Article

Prediction of Swelling Index Using Advanced Machine Learning Techniques for Cohesive Soils

Mohammed Amin Benbouras [1,2,*] and **Alexandru-Ionut Petrisor** [3]

1   École Normale Supérieure d'Enseignement Technologique de Skikda (ENSET), Skikda 21000, Algeria
2   Central Public Works Laboratory (LCTP), Algiers 16008, Algeria
3   Doctoral School of Urban Planning, Ion Mincu University of Architecture and Urbanism, 010014 Bucharest, Romania; alexandru.petrisor@uauim.ro
*   Correspondence: mouhamed_amine.benbouras@g.enp.edu.dz

Abstract: Several attempts have been made for estimating the vital swelling index parameter conducted by the expensive and time-consuming Oedometer test. However, they have only focused on the neuron network neglecting other advanced methods that could have increased the predictive capability of models. In order to overcome this limitation, the current study aims to elaborate an alternative model for estimating the swelling index from geotechnical physical parameters. The reliability of the approach is tested through several advanced machine learning methods like Extreme Learning Machine, Deep Neural Network, Support Vector Regression, Random Forest, LASSO regression, Partial Least Square Regression, Ridge Regression, Kernel Ridge, Stepwise Regression, Least Square Regression, and genetic Programing. These methods have been applied for modeling samples consisting of 875 Oedometer tests. Firstly, principal component analysis, Gamma test, and forward selection are utilized to reduce the input variable numbers. Afterward, the advanced techniques have been applied for modeling the proposed optimal inputs, and their accuracy models were evaluated through six statistical indicators and using K-fold cross validation approach. The comparative study shows the efficiency of *FS-RF* model. This elaborated model provided the most appropriate prediction, closest to the experimental values compared with other models and formulae proposed by the previous studies.

---

1. Introduction

The geotechnical study is the first phase of any structural project. It addresses the morphological, geological, local, and regional site conditions through a set of successive steps, aimed mainly at providing efficiently the necessary and sufficient information on the natural characteristics of a site. Moreover, it allows for determining all information and parameters necessary for the computation of foundations, leading to the stability of the structures under different risks, such as swelling [1,2]. The latter is considered a notoriously problematic phenomenon for any infrastructure foundation because of the considerable volume changes experienced by soils upon drying and wetting [3]. However, the integrated components of the geotechnical process for detecting swelling risk are very diverse. Moreover, these experimental processes are often time-consuming and, in many cases, the computation costs are very high. On the other hand, the Oedometer test suffers from several disadvantages, such as the requirement of sophisticated and expensive equipment, being very time-consuming, and requiring highly qualified laboratory workers [4,5]. A well-known example is the consolidated clay, which could take up to 30 days when the Oedometer test is conducted [6]. However, the mechanical soil properties are sometimes inexistent or insufficient, in circumstances where the engineer must provide a quick and important decision to rapidly handle the critical situation. These disadvantages

have been among the major obstacles for research, dissemination, and implementation of soil mechanics in engineering practice [4]. Furthermore, classical constitutive modelling based on the elasticity and plasticity theories has a limited ability to correctly simulate the real behavior of soils. This is attributed to reasons associated with the complexity of soil formation and its heterogeneity [7]. For these reasons, geotechnical researchers have tried to develop new safe economical models using systems influenced by many factors, like the technical, economic, environmental, and geometrical parameters. These systems are called, all together, Artificial Intelligence (AI) or Machine Learning (ML).

Many papers have been published in order to derive a different relationship and advanced models between the compression index (C_c) and the basic soil properties [6,8–11]. However, only a few widely accepted empirical equations have been proposed in the literature to estimate the swelling index (C_s) from physical soil parameters, such as the natural water content (*W*), the plasticity index (*PI*), the liquid limit (*WL*), the specific gravity, and others. Table 1 summarizes some proposed formulae for estimating C_s. However, major shortcomings have been observed. The large amount of correlations published with respect to the same parameters point to an inherent variability of its usage. Therefore, the application of correlation analysis in other conditions or sites could yield incorrect results [12,13]. In addition, these approaches generally depend on simplified assumptions, such as a linear behavior or production heuristics, which make regression analysis methods less effective when they are used for simulating the complex heterogeneous behavior of soil [14–17].

Table 1. Correlations proposed in the literature to estimate the swelling index C_s.

Variables	Correlations		Comments	References
$C_s(W_L,\ G_s)$	$C_s = 0.0463\left(\frac{W_L}{100}\right)G_s$	(1)	fine-grained soils	[18]
$C_s(I_P)$	$C_s = 0.00194(I_P - 4.6)$	(2)	fine-grained soils	[19]
$C_s(W_n)$	$C_s = 0.0133e^{0.036W_n}$	(3)	fine-grained soils	Isik 1 [20]
$C_s(e_0)$	$C_s = 0.0121e^{1.3131e_0}$	(4)	fine-grained soils	Isik 2 [20]
$C_s(Y_h)$	$C_s = 0.1257Y_h^{-2.8826}$	(5)	fine grained soils	Isik 3 [20]

On the other hand, the use of the *ANN* method in geotechnical engineering has witnessed a major development since the early nineties [21]. Several attempts using neural networks have led to impressive results. Between the important research works dealing with the swelling of soil, Işık has utilized one hidden layer of the *ANN* model by analyzing a database consisting of 42 test data for fine-grained soil. The selected input parameters included the initial void ratio (e_0) and *W*. The (2-8-1) *ANN* model (meaning two inputs, eight nodes in the hidden layer and one output) efficiently predicted C_s [20]. Das et al. have predicted the swelling pressure using an *ANN* model with an input layer containing *W*, dry density (Y_d), *WL*, *PI*, and clay fraction [22]. Kumar and Rani have developed an *ANN* model with one hidden layer to predict C_s and the C_c of clay by learning from 68 samples. They used the *FC*, *WL*, *PI*, maximum dry density, and optimum moisture content as an input layer. The suggested *ANN* model (5-8-1) provided better predictability in comparison with the multiple regression analysis (*MRA*) model [23]. Kurnaz et al. have used *ANN* models to estimate C_s and C_c from input layer including the *W*, e_0, *WL*, and *PI*. The proposed *ANN* model (4-6-2) has proven its efficiency in the prediction of C_c. Nevertheless, the predicted C_s values were not satisfactorily compared to the compression index [24]. Table 2 recapitulates the aforementioned proposed *ANN* models in the literature to estimate the swelling index C_s.

Table 2. Proposed *ANN* models in the literature to estimate the swelling index C_s.

Authors	Inputs	Targets	Architecture (Inputs–Nodes–Outputs)	Database	References
Işık (2009)	e_0 and W	C_s	2-8-1	42	[20]
Das et al. (2010)	W, Y_d, WL, PI, and FC	C_s	5-3-1	230	[22]
Kumar and Rani (2011)	FC, WL, PI, Y_{opt}, and W_{opt}	C_s and C_c	5-8-2	68	[23]
Kurnaz et al. (2016)	W, e_0, WL, and PI	C_s and C_c	4-6-2	246	[24]

The quality, learning ability, and effectiveness have made the use of the *ANN* method very useful. According to the author's knowledge, the previous studies have used only the *ANN* method for estimating C_s, although recent studies have showed that other techniques could have yielded more effective and accurate results than the *ANN* method in geotechnical applications [25–27]. Furthermore, the aforementioned studies have modeled C_s using a few input parameters and, therefore, ignored the different soil parameters that could increase the learning capacity of the network. Consequently, the complicated mechanism of the swelling phenomena has been oversimplified. Moreover, few samples have been used, meaning that the proposed models have a limited capacity to generalize new data not used in the few training data. Furthermore, they evaluated the predictive capacity of proposed models based on only one split to validate data learning. Therefore, the capacity of their model to overcome the over-fitting and under-fitting problems cannot be confirmed.

In view of the aforementioned shortcoming, this research sheds the light on the capability of the advanced machine learning methods to generate a reliable model, contributing to easily and effectively predict C_s, filling a gap in the literature where there is a lack in the use of the advanced machine learning methods in modeling swelling phenomenal. Consequently, the elaborated model offers plenty of benefits such as its reliability, and lowering the budget used to predict C_s from the easily obtained soil parameters and without the need to operate the odometer test.

2. Materials and Methods

2.1. Overview of the Methodology

Several-advanced machine learning techniques like Extreme Learning Machine (*ELM*), Deep Neural Network *(DNN)*, Support Vector Regression (*SVR*), Random Forest (*RF*), LASSO regression (*LASSO*), Partial Least Square Regression (*PLS*), Ridge Regression (*Ridge*) Kernel Ridge (*KRidge*), Stepwise Regression (*Stepwise*), Least Square Regression (*LS*), and Genetic Programming (*GP*) have been applied for modeling 875 samples. Multiple input parameters, including the wet density (Y_h), the dry density (Y_d), the degree of saturation (Sr), the plasticity index (PI), the water content (w), the void ratio (e_0), the liquid limit (WL), sample depth (Z), and the fine contents (FC) have been used. Firstly, from an effective viewpoint, the suitable input variables and nonlinear components are of considerable importance for efficient prediction. Thus, the Principal Component Analysis (*PSA*), Gamma Test (*GT*) and Forward selection (*FS*) methods have been used to select the optimal set of input variables. Afterward, the advanced machine learning techniques have been applied for modeling optimal inputs, and their accuracy models were evaluated through numerous statistical indicators. To assess the predictive capability of the best model, the k-fold cross-validation approach based on 10 splits has been utilized. Finally, to answer the question "Which input variables have the most or less influence on C_s through the proposed model?", a sensitivity analysis has been carried out using the step-by-step selection method.

2.2. Oedometer Test

The compressibility of soils is one of the most important phenomena in civil engineering. The Oedometer test is used to determine the compressibility properties of soil, which are usually described using C_c, C_s, and the coefficient of consolidation (C_v) [20,24]. The compressibility properties are used to predict how the settlement and the swelling will be held. A number of parameters influencing the swelling behavior have been reported in the

past, such as *W*, *e*, *WL*, *PI*, the type and amount of clay (*FC*), and others [8]. C_s is usually determined using the graphical analysis of compression and recompression curves in void ratio effective stress *(e = f(log(σ)))* plots [20,28] (see Figure 1), and used typically to estimate the consolidation settlement for soil layers using these formulae:

$$S_c = \frac{C_c H}{1 + e0} \log\left(\frac{\sigma'_{v0} + \Delta\sigma_v}{\sigma'_{v0}}\right) \text{ in normally consolidated soil if } \sigma'_{v0} = \sigma'_c \quad (6)$$

$$S_c = \frac{C_s H}{1 + e0} \log\left(\frac{\sigma'_{v0} + \Delta\sigma_v}{\sigma'_{v0}}\right) \text{ in over-consolidated soil if } \sigma'_{v0} + \Delta\sigma_v \leq \sigma'_c \quad (7)$$

$$S_c = \frac{C_s H}{1 + e0} \log\left(\frac{\sigma'_c}{\sigma'_{v0}}\right) + \frac{C_c H}{1 + e0} \log\left(\frac{\sigma'_{v0} + \Delta\sigma_v}{\sigma'_c}\right) \text{ if } \sigma'_{v0} + \Delta\sigma_v > \sigma'_c \quad (8)$$

where e_0: initial void ratio, $\Delta\sigma_v$: load increment, σ'_c: pre-consolidation pressure, σ'_{v0}: initial vertical effective stress, C_c: compression index, and C_s: swelling index.
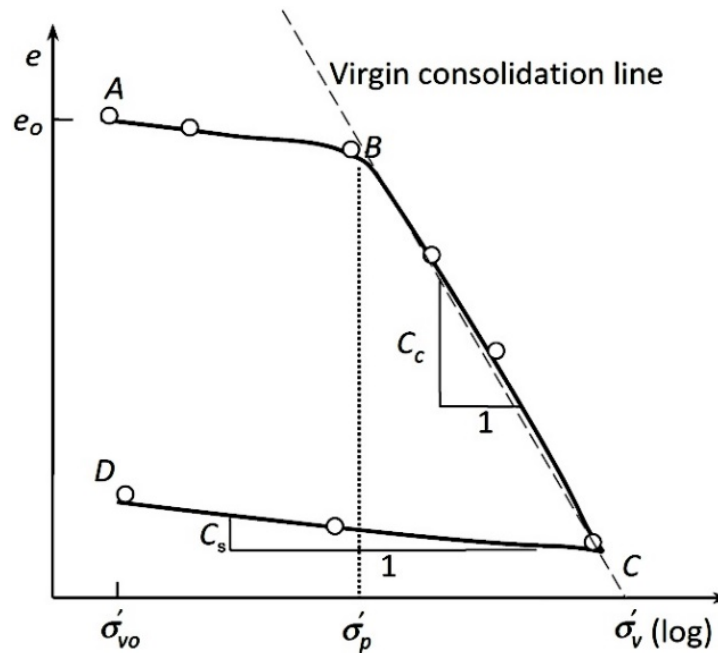

Figure 1. Determination of the swelling index from compression curves using the CD line.

2.3. Case Study

In the current study, a database of 875 Oedometer and other geotechnical tests has been collected from 570 boreholes. The reason for choosing the Northeast Algerian area is attributed to the widespread clay and marl geological formations, which generally suffered from shrinkage-swelling phenomena, resulting into a database including a wide range of data, sufficient for a reliable study. The sample depth (*Z*) ranges between 0.5 to 45 m with an average of 6.66 m. Figure 2 illustrates the distribution of the collected boreholes in the study area. The soil parameters used in this study were measured in the lab depending on the international or European standards.
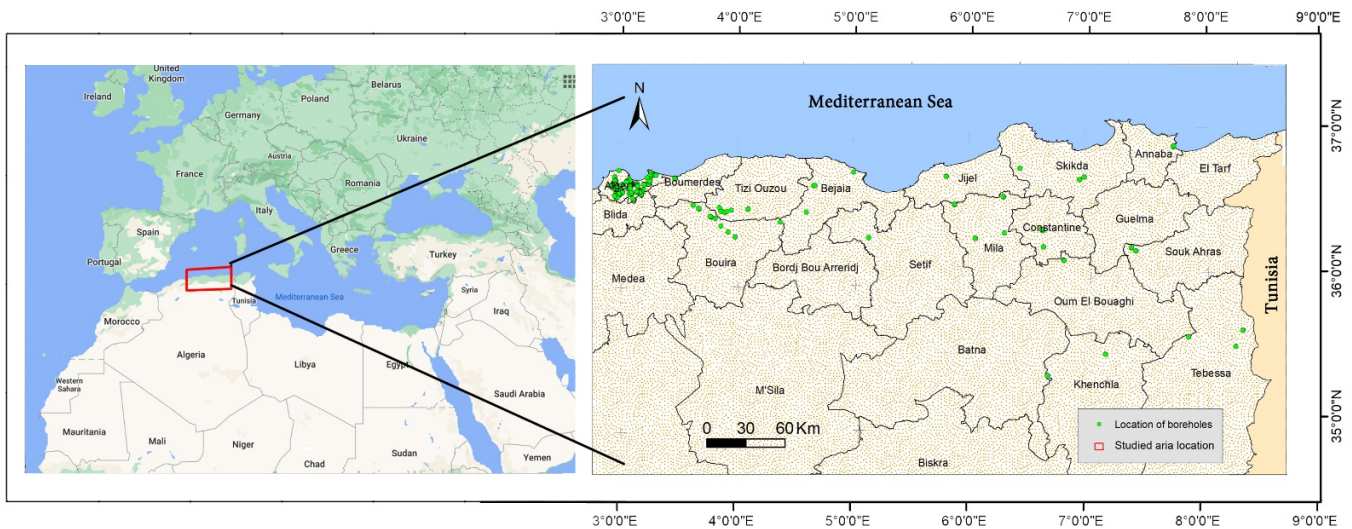
Figure 2. Map of the study area and collected boreholes sites.

