Geomechanics and Engineering, *Vol. 33, No. 3 (2023) 231-242*
https://doi.org/10.12989/gae.2023.33.3.231

Proposing new models to predict pile set-up in cohesive soils

Sara Banaei Moghadam[a] and Mohammadreza Khanmohammadi*

Department of Civil Engineering, Isfahan University of Technology, Isfahan, 84156-83111, Iran

(*Received September 4, 2021, Revised December 30, 2022, Accepted January 13, 2023*)

Abstract. This paper represents a comparative study in which Gene Expression Programming (GEP), Group Method of Data Handling (GMDH), and multiple linear regressions (MLR) were utilized to derive new equations for the prediction of time-dependent bearing capacity of pile foundations driven in cohesive soil, technically called pile set-up. This term means that many piles which are installed in cohesive soil experience a noticeable increase in bearing capacity after a specific time. Results of researches indicate that side resistance encounters more increase than toe resistance. The main reason leading to pile setup in saturated soil has been found to be the dissipation of excess pore water pressure generated in the process of pile installation, while in unsaturated conditions aging is the major justification. In this study, a comprehensive dataset containing information about 169 test piles was obtained from literature reviews used to develop the models. to prepare the data for further developments using intelligent algorithms, Data mining techniques were performed as a fundamental stage of the study. To verify the models, the data were randomly divided into training and testing datasets. The most striking difference between this study and the previous researches is that the dataset used in this study includes different piles driven in soil with varied geotechnical characterization; therefore, the proposed equations are more generalizable. According to the evaluation criteria, GEP was found to be the most effective method to predict set-up among the other approaches developed earlier for the pertinent research.

Keywords: cohesive soils; gene expression programming; group method of data handling; pile foundations; set-up

1. Introduction

Pile set-up is a term that refers to an increase in bearing capacity of pile foundations driven in cohesive soil. The results of many studies indicate that the side resistance experiences a significant increase in bearing capacity due to set-up (Titi and Wije Wathugala 1999, Fattah *et al.* 2013, Tarawneh 2013, Tarawneh and Imam 2014, Khanmohammadi and Fakharian 2019, Abu-Farsakh and Haque 2018, Haque and Steward 2020); however, tip resistance has been found to exhibit a negligible change, or in some cases a decrease as a result of relaxation (Xie 2011, Abu-Farsakh *et al.* 2017). Although soil set-up results in an eventual increase in the pile capacity, the bearing capacity decreases with time in the soil relaxation condition (Khanmohammadi and Fakharian 2019). The pertinent set-up phenomenon mostly emerges as a result of dissipation of excess pore water pressure generated during pile installation in saturated soil (Khanmohammadi and Fakharian 2018). It has been observed that the greatest amount of set-up mostly occurred during 2 weeks (Wang 2017). The main mechanism of set-up is usually implemented in three phases as follows: 1) the non-uniform dissipation of excess pore water pressure 2) the uniform dissipation of excess pore water pressure 3) aging (Komurka *et al.* 2003). Several studies have been conducted to investigate set-up influenced by factors such as pile diameter, pile length, soil type, effective stress, undrained shear strength, and time (Ng *et al.* 2013, Haque *et al.* 2014, Fakharian and Khanmohammadi 2016, Khanmohammadi and Fakharian 2018, Ng and Ksaibati 2018, Haque and Abu-Farsakh 2019, Banaei Moghadam and Khanmohammadi 2021). Even some researchers have tried to introduce an empirical equation using regression and analytical methods to estimate the time-dependent bearing capacity of piles (Randolph *et al.* 1979, Skov and Denver 1988, Svinkin 1996, Camp III and Parmar 1999, Svinkin and Skov 2000, Svinkin 2002, Haque and Abu-Farsakh 2019, Gong *et al.* 2020). One of the most important and practical equations is the semi-empirical equation presented by Skov and Denver (1988), in which the ultimate bearing capacity is a function of variables such as initial bearing capacity at the end of driving (EOD), the logarithm of interval time, and set-up parameter (A). This parameter varies between 0.5 to 0.7 which depends on the type of soil and has been subject to further studies as a function of soil features. For example, Haque and Abu-Farsakh (2019) presented a study to develop a non-linear multivariable regression model to predict the increase in pile bearing capacity with time considering the effects of soil properties on the set-up parameter of Skov and Denver equation. From this study, it has been understood that undrained shear strength, sensitivity, coefficient of consolidation, and plasticity index show a significant influence on A as it is proportional to soil sensitivity and plasticity index but inversely proportional to undrained shear strength and coefficient of consolidation.

*Corresponding author, Ph.D.
E-mail: mkhanmohammadi@iut.ac.ir
[a]Ms.
E-mail: Sara.banaeimoghadam@gmail.com

ISSN: 2005-307X (Print), 2092-6219 (Online)

However, as reported by Gardner and Dorling (1998), an extensive range of engineering problems may encounter conditions with non-linear and complex natures where traditional regression analyses are inadequate. In the other words, when the physical meaning of the system is difficult to understand or the underlying relationships among the data are unknown, empirical and statistical methods reveal incapability (Fatehnia and Amirinia 2018). Therefore, some new computational techniques with higher capabilities and more accuracy must be employed for function finding purposes. The successful application of artificial intelligence (AI) and machine learning (ML) have been performed in several recent studies (Jeon and Rahman 2007, Razavi *et al.* 2009, Razavi *et al.* 2018, Tarawneh 2018, Harandizadeh *et al.* 2019, Koopialipoor *et al.* 2019, Armaghani *et al.* 2020, Armaghani *et al.* 2020, Armaghani *et al.* 2020, Harandizadeh 2020, Harandizadeh *et al.* 2020, Harandizadeh and Toufigh 2020, Li *et al.* 2020, Li *et al.* 2020, Pham *et al.* 2020, Banaei Moghadam and Khanmohammadi 2021, Khanmohammadi *et al.* 2022). For instance, Tarawneh (2018) conducted a study to evaluate the efficiency of gene expression programming (GEP) to find a functional equation. In this study, the soil and pile properties were used to estimate the time-dependent bearing capacity of 104 pile tests. The results indicated that GEP is an effective approach for prediction areas. In another study, a back-propagation neural network was developed by Jeon and Rahman (Jeon and Rahman 2007) to examine the feasibility of ANNs to predict the pile bearing capacity owing to set-up. The result of this study demonstrated that the ANN model predicted pile set-up with a high amount of accuracy. Harandizade and Toufigh (Harandizadeh and Toufigh 2020) used the combination of group method of data handling (GMDH) and Neural-Fuzzy (NF) to predict the axial bearing capacity of driven piles. To optimize the designed system, the particle swarm optimization (PSO) and gravitational search algorithm (GSA) were performed as well. In comparison with their previous achieved results, the performance of NF-GMDH was better than Genetic Program (GP) which showed GMDH efficiency. The main purpose of the current research is to conduct a comparative study to evaluate the efficiency of three approaches including the multiple linear regression (MLR), GMDH, and GEP for estimating the time-dependent bearing capacity of piles driven in cohesive soil. To investigate the correlation between variables, MLR has been performed at the initial stage of the study

2. Database

The current dataset which has been obtained from the literature reviews containing information about 169 test piles driven in clay and mixed soil. The information includes pile and soil properties such as length of the pile (L), pile diameter (PD), pile type, soil type, plasticity index (PI), undrained shear strength (S_u), effective stress (ES), initial bearing capacity (Q_i), time interval after driving (T), and ultimate bearing capacity after a period of specific time (Q_u). Data references are listed in Table 1.

