Hindawi
Complexity
Volume 2022, Article ID 9415863, 11 pages
https://doi.org/10.1155/2022/9415863

Research Article

Extreme Gradient Boosting Algorithm for Predicting Shear Strengths of Rockfill Materials

Mahmood Ahmad,[1,2] **Ramez A. Al-Mansob**,[1] **Kazem Reza Kashyzadeh**,[3] **Suraparb Keawsawasvong**,[4] **Mohanad Muayad Sabri Sabri**,[5] **Irfan Jamil**,[6] **and Arnold C. Alguno**[7]

[1]*Department of Civil Engineering, Faculty of Engineering, International Islamic University Malaysia, Jalan Gombak, Selangor 50728, Malaysia*
[2]*Department of Civil Engineering, University of Engineering and Technology Peshawar (Bannu Campus), Bannu 28100, Pakistan*
[3]*Department of Transport, Academy of Engineering, Peoples' Friendship University of Russia (RUDN University), 6 Miklukho-Maklaya Street, Moscow 117198, Russia*
[4]*Department of Civil Engineering, Thammasat School of Engineering, Thammasat University, Pathumthani 12120, Thailand*
[5]*Peter the Great St. Petersburg Polytechnic University, Saint Petersburg 195251, Russia*
[6]*Department of Civil Engineering, University of Engineering and Technology Peshawar, Peshawar 25000, Pakistan*
[7]*Department of Physics, Mindanao State University-Iligan Institute of Technology, Iligan City 9200, Philippines*

Correspondence should be addressed to Mahmood Ahmad; ahmadm@uetpeshawar.edu.pk

Received 3 June 2022; Revised 19 July 2022; Accepted 20 July 2022; Published 24 August 2022

Academic Editor: Andrea Murari

For the safe and economical construction of embankment dams, the mechanical behaviour of the rockfill materials used in the dam's shell must be analyzed. The characterization of rockfill materials with specified shear strength is difficult and expensive due to the presence of particles greater than 500 mm in diameter. This work investigates the feasibility of using an extreme gradient boosting (XGBoost) computing paradigm to estimate the shear strength of rockfill materials. To train and validate the proposed XGBoost model, a total of 165 databases obtained from the literature are chosen. The XGBoost model was compared against support vector machine (SVM), adaptive boosting (AdaBoost), random forest (RF), and K-nearest neighbor (KNN) models described in the literature. XGBoost beats SVM, RF, AdaBoost, and KNN models in terms of performance evaluation metrics such as coefficient of determination (R^2), Nash–Sutcliffe coefficient (NSE), and error in the root mean square ratio (RMSE) to the standard deviation of the measured data (RSR). The results demonstrated that the XGBoost model has the highest prediction performance with ($R^2 = 0.9707$, NSE = 0.9701, and RSR = 0.1729), followed by the SVM model with ($R^2 = 0.9655$, NSE = 0.9639, and RSR = 0.1899), RF ($R^2 = 0.9545$, NSE = 0.9542, and RSR = 0.2140), the AdaBoost model with ($R^2 = 0.9390$, NSE = 0.9388, and RSR = 0.2474) and the KNN model with ($R^2 = 0.6233$, NSE = 0.6180, and RSR = 0.6181). A sensitivity analysis has been conducted to ascertain the impact of each investigated input parameter. This study demonstrates that the established XGBoost model for estimating the shear strength of rockfill materials is reliable.

1. Introduction

Rockfill materials (RFM) are commonly used in the construction of high embankment dams in order to harness natural water resources. RFM is comprised of gravels, cobbles, and boulders obtained by blasting rock quarries or natural riverbeds. Material from riverbeds is rounded to subrounded, and material from quarries is angular to subangular. Mineral composition, particle size, shape, gradation, individual particle strength, void content, relative density (RD), and particle surface roughness all influence the behaviour of these RFMs used in the construction of rockfill

dams. Therefore, it is essential to comprehend and characterise the behaviour of these materials for the study and safe construction of rockfill dams.

In engineering practice, the particle size of rockfill materials ranges from 400 to 600 millimetres and can exceed 1000 millimetres. Due to the constraints of laboratory testing equipment, rockfill materials that exceed the maximum permissible particle size must be scaled. To determine the mechanical properties of rockfill materials on-site, analog simulation is used in laboratory testing to build test specimens with the same internal structure as the prototype rockfill materials, thus determining the engineering characteristics of the prototype rockfill materials. Several research studies have investigated the behaviour of the RFM such as Abbas et al. [1], Gupta [2], Venkatachalam [3], Marsal [4], Mirachi [5], and Honkanadavar and Sharma [6] and carried out laboratory experiments on different RFMs, and it was revealed that their stress-strain behaviour is dependent on the stress level, but nonlinear and inelastic. They also reported that the angle of internal friction increases as the maximum particle size of riverbed RFM increases, while the opposite trend is true for quarry RFM. Frossard et al. [7] proposed a rational approach for estimating RFM shear strength based on size effects; Honkanadavar and Gupta [8] developed a power law for the relationship between the shear strength parameter and various riverbed RFM index features due to the difficulty of conducting large-scale strength testing and defining the mechanical behaviour of RFMs. Numerous methodologies have been developed to anticipate the behaviour of such soils. Large particle size RFM cannot be tested under laboratory circumstances as maximum large-scale shear tests are time-consuming and complicated, and it is hard to predict the nonlinear shear strength function without an analytical method (particle size 1200 mm) [8].

Over the last ten years, a newly developed approach based on machine learning (ML) algorithms has been widely applied to solve real-world problems, particularly civil engineering. Numerous practical problems have been effectively addressed using ML techniques, paving the way for many promising opportunities in civil engineering and other fields such as environmental [9] and geotechnical [10–15] including prediction of RFM shear strength [16–18]. In this context, the artificial neural network (ANN) approach is utilized by Kaunda [16] for estimating RFM shear strength. Cubist and random forest regression techniques are used by Zhou et al. [17], and they found that both models are accurate for RFM shear strength estimations than ANN and traditional regression models. Ahmad et al. [18] used support vector machine (SVM), random forest (RF), AdaBoost, and K-nearest neighbor (KNN) algorithms to estimate the shear strength of RFM and concluded that the SVM model achieved a better prediction performance compared to the RF, AdaBoost, and KNN models. This field, however, is currently being investigated. The article aims to provide the following contributions in the research field:

(i) To evaluate the predictive capacity of the XGBoost algorithm for the shear strength of RFM

(ii) To compare the proposed model to the reference models used in the published literature

(iii) Conduct sensitivity analysis to assess the influence of each input parameter on the RFM's shear strength

The structure of the paper is as follows: The theory of extreme gradient boosting is explained in Section 2. Data collection and correlation analysis are presented in Section 3. Section 4 explains the performance measurement employed. Section 5 presents the obtained results and a discussion of them. Finally, conclusions based on the achieved results are provided.

