (2009) Predicting hydraulic conductivity (k) of tropical soils by artificial neural network (ANN). UNIMASS E-J Civil Eng 1(1):1–6
11. Shamshirband S, Rabczuk T, Chau KW (2019) A survey of deep learning techniques: application in wind and solar energy resources. IEEE Access 7:164650–164666
12. Faizollazadeh Ardabili S, Najafi B, Shamshirband S, Minaei Bidgoli B, Deo RC, Chau KW (2018) Computational intelligence approach for modeling hydrogen production: a review. Eng Appl Comput Fluid Mech 12(1):438–458
13. Shahin MA, Jaksa MB, Maier HR (2001) Artificial neural network applications in geotechnical engineering. Aust Geomech 36(1):49–62
14. Parlak A, Islamoglu Y, Yasar H, Egrisogut A (2006) Application of artificial neural network to predict specific fuel consumption and exhaust temperature for a diesel engine. Appl Therm Eng 26(8–9):824–828
15. Faghri A, Hua J (1992) Evaluation of artificial neural network applications in transportation engineering. Transp Res Rec 1358:71
16. Ardabili SF, Mahmoudi A, Gundoshmian TM (2016) Modeling and simulation controlling system of HVAC using fuzzy and predictive (radial basis function, RBF) controllers. J Build Eng 6:301–308
17. Abraham A (2005) Handbook of measuring system design. John Wiley & Sons, United State
18. Boroumand A, Baziar MH (2005) Determination of compacted clay permeability by artificial neural networks. Sharm El-Sheikh, Egypt, s.n., pp 515–526
19. Ciaburro G, Venkateswaran B (2017) Neural network with R: smart models using CNN, RNN, deep learning, and artificial intelligence principles. Packt Publishing Ltd, Birmingham-Mumbai
20. Sinha SK, Wang MC (2007) Artificial neural network prediction models for soil compaction and permeability. Sprnger Sci 26:47–64
21. Ul-Saufie AZ, Yahya AS, Ramli NA (2011) Improving multiple linear regression model using principal component analysis for predicting PM10 concentration in Seberang Prai, Pulau Pinang. Int J Environ Sci 2(2):415–422
22. Goh ATC (1994) Nonlinear modelling in geotechnical engineering using neural networks. Aust Civil Eng Trans CE 36(4):293–297
23. Chan WT, Chow YK, Liu LF (1995) Neural network: An alternative to pile driving formulas. J Comput Geotech 17:135–156
24. Shahin MA, Jasksa MB, Maier HR (2000) Predicting the settlement of shallow foundations on cohesionless soils using back-propagation neural networks. Adelaide
25. Goh ATC (1994) Seismic liquefaction potential assessed by neural network. J Geotech Geoenviron Eng 120(9):1467–1480
26. Najjar YM, Ali HE (1998) CPT-based liquefaction potential assessment: a neuronet approach. Geotechn Special Publi ASCE 1:542–553
27. Bahmed IT, Harichane K, Ghrici M, Boukhatem B, Rebouh R, Gadouri H (2017) Prediction of geotechnical properties of clayey soils stabilised with lime using artificial neural networks (ANNs). Int J Geotech Eng 13(2):191–203
28. Sabat AK (2013) Prediction of california bearing ratio of a soil stabilized with lime and quarry dust using artificial neural network. Electron J Geotech Eng 18:3261–3272

29. Salarashayeri AF, Siosemarde M (2012) Prediction of soil hydraulic conductivity from particle-size distribution. Int J Geol Environ Eng 6(1):16–20

30. Kalkhajeh YK, Arshad RR, Amerikhah H, Sami M (2012) Comparison of multiple linear regressions and artificial intelligence-based modeling techniques for prediction the soil cation exchange capacity of Aridisols and Entisols in a semi-arid region. Aust J Agricul Eng 3(2):39–46

31. Shabani A, Norouzi M (2015) Predicting cation exchange capacity by artificial neural network and multiple linear regression using terrain and soil characteristics. Indian J Sci Technol 8(28):1–10

32. Das SK, Samui P, Sabat AK (2012) Prediction of field hydraulic conductivity of clay liners using an artificial neural network and support vector Machine. Int J Geomech 12(5):606–611

33. Minasny B, Hopmans JW, Harter T, Eching SO, Tuli A, Denton MA (2004) Neural network prediction of soil hydraulic functions for Alluvial soils using multistep outflow data. Soil Sci Soc Am J 68:417–429