Table 1 Datapoints references

References	Num. of data	Soil	Pile	PD
(Samson and Authier 1986)	3	Silty sand	H	0.305
(Fellenius *et al.* 2004)	9	Silty sand	H	0.305
(Svinkin *et al.* 1994)	9	Silty sand	SC	0.457
(Axelsson 1998)	16	Silty sand	SC	0.235
(Dover and Howard 2002)	30	Bay mud	PP	0.610
(Komurka 2004)	5	clay	PP	0.320
(Haque, Chen *et al.* 2014)	85	clay	SC	0.354
(Haque and Abu-Farsakh 2019)	12	mix	SC	0.457

*PD= Pile Diameter, H= H pile, SC= Square Pile, PP= Pipe Pile

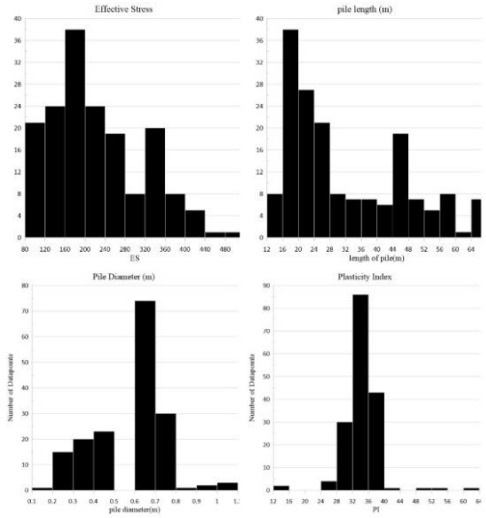


Fig. 1 Data frequency distribution histogram

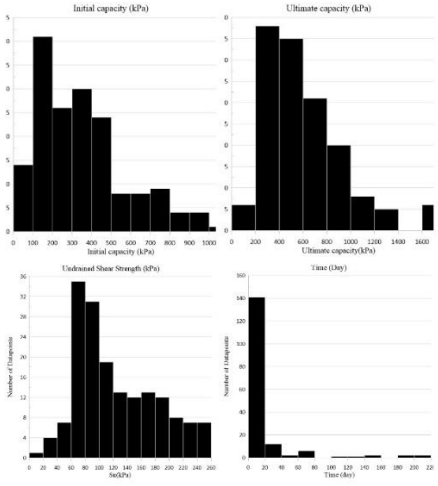


Fig. 2 Data frequency distribution histogram

In all the models, the ultimate bearing capacity was considered as a target, while the other ones were independent variables. Table 2 shows the descriptive statistics of the data. It is worth mentioning that to avoid overfitting, the data were randomly divided into testing and training datasets. In this study, 118 data points were used

Table 2 Descriptive statistics of the data

variable	Min	Max	Mean	SD	Range
PD (m)	0.23	0.76	0.54	0.16	0.52
L (m)	13.99	64	36.11	15.78	50.01
Qu (KN)	440	6023	2807.8	1365	5583.6
Qi (KN)	331.3	4263	1577.14	1021.9	3932.2
ES (kPa)	98.39	427.5	240.35	107.4	329.1
PI	30	40	34.55	2.57	10
Su (kPa)	50.65	80.14	67.71	7.45	29.49
T (Day)	2	143	16.95	28.07	141

Table 3 The correlation coefficient between independent variables

	Qi	Su	PI	L	PD	T	ES
Qi	1	0.11	0.007	0.24	0.39	-0.09	0.18
Su	0.11	1	0.15	-0.20	-0.16	0.03	-0.21
PI	0.007	0.15	1	-0.01	0.05	-0.14	-0.01
L	0.24	-0.20	-0.013	1	0.33	-0.20	0.93
PD	0.39	-0.16	0.05	0.33	1	-0.2	0.27
T	-0.09	0.03	-0.14	-0.15	-0.20	1	-0.07
ES	0.18	-0.21	-0.019	0.93	0.27	-0.07	1

*PD= Pile Diameter, T= Time, ES= Effective Stress, L=Pile Length, PI= Placticity Index, Su= Undrained Shear Strength, Qi= Initial bearing capacity

for training while 51 data points were considered as testing data. To compare the performances of the developed models, similar training and testing data were applied to them. To provide more information on distributions of numerical variables, Figs. 1 and 2 represent the frequency distribution of the data used in this research.

2.1 Data pre-processing

To develop artificial intelligence algorithms and multivariate regressions, it is necessary to perform statistical pre-analysis before applying the data to the algorithms. In this research, Rapidminer® which is a data mining software has been used to perform pre-processing of data. The following steps were accomplished with the database used in this study. Rapidminer uses a data detector operator to find the outlier data in the dataset that follows the normal distribution. This operator starts searching to find outliers based on the approach proposed by Ramzwami *et al.* (2000). In this approach, the available data is categorized based on their distance to a specific point (Ramaswamy *et al.* 2000).

2.1.1 Outliers and missing data detection

By applying this operator to the dataset, it was determined that 7 data were outliers and removed from the set. Rapidminer software can detect missing data in a database by using different operators and applying the appropriate filter as well. Using this method, the existence of missing data in the database was investigated, but no case was observed.

2.1.2 Investigation of collinearity

Collinearity means that there is a correlation between two independent variables of a dataset. This correlation may complicate the prediction process or cause noticeable errors. Rapidminer uses different approaches to investigate this phenomenon. One of the most common methods is using the correlation coefficient matrix. In this method, the correlation coefficients between each pair of independent variables must be calculated. If the correlation coefficient is more than 70%, the designer can choose one of the variables and remove the other one. The matrix of correlation coefficients between the independent variables in this study can be seen in Table 3. As it is shown, there is noticeable collinearity between effective stress and pile length; therefore, one of them must be removed from the prediction process. Because of the importance of effective stress in the set-up process, this variable was considered for the rest of the study.

3. Material and methods

3.1 Group method of data handling (GMDH)

GMDH is a computational technique proposed by (Ivakhnenko and Ivakhnenko 1995) in the 1960s in which a layered structure creates connections between input and output variables through the mathematical description of Kolmogorov-Gabour polynomials. GMDH is a practical algorithm that employs different operations such as seeding, rearing, crossbreeding, and selection and rejection to determine the inputs, general structure, and parameters of the model as well as selection of the desired model based on error criterion (Amanifard *et al.* 2008). Due to the layered structure of GMDH, the selected polynomials from the first layer are the inputs of the second layer. In the other words, this algorithm connects various pairs of neurons to produce new neurons for the next layers. This repetitive approach continues until acquiring the most valid polynomial which predicts the desired target effectively. GMDH is a combination of quadratic and higher neurons in a certain number of variable layers that map a vector of input features to the expected response by creating a multistage nonlinear pattern; it is mainly based on decomposition and dominance. In every layer of this network, a different subset of possible combinations in each neuron among the existing features is mapped to the expected response using polynomial functions (Mehra 1977). Based on the accuracy achieved for each combination, some weaker combinations are removed in favor of stronger ones (Anastasakis and Mort 2001). To create these sets of polynomials, the first combination of input variables in the form of Kolmogorov-Gabor will be randomly chosen by the algorithm itself. Generally, the GMDH network is built to explore a function $\hat{f}$ that can be employed approximately in place of the actual one f to predict output $\hat{y}$ for a given input vector

$X = (x_1.x_2.x_3 \ldots x_M)$ very close to its actual output y so that

$$y_i = f(x_{i1}, x_{i2,} x_{i3}, ..., x_{in})(i = 1,2,...M) \quad (1)$$

$$\hat{y}_i = \hat{f}(x_{i1}, x_{i2,} x_{i3}, ..., x_{in})(i = 1,2,...M) \quad (2)$$