2.4. Optimal Input Selections

2.4.1. Overview of Principal Component Analysis (*PCA*)

PCA is an exploratory statistics approach [29] considered between the multivariate statistical methods, which is generally utilized to decrease the complexity of the input variables. *PCA* is used in the practical case when we have a great number of information, and, in many fields, sometimes in conjunction with other methods [30]. Furthermore, this method reduces the input variables into a better interpretation of independent principal components (*PCs*) [31,32]. Instead of the direct use of input variables, we reduce them into *PCs* in order to use them as lower-dimensional inputs. Details for mastering the art of applying the *PCA* are published by other studies [33–36].

2.4.2. Overview of Gamma Test (*GT*)

The Gamma Test (*GT*) is an advanced approach for evaluating the variance of the noise, or the mean square error (*MSE*), which could be fulfilled without over-fitting. This method is very advantageous for assessing the nonlinear relationship between two random variables (input and target). Koncar (1997) [37] and Agalbjörn et al. (1997) [38] were the first reporting gamma test method and later many investigators have developed and detailed [35,36,39]. Only summarized information about *GT* is discussed here. Details for mastering the art of *GT* are presented in the aforementioned articles for the interested readers. By using necessary conditions detailed in the above-mentioned papers, the variance of the noise is specified by the bias of the regression between $\gamma(k)$ and $\delta(k)$, where $1 \leq k \leq p$. This variance is dubbed Γ. $\gamma(k)$ and $\delta(k)$ are presented in Equations (9) and (10):

$$\delta(k) = \frac{1}{M} \sum_{i=1}^{M} \left| x_{N[i,K]} - x_i \right|^2 \tag{9}$$

$$\gamma(k) = \frac{1}{2M} \sum_{i=1}^{M} \left| y_{N[i,K]} - y_i \right|^2 \tag{10}$$

$x_{N[i,K]}$ is the kth nearest neighbor of x_i, and $y_{N[i,K]}$ is the corresponding target. In order to calculate Γ, a least squares fit line should be carried out for the p points ($\delta(k)$, $\gamma(k)$). Afterward, the bias of the regression line will be easily estimated, which represents the gamma statistics parameter Γ. It is worth mentioning that Γ provides useful findings for building an accurate model. The smaller the value of Γ, the more appropriate is the input

set. In addition, V_{ratio} presents an important indication to assess the predictability of the selected target depending on utilized inputs, and illustrated as:

$$V_{ratio} = \frac{\Gamma}{\sigma^2(\mathrm{y})} \tag{11}$$

where $\sigma^2(\mathrm{y})$ is the output variance. Overall, in the current paper, by using the least value of Γ and V_{ratio}, the best-input combinations were chosen.

2.4.3. Overview of Forward Selection (*FS*)

The forward selection method is based on regression modelling. Firstly, input variables are considered in ascending order according to their correlation with the target variable (from the best to the weak correlated variable). Then, the best correlated variable with the target is chosen as the first one. The other variables are afterward added one by one as the second input, and we assess the predictive capability of the set according to their correlation development. Generally, the most significant variable increases the determination coefficient (R^2) is selected and added to the input set. In other words, if the R^2 is increased more than 5%, the new variable is accepted and added to the optimal input set. This step is repeated $N - 1$ times for assessing the impact of each parameter on modelling the target. Finally, among N tested inputs, the ones with optimum R^2 are accepted as the model input subset.

2.5. Machine Learning Methods

In the current study, several machine-learning methods have been used in order to conduct a reliable study and to propose a high performance model. Only the methods actually used are cited, followed by relevant references, which can be examined by the interested readers to better understand each one. The used methods are: Extreme Learning Machine (*ELM*) [40], Deep Neural Network (*DNN*) [6,41], Support Vector Regression (*SVR*) [42], Random Forest (*RF*) [43], LASSO regression (*LASSO*) [44], Partial Least Square Regression (*PLS*) [45], Ridge Regression (*Ridge*) [46], Kernel Ridge Regression (*KRidge*) [47], Stepwise Regression (*Stepwise*) [48], and Genetic Programing (*GP*) [6]. Matlab software has been used for programming the algorithms corresponding to each method, except *GP* when the HeuristicLab Interface has been utilized [49]. The controlling parameters of the *ELM*, *DNN*, *SVR*, *RF*, *LASSO*, *PLS*, *Ridge*, *KRidge*, *Stepwise*, and *PG* algorithms used in this study are listed in Table 3.

Table 3. Initial parameter settings for the algorithms.

Algorithms	Algorithm Parameters	Value
ELM	Hidden layers	$H = 1$
	hidden neurons	$N = 12$
	activation function	'linear'
	regulation parameter	$C = 0.02$
DNN	Hidden layers	$H = 2$
	hidden neurons in the first layer	$N1 = (1–20)$
	hidden neurons in the second layer	$N2 = (1–20)$
	activation function in the first layer	'Tansg'
	activation function in the second layer	'Tansg'
SVR	regulation parameter C	Series of C
	regulation parameter lambda	Series of lambda
	kernel function	'rbf'

Table 3. *Cont.*

Algorithms	Algorithm Parameters	Value
RF	nTrees mTrees	nTrees = 100 mTrees = 26
LASSO	lambda	series of lambda
PLS	PLS components	NumComp = 3 for PSO NumComp = 4 for GT and FS
Ridge	regularization parameter lambda	lambda = 1
KRidge	regularization parameter lambda kernel function parameter for kernel	lambda = 1 'linear' sigma = 2×10^{-7}
PG	Function set Population size Number of generations Genetic operators	$+, -, \times, \div$, power, ln, sqrt, sin, cos, tan 100 up to 500 1000 Reproduction, crossover, mutation

2.6. Statistical Performance Indicators

The prediction accuracy of the proposed models was evaluated through various statistical performance indicators and using graphical presentation. The statistical performance indicators are Mean absolute error (*MAE*), Root mean square error (*RMSE*), Index of scattering (*IOS*), Nash–Sutcliffe efficiency (*NSE*), Pearson correlation coefficient (*R*), and Index of agreement (*IOA*). They are expressed as follows [50]:

1. Mean absolute error (*MAE*):

$$MEA = \frac{1}{N} \sum_{i=1}^{N} |Y_{tar,i} - Y_{out,i}| \; (0 < MAE < \infty) \tag{12}$$

2. Root mean square error (*RMSE*):

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (Y_{tar,i} - Y_{out,i})^2} \; (0 < RMSE < \infty) \tag{13}$$

3. Index of scattering (*IOS*):

$$IOS = \frac{\sqrt{\frac{1}{N} \sum_{i=1}^{N} (Y_{tar,i} - Y_{out,i})^2}}{\overline{Y_{tar}}} \; (0 < RMSE < \infty) \tag{14}$$

4. Nash–Sutcliffe efficiency (*NSE*):

$$NSE = 1 - \frac{\sum_{i=1}^{N} (Y_{tar,i} - Y_{out,i})^2}{\sum_{i=1}^{N} (Y_{tar,i} - \overline{Y_{tar}})^2} \; (-\infty < NSE < 1) \tag{15}$$

5. Pearson correlation coefficient (*R*):

$$R = \frac{\sum_{i=1}^{N} ((Y_{tar,i} - \overline{Y_{tar}})(Y_{out,i} - \overline{Y_{out}}))}{\sqrt{\sum_{i=1}^{N} ((Y_{tar,i} - \overline{Y_{tar}})^2 (Y_{out,i} - \overline{Y_{out}})^2)}} \; (-1 < R < 1) \tag{16}$$

6. Index of agreement (*IOA*):

$$IOA = 1 - \frac{\sum_{i=1}^{N} (Y_{tar,i} - Y_{out,i})^2}{\sum_{i=1}^{N} \left(\sum_{i=1}^{N} |Y_{out,i} - \overline{Y_{tar}}| + \sum_{i=1}^{N} |Y_{tar,i} - \overline{Y_{tar}}| \right)^2} \; (0 < IOA < 1) \tag{17}$$

where $Y_{tar,i}$, $Y_{out,i}$, $\overline{Y_{tar}}$, and $\overline{Y_{out}}$ present the target, output, mean of target, and mean of output swelling index values for N data samples, respectively. Furthermore, the proposed machine learning model having the lowest value of *RMSE*, *IOS*, and *MAE* and the highest value of *IOA*, *NSE*, and *R* presents the better one and the closest to the experimental data.

Subsequently, after selecting the best model using statistical performance indicators, its predictive capacity is assessed using the *K*-fold cross-validation approach. This shows more accuracy and robustness for evaluating the capability of performance predictive of the best model by testing the existence of over-fitting and under-fitting problem in data learning [51,52]. The method consists of separating the data set into *k* equal sizes splits. Hence, for each one, $K-1$-folds are used for training and the last one for validation. This operation is repeated successively until the use of all split for the validation step [53]. The main advantage of this method is that all collected data are utilized in both the training and validation steps [52]. Breiman and Spector (1992) have demonstrated that $K = 10$ or $K = 5$-fold cross validation is the best choice for model evaluation [51]. In this work, we chose *K*-fold cross-validation with $K = 10$ for comparing the predictive capacity of each model.

2.7. Methodology

In order to find the most appropriate model to predict C_s of soil using previously mentioned geotechnical parameters as an input, the methodology consisted of the following steps:

1. Creation of a geotechnical database of Algerian soil, collected from different laboratories around the geotechnical constructions projects in progress or completed before.
2. Selecting the optimal input variables using Principal component analysis (*OSA*), Gamma Test (*GT*), and Forward selection (*FS*) has been used.
3. Analyzing selected optimal inputs using several machine learning methods. The *ELM*, *DNN*, *SVR*, *RF*, *LASSO*, *PLS*, *Ridge*, *KRidge*, *Stepwise*, and *PG* methods have been used in this step for proposing 30 models.
4. Determine the most appropriate model for predicting the C_s value between the thirty proposed models using important statistical performance indicators as *MAE*, *RMSE*, *IOS*, *NSE*, *R*, and *IOA*.
5. Assessing the predictive capacity of the best model to overcome under-fitting and over-fitting problem by using the K-fold cross validation approach with $K = 10$.
6. Doing a sensitivity analysis by utilizing the step-by-step method to know the most or less influenced input on C_s through the proposed model.

The research methodological for determining the most appropriate model to estimate C_s is systematically described in Figure 3.
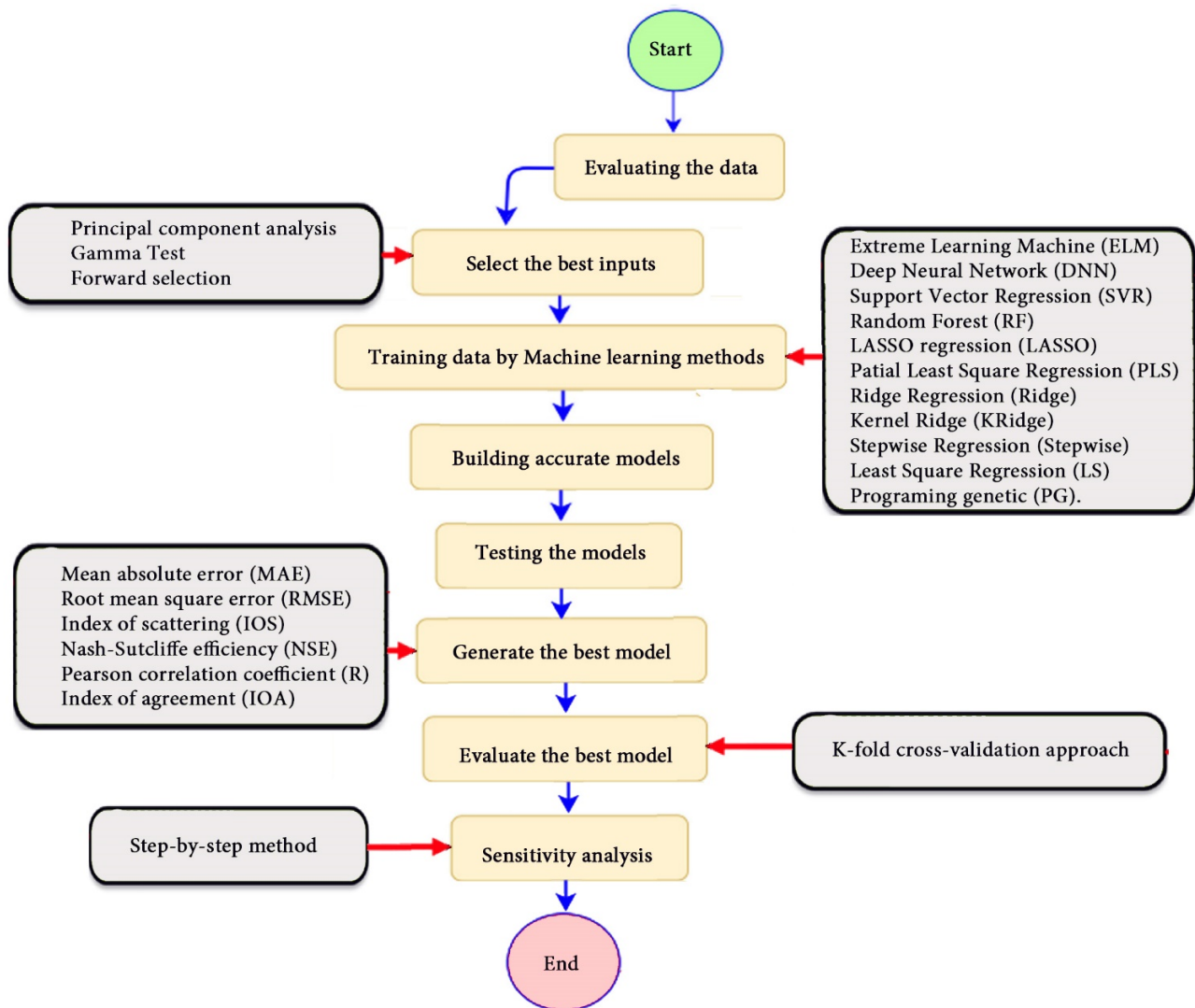
Figure 3. Flowchart of key steps for the research methodology to estimate C_s.

3. Results

3.1. Database Compilation

In the current study, a database of 875 Oedometer and other geotechnical tests has been collected resulting into a database including a wide range of data, sufficient for a reliable study. Table 4 displays the descriptive statistics of collected samples, determined using SPSS such as the mean, median, mode, standard deviation, variance, skewness, error of skewness, kurtosis, kurtosis error, range, minimum, and maximum. The skewness values indicate that all variables are evenly distributed. Moreover, the results point out that the database includes a wide range of data. Therefore, the collected dataset can be used to enhance novel empirical models and assess the predictive capacity of existing formulae. North East Algerian soil can be characterized as a dense soil with an average wet density of 1.67 according to the European norms [54]. Moreover, the soil appears to be plastic with an average plastic index of 26.09 according to SONGLIRAT's classification [55]. In the classification based on swelling, according to the European norms, the soils could be classified as a swelling soil, with an average C_s equal to 0.044 [55].

Table 4. Descriptive statistics of collected samples.