2. Extreme Gradient Boosting (XGBoost)

Chen and Guestrin [19] proposed the sophisticated supervised technique extreme gradient boosting (XGBoost) under the gradient boosting framework which has received widespread recognition in Kaggle machine learning contests due to its advantages of high efficiency and considerable flexibility. XGBoost's loss function adds a regularization term to the objective function, which helps to smoothen the final learning weights and avoid over-fitting [19]. It also optimizes the loss function using first and second-order gradient statistics. XGBoost also supports row and column sampling to address this issue in addition to providing regular terms to prevent over-fitting. As a result of the parallel and distributed computation, faster model exploration is possible.

The following is a description of the XGBoost algorithm [20]: given a dataset with n examples and m features $D = \{(x_i, y_i)\}$ $(|D| = n, x_i \in R^m, y_i \in R)$,$K$ additive functions will be used to predict the output values of a tree ensemble model as follows:

$$\widehat{y}i = \sum_{k=1}^{K} f_k(x_i), f_k \in F, \tag{1}$$

where F is the regression trees space. It is calculated as

$$F = \left\{f(x) = \omega_q(x)\right\}\left(q: R^m \longrightarrow T, \omega_q \in R^T\right), \tag{2}$$

where q represents for the structure of each tree, T represents for the number of leaves in the tree, and f_k is a function that corresponds to an independent tree structure q and leaf weights ω. To reduce errors of ensemble trees, the objective function is found in the XGBoost model:

$$L^{(t)} = \sum_{i=1}^{n} l\left(\left(y_i, \widehat{y}_j^{(t-1)}\right) + f_t(x_i)\right) + \Omega(f_k), \tag{3}$$

where l is a differentiable convex objective function to calculate the error between predicted and measured values; y_i and $\widehat{y}_i$ are regulated and predicted values, respectively; t shows the repetitions in order to minimize the errors; and Ω is the complexity penalized with the regression tree functions:

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda\|\omega\|^2, \quad (4)$$

ω is the vector of the score for the blades, and γ the minimal loss required for the further isolation of a blade node. λ is the regularization function. In addition, γ and λ are parameters which are able to control the complexity of the tree, and the regularization term helps to avoid overfitting by smoothening the final learnt weights. Taylor expansion is applied to the objective function in order to further simplify it as

$$F = \sum_{i=1}^{m}\left[f_t(x_i)g_i + \frac{1}{2}(f_t(x_i))^2 h_i\right] + \gamma T + \frac{1}{2}\lambda\sum_{j=1}^{T}\omega_j^2, \quad (5)$$

where g_i and h_i are the first and second derivatives obtained on the loss function, respectively. More detailed explanations of the XGBoost algorithm can be found in Chen and Guestrin's [19] research paper.

3. Dataset Collection and Correlation Analysis

In this study, a database of 165 samples of RFM shear strength reports was collected from Kaunda [16] and is presented in Appendix A and Table A1 in supplementary file. All input parameters that might influence the shear strength results of RFM were considered. The included parameters are D_{10}, D_{30}, D_{60}, and $D_{90,}$ corresponding to the 10%, 30%, 60%, and 90% sieve sizes passing, respectively. C_c and C_u refer to the curvature uniformity coefficients (C_c), respectively; FM and GM describe fineness modulus and gradation modulus, respectively; R represents International Society of Rock Mechanics (ISRM) hardness rating; $UCS_{min,}$ and UCS_{max} (MPa) signify the uniaxial compression strengths boundaries (MPa); and γ represents the dry unit weight (kN/m^3), while σ_n is the normal stress (MPa). The considered output is the shear strength of RFM (MPa) (denoted as τ (MPa)). The summary of the database statistics is presented in Table 1, which includes the boundary and standard deviation values of all parameters used in this study.

Correlation (ρ) was used to verify the intensity of correlation between different parameters (see Figure 1). For a given pair of random variables (m, n), the following equation for ρ is used:

$$\rho(m,n) = \frac{\mathrm{cov}(m,n)}{\sigma_m\sigma_n}, \quad (6)$$

where cov denotes covariance, σ_m denotes the standard deviation of m, and σ_n denotes the standard deviation of n. $|\rho| > 0.8$ represents a strong correlation between m and n, values between 0.3 and 0.8 represents a moderate relationship, and $|\rho| > 0.30$ represents a weak relationship [21]. As per Song et al. [22], correlation is considered as "strong" if $|\rho| > 0.8$. In the order of strong to weak, the relationships between input and output parameters are represented in Figure 1. Consequently, no factors from the estimation model's τ were deleted. The correlation coefficient has a maximum absolute value of 0.97, as shown in Figure 1.

4. Evaluation and Prediction

To evaluate the predictive capacity of the XGBoost algorithm, we compared it with some other machine learning methods developed in literature using performance measures.

4.1. Compared Machine Learning (ML) Methods. The XGBoost model was compared with other prediction methods such as support vector machine, adaptive boosting, random forest, and K-nearest neighbor proposed in literature. A brief description of each technique is presented. For a more in-depth discussion, the reader is referred to the relevant references.

4.1.1. Support Vector Machine (SVM). The Support Vector Machine (SVM) regression technique relies on feature classification and generates an interclass hyperplane and minimizes the vector lengths and variance between the features and the plane. The SVM is compatible with the majority of kernel types, including Euclidean, Gaussian, Exponential, and Dirichlet kernels [23]. The objective function for SVM regression contains a coefficient generated from the cost analysis that aids in determining the flatness of the created hyperplane [24]. This allows the user to change the SVM technique to fit unique datasets.