The most important purpose of the trained system is to minimize the square difference between the actual value of the target and the predicted value by the GMDH system according to Eq. (3).

$$\sum_{i=1}^{M} [\hat{f}(x_{i1}, x_{i2,} x_{i3}, ..., x_{in}) - y_i]^2 \quad (3)$$

As mentioned earlier, in the general network of GMDH, input and output variables can be connected using the discrete form of Volterra function called Kolmogorov-Gabour polynomials which is as follows

$$y = a_0 \sum_{i=1}^{M} a_i x_i + \sum_{i=1}^{M} \sum_{j=1}^{M} a_{ij} x_i x_j + \sum_{i=1}^{M} \sum_{j=1}^{M} \sum_{k=1}^{M} a_{ijk} x_i x_j x_k \quad (4)$$

In Eq. (4), $X = (x_1.x_2.x_3 \ldots x_M)$ is the vector of input variables and $A = (\mathrm{a}_1.\mathrm{a}_2.\mathrm{a}_3 \ldots \mathrm{a}_M.)$ is the vector of the summand coefficients. The usage of tri-quadratic and 3rd order polynomial for modeling some complex networks has been reported by some researchers. However, in this study, the second-order polynomial which is the general form proposed by Ivakhnenko was utilized in the form of

$$\hat{y} = G(x_i, x_j) = a_0 + a_1 x_i + a_2 x_j + a_3 x^2{}_i + a_4 x^2{}_j + a_5 x_i x_j \quad (5)$$

3.2 Gene expression programming (GEP)

GEP is an algorithm considered as a natural development of Genetic Algorithms (GA) and Genetic Program (GP). The main components of this algorithm which was invented by Ferreira (2001) include a function set containing mathematical operations, a terminal set containing input variables and numerical constant, fitness function, control parameters, and termination conditions (Gandomi *et al.* 2011). In GEP the computer programs are encoded as linear strings of fixed length (chromosomes or genomes) which are then expressed as non-linear entities of various shapes and sizes (expression trees or phenoms) (Ferreira 2002). GEP chromosomes are composed of more than one gene so that for each problem the number of genes is chosen. Because of this unique structure and the implementation of genetic operators at the chromosome level, GEP leads to the evolution of more complex programs and the creation of a varied genetic diversity. GEP genes are divided into head and tail. The head contains functions and terminals; however, the tail represents only terminals. In gene expression programming, the genome or chromosome consists of a linear, symbolic string of fixed length composed of one or more genes. In the other words, genes are created when input variables and mathematical operations are combined, and chromosomes can be created when genes are combined. Gene expression trees are

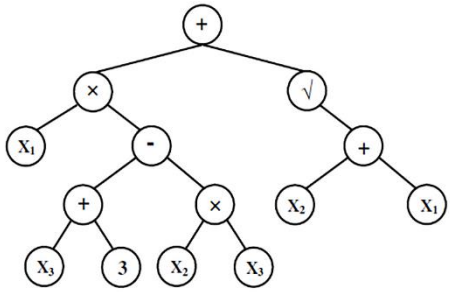


Fig. 3 Typical representation of GEP expression tree

generated after genetic operators are implemented to determine which terminals and functions (genes and chromosomes) are best combined according to their fitness function and other error indicators. Translation of expression trees and genes can be accomplished using the Karva language (a GEP language developed by Ferreira) (Ferreira 2002).

GEP genes with selected function and terminal set and the corresponding expression tree (Fig. 3) are as follows in which X_1, X_2, X_3 are input variables creating terminal set (Gandomi *et al.* 2011)

$$\pm.\times.\sqrt{}.X_1.-.+.+.\times.X_2.X_1.X_3.3.X_2.X_3 \quad (6)$$

The GEP gene can also be expressed in a mathematical form as

$$y = X_1[(X_1 + 3) - (X_2 \times X_3)] + \sqrt{(X_2 + X_1)} \quad (7)$$

GEP process involves six main steps repeated for a certain number of generations to find the best solution for the problem. The first step is the creation of an initial population of the chromosome, while the second stage deals with the expression of chromosomes and evaluation of individuals' fitness followed by the process of selection according to the fitness by roulette-wheel. Then the selected individuals start to reproduce while being subjected to genetic operators such as mutation, different types of transposition, and recombination. Finally, the new generation is created incorporating the same process until the best solution for the problem is found.

4. Modeling of pile set-up

4.1 Multiple linear regression (MLR) modeling

To evaluate the correlation between variables and determine the significant independent variables affecting the target, the least square multiple linear regression has been implemented using the SAS® program. This method is a statistical technique that provides the study to analyze the data and predict the target based on a simple linear relationship among independent variables. To determine the insignificant independent variables, the significance level ($P < 0.05$) was considered. The results showed that all the

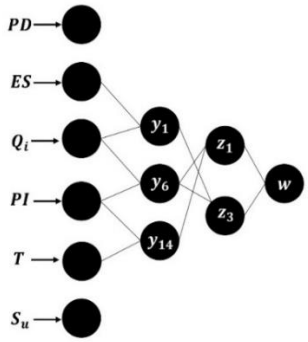


Fig. 4 GMDH structure

Table 4 Parameter setting for the GMDH algorithm

Parameters setting	Optimal value
Maximum number of layers	3
Maximum number of neurons	14
α	0.7

Table 5 Parameter setting for GEP algorithm

Parameter setting		Optimal value
General	Chromosomes	120
	Gene	3
	Head size	5
	Tail size	6
	Gene size	11
	Linking function	+
	Function set	{+, -, *, /, exp, ln, pwr, log)
Genetic operators	Mutation rate	0.06
	Inversion rate	0.1
	IS transposition rate	0.1
	RIS transpostion rate	0.1
	One-point recombination rate	0.2
	Two-point recombination rate	0.3
	Gene recombination rate	0.2
	Gene transposition rate	0.1

independent variables had a P value less than 0.05. The following equation in which PD is pile diameter (m), ES represents effective stress (kPa), PI is plasticity index, T is time (Day), S_u is considered as undrained shear strength (kPa) and Q_i and Q_u represent initial and time-dependent bearing capacity (KN), respectively, shows the MLR outcome that with R^2 of 0.76 predict the time-dependent ultimate bearing capacity.

$$Q_u = -1562.74 + 1421.04 * PD + 3.23 * ES + 54.39 * PI + 5.52T - 13.04 * S_u + 1.06 * Q \quad (8)$$

4.2 GMDH modeling

To develop a GMDH model, some important parameters such as the number of layers, neurons, and selection pressure ratio (α) must be determined at the initial stage. With the help of the selection pressure, the system will be capable of selecting the optimal fits at each step and transferring them to subsequent layers (Armaghani *et al*. 2020). The selection pressure in this research has been considered 0.7 which means 70% of the better and stronger polynomials will be transferred to the next layer. To find the optimal network, different architectures in which the number of layers and neurons varied between 2 and 10 and 2 and 20, were trained and tested, respectively. The final parameters of the GMDH network are shown in Table 4.

The R^2 results of 0.77, 0.74, and 0.76 are for training, testing, and all data, respectively, which indicate that the performance of the GMDH model in predicting the time-dependent increase in bearing capacity of the pile has been considerable. As it can be observed in Fig. 4 which shows the structure of the GMDH model developed in this study, there are three layers including different numbers of neurons forming the second-order polynomials connected to find the best combination of input variables to predict the time-dependent bearing capacity.