		Sr	Y_h	Y_d	W	e_0	FC	W_L	PI	C_s
N	Valid	875	875	875	875	875	875	875	875	875
	Missing	0	0	0	0	0	0	0	0	0
Mean		89.45	2.01	1.67	20.61	0.63	86.55	50.11	26.09	0.0443
Std. Error of Mean		0.397	0.003	0.004	0.164	0.005	0.572	0.338	0.233	0.00072
Median		94.00	2.01	1.67	20.00	0.62	94.00	50.00	26.00	0.0399
Mode		100.00	2.04	1.69	20.00	0.61	98.00	58.00	29.00	0.04
Std. Deviation		11.77	0.09	0.13	4.86	0.13	16.92	10.00	6.89	0.01910
Variance		138.54	0.01	0.02	23.64	0.02	286.23	100.09	47.43	0.000
Skewness		−1.32	0.10	0.29	0.36	0.21	−1.75	−0.08	−0.12	0.686
Std. Error of Skewness		0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.092
Kurtosis		1.09	−0.20	−0.09	−0.08	−0.29	2.41	−0.28	−0.43	0.073
Std. Error of Kurtosis		0.165	0.165	0.165	0.165	0.165	0.165	0.165	0.165	0.183
Range		64.45	0.57	0.73	26.00	0.79	78.00	64.31	38.00	0.10
Minimum		41.00	1.70	1.34	8.00	0.23	22.00	19.00	7.00	0.01
Maximum		100.00	2.27	2.07	34.00	1.02	100.00	83.31	45.00	0.11
Percentiles	25	84	1.95	1.58	17.10	0.53	81.82	42.81	21.50	0.03
	50	94	2.01	1.67	20.00	0.62	94.00	50.00	26.00	0.041
	75	99	2.075	1.75	23.85	0.71	98.00	58.00	31.38	0.057

3.2. Correlation between C_s and Geotechnical Parameters

To statistically predict the correlation between C_s and soil properties, the SPSS software has used. The cross-correlation between C_s and soil parameters is presented in Figure 4, which could provide a descriptive overview of the data distribution. The results indicate a positive relationship between C_s and others parameters, except for Y_d and Y_h, which seem to have a negative relationship (see Figure 4); this indicates that an increase in these parameters tends to proportionally increase C_s. The Spearman correlation coefficient R and its significance between C_s and other geotechnical parameters are presented in Table 5. The results show that the significance is less than 0.05 except Z, meaning that the majority of correlations are statistically significant. On the other hand, according to Smith' classification (1986), the C_s is moderately correlated to the aforementioned soil parameters, except Sr and Z which are poorly correlated. The results indicate that these parameters could have a complex nonlinear correlation with C_s. Moreover, in order to mathematically simulate the complex swellings phenomena, new advanced machine learning methods should be applied.

Table 5. Matrix of correlation between the geotechnical parameters (* Correlation significant at $\alpha = 0.05$; ** Correlation significant at $\alpha = 0.01$).

		Sr	Z	Y_h	Y_d	W	e_0	FC	WL	PI	C_s
Sr	R	1	0.199 **	0.170 **	−0.197 **	0.582 **	−0.06	0.194 **	0.082 *	−0.01	0.138 **
	Sig. (2-tailed)		0	0	0	0	0.06	0	0.02	0.78	0
Z	R	0.199 **	1	0.281 **	0.164 **	0.03	0.02	0.127 **	0	−0.06	0.02
	Sig. (2-tailed)	0		0	0	0.33	0.54	0	1	0.07	0.58
Y_h	R	0.170 **	0.281 **	1	0.877 **	−0.481 **	−0.579 **	−0.317 **	−0.275 **	−0.267 **	−0.230 **
	Sig. (2-tailed)	0	0		0	0	0	0	0	0	0
Y_d	R	−0.197 **	0.164 **	0.877 **	1	−0.803 **	−0.659 **	−0.384 **	−0.348 **	−0.292 **	−0.324 **
	Sig. (2-tailed)	0	0	0		0	0	0	0	0	0
W	R	0.582 **	0.03	−0.481 **	−0.803 **	1	0.633 **	0.385 **	0.372 **	0.264 **	0.349 **
	Sig. (2-tailed)	0	0.33	0	0		0	0	0	0	0

Table 5. *Cont.*

		Sr	Z	Y_h	Y_d	W	e_0	FC	WL	PI	C_s
e_0	R	−0.06	0.02	−0.579 **	−0.659 **	0.633 **	1	0.227 **	0.321 **	0.260 **	0.216 **
	Sig. (2-tailed)	0.06	0.54	0	0	0		0	0	0	0
FC	R	0.194 **	0.127 **	−0.317 **	−0.384 **	0.385 **	0.227 **	1	0.429 **	0.412 **	0.387 **
	Sig. (2-tailed)	0	0	0	0	0	0		0	0	0
WL	R	0.082 *	0	−0.275 **	−0.348 **	0.372 **	0.321 **	0.429 **	1	0.914 **	0.553 **
	Sig. (2-tailed)	0.02	1	0	0	0	0	0		0	0
PI	R	−0.01	−0.06	−0.267 **	−0.292 **	0.264 **	0.260 **	0.412 **	0.914 **	1	0.512 **
	Sig. (2-tailed)	0.78	0.07	0	0	0	0	0	0		0
C_s	R	0.138 **	0.02	−0.230 **	−0.324 **	0.349 **	0.216 **	0.387 **	0.553 **	0.552 **	1
	Sig. (2-tailed)	0	0.58	0	0	0	0	0	0	0	
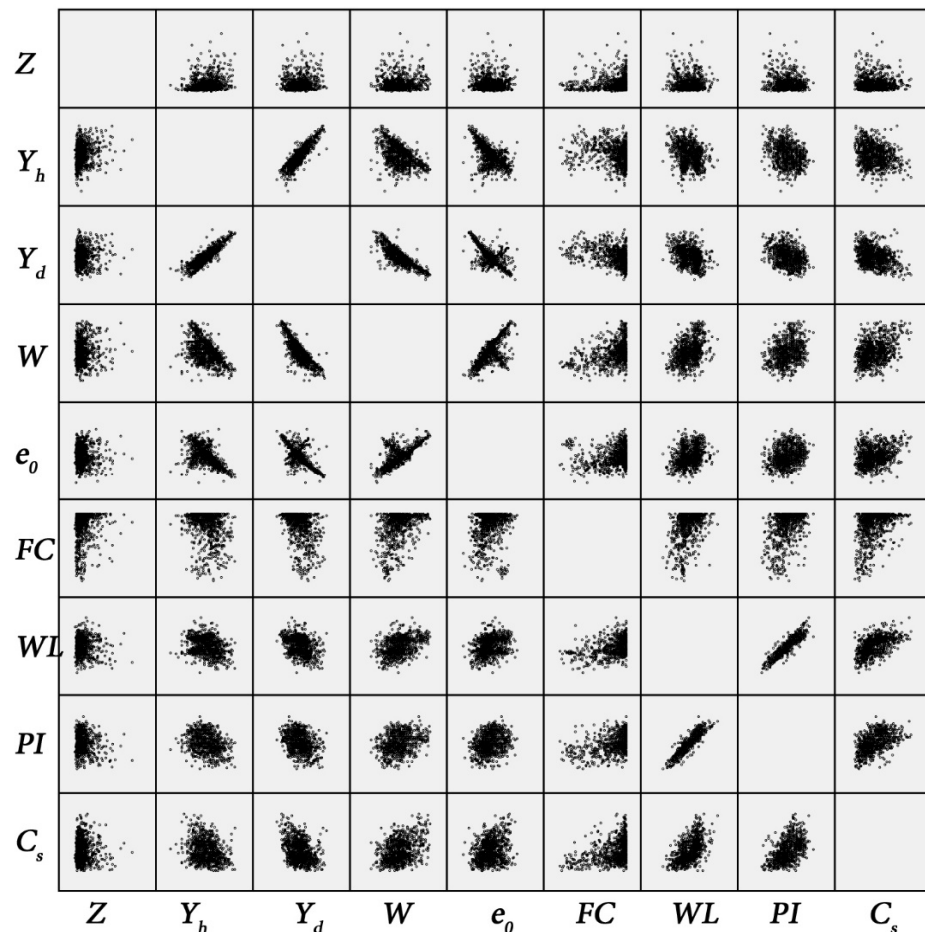
Figure 4. The cross-correlation between C_s and soil parameters.

3.3. Optimal Input Selection

3.3.1. Optimal Input Selection Using Principal Component Analysis

According to our knowledge, several methods were utilized in the literature to characterize the proper factors that affect the modeling precision. Therefore, the current study used the approach of selecting eigenvalues equal to or greater than 1, as illustrated in Table 6. Holland (2008) stated according to [56] that the eigenvalues are applied to condense the variance where the highest eigenvalues (1 and above) are deemed for any analysis by eigenvectors ranking in any correlation matrix. Figure 5 presents eigenvalues of each factor, which show that nine input variables correspond to nine eigenvalues. Likewise,

Table 6 displays the Eigenvalue and percentage of data in each factor. It is obvious from the table that the main first three variables explain more than 75% of the factors. Similarly, the results showed that five factors have a significant percentage contribution of more than 94. The findings indicate that the best model includes three *PCs* as input variables according to principal component analysis.

Table 6. Eigenvalue and percentage of data explained by each factor.

Number	Eigenvalue	% Variance	% Cumulative Variance
1	3.81	42.34	42.34
2	1.61	17.85	60.19
3	1.48	16.44	76.63
4	0.92	10.21	86.84
5	0.65	7.23	94.07
6	0.42	4.64	98.71
7	0.08	0.91	99.62
8	0.03	0.29	99.91
9	0.01	0.09	100.00
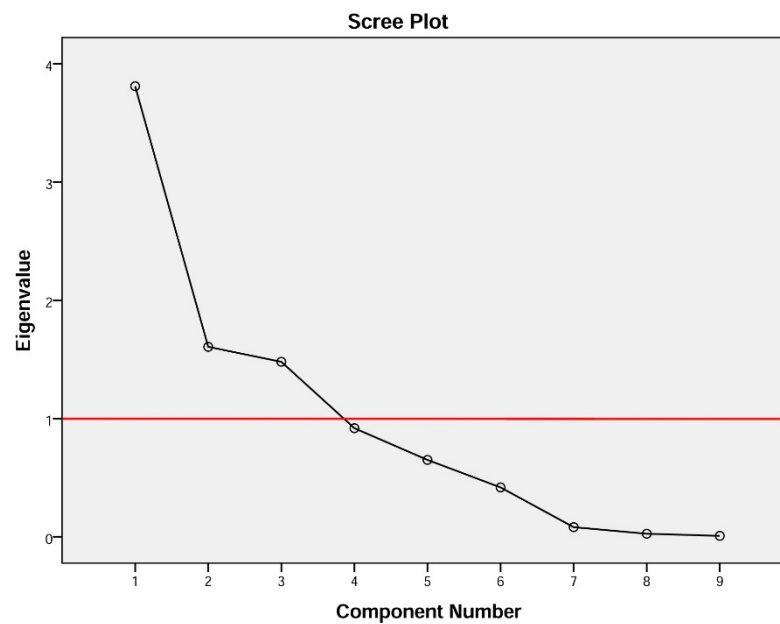
Figure 5. The eigenvalue versus the number of factors (The black line represents the eigenvalues of each factor and the red one shows the eigenvalues equal to or greater than 1).

3.3.2. Optimal Input Selection Using the Gamma Test

In this part, the influence of the separated inputs was evaluated by building the ten various combinations (Table 7) through diverse input parameters *(Sr, Z, Y_h, Y_d, W, e_0, FC, WL* and *PI)*. It was noticed from Table 7 that the first combination contains all nine inputs (dubbed an initial set). Similarly, the second one included eight input parameters (*All-Sr*), which means omitted *Sr* from the initial set; the fourth combination comprises all inputs except Y_h, and so on for rest of the combinations as presented in Table 7. The results of *GT* analysis shown in Table 7 prove that the parameters *W, FC, WL,* and *PI* have an important effect on the target (C_s). The four input parameters are chosen according to the maximum value of gamma statistics (Γ), and V_{ratio}. Based on this finding, four new combinations were tested in Table 8 for the sake of determining the optimal input set. In this case, the best set was designated based on the minimum of Γ and V_{ratio}. The outcomes of *GT* on four diverse combinations are illustrated in Table 8. The findings indicate that the *WL, PI,*

FC, and *W* set had the lowest value of gamma statistics ($\Gamma = 1.3524 \times 10^{-4}$, $V_{ratio} = 0.3714$, and Mask = 1111), and used as input variables for modelling C_s according to the Gamma Test method.

Table 7. Optimal input variable nomination using GT.

Input Parameters	Gamma Test Statistics		
	Γ	V_{ratio}	Mask
All	0.00014759	0.4054	111111111
All-Sr	0.00014653	0.4025	011111111
All-Z	0.00015130	0.4156	101111111
All-Y_h	0.00014672	0.4030	110111111
All-Y_d	0.00014689	0.4034	111011111
All-W	0.00017471	0.4798	111101111
All-e_0	0.00014712	0.4041	111110111
All-FC	0.00019292	0.5299	111111011
All-WL	0.00017584	0.4829	111111101
All-PI	0.00016223	0.4456	111111110

Table 8. GT statistics of different input models.

Input Parameters	Gamma Test Statistics		
	Γ	V_{ratio}	Mask
WL	0.00020688	0.5944	1000
WL, PI	0.00018845	0.5176	1100
WL, PI, FC	0.00017979	0.4938	1110
WL, PI, FC, W	0.00013524	0.3714	1111

3.3.3. Optimal Input Selection Using Forward Selection

In this part, the *FS* technique is utilized as a nonlinear input selection approach in order to choose the optimal input set between nine parameters. The *ANN* method with one layer is used to implement the nonlinear analysis of each set according to the empirical rule proposed by Kanellopoulos and Wilkinson (1997). The authors have demonstrated that the optimal node number in a simple network is twice the number of input parameters [57]. Firstly, correlation between target and other geotechnical parameters is estimated. Secondly, the parameter having the highest correlation coefficient, (i.e., *WL* with $R = 0.55$) is chosen as the first and the most significant parameter. Then, the other inputs are added into the model one by one and launch a new *ANN* modelling. The one providing the best modeling result (high determination coefficient (R^2)) is selected as a new input and gathered into the formerly selected parameters. This work is repeated several times until that appending other parameter to the input set does not improve the modeling performance. Hence, if the determination coefficient increases more than 5%, the novel parameter is selected. Finally, input parameters having the most heavy influence on the target are selected and others ones are rejected. Table 9 shows the results of the forward selection procedure of different input models. The findings indicate that *WL*, *Yd*, *W*, and *PI* are selected as inputs for modeling C_s according to the forward selection procedure, and the other parameters are eliminated.

Table 9. Results of forward selection procedure of different input models.

Input Subset	ANN Architecture	R^2	Decision
WL	1-2-1	0.327	*WL* selected
WL, Sr	2-4-1	0.332	*Sr* rejected
WL, Z	2-4-1	0.328	*Z* rejected
WL, Y_d	2-4-1	0.38	*Y_d* selected
WL, Y_d, Y_h	3-6-1	0.41	*Y_h* rejected
WL, Y_d, W	3-6-1	0.444	*W* selected
WL, Y_d, W, e_0	4-8-1	0.47	*e_0* rejected
WL, Y_d, W, FC	4-8-1	0.46	*FC* rejected
WL, Y_d, W, PI	4-8-1	0.498	*PI* selected.