4.1.2. Adaptive Boosting (AdaBoost). Adaptive Boosting is a boosting machine learning technique in which strong learning algorithms augment weak learning algorithms. AdaBoost must define the number of beginning students (n) as a parameter [25]. During the training phase, AdaBoost develops learners with low accuracy who improve based on their predecessors [26]. Using this method, the AdaBoost dynamically modifies the training weight based on the performance of the fundamental learning algorithms [27].

4.1.3. Random Forest (RF). Random Forests are ensemble models that use many decision trees as base-learners to obtain more precise outcomes. Individual trees are generated from training data using random parameters as their roots and nodes using the bootstrap sampling method [28]. Multiple decision trees are more stable than a single tree because they reduce overfitting and average the outcomes [26]. The number of trees in the forest at each binary node, the number of randomly selected predictors, and the lowest number of observations at the nodes of the trees are the three primary parameters for random forests [29].

4.1.4. K-nearest Neighbor (KNN). The supervised KNN is a machine learning algorithm that can be used to tackle both classification and regression problems. In regression problems, the input data set is comprised of k that is most similar to the training data sets utilized in the highlighted set. The outcome of KNN regression is the object's characteristic value, which is the mean value of k's nearest neighbors. As

Table 1: Statistics of parameters of the training and testing datasets.

Statistical parameter	Dataset	Input variable													Output variable
		D_{10} (mm)	D_{30} (mm)	D_{60} (mm)	D_{90} (mm)	C_C	C_U	GM	FM	R	UCS_{min} (MPa)	UCS_{max} (MPa)	γ (kN/m^3)	σ_n (MPa)	τ (MPa)
Minimum	Total data	0.010	0.560	1.200	2.600	0.100	1.360	0.200	3.000	1.000	1.000	5.000	9.320	0.002	0.005
	Training	0.010	0.560	1.200	2.600	0.100	1.360	0.200	3.000	1.000	1.000	5.000	9.320	0.002	0.005
	Testing	0.010	0.560	1.200	2.600	0.100	1.470	0.200	3.000	1.000	1.000	5.000	9.320	0.021	0.024
Maximum	Total data	33.900	42.400	80.100	100.000	22.270	1040.000	6.000	8.800	6.000	250.000	400.000	38.900	4.205	3.921
	Training	33.900	42.400	80.100	100.000	22.270	1040.000	6.000	8.800	6.000	250.000	400.000	38.900	4.205	3.921
	Testing	33.900	42.400	50.000	99.000	22.270	1040.000	6.000	8.800	5.000	100.000	250.000	38.900	3.223	2.492
Mean	Total data	4.463	7.860	18.280	39.927	2.404	69.561	2.903	6.142	4.327	73.691	168.455	20.799	0.734	0.662
	Training	4.867	8.465	19.287	40.386	2.199	53.324	2.788	6.250	4.364	75.045	170.682	20.766	0.729	0.660
	Testing	2.887	5.442	14.252	38.091	3.226	134.510	3.365	5.709	4.182	68.273	159.545	20.932	0.756	0.668
Standard deviation	Total data	8.875	10.335	14.420	22.432	3.414	193.628	1.278	1.298	0.957	37.975	87.844	4.861	0.785	0.652
	Training	9.179	10.577	15.135	22.018	3.075	156.064	1.243	1.261	0.910	39.230	88.010	4.605	0.780	0.662
	Testing	7.453	9.050	10.349	24.289	4.492	194.958	1.331	1.374	1.131	32.444	87.967	5.854	0.816	0.619

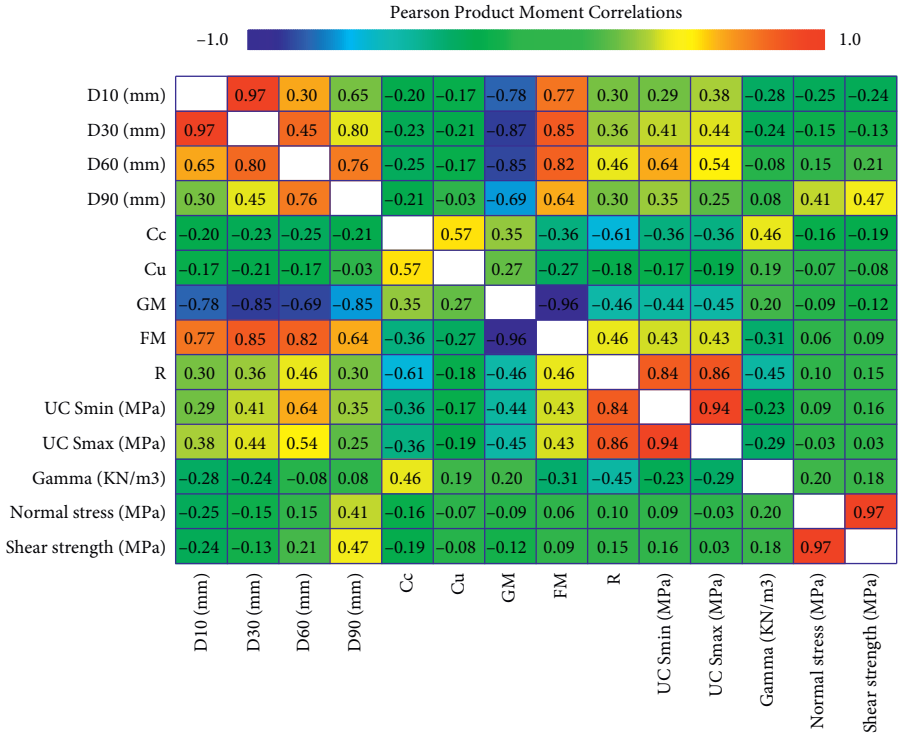

Figure 1: Correlation coefficient between parameters.

the distance metric, a parameter such as Euclidean or Mahalanobis distance can be utilized to locate the k of a data point [30].