In the last layer of the proposed GMDH model, there is one polynomial determined by taking into account two polynomials of the previous (second) layer. The details of this equation is as follows

$$Q_u = w = -28.27 - 0.16z_1 + 1.19z_3 - 1.32\times10^{-5}z_1^2 - 1.19\times10^{-4}z_3^2 + 2\times10^{-4}z_1z_3 \quad (9)$$

Regarding the second layer, there are two polynomials each of which is determined considering three polynomials of the previous layer. The mathematical description of these equations is as follows

$$z_1 = -2.5\times10^3 - 0.44y_6 + 2.28y_{14} + 1.12\times10^{-5}y_6^2 - 6.59\times10^{-4}y_{14}^2 + 4.82\times10^{-4}y_6y_{14} \quad (10)$$

$$z_3 = 98.98 + 1.53y_1 - 0.54y_6 - 4.34\times10^{-4}y_1^2 - 1.45\times10^{-4}y_6^2 + 5.82\times10^{-4}y_1y_6 \quad (11)$$

Turning to the first layer, the dual combinations of the input variables have been determined by employing the second-order polynomials. The following equations show the contribution of each input in the prediction process

$$y_1 = -229.97 + 1.38Q_i + 4.03ES + 1.4\times10^{-5}Q_i^2 + 46\times10^{-4}ES^2 - 11\times10^{-4}Q_iES \quad (12)$$

$$y_6 = 1.09\times10^4 - 653.28PI + 1.67Q_i + 10.20PI^2 - 2.07\times10^{-5}Q_i^2 - 94\times10^{-4}PIQ_i \quad (13)$$

$$y_{14} = 1.41\times10^4 - 763.16PI + 46.55T + 12.3PI^2 - 0.12T^2 - 0.79PIT \quad (14)$$

Eventually, Eqs. (9)-(14) can be used in the mentioned order to predict Q_u.

4.3 GEP modeling

Table 6 Statistical indicators of the developed models

Formula	Condition	Dataset	MLR	GMDH	GEP
$R^2 = 1 - \frac{\sum_{i=1}^{n}(t_i - h_i)}{\sum_{i=1}^{n}(h_i - \bar{h}_i)}$	$0.7 < R^2$	Training	0.76	0.77	0.81
		Testing	0.73	0.74	0.81
$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(t_i - h_i)^2}{N}}$		Training	773.63	753.98	676.67
		Testing	717.02	709.40	602.75
$k = \frac{\sum_{i=1}^{n}(h_i \times t_i)}{h_i^2}$	$0.85 < k < 1.15$	Testing	0.94	0.96	0.95
$k' = \frac{\sum_{i=1}^{n}(h_i \times t_i)}{t_i^2}$	$0.85 < k' < 1.15$	Testing	0.99	0.97	1.01
$Ro^2 = 1 - \frac{\sum_{i=1}^{n}(t_i - h_i^o)^2}{\sum_{i=1}^{n}(t_i - \bar{t}_i)^2}, h_i^o = k \times t_i$		Testing	0.99	0.99	0.99
$Ro'^2 = 1 - \frac{\sum_{i=1}^{n}(h_i - t_i^o)^2}{\sum_{i=1}^{n}(h_i - \bar{h}_i)^2}, t_i^o = k' \times h_i$		Testing	1	1	1

Note: $\boldsymbol{t_i}$ is the predicted value, $\boldsymbol{h_i}$ is the actual value, $\bar{\boldsymbol{h}}_{\boldsymbol{i}}$ and $\bar{\boldsymbol{t}}_{\boldsymbol{i}}$ represent the mean value of the actual and predicted time-dependent bea capacity, respectively, and N is the number of samples.

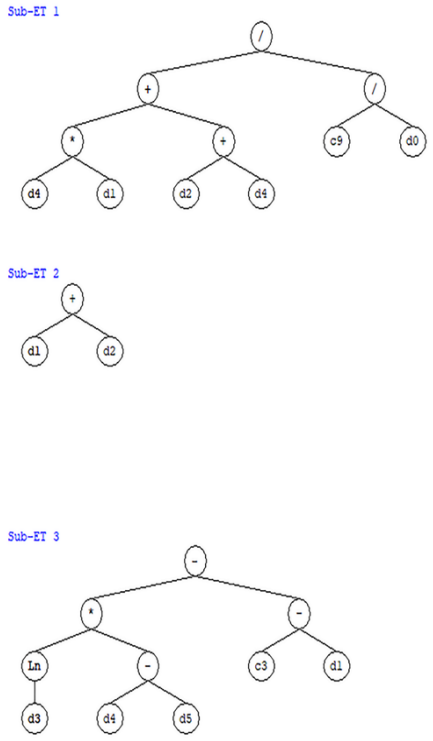


Fig. 5 GEP expression tree

To develop the GEP model, five major steps must be followed. The first step is determining a fitness function which in this study RMSE function was used to evaluate the overall fitness of the developed program. The second step is to determine the terminal and function sets followed by the third and fourth ones which are the determination of chromosomal architecture and genetic operation rates. Finally, the last step is to choose an appropriate linking function. Table 5 indicates the parameter setting of GEP. Subsequently, the program was run until there was no longer noticeable progress in the performance of the models. The GEP-based formulation of the pile set-up is as follows (Eqs. (15)-(18)) where d_0 represents the pile diameter in m (PD), d_1 is plasticity index (PI), d_2 is the initial bearing capacity in kPa (Q_i), d_3 is time in day (T), d_4 is the effective stress in kPa (ES), and d_5 is undrained shear strength in kPa (S_u). The summation of the three sub-trees is presented in an equation that predicts the time-dependent bearing capacity of piles (Q_u). The developed GEP model is also presented in Fig. 5 in form of expression trees.

