Kochi Chapter

Indian Geotechnical Conference
IGC 2022
15th – 17th December, 2022, Kochi

Application of machine learning algorithms for estimation of filter dimensions for an earthen embankment dam

Krishna Prajapati[1] and Arindam Dey[2]

[1]Research Scholar, Department of Civil Engineering, Indian Institute of Technology, Madras-600036
[2]Associate Professor, Department of Civil Engineering, Indian Institute of Technology, Guwahati-781039
arindam.dey@iitg.ac.in

Abstract. Seepage through an embankment must be controlled to prevent concealed internal erosion and migration of fine materials. It is extremely important to control the seepage flow and inhibit removal of the soil particles comprising the dam body. Modern design practise incorporates this control into the dam design through the use of internal filters and adequate drainage provisions. The seepage analysis theories proposed by researchers like Casagrande, Schaffernak, Dupuit, and Pavlovsky for homogenous earthen dam resting on impervious base finds their application in estimation of filter dimension. This paper reports the utilization of machine learning in this regard. A large dataset is generated by using Schaffernak's theory for phreatic surface assessment and by varying the governing parameters in all possible range that affects the filter dimension. The dataset is used for training in suitable algorithms related to Multilayer perceptron (MLP), Random Forest (RF), Support Vector Regression (SVR), Ridge Regression (RR) and Xtreme Gradient Boosting (XGBoost) algorithms. The results illustrated that XGBoost algorithm could potentially be used to estimate filter dimension and exit discharge. These trained models can be used as a ready reference solution by the practising engineers, which provides them a preliminary idea for designing toe filters.

Keywords: Filter dimension; Machine learning; Random Forest; XGBoost, Nonlinear regression; Multilayer perceptron

1 Introduction

Seepage is the slow percolation of liquid through porous medium (soil) when water moves from higher head to lower head. The water held in the reservoir of earthen dam may often lead to the seepage within the body and below the dam. It is mandatory to control the content of seeping water which otherwise may lead to the catastrophic failure of dam by means of piping and sloughing. Grouting or installation of upstream blanket, cut offs and internal filters are some of the methodologies adopted in modern practise to control the content of seepage. In the present study, internal filters are taken into consideration because of their effectiveness and economy.

The conventional seepage analysis theory forms the basis for the design of toe-filter. The point of intersection of phreatic surface with the downstream slope of the dam or the toe of the dam aids in the assessment of filter dimension. The difference in the consideration of phreatic surface and hydraulic gradient forms the basis for proposal of several solution for the seepage analysis through body of earthen dam by researcher like Casagrande, Schaffernak, Dupuit, and Pavlovsky. The hydraulic gradient *i* is assumed as the slope of free surface (*dz/dx*) in Dupit and Schaffernak solution while Casagrande solution is based on the arc length (*ds*) and height of elemental strip (*dz*) (i.e., *i = dz/ds*) which increases the complexity of solution as its difficult to estimate the arc length. Further, the phreatic surface assumed by Dupuit's solution is different from that depicted by its mathematical expression i.e., parabolic phreatic surface at entry and exit point. Casagrande and Schaffernak solutions make use of parabolic phreatic surface throughout the dam cross section. In Palvovsky's solution, the dam cross section has been divided into three zones such that the phreatic surface for first and last zone is assumed as linear while the middle zone has parabolic profile. This limits the applicability of Palvovsky's solution for most of the practical cases. In the present study, Schaffernak's solution has been adopted for estimating the filter dimension because of its wider applicability and easier analysis, the schematic representation of which is shown in Fig. 1.