3.4. Swelling Index Prediction through AI Models

The controlling parameters of the *ELM, DNN, SVR, RF, LASSO, PLS, Ridge, KRidge, Stepwise,* and *GP* algorithms used in this study are listed in Table 3. The performance of each models for selected optimal inputs (*PSO, GT* and *FS*) during training and validation phases is presented in Table 10. Six statistical performance indicators have been used to compare between proposed models in order to select the best one. We mention the Mean Absolute Error (*MAE*), Root Mean Square Error (*RMSE*), Index of Scattering (*IOS*), Nash–Sutcliffe Efficiency (*NSE*), Pearson correlation coefficient (*R*), and Index of Agreement (*IOA*). The data were divided into two parts, i.e., 80% for training and 20% for validation (700 samples for training and 175 for validation). It was indicated from Table 10 that the C_s modeled with several machine learning methods produced *MAE* (5.6×10^{-3} to 12.4×10^{-3}), *RMSE* (0.007 to 0.0154), *IOS* (0.165 to 0.355), *NSE* (-0.88 to 0.75), *R* (0.59 to 0.94), and *IOA* (0.69 to 0.95) in the training phase. Similarly, in the validation phase, we obtain *MAE* (10.6×10^{-3} to 13.6×10^{-3}), *RMSE* (0.013 to 0.017), *IOS* (0.298 to 0.363), *NSE* (-1.68 to 0.13), *R* (0.53 to 0.71), and *IOA* (0.69 to 0.82). Furthermore, the finding clearly indicates that the *FS-RF* (the optimal inputs of FS method trained by RF method) presents the most appropriate model that gives the highest accuracy in terms of *MAE* ($5.6 \times 10^{-3}/10.6 \times 10^{-3}$), *RMSE* (0.007/0.013), *IOS* (0.165/0.298), *NSE* (0.75/0.13), *R* (0.94/0.71), and *IOA* (0.95/0.82) during the training/validation phase. In addition, the most appropriate *FS-RF* model clearly follows the criteria of minimum values of error metrics (*MEA, RMSE* and *IOS*) and higher values of *NSE, R,* and *IOA* for the phase of training and validation. Furthermore, this model is closely followed by the *FS-DNN* model, which gives an acceptable accuracy and ranked second. Moreover, the results reveal the poor performance of the *PSO-Step* model in predicting the swelling index. With respect to the performance of machine learning models, during the training phase, the hierarchy follows the order of *FS-RF, GT-RF, PRO-RF, FS-DNN, GT-DNN, PSO-DNN, PSO-SVR, FS-GP, GT-GP, PSO-GP, GT-SVR, GT-Step, FS-SVR, GT-Lasso, GT-PLS, FS-PLS, GT-Kridge, FS-Kridge, PSO-LS, GT-LS, PSO-ELM, FS-ELM, PSO-Lasso, PSO-Kridge, PSO-Ridge, GT-Ridge, FS-Ridge, FS-LS, FS-Step, FS-Lasso, PSO-PLS, GT-ELM,* and *PSO-Step.* Finally, the scatter plots between target and output swelling index values of each model are presented in Appendix A.

3.5. Evaluating the Best Fitted Model Using the K-fold Cross Validation Approach

The 10-fold cross validation approach was efficiently used to assess the predictive capacity of the proposed model. We must stress out the fact that previous studies interested in estimating C_s [20,22–24] have evaluated the predictive capacity of their proposed models depending on a single split. Subsequently, the capacity of their models in overcoming the over-fitting and under-fitting problems could not be ascertained. Figure 6 shows the performance measures of the best *FS-RF* models using 10-fold cross validation depending on training and validation data for each split. The findings prove the performance of the proposed model. The fact that *R* ranges between 0.92 and 0.94 for training data, and between 0.62 and 0.75 for validation data in the 10 splits, indicates the predictive capacity

of the most appropriate *FS-RF* model for learning data, generating new validation data, and overcoming the over-fitting or under-fitting problems.

Table 10. Performance indicator values of AI models for predicting the Swelling index.

	PSO						GT						FS					
	MAE $\times 10^{-3}$	*RMSE*	*IOS*	*NSE*	*R*	*IOA*	*MAE* $\times 10^{-3}$	*RMSE*	*IOS*	*NSE*	*R*	*IOA*	*MAE* $\times 10^{-3}$	*RMSE*	*IOS*	*NSE*	*R*	*IOA*
Training																		
DNN	9.5	0.013	0.283	0.56	0.75	0.85	8.3	0.0113	0.251	0.64	0.80	0.88	8.4	0.011	0.245	0.67	0.82	0.89
ELM	12	0.015	0.355	−0.67	0.61	0.72	12	0.0153	0.340	−1.33	0.61	0.69	12,2	0.015	0.34	−0.88	0.61	0.72
Lasso	12.2	0.0154	0.344	−0.76	0.60	0.72	12.1	0.0151	0.335	−0.65	0.61	0.73	12.1	0.015	0.34	−0.84	0.59	0.71
PLS	11.9	0.015	0.338	−0.81	0.6	0.71	12.1	0.0152	0.339	−0.66	0.61	0.73	12	0.015	0.34	−0.69	0.61	0.73
RF	5.8	0.0075	0.168	0.72	0.94	0.95	5.7	0.0075	0.167	0.72	0.94	0.95	5.6	0.007	0.165	0.75	0.94	0.95
Kridge	12	0.015	0.342	−0.73	0.61	0.72	12	0.015	0.343	−0.71	0.61	0.73	12	0.015	0.334	−0.67	0.61	0.73
Ridge	12.2	0.015	0.341	−0.77	0.60	0.72	11.9	0.015	0.337	−0.71	0.61	0.72	11.9	0.015	0.343	−0.74	0.60	0.72
LS	12	0.0152	0.343	−0.63	0.62	0.73	12.1	0.015	0.34	−0.64	0.61	0.73	12	0.015	0.341	−0.74	0.60	0.72
Step	12.1	0.0153	0.346	−0.86	0.59	0.71	11.9	0.015	0.33	−0.61	0.62	0.74	12.4	0.015	0.343	−0.76	0.60	0.72
SVR	10.3	0.014	0.32	0.12	0.7	0.8	11.8	0.015	0.33	−0.57	0.64	0.75	11.8	0.015	0.331	−0.63	0.63	0.74
GP	11.3	0.014	0.305	−0.22	0.67	0.78	11.1	0.014	0.302	0.46	0.68	0.79	11	0.014	0.299	0.47	0.69	0.8
Validation																		
DNN	10.8	0.0135	0.304	0.47	0.69	0.82	11.2	0.0149	0.347	0.41	0.66	0.80	10,3	0.014	0.312	0.47	0.70	0.82
ELM	11.4	0.014	0.346	−0.53	0.6	0.73	12.5	0.015	0.35	−1.68	0.64	0.69	11.7	0.015	0.331	−0.93	0.62	0.72
Lasso	11.5	0.014	0.318	−0.65	0.64	0.74	12	0.015	0.325	−0.55	0.64	0.75	11.6	0.015	0.346	−0.85	0.67	0.74
PLS	12.3	0.016	0.354	−0.66	0.65	0.74	11.7	0.0146	0.312	−0.59	0.64	0.74	12.2	0.015	0.339	−0.59	0.61	0.73
RF	11	0.0138	0.308	−0.29	0.70	0.79	11.1	0.0143	0.32	−0.17	0.70	0.8	10.6	0.013	0.298	0.13	0.71	0.82
Kridge	12	0.0156	0.335	−1.14	0.61	0.7	11.7	0.015	0.316	−0.86	0.65	0.73	12.4	0.015	0.34	−0.93	0.60	0.71
Ridge	11.8	0.0144	0.322	−0.37	0.63	0.753	12	0.015	0.34	−0.75	0.66	0.74	12.1	0.014	0.333	−0.79	0.63	0.73
LS	12	0.015	0.334	−0.49	0.57	0.72	12	0.015	0.33	−0.49	0.62	0.75	12.2	0.015	0.33	−0.55	0.63	0.75
Step	11.8	0.0145	0.315	−0.47	0.67	0.76	12.3	0.015	0.34	−0.67	0.61	0.73	11	0.014	0.31	−0.54	0.64	0.75
SVR	12.1	0.016	0.34	−0.76	0.53	0.69	12.6	0.015	0.34	−0.51	0.59	0.71	12.6	0.015	0.344	−0.49	0.58	0.71
GP	12.8	0.016	0.36	−0.95	0.53	0.62	13.6	0.0162	0.357	−1.14	0.55	0.60	13.5	0.017	0.363	−0.72	0.55	0.62
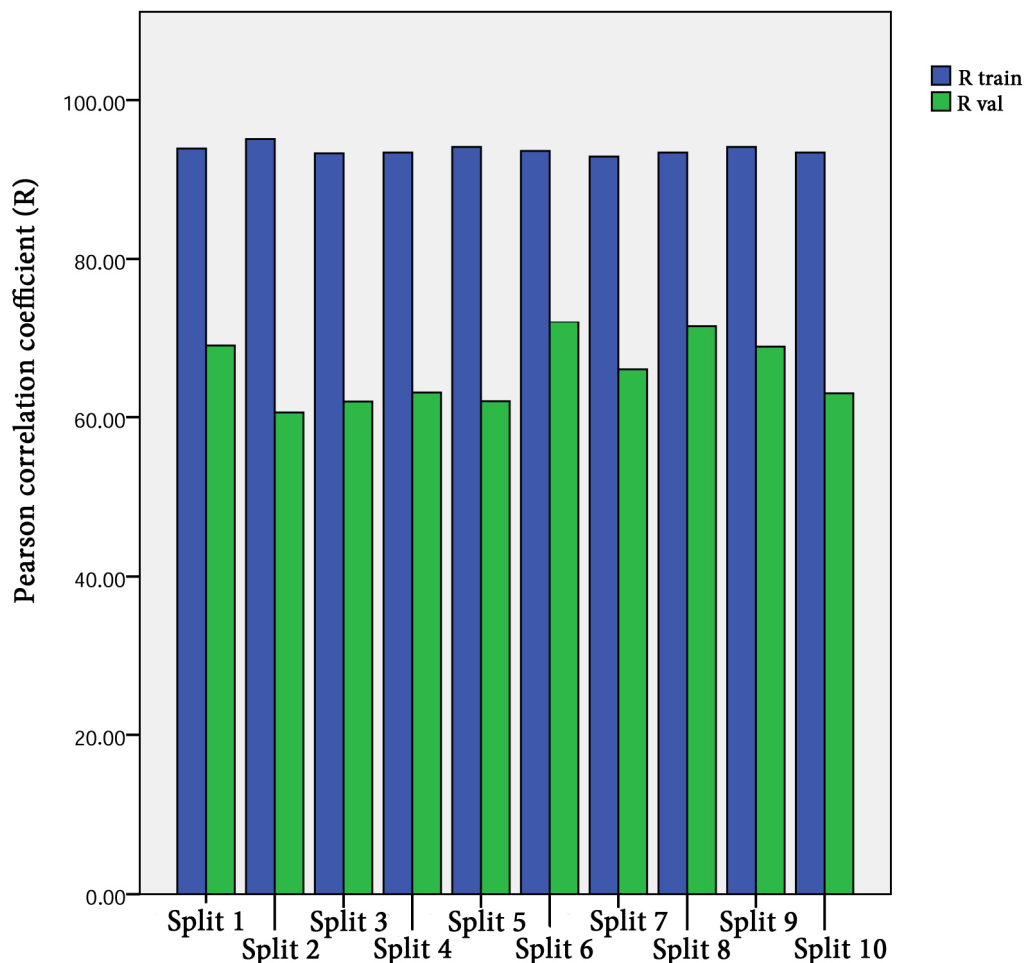
Figure 6. Performance measures of *FS-RF* models using K-fold cross validation with K = 10.

3.6. Comparison between the Proposed Models and Empirical Formulae

In order to test the effectiveness of the most appropriate *FS-RF* model, a comparative study was performed with some empirical formulae suggested by the previous studies for predicting C_s. These formulae were presented in Table 1. Table 11 illustrates the statistical results of $C_{s,estimated}/C_{s,measured}$ for aforementioned empirical equations in comparison with the proposed *FS-RF* and *FS-DNN* models. The mean and standard deviation of the ratio $C_{s,estimated}/C_{s,measured}$ could be useful evidence for evaluating the predictive capacity. The closer the mean value to one and the standard deviation to zero, the better is the model. The results show that the most appropriate *FS-RF* model is the one with the minimum standard deviation $\sigma_{FS-RF} = 0.25$, in addition to being the closer mean value to 1 average($FS-RF$) = 1.07. The other equations indicate a poor predictive capacity, yielding a mean value in the range of 0.45–1.7, and standard deviation value between 0.26–0.99. Equally, the box plot of $C_{s,estimated}/C_{s,measured}$ of aforementioned formulae is displayed in Figure 7. This graphical representation provides an overview of the dispersion and skewness of every model. The scattering of the most appropriate *FS-RF* model appears only to be slightly regular and close to one, and is described by a shorter box than the others. The large box distant from 1, increasing to five in certain models, shows little variation in predicting C_s. Data characterized by circles and stars in the figure denote the extreme and extra-extreme value.

Table 11. Statistical results of the ratio $C_{s,estimated}/C_{s,measured}$ for some proposed empirical formulae.

Equations No.	Study	Average	Standard Deviation
	FS-RF (in the current study)	1.07	0.25
	FS-DNN (in the current study)	1.08	0.34
(2)	Cozzolino 1961	1.096	0.761
(1)	Nagaraj and Srinivasa 1986	1.695	0.989
(3)	Işık1 2009	0.81	0.497
(4)	Işık2 2009	0.766	0.461
(5)	Işık3 2009	0.446	0.259
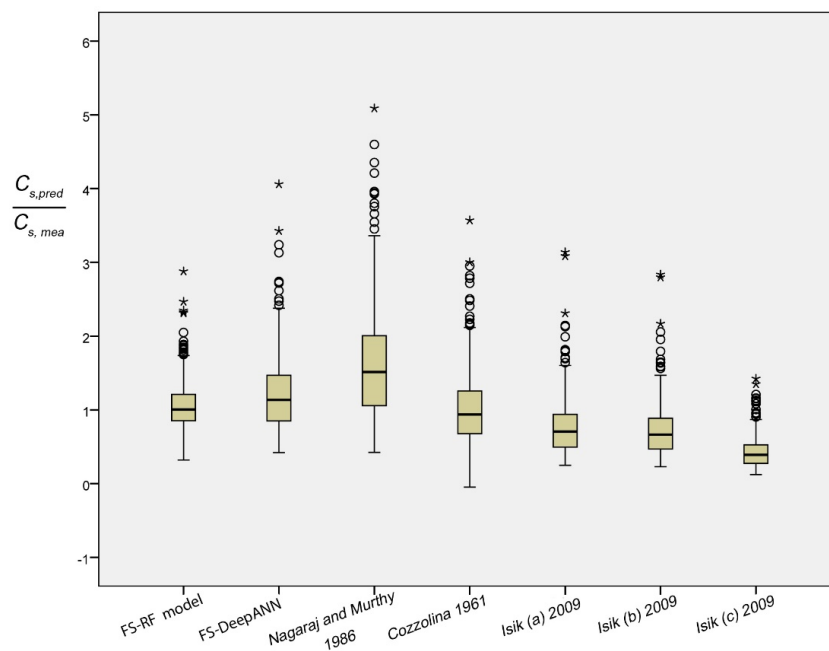
Figure 7. Box plot of the ratio $C_{s,estimated}/C_{s,measured}$ for some empirical formulas and the most appropriate model (Hollow and asterisks circles indicate respectively the extreme and extra-extreme value).