$$G_1 = \frac{[(d_4 \times d_1) + (d_2 + d_4)] \times d_0}{2.26} \quad (15)$$

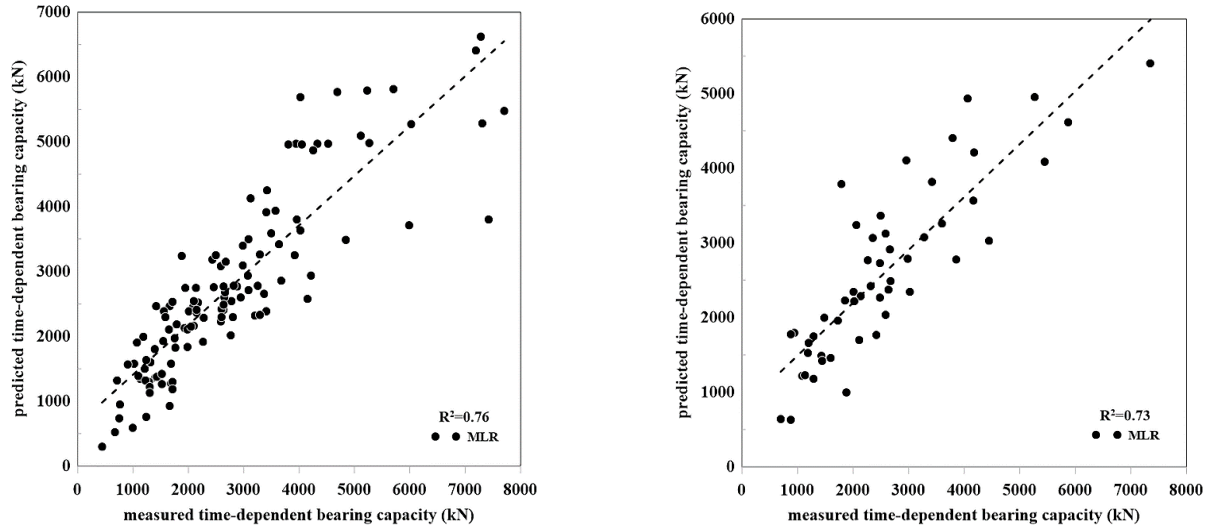
$$G_2 = d_1 + d_2 \quad (16)$$

$$G_3 = [\ln d_3 \times (d_4 - d_5)] - [-5.07 - d_1] \quad (17)$$

$$Q_u = G_1 + G_2 + G_3 \quad (18)$$

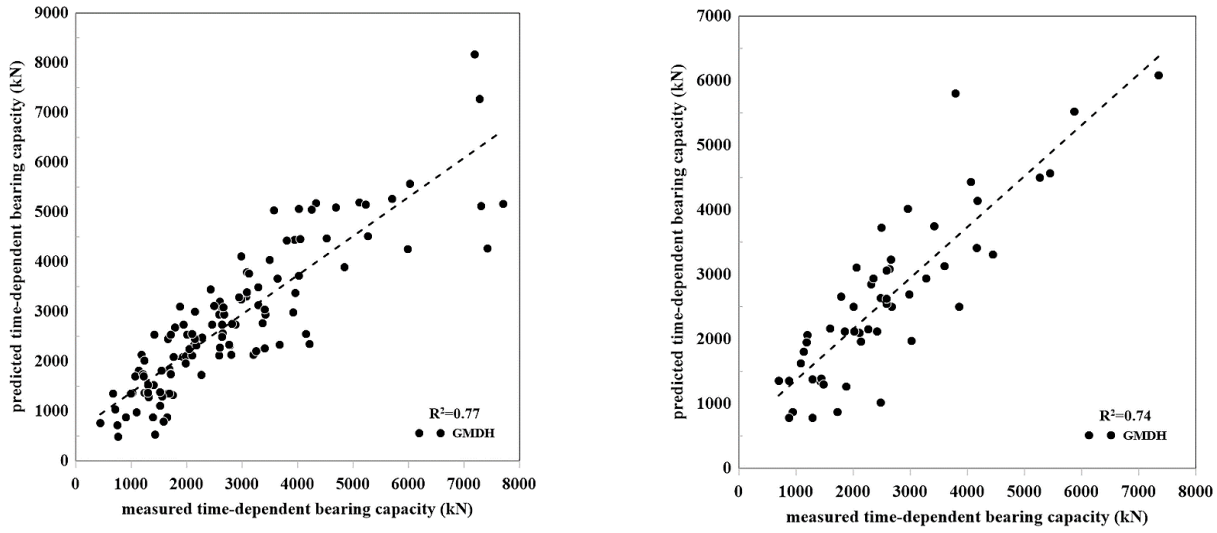
5. Models assessment

To evaluate the validity of the developed models, the statistical indicators such as Root-Mean-Squared Error ($RMSE$), and coefficient of determination (R^2) are considered. To judge the performances of the models, Smith



(a) Training data (b) Testing data

Fig. 6 Measured versus predicted pile set-up using MLR model

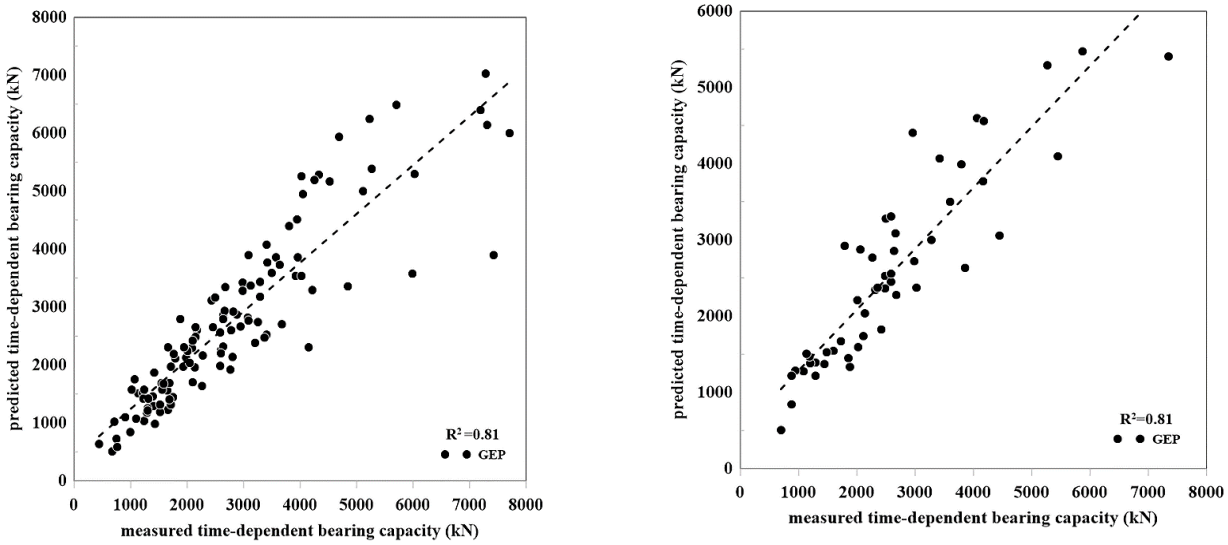


(a) Training data (b) Testing data

Fig. 7 Measured versus predicted pile set-up using GMDH model

(Smith 1986) suggested that if a model gives R^2>0.7, a noticeable correlation exists between independent and target variables. Furthermore, to guarantee the validity of the developed models, the error values (e.g., RMSE) should be minimum and as similar as possible for the training and testing datasets. This suggests that the proposed model has a very good predictive ability (low values) with excellent generalization performance (similar value) (Pan *et al.* 2009).

They suggested that to ensure the accuracy of the models and their prediction powers, at least one slope of the regression lines (k or $\acute{k}$) as well as the squared correlation coefficient (through the origin) between predicted and measured values (R_0^2) or the coefficient between measured and predicted values ($\acute{R}_0^2$) should be close to 1 (Gandomi *et al.* 2011, Gandomi *et al.* 2011). Table 6 shows these indicators for the training and testing datasets for the developed models in this research. Regarding all the indexes and the value of RMSE and R^2 obtained for GMDH, GEP, and MLR, respectively, it could be confirmed that all the models have exhibited good performance; however, GEP predicted pile set-up more accurately. The R^2 results of 0.81 for training and testing indicate that the performance of GEP has been considerable. Figs. 6-8 illustrate the measured ultimate bearing capacity versus predicted value by the MLR, GMDH, and GEP models. Considering the figures, it is observed that all of the models present an acceptable performance; however, the highest amount of accuracy belongs to the GEP model with the coefficient determination of 0.81 for both testing and training datasets, and the RMSE of 676.67 and 602.75 for training and



(a) Training data (b) Testing data

Fig. 8 Measured versus predicted pile set-up using GEP model

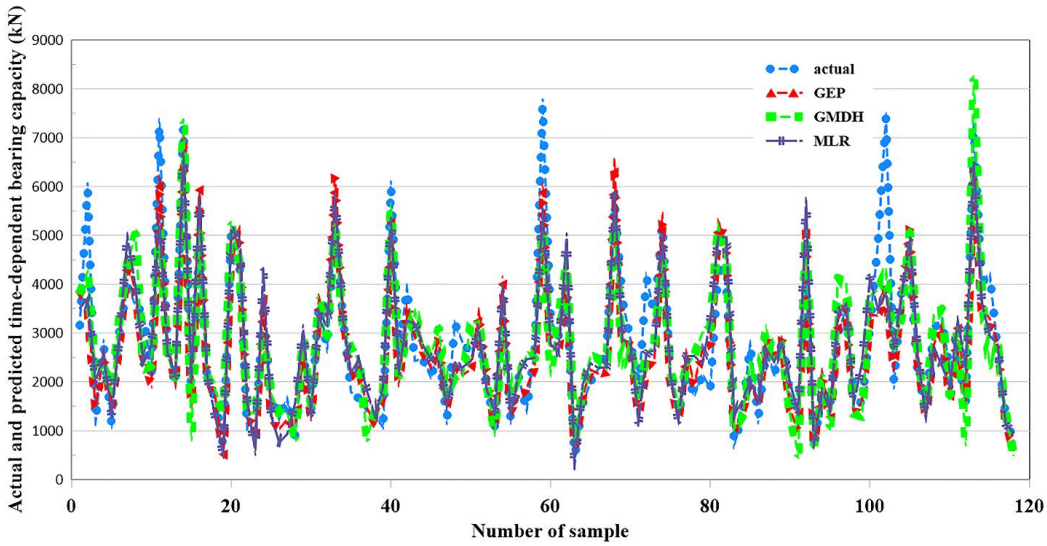


Fig. 9 Comparison between actual and predicted time-dependent bearing capacity, training data