4.2. Evaluation Measures. Three quantitative statistical indices, i.e., coefficient of determination (R^2), error in the root mean square ratio to the measured data standard deviation (RSR), and Nash–Sutcliffe coefficient (NSE) were employed to validate and compare the XGBoost model. The following equations characterise the supplied indices:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \hat{y})^2}, \tag{7}$$

$$RSR = \frac{\sqrt{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}}{\sqrt{\sum_{i=1}^{n} (y_i - \overline{y})^2}}, \tag{8}$$

$$NSE = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2}, \tag{9}$$

where n is the total number of data; y_i and $\hat{y}_i$ are the actual shear strength and the predicted shear strength, respectively; and $\overline{y}$ is the mean of the actual shear strength.

Values of the coefficient of determination (R^2) that are closer to 1 imply that this model better fits the data. When R^2 is greater than 0.8 and close to 1, the model is deemed robust [31]. The NSE is a normalized statistic that regulates the level of residual variance compared to the variance of the data being measured [32]. The NSE scale ranges from $-\infty$ to 1, with 1 denoting an ideal match. If the NSE value is greater than 0.65, a strong correlation exists [32, 33]. The root mean square error (RMSE)-standard deviation ratio (RSR) is computed by dividing the RMSE by the standard deviation of the observed data. The RSR varies from 0, representing the optimal value, to a significant positive value. The RSR ranges from the optimal value of 0 to a substantial positive number. Classification ranges are expressed as very good, good, acceptable, and unacceptable. The RSR ranges are $0.000 \le \text{RSR} \le 0.500$, $0.500 \le \text{RSR} \le 0.600$, $0.600 \le \text{RSR} \le 0.700$, and $\text{RSR} > 0.700$, respectively [34].

5. Methodology

The present study is carried out based on the proposed framework that involves four main steps as follows: (1) data preparation and correlation analysis, (2) development of the model, (3) validation of the proposed model, and (4) sensitivity analysis (Figure 2):

(1) Data preparation and correlation analysis: In this first step, the data of samples from the laboratory were utilized to build the training and testing datasets. The training dataset was constructed using 80% of the total data, while the testing dataset was built from the remaining 20%.

(2) Development of the model: In this second step, the training dataset was applied for training the model based on the XGBoost algorithm. The optimization of user defined parameters is undertaken by carrying out multiple runs with these parameters on the training data and analyzing the performance of the

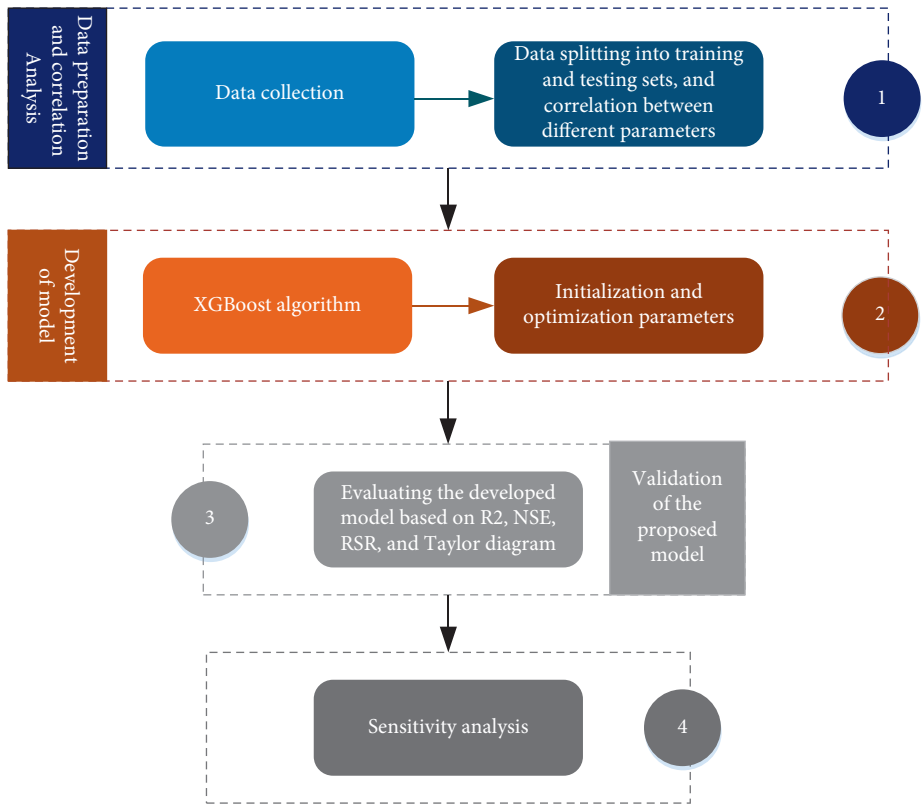

Figure 2: Flowchart illustrates the proposed methodology for present study.

resulting models on testing data. All training and testing operations were conducted out in Orange software.

(3) Validation of the proposed models: In this third step, the testing dataset was adopted for validating the proposed models. Statistical indices including R^2, NSE, and RSR were applied to validate the models. The proposed model is compared to the reference models used in the published literature. Furthermore, Taylor diagram is utilized to illustrate how similar the models (including the proposed XGBoost) are to the reference/observed point position.

(4) Sensitivity analysis: In the last step, sensitivity analysis is used for evaluating the influence of input factors on the shear strength of rockfill material.

6. Results and Discussion

The proposed model that estimates the RFM shear strength is developed using orange software. The predictor variables were provided via an input set (x) defined by $x = [D_{10}, D_{30}, D_{60}, D_{90}, C_c, C_u$, GM, FM, R, $\mathrm{UCS}_{\min}$, $\mathrm{UCS}_{\max}$, γ, and $\sigma_n]$, while the target variable (y) is shear strength (τ) of the rockfill material. Every modelling stage requires the selection of the suitable size of training and testing datasets. Consequently, 80% (132 cases) of the total data were employed to generate models while the remaining 20% (33 cases) of the data were used to test the developed models in this study. The XGBoost model was tuned through trial and error to get an optimal hyperparameters values owing to accurate estimate of the shear strength of rockfill materials. This study optimizes some essential XGBoost parameters and clarifies the definitions of these hyperparameters. The tuning parameters for the model were selected and then changed during the trials until the best metrics from Table 2 were obtained.