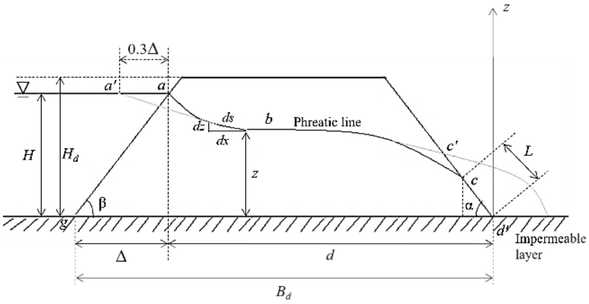

Fig. 1. Schematic diagram for Schaffernak's analysis

The result obtained from Schaffernak's analysis was found to be functions of various dam parameters such as the slopes of upstream face (β) and the downstream face (α), the ratio of crest width to height of dam (B/H_d), and the ratio of upstream water level to the height of dam (H/H_d). A set of total 6860 data has been generated by varying the governing parameters in all possible ranges. The present paper describes the application of Machine Learning (ML) algorithms to estimate the filter dimension obtained by Schaffernak's analyses. Various ML algorithms namely, Multilayer perceptron (MLP), Random Forest (RF), Support Vector Regression (SVR), Ridge Regression (RR) and Xtreme Gradient Boosting (XGBoost) were trained using the generated dataset. These models accept the dam-related parameters as inputs (as described in the the previous paragraph). The outputs obtained from the ML-base analyses comprise the filter dimensions and exit discharge from filter, namely the ratio of filter height to the height of dam [$F_{HN} = (L^*\sin\alpha)/H_d$], the ratio of filter width to the base width of dam [$F_{WN} = (L^*\cos\alpha)/b_d$] and the total exit discharge per unit dam length divided by the product of permeability coefficient and height of the dam [$q_{nD} = q/(k^*H_d)$].

2 Machine Learning Algorithms

Machine learning is the subset of artificial intelligence which makes use of complex set of algorithms to simulate human learning process and automatically update their knowledge from experiences to optimize their functionality. Machine learning algorithms generally performs two sets of operation namely, classification and regression. In recent years, the usage of Machine learning and data-driven approaches increased significantly as they had given/shown state-of-the-art results in various field such as weather forecasting, image recognition, traffic prediction, medical diagnosis and more.

The state of art result produced by ML algorithms such as artificial neural nets (ANN), Bayesian network (BN), support vector machine (SVM), and RF proves them to be best alternative solution for complex iterative geotechnical problems (Zhou et al. 2019). Chen and Jia (2016) have used logistic model tree (LMT), random forest (RF), and classification and regression tree (CART) models for prediction of landslide susceptibility in which RF model shows best result. Zhou et al. (2017) used RF algorithm for the prediction of surface movement due to construction of tunnel. Mitu et al. (2021) used ML algorithms for automation of inversion procedure in SASW to predict shear wave velocity profile from dispersion curve.

The training process offered by conventional computational models such as ANN is tedious as the optimal configuration were not previously known and is attained after too many iterations (Zhang and Goh 2013). Moreover, it is evident from previous researches that the machine learning algorithms outperform neural networks when it comes to prediction capability. Machine learning offers the acceptance of ANN in terms of MLP algorithm. Thus, the MLP model has also been used in the present paper to compare its predictive performance relative to other algorithms. The dataset obtained from Schaffernak's solution consist of both input and output; therefore, only the supervised ML algorithms are chosen for training purpose.

2.1 Multilayer Perceptron (MLP)

Machine learning offers the usage of neural networks in terms of MLP algorithms. The working principle of MLP model is similar to that of feed-forward backpropagation neural network. The neurons are the basic building block of MLP model. The signals received at each input channels of a neuron are multiplied with the corresponding connection weight. These weighted inputs are summed up and a bias is added to it. The resultant sum is then filtered by activation function to obtain the output for that neuron (Fig. 2a).

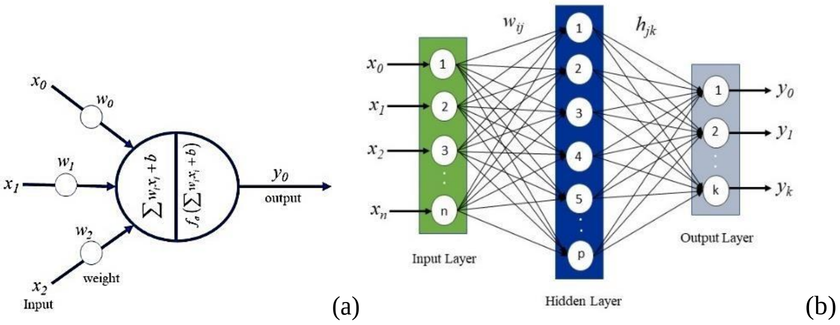

Fig. 2. (a) Schematic representation of an artificial neuron (b) Structure of MLP model with single hidden layer