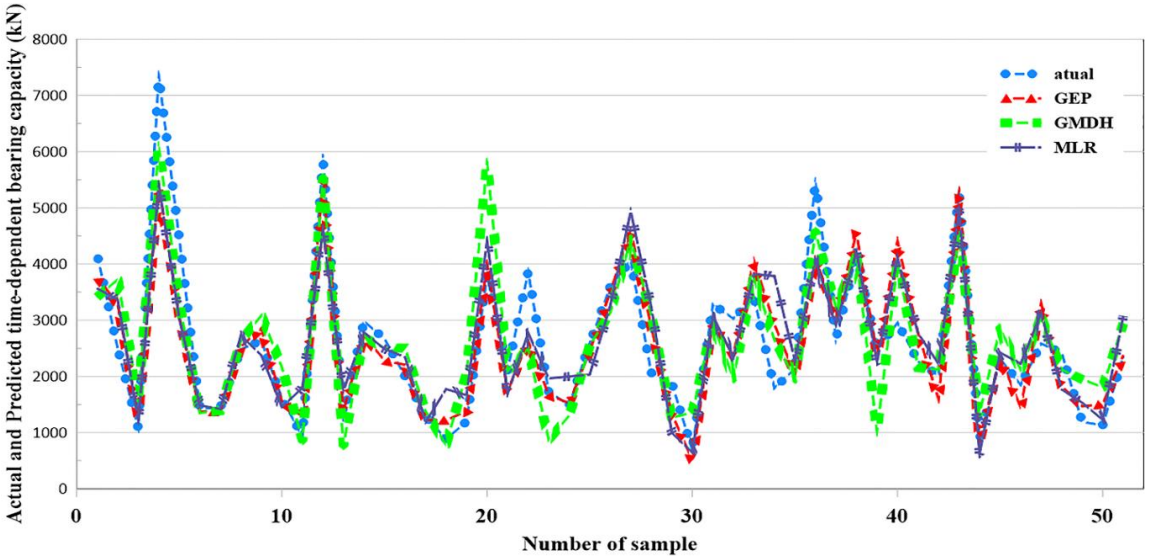


Fig. 10 Comparison between actual and predicted time-dependent bearing capacity, testing data

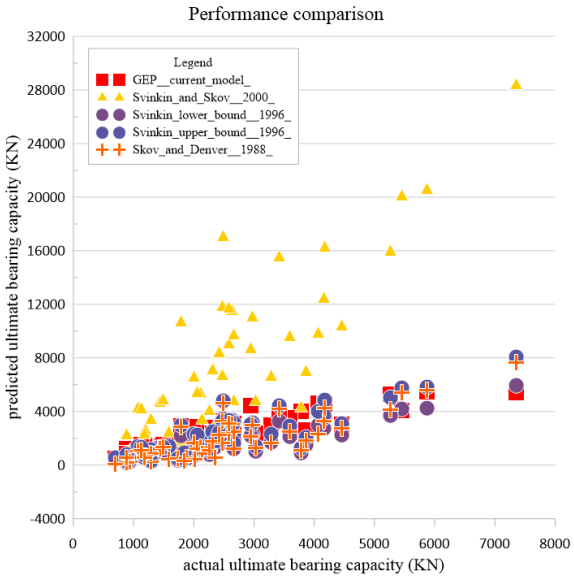


Fig. 11 Performance of different models in the prediction of pile set-up

validation datasets, respectively, which confirm higher prediction ability and accuracy level of GEP. Figs. 9 and 10 exhibit a comparison between the actual values of Q_u and the predicted ones by MLR, GMDH, and GEP models for training and testing data, respectively. As can be seen from the graphs, although all the models performed well, the values of Q_u being predicted by GEP seem to be closer to the actual values obtained from in situ tests In order to compare the performance of the developed models in this research with other models from the literature, the prediction ability of the equation obtained from GEP has been compared with the equations proposed by Skov and Denver, Svinkin And Svinkin and Skov. For this purpose, the same testing dataset which was used for the validation of the GEP model in this research, was introduced to these equations (Skov and Denver 1988, Svinkin 1996, Svinkin and Skov 2000). Fig. 11 Shows the performance of all the equations in estimating the time-dependent bearing capacity of pile. As it can be seen from the graph and based on the R-Squared value obtained for the other equations, the GEP model developed in this research has exhibited a good performance compared to the others. According to the graph, the R^2 for Svinkin and Skov, Svinkin (upper and lower bound) and Skov and Denver are equal to 0.69, 0.74, 0.74 and 0.70 respectively. This indicates that the GEP model developed in this research with the R^2 value of 0.81 has exhibited better performance.

5.1 Sensitivity analysis

To investigate the contribution of the independent variables in the process of prediction, the frequency values of the input variables introduced to the GEP model were considered. Regarding this methodology which is common in evolutionary algorithms development, the input with the value of 1 would be considered as the most effective parameter in the prediction process. The frequency values

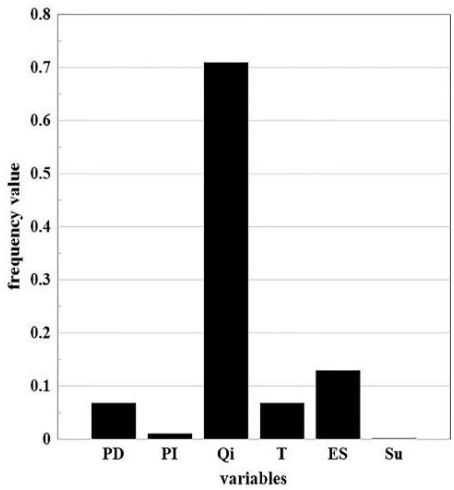


Fig. 12 Contribution of input variables in the GEP analysis

of the input variables in this study are presented in Fig. 12. From the results, it can be understood that initial bearing capacity (Q_i) and effective stress (ES) are two important variables in the prediction process of pile set-up.

6. Conclusions

In this study, a comparative approach was conducted to formulate the time-dependent bearing capacity of piles driven in cohesive soil using multiple linear regression (MLR), group method of data handling (GMDH), and gene expression programming (GEP). To develop the prediction models, a comprehensive database obtained from the literature review was used. Regarding the results of this research, the following concluding remarks can be expressed.