3.7. Sensitivity Analysis

To answer the question "Which input variables have the most or less influence on C_s in the proposed model?", a sensitivity analysis has been carried out using the step-by-step method [58]. In this approach, the normalized input neurons vary at a constant rate, one at a time, while the other variables are held constant. Different constant rates (0.3, 0.6 and 0.9) are selected in the current study. For every input, the percentage of change in the output, as a result of the change in the input, is recorded. The sensitivity of each input is computed based on Equation (18):

$$\text{Sensitivity level of } X_j(\%) = \frac{1}{K} \sum_{i=1}^{K} \left(\frac{\% \ change \ in \ output}{\% \ change \ in \ input} \right)_i \tag{18}$$

where K is the number of data sets used in the study ($K = 875$). The results of the sensitivity analysis of proposed *FS-RF* model are shown in Figure 8. It can be noticed that C_s is significantly influenced by *WL*, and its sensibility ratio is between 32–38%. This parameter is closely followed by the *PI*, which gives a moderate sensitivity and is ranked second. In addition, *W* and Y_d have little effect on C_s.
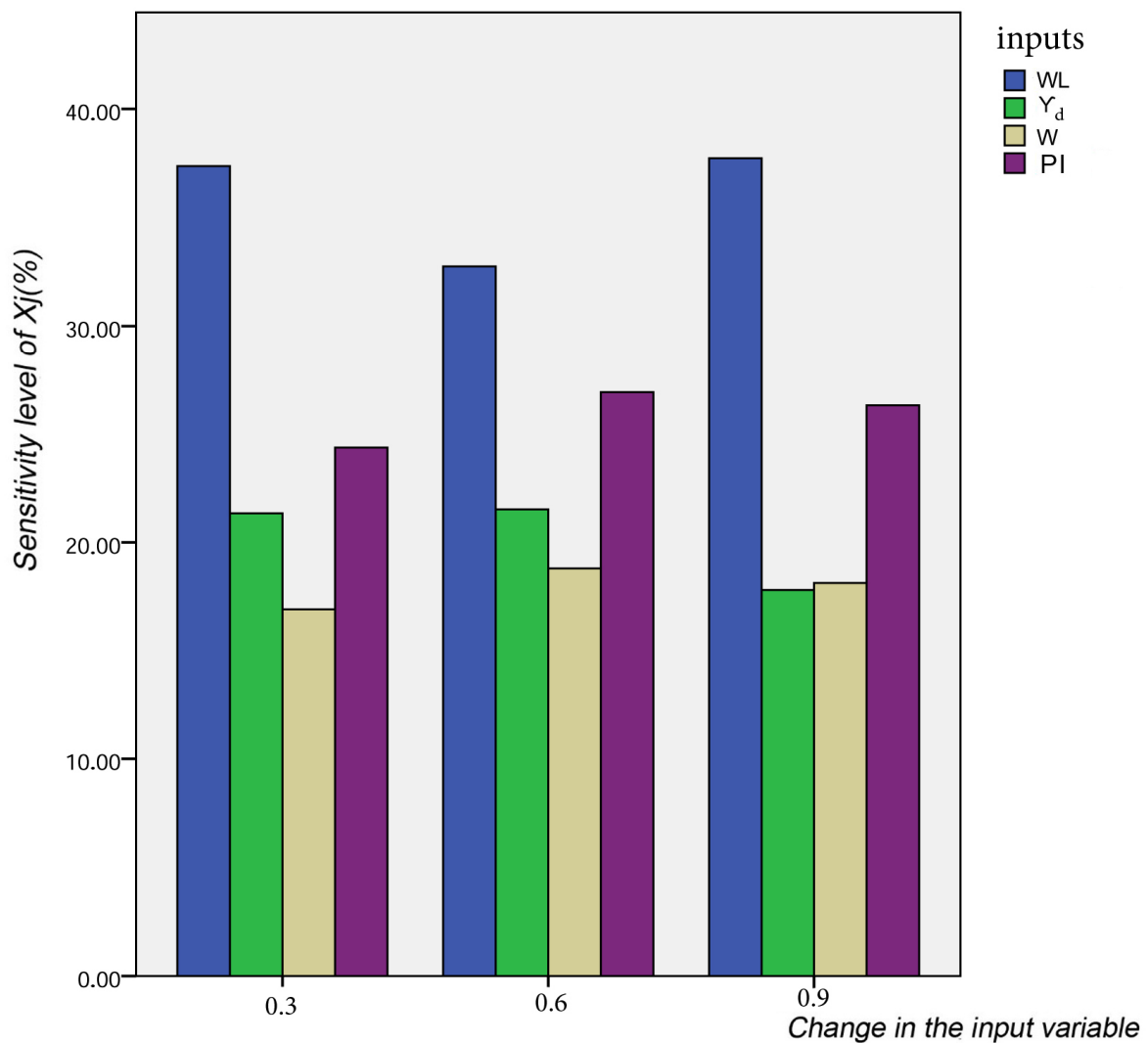
Figure 8. Results of the sensitivity analysis of the proposed model.

4. Discussion

4.1. Significance of the Findings and Cross-Validation of the Results

The main motivation of this study is to explore the capability of advanced machine learning methods to generate a reliable model aimed at easily predicting C_s. Needless to say, C_s is one of the most indispensable geotechnical parameters required to estimate the settlement and the swelling degree in the every site. Firstly, in order to identify the optimal input parameters, which have the ideal influence on C_s three advanced methods have been used. *PSO* proposed three lower-dimensional parameters as an optimal input. *GT* indicated that *WL*, *PI*, *FC*, and *W* could formalize the optimal input set. However, *FS* showed that *WL*, Y_d, *W*, and *PI* are the best ones. The reason behind the difference between the three approaches lies in the philosophy of each one in handling the data. Based on that, ten advanced machine learning methods (*ELM, DNN, SVR, RF, LASSO, PLS, Ridge, KRidge, Stepwise,* and *GP*) have applied for modeling the three selected optimal input set (*PSO, GT,* and *FS*). The findings clearly indicate that the optimal input is the one chosen by *FS* and trained by the *RF* method (*FS-RF*). The latter presents the most appropriate model, which gave the minimum values of error metrics (*MEA, RMSE,* and *IOS*) and higher values of *NSE, R,* and *IOA* compared to other models. Furthermore, the emerging model was evaluated by the K-fold cross validation approach and compared with other proposed formulae. The conclusion is that the *FS-RF* model could generate new data without over-fitting or under-fitting, and being more effective than the empirical formulae. The other most interesting aspect is the optimal input set related to the best *FS-RF* model (*WL*, Y_d, *W*, *PI*). Interestingly, this also accords with several studies, which have showed that physical parameters indirectly affected the swelling phenomena [59]. It is known that the micro-scale features of swelling soils comprise the mineral composition of clay particles, their reaction with the water chemistry, and the cations attracted to the clay particle by electrical forces. These micro-scale factors influence macro-scale physical factors, such as density, plasticity, and water content to control the engineering comportment of soil [59]. Additionally, the last part consists of the sensitivity analysis, which gives an overview about the more influenced parameters on C_s according to the proposed model. The findings indicate that *WL* and *PI* are respectively the most affected factors on C_s, meaning that swelling phenomena are primarily influenced by plasticity parameters. In addition, water content and dry density have little effect on C_s.

4.2. Scientific Importance of the Findings and Novelty of the Research

Our findings represent a crucial contribution to the geotechnical field. The elaborated model building in our study represents a reliable tool for estimating C_s without doing an Oedomter test. The performance of the estimation has been highly developed compared with other models and formulae proposed in the literature, which are based just on simple regression or neural networks. According to these data, we can infer that the Random Forest method, which is applied in this study for the first time for modeling swelling index, could yield more effective and accurate results than the DeepANN and ANN method in modelling geotechnical phenomena. These results provide further support for the hypothesis that macroscale physical factors, such as density, plasticity, and water content, are the parameters that affected the swelling phenomena. Moreover, the sensitivity analysis of the proposed model revealed the most influenced parameters between them for better understanding the complex behavior of the swelling. This investigation enhances our understanding that the plasticity of soil consisting of the *WL* and *PI* are the most affected factors on swelling phenomena. In addition to these conceptual advantages, for enhancing the training phase, a large number of samples (875 tests) and multiple input parameters have been used in this study. The sophisticated k-fold cross-validation approach was utilized to test the capability of the best model to overcome under-fitting and over-fitting problems.

4.3. Limitations of the Study and Future Research Directions

Despite the impressive multiple results presented in this study, a number of important limitations need to be considered. The most important limitation lies in the fact that the proposed machine learning models suffered from the hard fitting used in the future study. Generally, to overcome this limitation, researchers have presented elaborated models in the form of programed Interface or simple script by a known programming language like Matlab and Python for generating the proposed model. Another limitation of using this kind of data is due to its inability to generalize new conditions or circumstances that are not used in training data. Investigators generally used big collected data by transferring knowledge between them. This is an important issue for future research to use more data gathered from multiple countries for better learning and more reliable results. A further study using meta-heuristic algorithms on estimating C_s is therefore suggested. We note, for example, Particle Swarm Optimization (*PSO*) and Gravitational Search Algorithm (*GSA*), bee colony algorithm (ABC), Bio-geography-Based Optimization (*BBO*), Whale Optimization Algorithm (*WOA*), Ant Colony Optimization (*ACO*), and Grey Wolf Optimizer (*GWO*). These algorithms have proved high-performance results combined with machine learning techniques leading to improving their learning and rapidly converging to the best solution. The application of these meta-heuristic algorithms combined with machine learning methods have shown very impressive results in the abroad fields [60–62].

5. Conclusions

This study set out to optimize the swelling index parameter conducted by the expensive and time-consuming Oedometer test, contributing to elaborating on a new accurate model for predicting the swelling index (C_s) from easily obtained geotechnical physical parameters. To achieve our aim, several advanced machine learning methods were used for a practical analysis aimed at modeling the physical parameters including the wet density (Y_h), the dry density (Y_d), the degree of saturation (*Sr*), the plasticity index (*PI*), the water content (*w*), the void ratio (*e*), the liquid limit (*WL*), sample depth (Z), and the fine contents (*FC*). Firstly, principal component analysis (*PCA*), Gamma test (*GT*), and the forward selection (*FS*) approach are utilized to reduce the input variable numbers and choose the optimal ones. The results indicate the reduction of nine input variables to four (using *FS* and *GT*) and three (using *PCA* techniques). Afterward, the advanced machine learning techniques have applied for modeling the proposed optimal inputs and their accuracy models were evaluated through six statistical indicators (MAE, RMSE, IOS, NSE, R, and IOA). The comparison of results assessment between different proposed models revealed the superiority of the *FS-RF* model, which gives the highest accuracy in terms of *MAE* ($5.6 \times 10^{-3}/10.6 \times 10^{-3}$), *RMSE* (0.007/0.013), *IOS* (0.165/0.298), *NSE* (0.75/0.13), *R* (0.94/0.71), and *IOA* (0.95/0.82) during the training/validation phase. For assessing the predictive capacity of the proposed *FS-RF* model, the K-fold cross validation approach with *K* = 10 has been carried out. The results show that this model has a high correlation coefficient, ranging between 0.92 and 0.94 for training data, and 0.62 to 0.75 for validation data in the 10 splits, meaning that any over-fitting or under-fitting have been found. Three criteria were used to compare the performances of the most appropriate *FS-RF* model with the proposed formulas in the literature: the mean, the standard deviation, and the Box plot of the ratio $C_{s,estimated}/C_{s,measured}$. The findings indicate that the aforementioned *FS-RF* model is more effective than the empirical formulae. Finally, a sensitivity analysis was carried out in order to assess the impacts of the soil parameter inputs on the model performance. The results proved that *WL* has the most important effect on the prediction of C_s. *PI* has a moderate influence and ranked second. In addition, *W* and Y_d have little effect on C_s.

In the future studies, it is recommended that meta-heuristic algorithms be undertaken, like Particle Swarm Optimization (*PSO*) and Gravitational Search Algorithm (*GSA*), Biogeography-Based Optimization (*BBO*), Whale Optimization Algorithm (*WOA*), Ant

Colony Optimization (*ACO*), and Grey Wolf Optimizer (*GWO*). These algorithms have proved high performance results combined with the machine learning techniques leading to improving their learning and rapidly converging to the best solution.

Author Contributions: Conceptualization, M.A.B.; methodology, M.A.B.; software, M.A.B.; validation, M.A.B. and A.-I.P.; formal analysis, M.A.B.; investigation, M.A.B.; resources, M.A.B.; data curation, M.A.B.; writing—original draft preparation, M.A.B. and A.-I.P.; writing—review and editing, M.A.B. and A.-I.P.; visualization, M.A.B. and A.-I.P.; supervision, M.A.B. and A.-I.P.; project administration, M.A.B.; funding acquisition, M.A.B. and A.-I.P. All authors have read and agreed to the published version of the manuscript.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Restrictions apply to the availability of these data. Data was obtained from the National Earthquake Engineering Research Center (CGS), central public works laboratory (LCTP), Algiers Metro, Algiers Tramway, Cosider Engineering, Cosider Construction, and SAPTA and are available from the providers by request.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