The predictive performance of the training and testing datasets is shown in regression form in Figure 3. In terms of training, the XGBoost model produced the best prediction results (i.e., $R^2 = 0.9707$, NSE = 0.9701 and RSR = 0.1729) compared to SVM (i.e., $R^2 = 0.9655$, NSE = 0.9639 and RSR = 0.1899), RF (i.e., $R^2 = 0.9545$, NSE = 0.9542, and RSR = 0.2140), AdaBoost (i.e., $R^2 = 0.9390$, NSE = 0.9388, and RSR = 0.2474), and KNN (i.e., $R^2 = 0.6233$, NSE = 0.6180, and RSR = 0.6181). It is also verified by the findings of R^2, NSE, and RSR in Figure 4 as XGBoost produced lesser RSR, higher R^2, and NSE values compared to SVM, RF, AdaBoost, and KNN models developed in the literature by Ahmad et al. [18] and the parameter optimization is presented in Table 2.

As depicted in Figure 4, the XGBoost model performed the best in terms of R^2, NSE, and RSR (i.e., $R^2 = 0.9676$, NSE = 0.9672, and RSR = 0.1812) compared to SVM (i.e., $R^2 = 0.9656$, NSE = 0.9654, and RSR = 0.1861), RF (i.e., $R^2 = 0.9656$, NSE = 0.9164, and RSR = 0.2891), AdaBoost (i.e., $R^2 = 0.9181$, NSE = 0.8835, and RSR = 0.3414), and KNN (i.e., $R^2 = 0.6304$, NSE = 0.6076, and RSR = 0.6264) in the testing phase. The outcomes of this and a prior study by Ahmad et al. [18] (see Figure 4) demonstrate that the ML method may accurately predict the shear strength of RFMs. The

Table 2: Parameter configuration.

Algorithm	Parameter optimization
XGBoost	n estimators = 40, learning rate = 0.250, maximum depth = 4
SVM	Cost = 8, regression loss epsilon = 0.1, kernel type = radial basis function
RF	Number of trees = 15, limit depth of individual trees = 3
KNN	Number of neighbors = 5, metric = euclidean, weight = uniform
AdaBoost	Number of estimators = 2, learning rate = 0.1, boosting algorithm = SAMME, regression loss function = linear

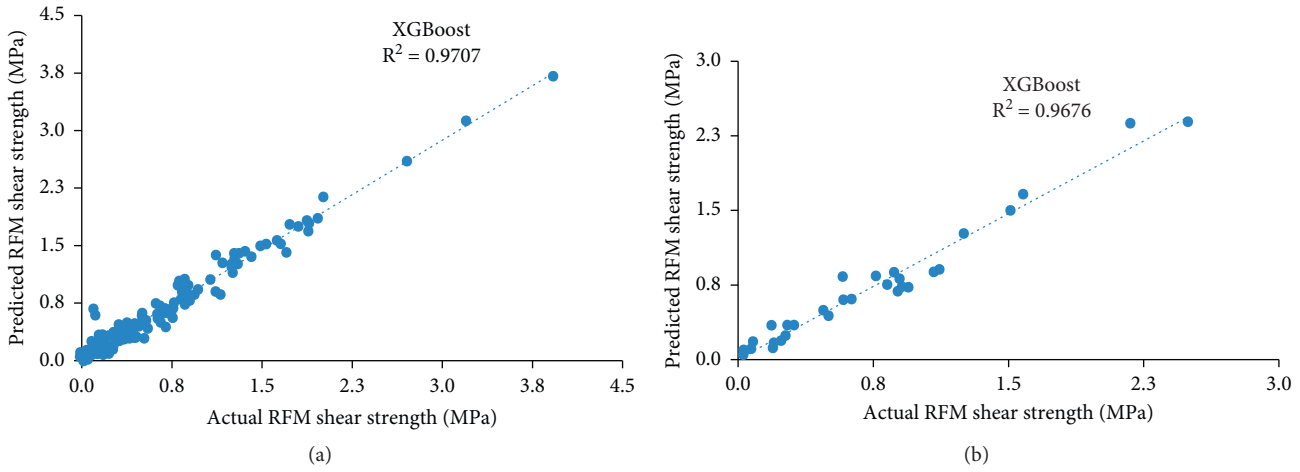

Figure 3: Regression graph of the XGBoost model for (a) training and (b) testing datasets.

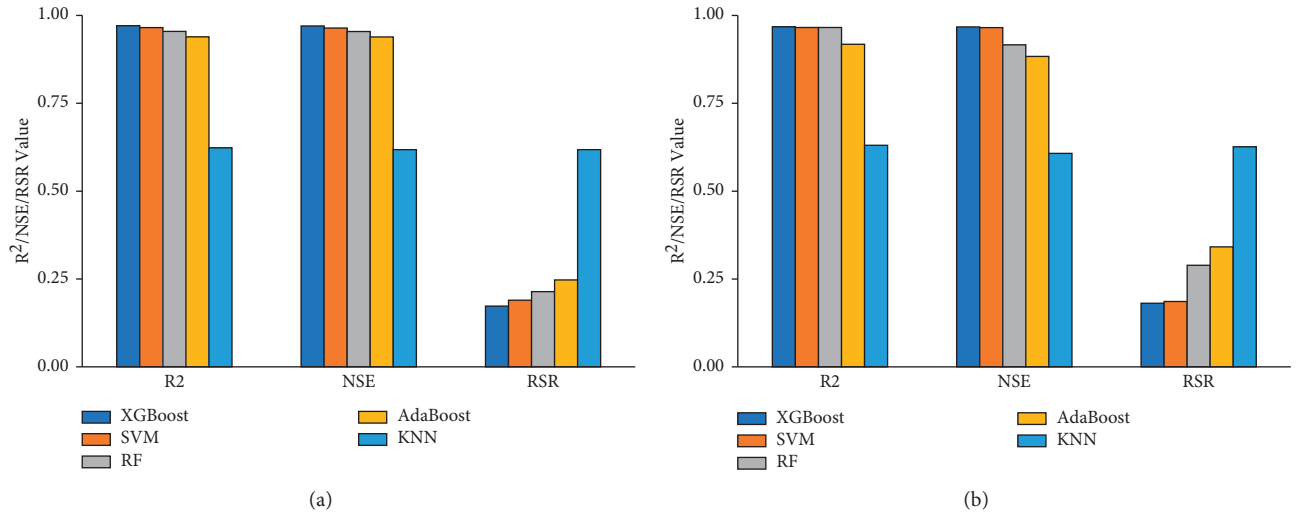

Figure 4: Comparison of R^2, NSE, and RSR values from the XGBoost, SVM, RF, AdaBoost, and KNN models in (a) training; and (b) testing phases.

comparison of study outcomes makes sense because the data sets and inputs are the same. In contrast, the XGBoost model beats the other models in terms of predictive performance and offered a balanced prediction throughout the training and testing data sets. In addition, due to the study's small data set, additional research on other data sets is necessary to establish the most generic model for predicting the shear strength of RFM.