- Although error values associated with MLR show that this model performed well, statistical regression techniques seem to be uncertain or may have some limitations in complex engineering systems modeling.
- As regards the error values and the coefficient of determination, it is obvious that GMDH and GEP, as AI methods, have exhibited better performances rather than the regression method.
- Considering the error values, GEP with R^2 of 0.81 seems to be the most effective approach predicting the time-dependent bearing capacity of the pile with a higher amount of accuracy; however, according to the RMSE value obtained for training and testing datasets, it can be concluded that GMDH is a bit more generalizable than GEP.
- As a result of the performance comparison between the GEP model developed in this study and a few well-known equations for estimating pile setup, the GEP-based equation proposed in this research has shown to be more accurate than the other equations; therefore, it can be regarded as an appropriate model for predicting pile setup.
- Regarding sensitivity analysis performed on models' outcomes, the input variables such as Q_i and ES had more contribution in the prediction process.
- Based on the foregoing discussion, the GEP model which has introduced in this study can be reliably considered for pre-designing purposes as it contains information of test piles driven in various locations with different geotechnical characterization.

References

Abu-Farsakh, M.Y. and Haque, M.N. (2018), "Estimation and incorporation of pile setup into LRFD design methodology", *Transportation Research Board 97th Annual Meeting*, Washington DC, United State.

Abu-Farsakh, M.Y., Haque, M.N., Tavera, E. and Zhang, Z. (2017), "Evaluation of pile setup from osterberg cell load tests and its cost–benefit analysis", *Transport. Res. Record*, **2656**(1), 61-70. https://doi.org/10.3141%2F2656-07.

Amanifard, N., Nariman-Zadeh, N., Farahani, M. and Khalkhali, A. (2008), "Modelling of multiple short-length-scale stall cells in an axial compressor using evolved GMDH neural networks", *Energ. Convers. Management*, **49**(10), 2588-2594. https://doi.org/10.1016/j.enconman.2008.05.025.

Anastasakis, L. and Mort, N. (2001), "The development of self-organization techniques in modelling: a review of the group method of data handling (GMDH)", *Research report - University of sheffield department of automatic control and systems engineering.*

Armaghani, D.J., Asteris, P.G., Fatemi, S.A., Hasanipanah, M., Tarinejad, R., Rashid, A.S.A. and Huynh, V.V. (2020), "On the use of neuro-swarm system to forecast the pile settlement", *Appl. Sci.*, **10**(6), 1904-1921. https://doi.org/10.3390/app10061904.

Armaghani, D.J., Mirzaei, F., Shariati, M., Trung, N.T., Shariati, M. and Trnavac, D. (2020), "Hybrid ANN-based techniques in predicting cohesion of sandy-soil combined with fiber", *Geomech. Eng.*, **20**(3), 191-205. https://doi.org/10.12989/gae.2020.20.3.191.

Armaghani, D.J., Momeni, E. and Asteris, P.G. (2020), "Application of group method of data handling technique in assessing deformation of rock mass", *Metaheuristic Comput. Appl.*, **1**(1), 1-18. https://doi.org/10.12989/mca.2020.1.1.018. Axelsson, G. (1998), "Long-term increase in shaft capacity of driven piles in sand".

Banaei Moghadam, S. and Khanmohammadi, M. (2021), "Prediction of time-dependent bearing capacity of pile driven in cohesive soil using group method of data handling", *Sharif J. Civil Eng.*, **37**(3.2), 27-35.

Camp III, W.M. and Parmar, H.S. (1999), "Characterization of pile capacity with time in the Cooper Marl: study of applicability of a past approach to predict long-term pile capacity", *Transport. Res. Record*, **1663**(1), 16-24. https://doi.org/10.3141%2F1663-03.

Dover, A.R. and Howard, J., Roger (2002), "High capacity pipe piles at san francisco international airport", *Deep Foundations 2002: An International Perspective on Theory, Design, Construction, and Performance*, Orlando, Florida.

Fakharian, K. and Khanmohammadi, M.R. (2016), "Numerical modeling of pile installation effects on stress state in clay", *Jpn. Geotech. Soc. Spec. Publication*, **2**(39), 1402-1406. 10.3208/jgssp.IRN-19.

Fatehnia, M. and Amirinia, G. (2018), "A review of genetic programming and artificial neural network applications in pile foundations", *Int. J. Geoeng.*, **9**(1), 1-20. https://doi.org/10.1186/s40703-017-0067-6.

Fattah, M.Y., Al-Mosawi, M.J. and Al-Zayadi, A.A. (2013), "Time dependent behavior of piled raft foundation in clayey soil", *Geomech. Eng.*, **5**(1), 17-36. https://doi.org/10.12989/gae.2013.5.1.017.

Fellenius, B.H., Harris, D.E. and Anderson, D.G. (2004), "Static loading test on a 45 m long pipe pile in Sandpoint, Idaho", *Can. Geotech. J.*, **41**(4), 613-628.

Ferreira, C. (2001), "Gene expression programming: a new adaptive algorithm for solving problems", *arXiv preprint cs/0102027*, https://arxiv.org/abs/cs/0102027.

Ferreira, C. (2002), "Gene expression programming in problem solving", *Soft Comput. Ind.*, 635-653. https://doi.org/10.1007/978-1-4471-0123-9_54.

Gandomi, A.H., Alavi, A.H., Mirzahosseini, M.R. and Nejad, F.M. (2011), "Nonlinear genetic-based models for prediction of flow number of asphalt mixtures", *J. Mater. Civil Eng.*, **23**(3), 248-263. https://doi.org/10.1061/(ASCE)MT.1943-5533.0000154.

Gandomi, A.H., Tabatabaei, S.M., Moradian, M.H., Radfar, A. and Alavi, A.H. (2011), "A new prediction model for the load capacity of castellated steel beams", *J. Constr. Steel Res.*, **67**(7), 1096-1105. https://doi.org/10.1016/j.jcsr.2011.01.014.

Gardner, M.W. and Dorling, S. (1998), "Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences", *Atmosph. Environ.*, **32**(14-15), 2627-2636. https://doi.org/10.1016/S1352-2310(97)00447-0.

Gong, W., Li, L., Zhang, S. and Li, J. (2020), "Long-term setup of a displacement pile in clay: An analytical framework", *Ocean Eng.*, **218**, 108143. https://doi.org/10.1016/j.oceaneng.2020.108143.

Haque, M.N. and Abu-Farsakh, M.Y. (2019), "Development of analytical models to estimate the increase in pile capacity with time (pile setup) from soil properties", *Acta Geotechnica*, **14**(3), 881-905. https://doi.org/10.1007/s11440-018-0654-5.

Haque, M.N., Chen, Q., Abu-Farsakh, M. and Tsai, C. (2014), "Effects of pile size on set-up behavior of cohesive soils", *Proceedings of the Geo-Congress 2014: Geo-characterization and Modeling for Sustainability*, Atlanta, Georgia.

Haque, M.N. and Steward, E.J. (2020), "Evaluation of Pile Setup Phenomenon for Driven Piles in Alabama", *Proceedings of the*

Geo-Congress 2020: Foundations, Soil Improvement, and Erosion, Minneapolis, Minnesota.

Harandizadeh, H. (2020), "Developing a new hybrid soft computing technique in predicting ultimate pile bearing capacity using cone penetration test data", *AI EDAM*, **34**(1), 114-126. https://doi.org/10.1017/S0890060420000025.

Harandizadeh, H., Armaghani, D.J. and Mohamad, E.T. (2020), "Development of fuzzy-GMDH model optimized by GSA to predict rock tensile strength based on experimental datasets", *Neural Comput. Appl.*, **32**, 14047-14067. https://doi.org/10.1007/s00521-020-04803-z.

Harandizadeh, H., Toufigh, M.M. and Toufigh, V. (2019), "Application of improved ANFIS approaches to estimate bearing capacity of piles", *Soft Comput.*, **23**(19), 9537-9549. https://doi.org/10.1007/s00500-018-3517-y.