The scatter plots between target and output swelling index value by advanced machine learning models:
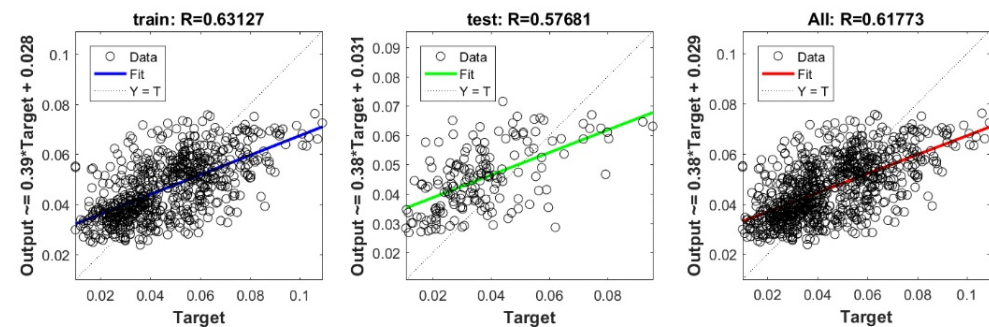
Figure A1. Scatter plots between target and output value by the *FS-SVR* model.
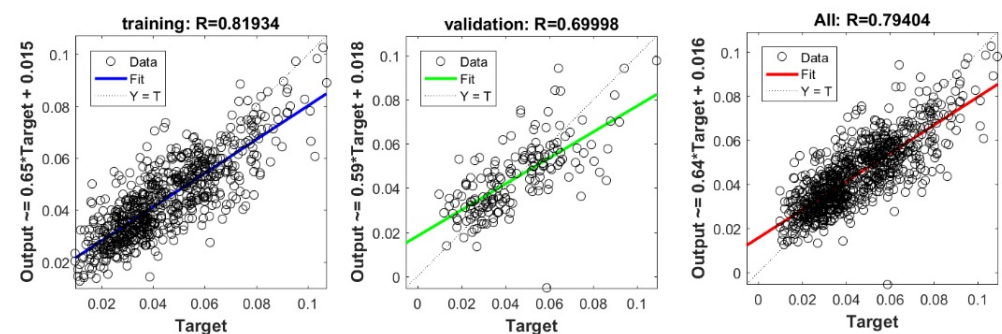
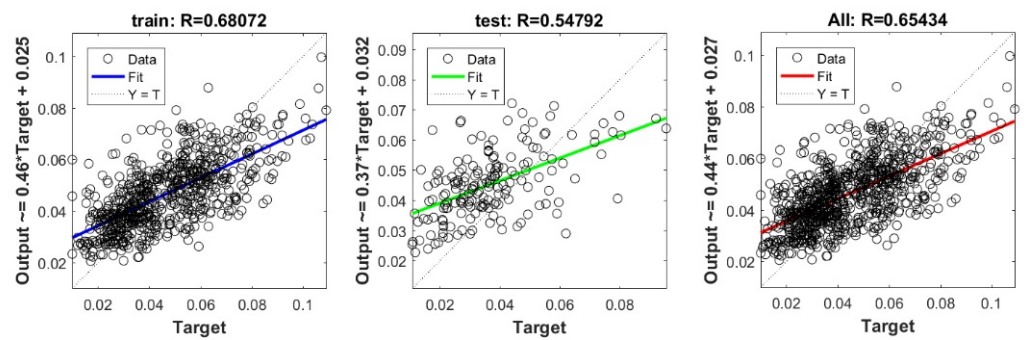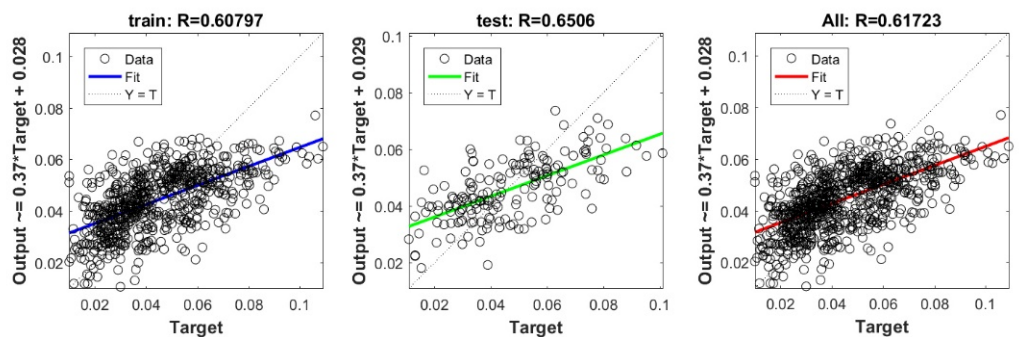**Figure A2.** Scatter plots between target and output value by the *FS-DNN* model.
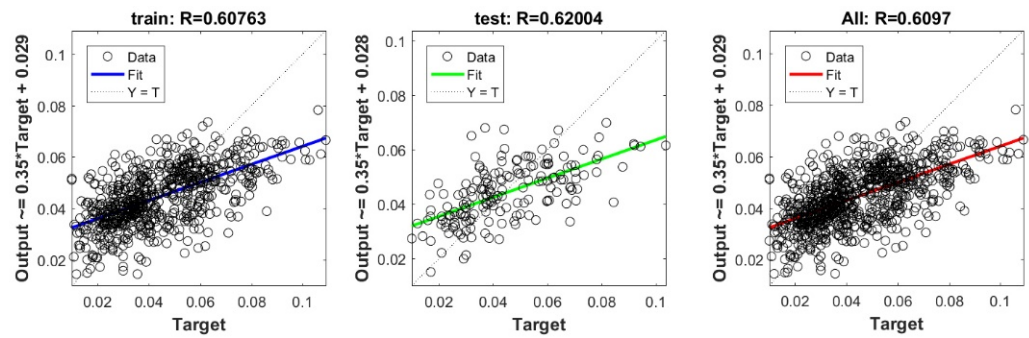
Figure A3. Scatter plots between target and output value by the *FS-ELM* model.
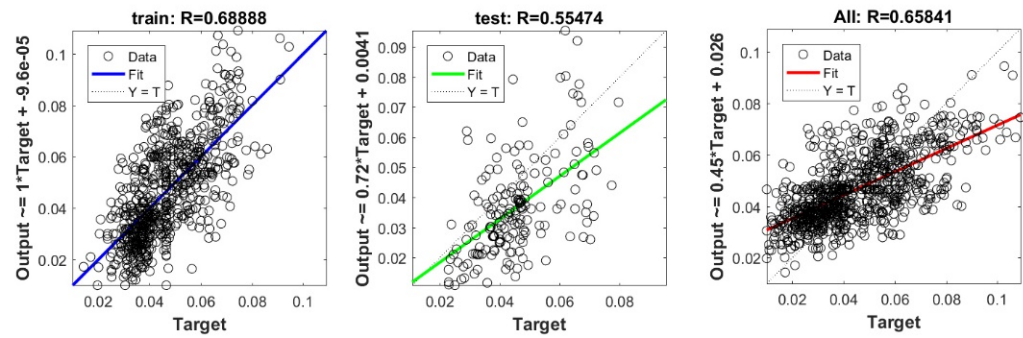
Figure A4. Scatter plots between target and output value by the *FS-GP* model.
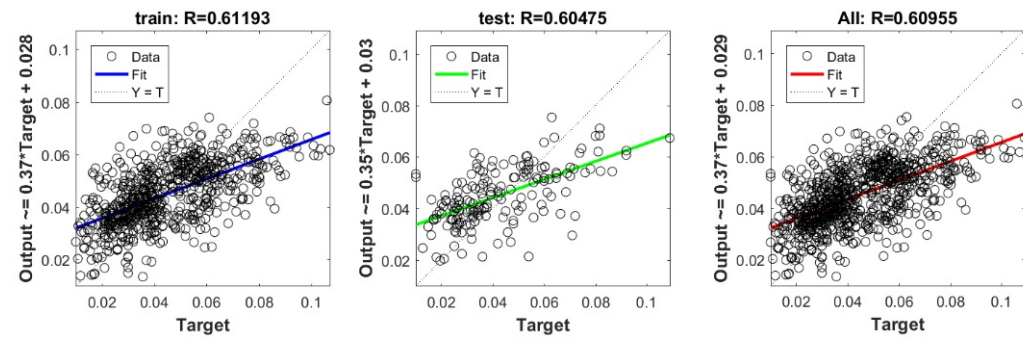
Figure A5. Scatter plots between target and output value by the FS-KRidge model.
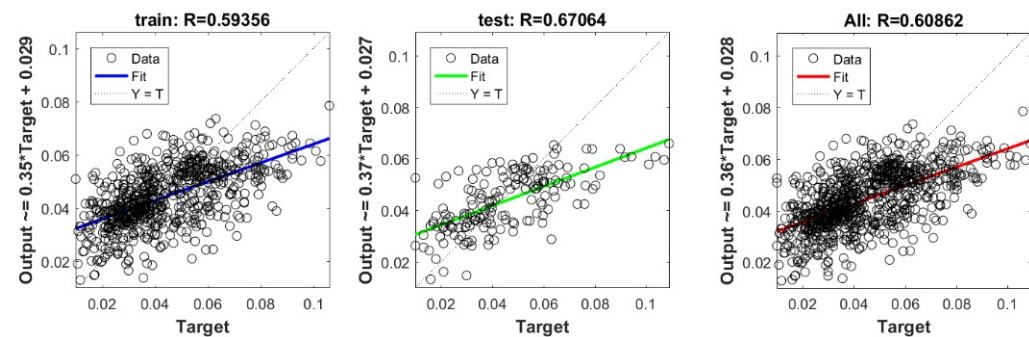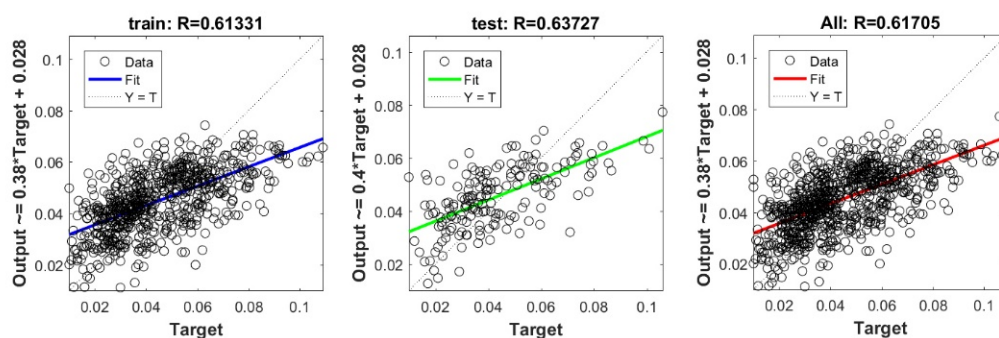

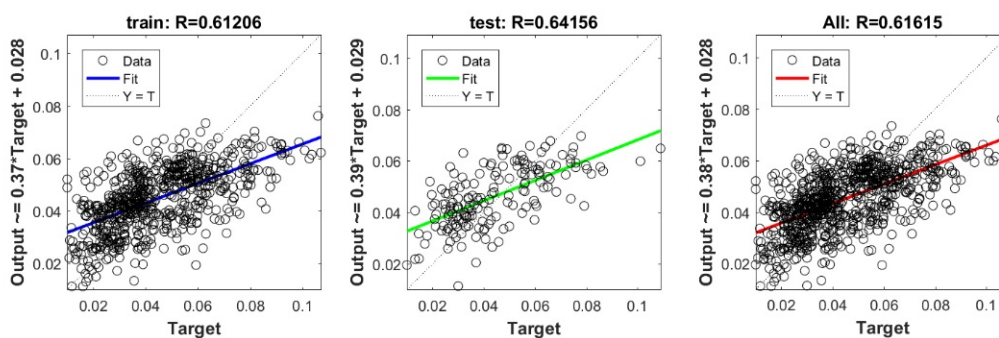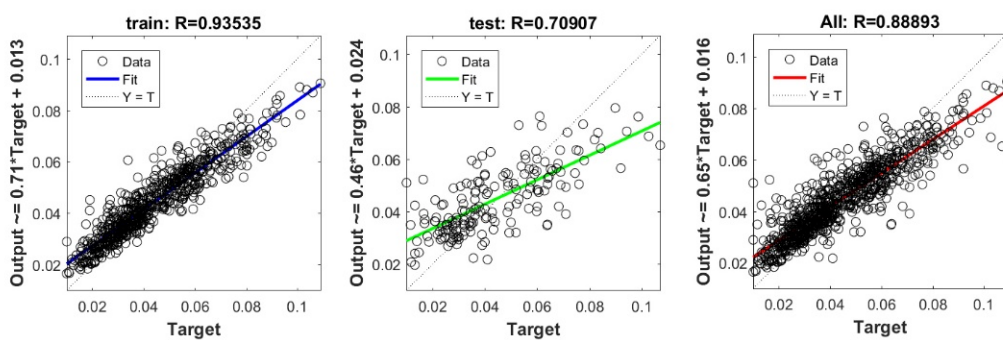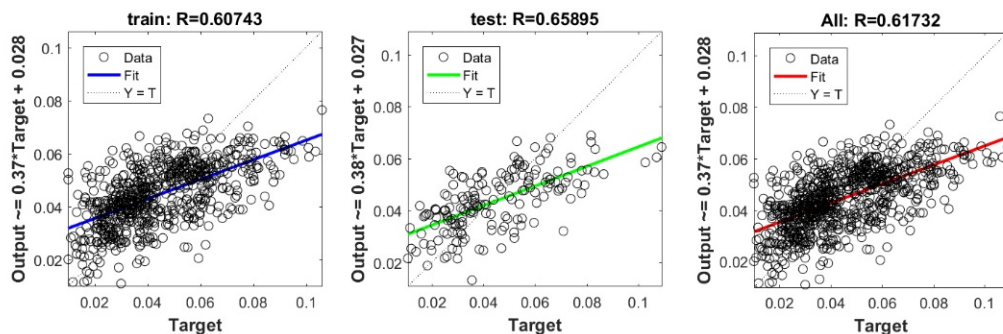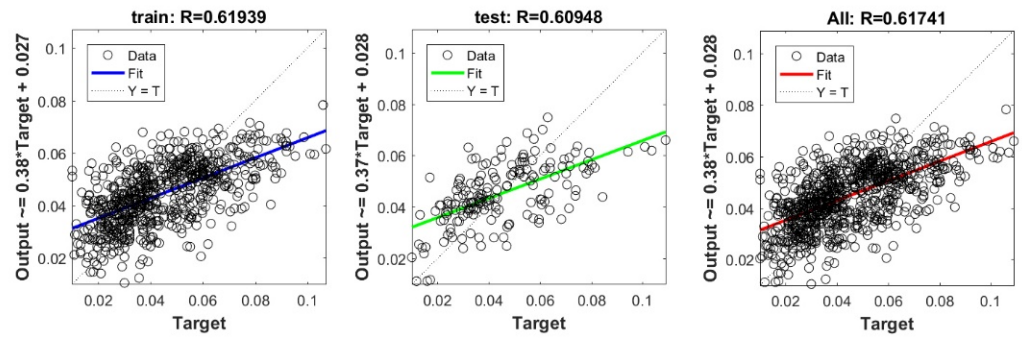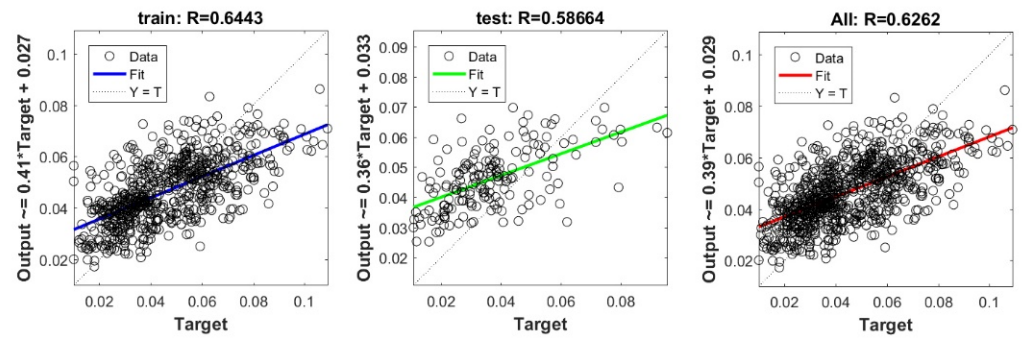**Figure A6.** Scatter plots between target and output value by the FS-lasso model.
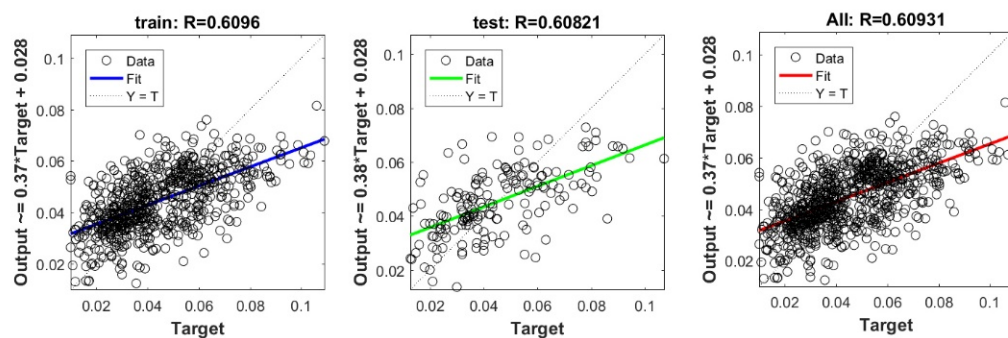
Figure A7. Scatter plots between target and output value by the *FS-PLS* model.
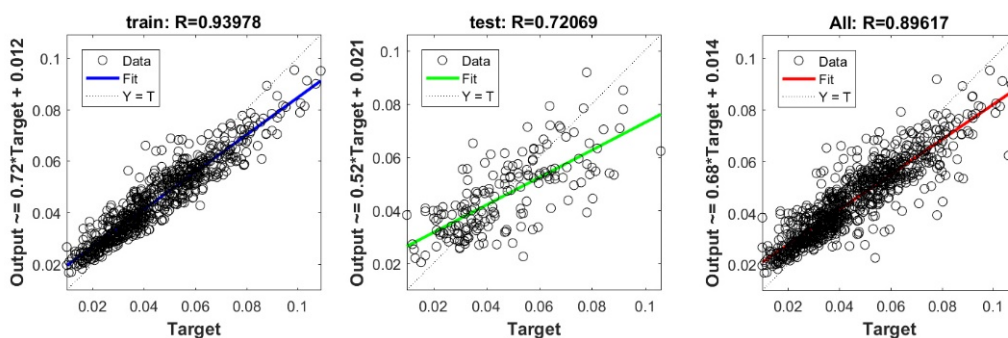
Figure A8. Scatter plots between target and output value by the *FS-RF* model.
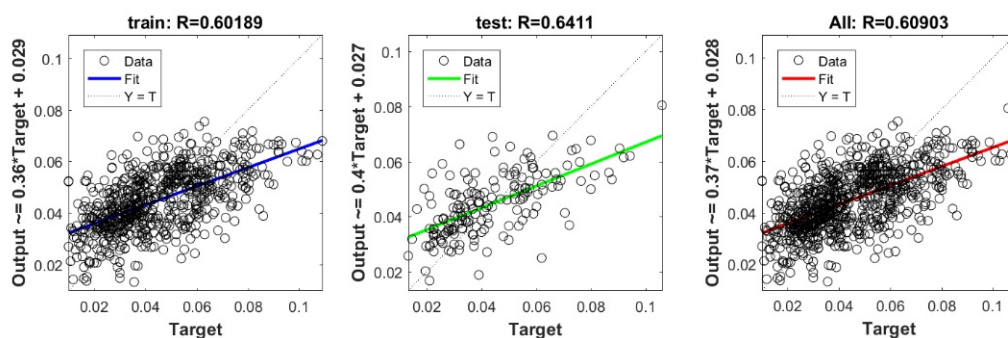
Figure A9. Scatter plots between target and output value by the FS-Ridge model.
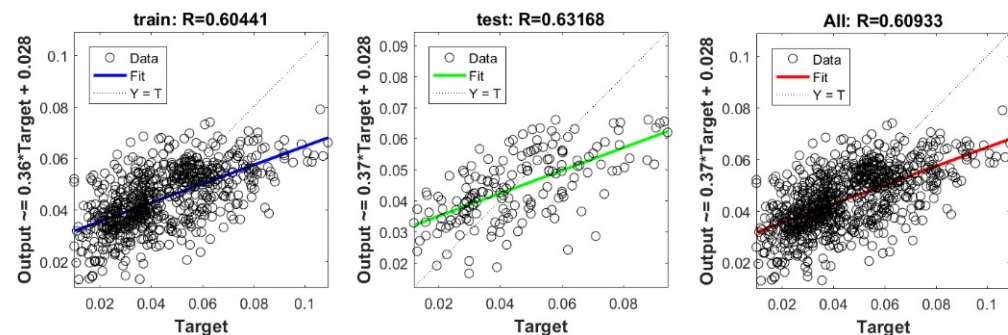
Figure A10. Scatter plots between target and output value by the FS-STEP model.
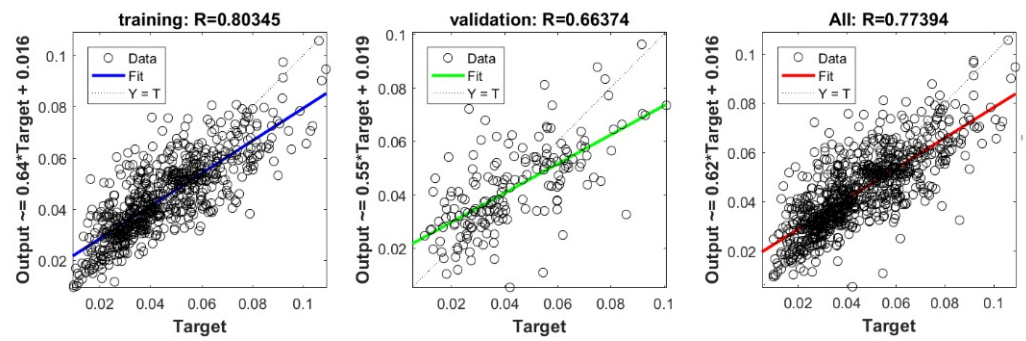
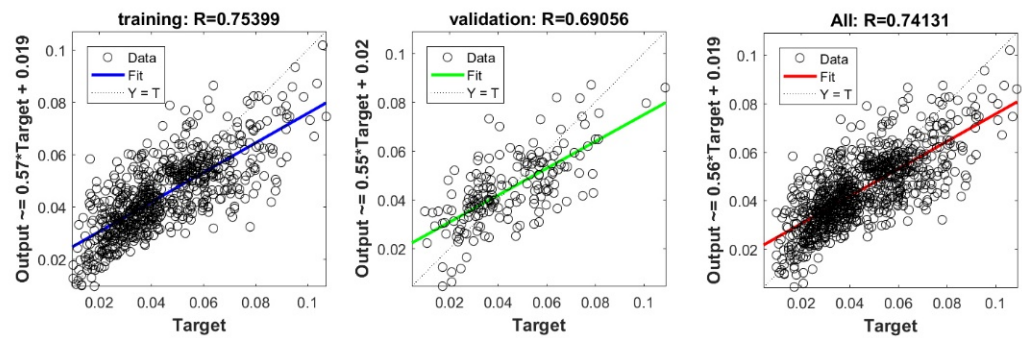**Figure A11.** Scatter plots between target and output value by the GT-DNN model.
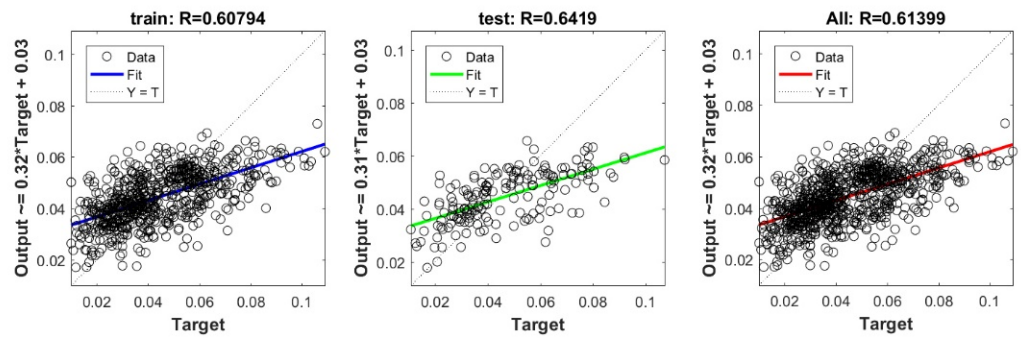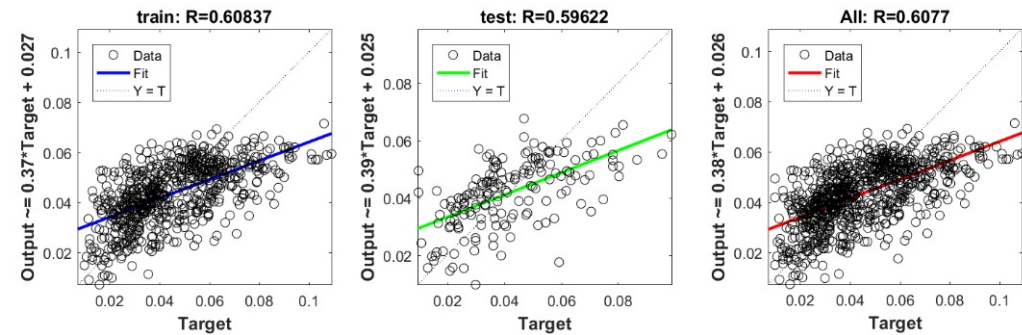