34. Nieva PM, Francisca FM (2007) On the permeability of compacted and stabilized loessical silts in relation to liner system regulations. In international congress on development, environment and natural resources: multi-level and multi-scale sustainability (pp. 69–77)

35. Lambe T (1954) The permeability of compacted fine-grained soils. Spec Tech Publi 163:56–67

36. R CoreTeam (2018) R: a language and environment for statistical computing R foundation for statistical computing Vienna

37. Akan R, Keskin SN, Uzundurukan S (2015) Multiple regression model for the prediction of unconfined compressive strength of jet grount coloumns. Procedia Earth Planet Sci 15:299–303

38. Kravchenko AN, Bullock DG (2000) Correlation of corn and soybean grain yield with topography and soil properties. Agron J 92(1):75–83

39. Arshad RR, Sayyad G, Mosaddeghi M, Gharabaghi B (2013) Predicting saturated hydraulic conductivity by artificial intelligence and regression models. ISRN Soil Sci 2013:1–8

40. Wu CL, Chau KW (2013) Prediction of rainfall time series using modular soft computing methods. Eng Appl Artif Intell 26(3):997–1007

41. Hui CLP (ed) (2011) Artificial neural networks: application. BoD–Books on Demand

42. Basu S, Paul FH, Nihar B, Hon K (2002) Prediction of gas-phase adsorption isotherms using neural nets. Can J Chem Eng 80(3):506–512

43. Lilja DJ (2016) Linear regression using R: an introduction data modeling. USA, University of Minesota Libraries Publishing, Minnesota

44. Wang MC, Huang CC (1984) Soil compaction and permeability prediction models. J Environ Eng 110(6):1063–1083

45. Benson CH, Zhai H, Wang X (1994) Estimating hydraulic conductivity of compacted clay liners. Journal of Geotech Eng 120(2):366–387

46. Anderson SA, Brandon HH (1995) Hydraulic conductivity of compacted lateritic soil with bentonite admixture. Environ Eng Geosci 1(3):299–312

47. Benson CH, John MT (1995) Hydralic conductivity of thirteen compacted clays. Clays Clay Miner 43(6):669–681

48. Othman MA, Benson CH (1992) Effect of freeze-Thaw on the hydraulic conductivity of three compacted clays from Wisconsin. Madison, p. 1369

49. Alhassan M (2008) Permeability of lateritic soil treated with lime and rice husk ash. Assumpt Univ Thail 12(2):115–120

50. Cuisinier O, Auriol JC, Le Borgne T, Deneele D (2011) Microstructure and hydraulic conductivity of a compacted lime-treated soil. Eng Geol 123(3):187–193

51. Amiralian S, Chegenizadeh A, Nikraz H (2012) Investigation on the effect of lime and fly ash on hydraulic conductivity of soil. Int J Biol Ecol Environ Sci 1(3):120–123

52. Aytekin M, Akcanca F (2013) Hydraulic conductivity of lime stabilized sand-bentonite mixtures for sanitary liners. Greece, Athens

53. Elsharief AM, Elhassan AA, Mohamed AE (2013) Lime stabilization of tropical soils from Sudan for road construction. Int J GEOMATE Geotech Constr Mater Environ 4(2):533–538

54. Mohammed AA, Elshariel AM (2015) Engineering properties of lime stabilized swelling soils from Sudan. Int J Sci Eng Technol Res 4(10):3595–3600

55. Umar SY, Elinwa AU, Matawal DS (2015) Hydraulic conductivity of compacted lateritic soil partially replsced with metakaolin. J Environ Earth Sci 5(4):53–64

56. Govindasamy P, Taha MR (2016) Hydraulic conductivity of residual soil-cement mix. IOP Conf Series Mater Sci Eng 136(1):012031

57. Maurya R, Umesh K, Gupta MK (2016) Hydraulic conductivity for silty soil added with plastic wastes. In: International seminar on sources of planet energy, environmental and disaster science: challenges and strategies SPEEDS-16. India

58. Umar SY, Slim MD, Uchechukwu EA (2016) Hydraulic conductivity of compacted laterite treated with iron ore tailings. Adv Civ Eng. https://doi.org/10.1155/2016/4275736

59. Merdun H, Cinar O, Meral R, Apan M (2006) Comparison of artificial neural network and regression pedotransfer functions for prediction of soil water retention and saturated hydraulic conductivity. Soil Tillage Res 90(1–2):108–116