The difference between the actual and predicted shear strength of RFM is represented in Figure 5 by comparing the results of the training and testing sets. The proposed XGBoost model is satisfactory for predicting the RFM shear strength, barring a few noise points.

Taylor diagram (see Figure 6) is utilized to illustrate how similar the models (including the proposed XGBoost) are to the reference/observed point position based on their correlation, root-mean-square error difference, and amplitude of their variations (represented by their standard deviations). The better the performance, the closer each model point is to the position of the reference/observed point. In terms of predictive ability, the proposed XGBoost model beats the SVM, RF, AdaBoost, and KNN models developed in the literature by Ahmad et al. [18].

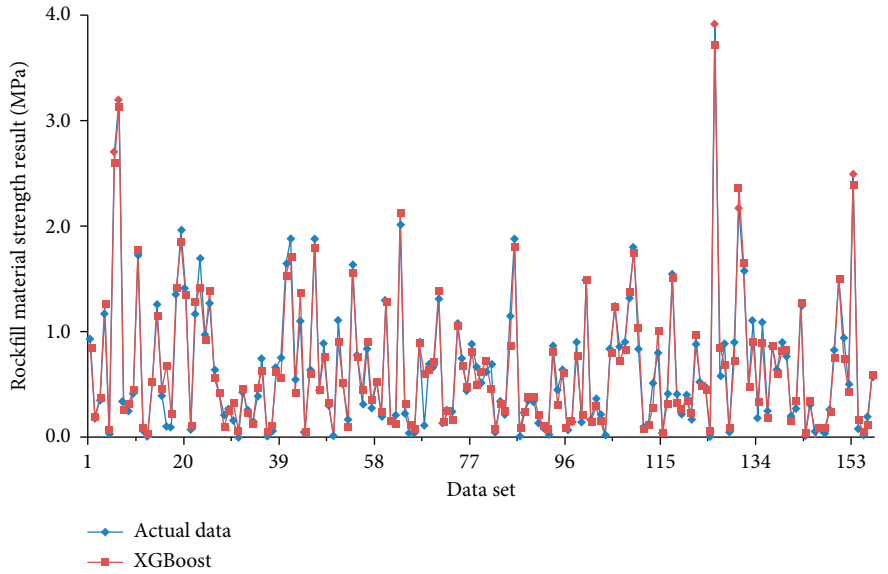

Figure 5: Results of XGBoost model training and testing phases for rockfill material shear strength.

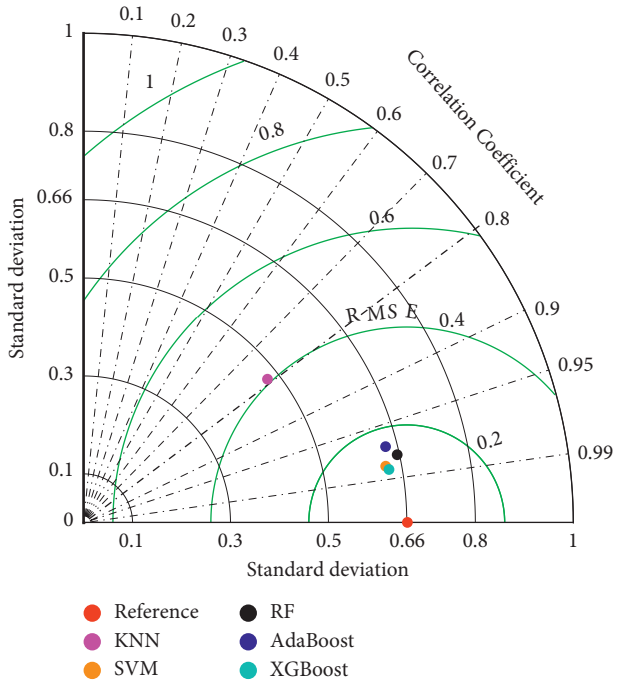

Figure 6: Taylor diagram of the models.

The sensitivity results of the XGBoost model were evaluated utilising Yang and Zang's [35] approach for evaluating the influence of input factors on the shear strength of rockfill material. This approach, which has been the topic of numerous studies [36–41], is as follows:

$$r_{ij} = \frac{\sum_{m=1}^{n} (y_{im} \times y_{om})}{\sqrt{\sum_{m=1}^{n} y_{im}^2 \sum_{m=1}^{n} y_{om}^2}}, \tag{10}$$

where n represents the number of values (i.e., 132); y_{im} and y_{om} denotes input and output variables, respectively. For each input parameter, the r_{ij} value ranges from zero to one, with the greatest r_{ij} values indicating the efficient output variable (i.e., τ). Figure 7 shows the r_{ij} scores for all input variables and demonstrates that σ_n ($r_{ij} = 0.99$) has the greatest effect on the shear strength of rockfill material. Furthermore, Figure 1 shows that the normal stress σ_n has the highest ρ of 0.97 in all other parameters validating the sensitivity analysis results.