The network architecture of MLP consists of input layer, output layer and hidden layer. The number of hidden layers is flexible and is decided on the basis of complexity of dataset. Thus, MLP can be considered as a tool for deep learning. The usage of feed-forward backpropagation technique for training purpose optimizes synaptic weight and minimizes mean square error in prediction during each iteration. Each neuron in a given layer is fully connected to the neurons of next layer. Figure 2b shows the architecture of a single hidden layer neural network with *n*-dimensional input, *p* hidden neurons and *k*-dimensional output. The w_{ij} and h_{jk} represents the connection weight of input layer-hidden layer and hidden layer-output layer respectively.

The input and output for neurons of hidden layer can be formulated mathematically as:

$$s_j^{input} = w_{0j}x_0 + w_{1j}x_1 + \ldots + w_{nj}x_n = \sum_{i=0}^{n} w_{ij}x_i \qquad (1)$$

$$s_j^{output} = f_a^{hidden}(s_j^{input} + b_j^{hidden}) \qquad (2)$$

The output for hidden layer is going to act as an input to the output layer. Thus, the input and output for neurons of output layer can be formulated as:

$$y_k^{input} = h_{0k}s_0^{output} + h_{1k}s_1^{output} + \ldots + h_{pk}s_p^{output} = \sum_{j=0}^{p} h_{jk}s_j^{output} \qquad (3)$$

$$y_k^{output} = f_a^{output}(y_k^{input} + b_k^{output}) \qquad (4)$$

The cost or loss function for MLP is defined in terms of mean squared error as:

$$L = \frac{1}{2}\sum_i\sum_k \left\| y_k^{\text{output}} - y_k^{\text{target}} \right\|^2 \qquad (5)$$

$$L = \frac{1}{2}\sum_i\sum_k \left\| f_a^{output}\left(\sum_{j=0}^{p} h_{jk}s_j^{output} + b_k^{output}\right) - y_k^{\text{Target}} \right\|^2 \qquad (6)$$

2.2 Random Forest (RF)

Random Forest is an ensemble learning technique that makes use of a set of decision tree for predictive analysis (Fig. 3a). It follows bootstrapping of sample in which the original dataset is split into subsets (with replacement). Each subset is then used to train a particular decision tree such that the total number of subset dataset is equal to the number of decision tree present in the RF model. The subset dataset consists of random number of features that is decided on the basis of hyperparameter max_features, which includes auto, sqrt(.), and log(.) functions. Due to this randomness of feature in a subset, each tree will be specialized for some particular region while, at the same time, predict inaccurately for other regions. The set represents subset of dataset where *i* varies from 1 to *N* (total number of decision tree) and *K* represents the total numbers of features in dataset and (X,Y) corresponds to the original dataset.

$$(X_i, Y_i) = \left\{ (x_j, y_j) \;\; for \;\; j = 1,2 \ldots K \;\; and \;\; (x_j, y_j) \sim (X,Y) \right\} \qquad (7)$$

Each tree is trained independently with subset dataset. Once the training is being done, the RF model predict the result by averaging the prediction of each tree.

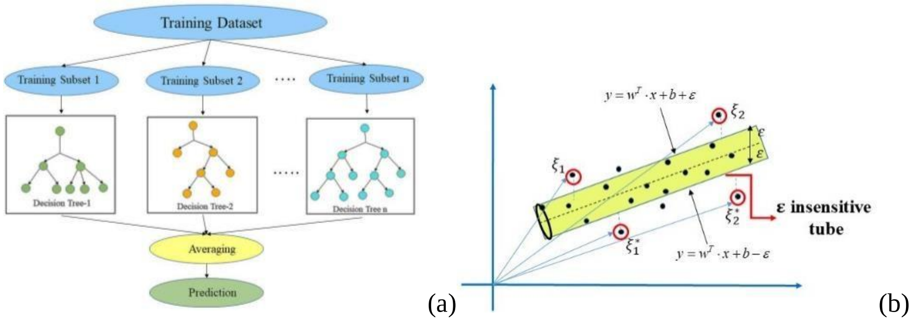

Fig. 3. (a) Workflow of RF model with n decision trees (b) Schematic diagram of SVR