Figure A12. Scatter plots between target and output value by the GT-ELM model.

Figure A13. Scatter plots between target and output value by the GT-GP model.

Figure A14. Scatter plots between target and output value by the GT-KRidge model.

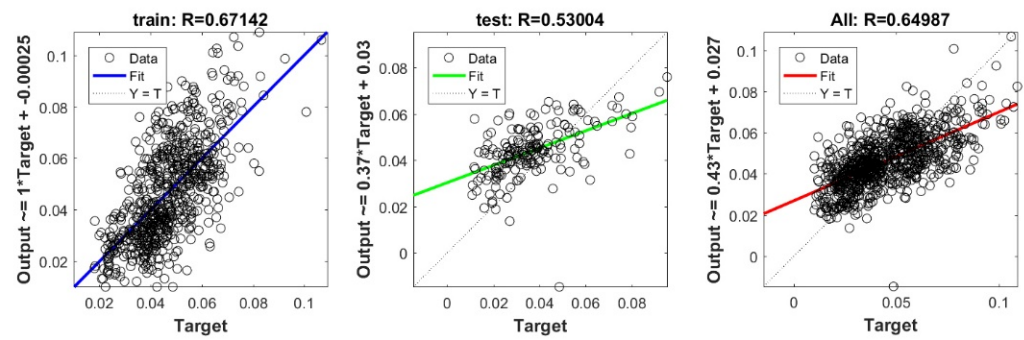**Figure A15.** Scatter plots between target and output value by the GT-Lasso model.

Figure A16. Scatter plots between target and output value by the GT-PLS model.

Figure A17. Scatter plots between target and output value by the GT-RF model.

Figure A18. Scatter plots between target and output value by the GT-Ridge model.

Figure A19. Scatter plots between target and output value by the GT-Step model.

Figure A20. Scatter plots between target and output value by the GT-SVR model.

Figure A21. Scatter plots between target and output value by the PSO-DNN model.

Figure A22. Scatter plots between target and output value by the PSO-ELM model.

Figure A23. Scatter plots between target and output value by the PSO-GP model.
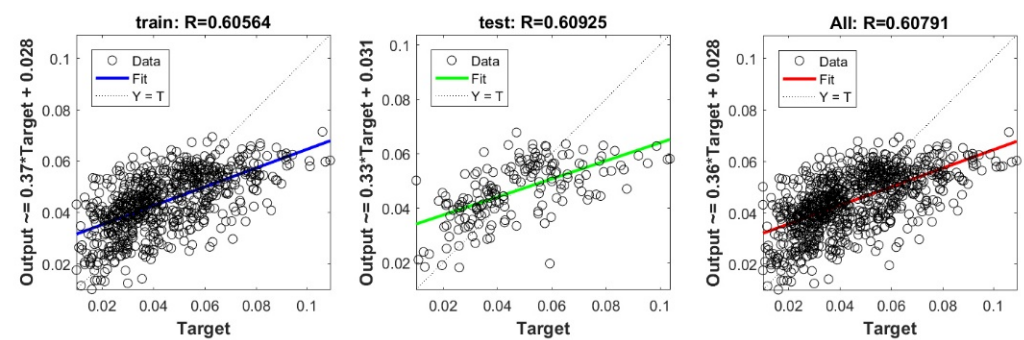

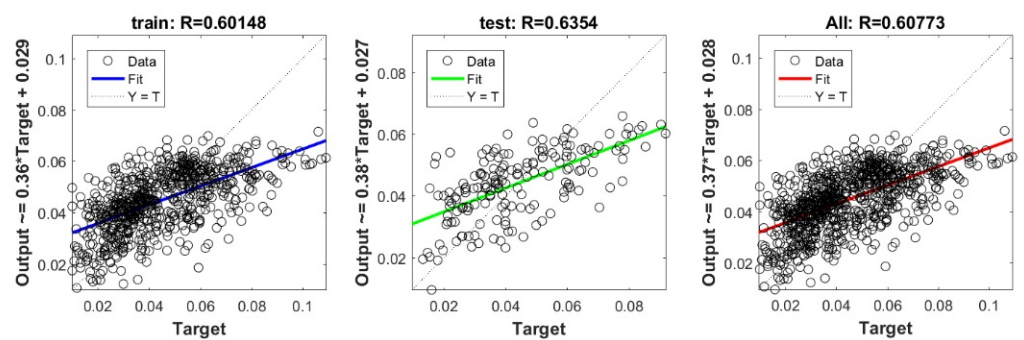**Figure A24.** Scatter plots between target and output value by the PSO-KRidge model.

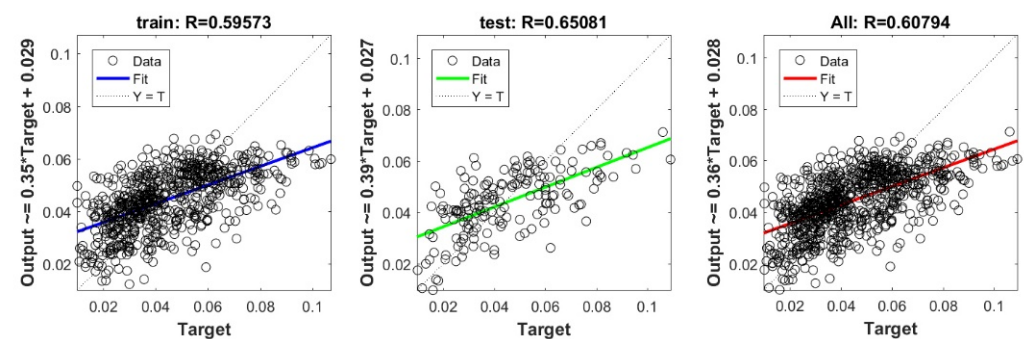**Figure A25.** Scatter plots between target and output value by the PSO-Lasso model.

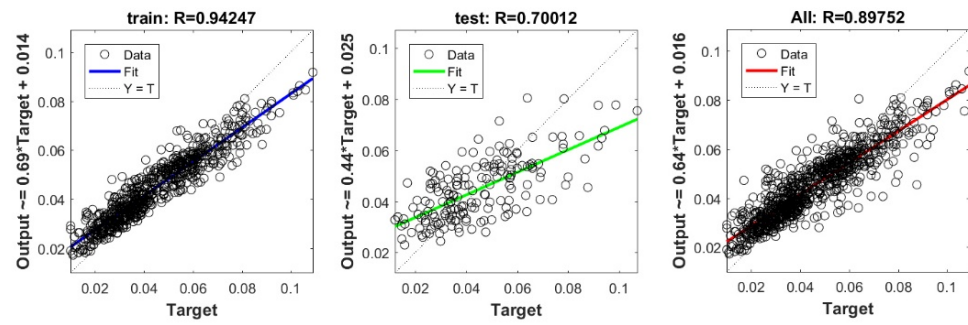**Figure A26.** Scatter plots between target and output value by the PSO-PLS model.

Figure A27. Scatter plots between target and output value by the PSO-RF model.
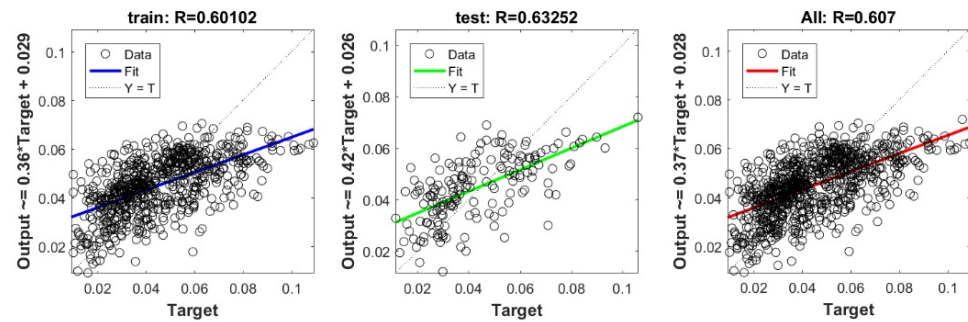

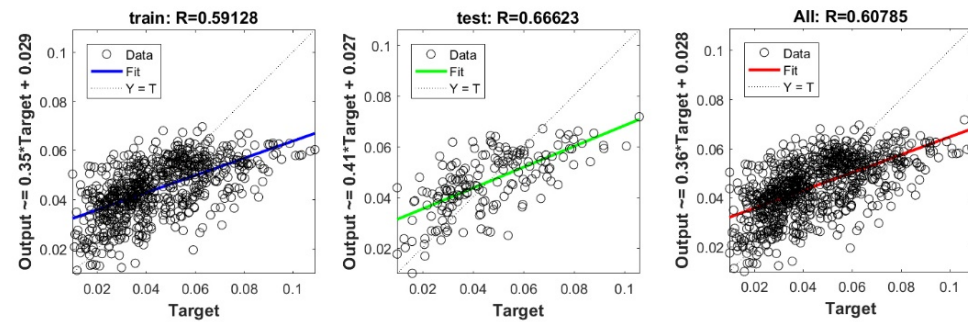**Figure A28.** Scatter plots between target and output value by the PSO-Ridge model.

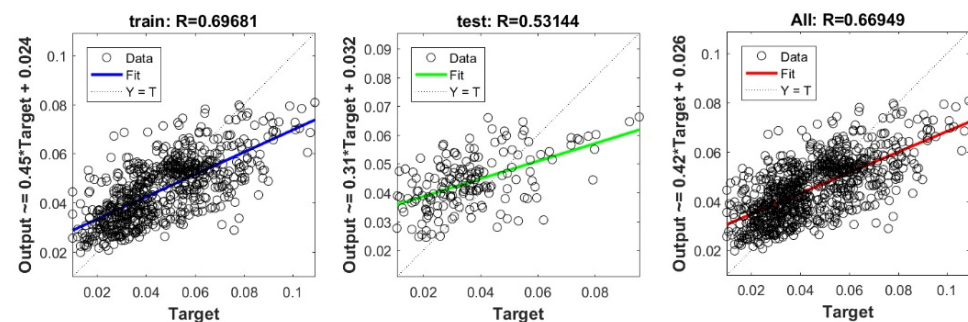**Figure A29.** Scatter plots between target and output value by the PSO-Step model.

Figure A30. Scatter plots between target and output value the by PSO-SVR model.

References

1. Hunt, R. *Geotechnical Investigation Methods: A Field Guide for Geotechnical Engineers*; CRC Press: Boca Raton, FL, USA, 2006; ISBN 978-1-4200-4274-0.
2. Benbouras, M.A.; Kettab, R.M.; Debiche, F.; Lagaguine, M.; Mechaala, A.; Bourezak, C.; Petrişor, A.-I. Use of Geotechnical and Geographical Information Systems to Analyze Seismic Risk in Algiers Area. *Rev. Şcolii Dr. de Urban* **2018**, *3*, 11.