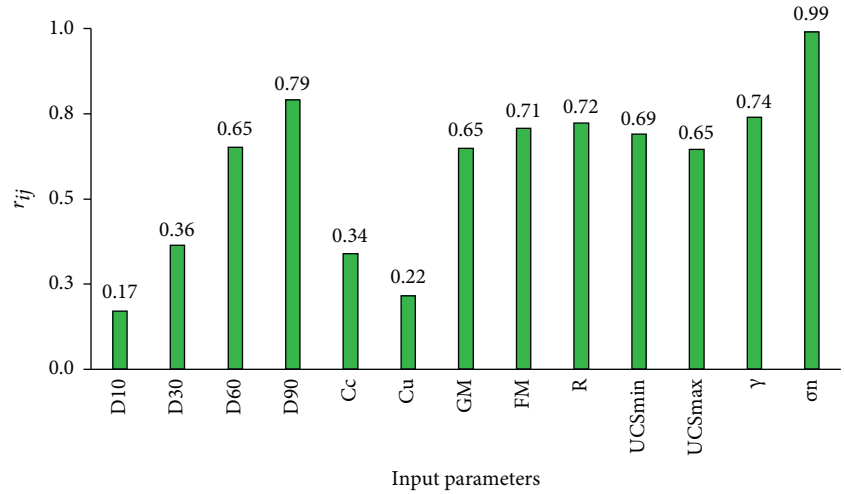

Figure 7: Sensitivity analysis results.

7. Conclusions

Using an XGBoost algorithm, a new prediction model for RFM shear strength is proposed in the current study. Comparisons reveal that the proposed XGBoost model provides the most accurate prediction of the RFM's shear strength when compared to the algorithms developed using the SVM, RF, AdBoost, and KNN model. Important findings found from this study include as follows:

(1) In the test phase, results showed that the XGBoost had the highest power performance ($R^2 = 0.9676$, NSE = 0.9672, and RSR = 0.1812) compared to other machine learning models. Furthermore, based on the scatter plots of actual and predicted values, the XGBoost model exhibited a better fit to the observed data, indicating that it has potential for broader applications in RFM material properties prediction.

(2) Compared to SVM, RF, AdaBoost, and KNN models in the literature, the proposed XGBoost model has a superior predictive capability. In addition, the proposed model is amenable to further modification so that the accumulation of further data will considerably enhance its predictive potential.

(3) The findings of the sensitivity analysis indicate that five parameters, namely, the normal stress, the 90% passing sieve diameters (D_{90}), the dry unit weight, and the ISRM hardness rating, are the most sensitive and important factors for estimating the shear strength of rockfill materials.

(4) The developed XGBoost model gives predictions with the same level of accuracy as existing soft computing methods.

Since the proposed XGBoost model produces predictions based on the input values, interpolation between the input variables is more accurate and reliable than extrapolation. Therefore, the model should not be used for input parameter values beyond the defined range of the study.

Data Availability

The data presented in this study are available in Appendix A, Table A1 (see supplementary file).

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgments

The research was partially funded by the Ministry of Science and Higher Education of the Russian Federation under the strategic academic leadership program "Priority 2030" (Agreement 075-15-2021-1333 dated 30.09.2021).

Supplementary Materials

Table A1. Dataset used in the development and validation of the model. (*Supplementary Materials*)

References

[1] S. Abbas, A. Varadarajan, and K. Sharma, "Prediction of shear strength parameter of prototype rockfill material," *IGC-2003, Roorkee*, vol. 1, pp. 5–8, 2003.

[2] A. K. Gupta, "Constitative Modelling of Rockfill Materials," vol. 6, 2000.

[3] K. Venkatachalam, "Prediction of Mechanical Behaviour of Rockfill Materials," vol. 15, 1993.

[4] R. J. Marsal, "Large scale testing of rockfill materials," *Journal of the Soil Mechanics and Foundations Division*, vol. 93, no. 2, pp. 27–43, 1967.

[5] N. D. Marachi, "Strength and deformation, Characteristics of Rockfill Materials," *Report No. TE-69-5 to State of California Department of Water Resources*, U S A, 1969.

[6] N. P. Honkanadavar and K. G. Sharma, "Testing and modeling the behavior of riverbed and blasted quarried rockfill materials," *International Journal of Geomechanics*, vol. 14, no. 6, 2014.

[7] E. Frossard, W. Hu, C. Dano, and P.-Y. Hicher, "Rockfill shear strength evaluation: a rational method based on size effects," *Géotechnique*, vol. 62, no. 5, pp. 415–427, 2012.

[8] N. Honkanadavar and S. Gupta, "Prediction of shear strength parameters for prototype riverbed rockfill material using index properties," *Proceedings of Indian Geotechnical Conf*, vol. 55, pp. 335–338, —2010.

[9] A. Froemelt, D. J. Dürrenmatt, and S. Hellweg, "Using data mining to assess environmental impacts of household consumption behaviors," *Environmental Science & Technology*, vol. 52, no. 15, pp. 8467–8478, 2018.

[10] A. Mahmood, X.-W. Tang, J.-N. Qiu, W.-J. Gu, and A. Feezan, "A hybrid approach for evaluating CPT-based seismic soil liquefaction potential using Bayesian belief networks," *Journal of Central South University*, vol. 27, no. 2, pp. 500–516, 2020.

[11] M. Ahmad, X.-W. Tang, J.-N. Qiu, and F. Ahmad, "Evaluating seismic soil liquefaction potential using bayesian belief network and C4. 5 decision tree approaches," *Applied Sciences*, vol. 9, no. 20, p. 4226, 2019.

[12] M. Ahmad, X. Tang, J. Qiu, F. Ahmad, and W. Gu, "LLDV-A comprehensive framework for assessing the effects of liquefaction land damage potential," in *Proceedings of the 2019 IEEE 14th International Conference on Intelligent Systems and Knowledge Engineering*, pp. 527–533, ISKE), Dalian, China, November 2019.

[13] M. Ahmad, X.-W. Tang, J.-N. Qiu, F. Ahmad, and W.-J. Gu, "A step forward towards a comprehensive framework for assessing liquefaction land damage vulnerability: exploration from historical data," *Frontiers of Structural and Civil Engineering*, vol. 14, no. 6, pp. 1476–1491, 2020.

[14] M. Ahmad, X. Tang, and F. Ahmad, "Evaluation of liquefaction-induced settlement using random forest and REP tree models: taking pohang earthquake as a case of illustration," *Natural Hazards-Impacts, Adjustments & Resilience, IntechOpen*, vol. 44, 2020.

[15] M. Ahmad, N. A. Al-Shayea, X.-W. Tang, A. Jamal, H. M Al-Ahmadi, and F. Ahmad, "Predicting the pillar stability of underground mines with random trees and C4. 5 decision trees," *Applied Sciences*, vol. 10, no. 18, p. 6486, 2020.