Harandizadeh, H. and Toufigh, V. (2020), "Application of developed new artificial intelligence approaches in civil engineering for ultimate pile bearing capacity prediction in soil based on experimental datasets", *Iranian J. Sci. Technol. T. Civil Eng.*, **44**, 545-559. https://doi.org/10.1007/s40996-019-00332-5.

Ivakhnenko, A. and Ivakhnenko, G. (1995), "The review of problems solvable by algorithms of the group method of data handling (GMDH)", *Pattern Recognition And Image Analysis C/C Of Raspoznavaniye Obrazov I Analiz Izobrazhenii*, **5**, 527-535.

Jeon, J. and Rahman, M.S. (2007), "A neural network model for prediction of pile setup", *Transport. Res. Record*, **2004**(1), 12-19. https://doi.org/10.3141%2F2004-02.

Khanmohammadi, M., Armaghani, D.J. and Sabri Sabri, M.M. (2022), "Prediction and optimization of pile bearing capacity considering effects of time", *Mathematics*, **10**(19), 3563.

Khanmohammadi, M. and Fakharian, K. (2019), "Numerical modelling of pile installation and set-up effects on pile shaft capacity", *Int. J. Geotech. Eng.*, **13**, 484-498. https://doi.org/10.1080/19386362.2017.1368185.

Khanmohammadi, M. and Fakharian, K. (2018), "Evaluation of performance of piled-raft foundations on soft clay: A case study", *Geomech. Eng.*, **14**(1), 43-50.

Khanmohammadi, M. and Fakharian, K. (2018), "Numerical simulation of soil stress state variations due to mini-pile penetration in clay", *Int. J. Civil Eng.*, **16**(4), 409-419. https://doi.org/10.1007/s40999-016-0141-z.

Komurka, V.E. (2004), "Incorporating set-up and support cost distributions into driven pile design", *Current Practices and Future Trends in Deep Foundations*, Los Angeles, California.

Komurka, V.E., Wagner, A.B. and Edil, T.B. (2003), "Estimating soil/pile set-up", *Wisconsin Highway Research Program Madison*, WI, USA.

Koopialipoor, M., Nikouei, S.S., Marto, A., Fahimifar, A., Armaghani, D.J. and Mohamad, E.T. (2019), "Predicting tunnel boring machine performance through a new model based on the group method of data handling", *Bull. Eng. Geol. Environ.*, **78**(5), 3799-3813. https://doi.org/10.1007/s10064-018-1349-8.

Li, D., Armaghani, D.J., Zhou, J., Lai, S.H. and Hasanipanah, M. (2020), "A GMDH predictive model to predict rock material strength using three non-destructive tests", *J. Nondestruct. Eval.*, **39**(4), 1-14. https://doi.org/10.1007/s10921-020-00725-x.

Li, D., Moghaddam, M.R., Monjezi, M., Jahed Armaghani, D. and Mehrdanesh, A. (2020), "Development of a group method of data handling technique to forecast Iron ore price", *Appl. Sci.*, **10**(7), 2364-2384. https://doi.org/10.3390/app10072364.

Mehra, R. (1977), "Group method of data handling (GMDH): review and experience", *Proceedings of the 1977 IEEE conference on decision and control including the 16th symposium on adaptive processes and a special symposium on fuzzy set theory and applications*.

Ng, K. and Ksaibati, R. (2018), "Effect of soil layering on shorter-term pile setup", *J. Geotech. Geoenviron. Eng. - ASCE*, **144**(5), 1-12.

Ng, K.W., Roling, M., AbdelSalam, S.S., Suleiman, M.T. and Sritharan, S. (2013), "Pile setup in cohesive soil. I: experimental investigation", *J. Geotech. Geoenviron. Eng.*, **139**(2), 199-209. https://doi.org/10.1061/(ASCE)GT.1943-5606.0000751.

Pan, Y., Jiang, J., Wang, R., Cao, H. and Cui, Y. (2009), "A novel QSPR model for prediction of lower flammability limits of organic compounds based on support vector machine", *J. Hazard. Mater.*, **168**(2-3), 962-969. https://doi.org/10.1016/j.jhazmat.2009.02.122.

Pham, T.A., Ly, H.B., Tran, V.Q., Giap, L.V., Vu, H.L.T. and Duong, H.A.T. (2020), "Prediction of pile axial bearing capacity using artificial neural network and random forest", *Appl. Sci.*, **10**(5), 1871-1892. https://doi.org/10.3390/app10051871.

Randolph, M.F., Carter, J. and Wroth, C. (1979), "Driven piles in clay—the effects of installation and subsequent consolidation", *Geotechnique*, **29**(4), 361-393. https://doi.org/10.1680/geot.1979.29.4.361.

Razavi, M., Dehghani, A. and Khanmohammadi, M. (2009), "Simulation of thermal stratification in cisterns using artificial neural networks", *J. Energ. Heat Mass Transf.*, **31**, 201-210.

Razavi, S., Goshtasbi, K., Noorzad, A. and Ahangari, K. (2018), "Proposing new relationships to estimate the pressuremeter modulus of cohesive and cohesionless media", *Innov. Infrastruct. Solutions*, **3**(1), 67-78. https://doi.org/10.1007/s41062-018-0172-1.

Samson, L. and Authier, J. (1986), "Change in pile capacity with time: case histories", *Can. Geotech. J.*, **23**(2), 174-180. https://doi.org/10.1139/t86-027.

Skov, R. and Denver, H. (1988), "Time-dependence of bearing capacity of piles", *Proceedings of the 3rd International Conference on the Application of Stress-Wave Theory to Piles.* , Ottawa.

Smith, G.N. (1986), "Probability and statistics in civil engineering", *Collins professional and technical books*, **244**. https://ci.nii.ac.jp/naid/10007808566/.

Svinkin, M.R. (1996), "Setup and relaxation in glacial sand-discussion", *J. Geotech. Eng. - ASCE*, **122**(4), 319-321.

Svinkin, M.R. (2002), "Engineering judgement in determination of pile capacity by dynamic methods", *Proceedings of the Deep Foundations 2002: An International Perspective on Theory, Design, Construction, and Performance*, Virginia, USA.

Svinkin, M.R., Morgano, C.M. and Morvant, M. (1994), "Pile capacity as a function of time in clayey and sandy soils", *Proceedings of the Deep Foundations Institute Fifth International Conference and Exhibition on Piling and Deep Foundations*, Bruges, Belgium.

Svinkin, M.R. and Skov, R. (2000), "Set-up effect of cohesive soils in pile capacity", *Proceedings of the 6th international conference on application of stress waves to piles*, Florid, USA.

Tarawneh, B. (2013), "Pipe pile setup: database and prediction model using artificial neural network", *Soils Found.*, **53**(4), 607-615.

Tarawneh, B. (2018), "Gene expression programming model to predict driven pipe piles set-up", *Int. J. Geotech. Eng.*, **14**(538-544. https://doi.org/10.1080/19386362.2018.1460964.

Tarawneh, B. and Imam, R. (2014), "Regression versus artificial neural networks: predicting pile setup from empirical data", *KSCE J. Civil Eng.*, **18**, 1018-1027.

Titi, H.H. and Wije Wathugala, G. (1999), "Numerical procedure for predicting pile capacity—setup/freeze", *Transport. Res. Record*, **1663**(1), 25-32. https://doi.org/10.3141%2F1663-04.

Wang, J.X. (2017), "Growth-rate-dependent prediction of pile set-up and its application in driven pile foundation construction",

Geomech. Geoeng., **12**(2), 86-106. https://doi.org/10.1080/17486025.2016.1177208.
Xie, Y. (2011), Observed tip resistance at EOD & BOR using bottom tip gages for driven piles, University of Florida.

CC