3. Yuan, S.; Liu, X.; Sloan, S.W.; Buzzi, O.P. Multi-Scale Characterization of Swelling Behaviour of Compacted Maryland Clay. *Acta Geotech.* **2016**, *11*, 789–804. [CrossRef]

4. Zhang, X.; Lu, H.; Li, L. Use of Oedometer Equipped with High-Suction Tensiometer to Characterize Unsaturated Soils. *Transp. Res. Rec.* **2016**, *2578*, 58–71. [CrossRef]

5. Teerachaikulpanich, N.; Okumura, S.; Matsunaga, K.; Ohta, H. Estimation of Coefficient of Earth Pressure at Rest Using Modified Oedometer Test. *Soils Found.* **2007**, *47*, 349–360. [CrossRef]

6. Benbouras, M.A.; Kettab, R.; Zedira, H.; Petrisor, A.; Mezouer, N.; Debiche, F. A new approach to predict the Compression Index using Artificial Intelligence Methods. *Mar. Georesour. Geotechnol.* **2018**, *6*, 704–720. [CrossRef]

7. Shahin, M.A. Artificial intelligence in geotechnical engineering: Applications, modeling aspects, and future directions. In *Metaheuristics in Water, Geotechnical and Transport Engineering*; Elsevier: Amsterdam, The Netherlands, 2013; pp. 169–204.

8. Nagaraj, H.; Munnas, M.; Sridharan, A. Swelling Behavior of Expansive Soils. *Int. J. Geotech. Eng.* **2010**, *4*, 99–110. [CrossRef]

9. Ameratunga, J.; Sivakugan, N.; Das, B.M. *Correlations of Soil and Rock Properties in Geotechnical Engineering*; Developments in Geotechnical Engineering; Springer India: New Delhi, India, 2016; ISBN 978-81-322-2627-7.

10. Samui, P.; Hoang, N.-D.; Nhu, V.-H.; Nguyen, M.-L.; Ngo, P.T.T.; Bui, D.T. A New Approach of Hybrid Bee Colony Optimized Neural Computing to Estimate the Soil Compression Coefficient for a Housing Construction Project. *Appl. Sci.* **2019**, *9*, 4912. [CrossRef]

11. Moayedi, H.; Tien Bui, D.; Dounis, A.; Ngo, P.T.T. A Novel Application of League Championship Optimization (LCA): Hybridizing Fuzzy Logic for Soil Compression Coefficient Analysis. *Appl. Sci.* **2020**, *10*, 67. [CrossRef]

12. Onyejekwe, S.; Kang, X.; Ge, L. Assessment of Empirical Equations for the Compression Index of Fine-Grained Soils in Missouri. *Bull. Eng. Geol. Environ.* **2015**, *74*, 705–716. [CrossRef]

13. Onyejekwe, S.; Kang, X.; Ge, L. Evaluation of the Scale of Fluctuation of Geotechnical Parameters by Autocorrelation Function and Semivariogram Function. *Eng. Geol.* **2016**, *214*, 43–49. [CrossRef]

14. Shahin, M.A.; Jaksa, M.B.; Maier, H.R. Recent Advances and Future Challenges for Artificial Neural Systems in Geotechnical Engineering Applications. *Adv. Artif. Neural Syst.* **2009**, *2009*, 1–9. [CrossRef]

15. Hossein Alavi, A.; Hossein Gandomi, A. A Robust Data Mining Approach for Formulation of Geotechnical Engineering Systems. *Eng. Comput.* **2011**, *28*, 242–274. [CrossRef]

16. Park, H.I.; Lee, S.R. Evaluation of the Compression Index of Soils Using an Artificial Neural Network. *Comput. Geotech.* **2011**, *38*, 472–481. [CrossRef]

17. Cai, G.; Liu, S.; Puppala, A.J.; Tong, L. Identification of Soil Strata Based on General Regression Neural Network Model from CPTU Data. *Mar. Georesour. Geotechnol.* **2015**, *33*, 229–238. [CrossRef]

18. Nagaraj, T.S.; Srinivasa Murthy, B.R. A Critical Reappraisal of Compression Index Equations. *Geotechnique* **1986**, *36*, 27–32. [CrossRef]

19. Cozzolino, V.M. Statistical Forecasting of Compression Index. In Proceedings of the Proceedings of the fifth international conference on soil mechanics and foundation engineering, Paris, France, 17–22 July 1961; Volume 1, pp. 51–53.

20. Işık, N.S. Estimation of Swell Index of Fine Grained Soils Using Regression Equations and Artificial Neural Networks. *Sci. Res. Essays* **2009**, *4*, 1047–1056.

21. Shahin, M.A.; Jaksa, M.B.; Maier, H.R. State of the Art of Artificial Neural Networks in Geotechnical Engineering. *Electron. J. Geotech. Eng.* **2008**, *8*, 1–26.

22. Das, S.K.; Samui, P.; Sabat, A.K.; Sitharam, T.G. Prediction of Swelling Pressure of Soil Using Artificial Intelligence Techniques. *Environ. Earth Sci.* **2010**, *61*, 393–403. [CrossRef]

23. Kumar, V.P.; Rani, C.S. Prediction of Compression Index of Soils Using Artificial Neural Networks (ANNs). *Int. J. Eng. Res. Appl.* **2011**, *1*, 1554–1558.

24. Kurnaz, T.F.; Dagdeviren, U.; Yildiz, M.; Ozkan, O. Prediction of Compressibility Parameters of the Soils Using Artificial Neural Network. *SpringerPlus* **2016**, *5*, 1801. [CrossRef]

25. Alavi, A.H.; Gandomi, A.H.; Mollahasani, A.; Bazaz, J.B. Linear and Tree-Based Genetic Programming for Solving Geotechnical Engineering Problems. In *Metaheuristics in Water, Geotechnical and Transport Engineering*; Elsevier: Amsterdam, The Netherlands, 2013; pp. 289–310. ISBN 978-0-12-398296-4.

26. Narendra, B.S.; Sivapullaiah, P.V.; Suresh, S.; Omkar, S.N. Prediction of Unconfined Compressive Strength of Soft Grounds Using Computational Intelligence Techniques: A Comparative Study. *Comput. Geotech.* **2006**, *33*, 196–208. [CrossRef]

27. Rezania, M.; Javadi, A.A. A New Genetic Programming Model for Predicting Settlement of Shallow Foundations. *Can. Geotech. J.* **2007**, *44*, 1462–1473. [CrossRef]

28. Yin, Z.-Y.; Zhu, Q.-Y.; Zhang, D.-M. Comparison of Two Creep Degradation Modeling Approaches for Soft Structured Soils. *Acta Geotech.* **2017**, *12*, 1395–1413. [CrossRef]

29. Stoica, I.-V.; Tulla, A.F.; Zamfir, D.; Petrişor, A.-I. Exploring the Urban Strength of Small Towns in Romania. *Soc. Indic. Res.* **2020**, *152*, 843–875. [CrossRef]

30. Petrişor, A.-I.; Ianoş, I.; Iurea, D.; Văidianu, M.-N. Applications of Principal Component Analysis Integrated with GIS. *Procedia Environ. Sci.* **2012**, *14*, 247–256. [CrossRef]

31. Lu, W.Z.; Wang, W.J.; Wang, X.K.; Xu, Z.B.; Leung, A.Y.T. Using Improved Neural Network Model to Analyze RSP, NO x and NO 2 Levels in Urban Air in Mong Kok, Hong Kong. *Environ. Monit. Assess.* **2003**, *87*, 235–254. [CrossRef]

32. Çamdevýren, H.; Demýr, N.; Kanik, A.; Keskýn, S. Use of Principal Component Scores in Multiple Linear Regression Models for Prediction of Chlorophyll-a in Reservoirs. *Ecol. Model.* **2005**, *181*, 581–589. [CrossRef]

33. Wackernagel, H. *Multivariate Geostatistics: An Introduction with Applications*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2013; ISBN 978-3-662-05294-5.

34. Tabachnick, B.G.; Fidell, L.S.; Ullman, J.B. *Using Multivariate Statistics*, 7th ed.; Pearson: New York, NY, USA, 2019; ISBN 978-0-13-479054-1.

35. Noori, R.; Karbassi, A.; Salman Sabahi, M. Evaluation of PCA and Gamma Test Techniques on ANN Operation for Weekly Solid Waste Prediction. *J. Environ. Manag.* **2010**, *91*, 767–771. [CrossRef]

36. Noori, R.; Karbassi, A.R.; Moghaddamnia, A.; Han, D.; Zokaei-Ashtiani, M.H.; Farokhnia, A.; Gousheh, M.G. Assessment of Input Variables Determination on the SVM Model Performance Using PCA, Gamma Test, and Forward Selection Techniques for Monthly Stream Flow Prediction. *J. Hydrol.* **2011**, *401*, 177–189. [CrossRef]

37. Koncar, N. Optimisation Methodologies for Direct Inverse Neurocontrol. Ph.D. Thesis, University of London, London, UK, 1997.

38. Stefánsson, A.; Končar, N.; Jones, A.J. A Note on the Gamma Test. *Neural Comput. Appl.* **1997**, *5*, 131–133. [CrossRef]

39. Kemp, S.; Wilson, I.; Ware, J. A Tutorial on the Gamma Test. *Int. J. Simul. Syst. Sci. Technol.* **2004**, *6*, 67–75.

40. Huang, G.-B.; Zhu, Q.-Y.; Siew, C.-K. Extreme Learning Machine: Theory and Applications. *Neurocomputing* **2006**, *70*, 489–501. [CrossRef]

41. Benbouras, M.A.; Kettab, R.M.; Zedira, H.; Debiche, F.; Zaidi, N. Comparing nonlinear regression analysis and artificial neural networks to predict geotechnical parameters from standard penetration test. *Urban. Archit. Constr.* **2018**, *1*, 275–288.

42. Awad, M.; Khanna, R. Support Vector Regression. In *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers*; Awad, M., Khanna, R., Eds.; Apress: Berkeley, CA, USA, 2015; pp. 67–80. ISBN 978-1-4302-5990-9.

43. Biau, G.; Scornet, E. A Random Forest Guided Tour. *TEST* **2016**, *25*, 197–227. [CrossRef]

44. Hebiri, M.; Lederer, J. How Correlations Influence Lasso Prediction. *IEEE Trans. Inf. Theory* **2013**, *59*, 1846–1854. [CrossRef]

45. Vinzi, V.E.; Chin, W.W.; Henseler, J.; Wang, H. Editorial: Perspectives on Partial Least Squares. In *Handbook of Partial Least Squares: Concepts, Methods and Applications*; Esposito Vinzi, V., Chin, W.W., Henseler, J., Wang, H., Eds.; Springer Handbooks of Computational Statistics; Springer: Berlin/Heidelberg, Germany, 2010; pp. 1–20. ISBN 978-3-540-32827-8.

46. Hoerl, A.E.; Kennard, R.W. Ridge Regression—1980: Advances, Algorithms, and Applications. *Am. J. Math. Manag. Sci.* **1981**, *1*, 5–83. [CrossRef]

47. Douak, F.; Melgani, F.; Benoudjit, N. Kernel Ridge Regression with Active Learning for Wind Speed Prediction. *Appl. Energy* **2013**, *103*, 328–340. [CrossRef]

48. Jennrich, R.I.; Sampson, P.F. Application of Stepwise Regression to Non-Linear Estimation. *Technometrics* **1968**, *10*, 63–72. [CrossRef]

49. Wagner, S.; Kronberger, G.; Beham, A.; Kommenda, M.; Scheibenpflug, A.; Pitzer, E.; Vonolfen, S.; Kofler, M.; Winkler, S.; Dorfer, V.; et al. Architecture and Design of the HeuristicLab Optimization Environment. In *Advanced Methods and Applications in Computational Intelligence*; Klempous, R., Nikodem, J., Jacak, W., Chaczko, Z., Eds.; Topics in Intelligent Engineering and Informatics; Springer International Publishing: Heidelberg, Germany, 2014; pp. 197–261. ISBN 978-3-319-01436-4.

50. Tikhamarine, Y.; Malik, A.; Pandey, K.; Sammen, S.S.; Souag-Gamane, D.; Heddam, S.; Kisi, O. Monthly Evapotranspiration Estimation Using Optimal Climatic Parameters: Efficacy of Hybrid Support Vector Regression Integrated with Whale Optimization Algorithm. *Environ. Monit. Assess.* **2020**, *192*, 696. [CrossRef]

51. Breiman, L.; Spector, P. Submodel Selection and Evaluation in Regression. The X-Random Case. *Int. Stat. Rev.* **1992**, *60*, 291–319. [CrossRef]

52. Oommen, T.; Baise, L.G. Model Development and Validation for Intelligent Data Collection for Lateral Spread Displacements. *J. Comput. Civ. Eng.* **2010**, *24*, 467–477. [CrossRef]

53. Goetz, J.N.; Brenning, A.; Petschko, H.; Leopold, P. Evaluating Machine Learning and Statistical Prediction Techniques for Landslide Susceptibility Modeling. *Comput. Geosci.* **2015**, *81*, 1–11. [CrossRef]

54. Benbouras, M.A.; Kettab, R.M.; Zedira, H.; Petrisor, A.-I.; Debiche, F. Dry Density in Relation to Other Geotechnical Proprieties of Algiers Clay. *Rev. Şcolii Dr. Urban* **2017**, *2*, 5–14.

55. Debiche, F.; Kettab, R.M.; Benbouras, M.A.; Benbellil, B.; Djerbal, L.; Petrisor, A.-I. Use of GIS systems to analyze soil compressibility, swelling and bearing capacity under superficial foundations in Algiers region, Algeria. *Urban. Arhit. Constr.* **2018**, *9*, 357–370.

56. Abba, S.I.; Pham, Q.B.; Usman, A.G.; Linh, N.T.T.; Aliyu, D.S.; Nguyen, Q.; Bach, Q.-V. Emerging Evolutionary Algorithm Integrated with Kernel Principal Component Analysis for Modeling the Performance of a Water Treatment Plant. *J. Water Process. Eng.* **2020**, *33*, 101081. [CrossRef]

57. Kanellopoulos, I.; Wilkinson, G.G. Strategies and Best Practice for Neural Network Image Classification. *Int. J. Remote Sens.* **1997**, *18*, 711–725. [CrossRef]

58. Liong, S.-Y.; Lim, W.-H.; Paudyal, G.N. River Stage Forecasting in Bangladesh: Neural Network Approach. *J. Comput. Civ. Eng.* **2000**, *14*, 1–8. [CrossRef]

59. Mawlood, Y.I.; Hummadi, R.A. Large-Scale Model Swelling Potential of Expansive Soils in Comparison with Oedometer Swelling Methods. *Iran. J. Sci. Technol. Trans. Civ. Eng.* **2020**, *44*, 1283–1293. [CrossRef]

60. Ly, H.-B.; Le, T.-T.; Le, L.M.; Tran, V.Q.; Le, V.M.; Vu, H.-L.T.; Nguyen, Q.H.; Pham, B.T. Development of Hybrid Machine Learning Models for Predicting the Critical Buckling Load of I-Shaped Cellular Beams. *Appl. Sci.* **2019**, *9*, 5458. [CrossRef]
61. Asteris, P.G.; Nikoo, M. Artificial Bee Colony-Based Neural Network for the Prediction of the Fundamental Period of Infilled Frame Structures. *Neural Comput. Appl.* **2019**, *31*, 4837–4847. [CrossRef]
62. Sarir, P.; Chen, J.; Asteris, P.G.; Armaghani, D.J.; Tahir, M.M. Developing GEP Tree-Based, Neuro-Swarm, and Whale Optimization Models for Evaluation of Bearing Capacity of Concrete-Filled Steel Tube Columns. *Eng. Comput.* **2019**. [CrossRef]