[16] R. Kaunda, "Predicting shear strengths of mine waste rock dumps and rock fill dams using artificial neural networks," *International Journal of Mining and Mineral Engineering*, vol. 6, no. 2, p. 139, 2015.

[17] J. Zhou, E. Li, H. Wei, C. Li, Q. Qiao, and D. J. Armaghani, "Random forests and cubist algorithms for predicting shear strengths of rockfill materials," *Applied Sciences*, vol. 9, no. 8, p. 1621, 2019.

[18] M. Ahmad, P. Kamiński, P. Olczak et al., "Development of prediction models for shear strength of rockfill material using machine learning techniques," *Applied Sciences*, vol. 11, no. 13, p. 6167, 2021.

[19] T. Chen and C. X. Guestrin, "A scalable tree boosting system," in *Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, California CA U S A, August 2016.

[20] H. Nguyen, X.-N. Bui, H.-B. Bui, and D. T. Cuong, "Developing an XGBoost model to predict blast-induced peak particle velocity in an open-pit mine: a case study," *Acta Geophysica*, vol. 67, no. 2, pp. 477–490, 2019.

[21] T. van Vuren, *Modeling of Transport Demand–Analyzing, Calculating, and Forecasting Transport Demand: By VA Profillidis and GN Botzoris*, p. 472, Elsevier, Amsterdam, 2018.

[22] Y. Song, J. Gong, S. Gao et al., "Susceptibility assessment of earthquake-induced landslides using Bayesian network: a case study in Beichuan, China," *Computers & Geosciences*, vol. 42, pp. 189–199, 2012.

[23] B. Schölkopf, A. J. Smola, and F. Bach, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and beyond*, MIT press, 2002.

[24] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and Computing*, vol. 14, no. 3, pp. 199–222, 2004.

[25] M. Moeini, A. Shojaeizadeh, and M. Geza, "Supervised machine learning for estimation of total suspended solids in urban watersheds," *Water*, vol. 13, no. 2, p. 147, 2021.

[26] M. Zounemat-Kermani, O. Batelaan, M. Fadaee, and R. Hinkelmann, "Ensemble machine learning paradigms in hydrology: a review," *Journal of Hydrology*, vol. 598, p. 126266, 2021.

[27] T. Yang, X. Liu, L. Wang, P. Bai, and J. Li, "Simulating hydropower discharge using multiple decision tree methods and a dynamical model merging technique," *Journal of Water Resources Planning and Management*, vol. 146, no. 2, 2020.

[28] S. Ardabili, A. Mosavi, and A. R. Várkonyi-Kóczy, "Advances in machine learning modeling reviewing hybrid and ensemble methods," *Proceedings of International Conference on Global Research and Education*, pp. 215–227.

[29] L. T. Pham, L. Luo, and A. Finley, "Evaluation of random forests for short-term daily streamflow forecasting in rainfall- and snowmelt-driven watersheds," *Hydrology and Earth System Sciences*, vol. 25, no. 6, pp. 2997–3015, 2021.

[30] V. Prasath, H. A. A. Alfeilat, A. Hassanat et al., "Distance and Similarity Measures Effect on the Performance of K-Nearest Neighbor Classifier--A Review," 2017, https://arxiv.org/abs/1708.04321.

[31] A. H. Gandomi, S. K. Babanajad, A. H. Alavi, and Y. Farnam, "Novel approach to strength modeling of concrete under triaxial compression," *Journal of Materials in Civil Engineering*, vol. 24, no. 9, pp. 1132–1143, 2012.

[32] J. Nash and J. V. Sutcliffe, "River flow forecasting through conceptual models part I—a discussion of principles," *Journal of Hydrology*, vol. 10, no. 3, pp. 282–290, 1970.

[33] D. N. Moriasi, J. G. Arnold, M. W. Van Liew, R. L. Bingner, R. D. Harmel, and T. L. Veith, "Model evaluation guidelines for systematic quantification of accuracy in watershed simulations," *Transactions of the ASABE*, vol. 50, no. 3, pp. 885–900, 2007.

[34] K. Khosravi, L. Mao, O. Kisi, Z. M. Yaseen, and S. Shahid, "Quantifying hourly suspended sediment load using data mining models: case study of a glacierized Andean catchment in Chile," *Journal of Hydrology*, vol. 567, pp. 165–179, 2018.

[35] Y. Yang and Q. Zhang, "A hierarchical analysis for rock engineering using artificial neural networks," *Rock Mechanics and Rock Engineering*, vol. 30, no. 4, pp. 207–222, 1997.

[36] R. Shirani Faradonbeh, D. Jahed Armaghani, M. Z. Abd Majid et al., "Prediction of ground vibration due to quarry blasting based on gene expression programming: a new model for peak particle velocity prediction," *International journal of Environmental Science and Technology*, vol. 13, no. 6, pp. 1453–1464, 2016.

[37] W. Chen, M. Hasanipanah, H. Nikafshan Rad, D. Jahed Armaghani, and M. M. Tahir, "A new design of evolutionary hybrid optimization of SVR model in predicting the blast-induced ground vibration," *Engineering with Computers*, vol. 37, no. 2, pp. 1455–1471, 2019.

[38] H. Nikafshan Rad, I. Bakhshayeshi, W. A. Wan Jusoh, M. M. Tahir, and L. K. Foong, "Prediction of flyrock in mine

blasting: a new computational intelligence approach," *Natural Resources Research*, vol. 29, no. 2, pp. 609–623, 2020.

[39] M. H. Ahmad, F. Ahmad, X.-W. Tang et al., "Supervised Learning Methods for Modeling Concrete Compressive Strength Prediction at High Temperature," *Materials*, vol. 14, 2021.

[40] M. Ahmad, M. Amjad, R. A. Al-Mansob et al., "Prediction of liquefaction-induced lateral displacements using Gaussian process regression," *Applied Sciences*, vol. 12, no. 4, p. 1977, 2022.

[41] M. Amjad, I. Ahmad, M. Ahmad, P. Wróblewski, P. Kamiński, and U. Amjad, "Prediction of pile bearing capacity using XGBoost algorithm: modeling and performance evaluation," *Applied Sciences*, vol. 12, no. 4, p. 2126, 2022.