2.3 Support Vector Regression (SVR)

Support vector Machines are the statistical learning-based ML tool that can be used for both classification and regression. SVM make use of kernel function that maps low dimension nonlinear function into higher dimensional space through nonlinear mapping. In higher dimension, SVM classifies the data using appropriate support vector classifier. The selection of adequate kernel function is the most crucial step in SVM. In general, these four types of kernel functions are most commonly used for practical applications: linear, polynomial, radial basis function (RBF), and sigmoid.

$$K(x_i, x_i) = \begin{cases} e^{-\gamma |x_i - x_i|^2} & \text{Radial Basis} \\ \left(\gamma x_i^T \cdot x_j + r\right)^c & \text{Polynomial} \\ \tanh\left(\gamma x_i^T \cdot x_j + r\right) & \text{Sigmoid} \\ x_i \cdot x_j & \text{Linear} \end{cases} \tag{8}$$

SVR is a regression algorithm whose working principle is similar to Support Vector machine (SVM). It is used to predict continuous ordered variable. The main goal in SVR is to confine error within a threshold called fit tube or ε-insensitive tube, for which a function *f(x)* is defined such that it has at most ε deviation from actual target output y_i for all training dataset (Fig. 3b). The data points lying outside the tube are known as slack variables and only these points are considered for error estimation. For a given set of observation samples $(x_1, y_1), (x_2, y_2)......(x_n, y_n) \subset R^n \times R$, let the model equation of regression function is:

$$f(x) = w^T \cdot x + b \tag{9}$$

The objective function for SVR can be mathematically formulated as:

$$Obj = \min\left[\frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{n}\left(\xi + \xi^*\right)\right] \tag{10}$$

where, *C* is a hyper-parameter that influences trade-off between an approximation error and the weights vector norm, and ξ and ξ^* are the slack variables that represent the distance from actual values to the corresponding boundary values of fit-tube.

2.4 Ridge Regression (RR)

Ridge Regression is regularization technique which analyses multiple regression when the data suffers from multicollinearity. The existence of linear relationship between data points can be understood as multicollinearity, due to which least square cost function estimate for training dataset becomes too less i.e., bias is low. However, the same becomes too high for new or test data i.e., variance becomes high, which can be reduced by adding a penalty term in the least square cost function to limit the squared L2 norm. For dataset (X, Y), let the independent variable be $X = \{x_1, x_2 \quad , x_n\}, x_i \in R$ and the dependent variable be $Y = \{y_1, y_2 \quad , y_n\}, y_i \in R$. The model equation for ridge regression is approximated as:

$$y_k = \alpha_0 + \sum_{i=1}^{n} \alpha_1 x_k^i \Rightarrow Y = X\theta + \alpha_0 \tag{11}$$

where, $\alpha_1 = \theta$ is regression coefficient and α_0 is the residual. The mathematical formulation for ordinary least square cost function is given as:

$$L = \|Y - X\theta\|^2 - \lambda \|\theta\|^2 \tag{12}$$

where, λ is the penalty term that is also known as alpha coefficient, which reduces the model complexity by shrinking the regression coefficient. The optimal value of regression coefficient in terms of λ is given as: $\theta_{opt} = \left(X^T X + \lambda I_m\right)^{-1} X^T Y$ (13)

2.5 Xtreme Gradient Boosting (XGBoost)

XGBoost is an ensemble learning technique which is based on decision tree. It involves sequential addition of trees in each iteration to the base learning decision tree in order to minimize the objective function using the Steepest Gradient Descent method. The model equation for the prediction of output by base learning decision tree to final regression tree can be sequentially expressed as:

$$\hat{y}_0 = 0 \tag{14}$$

$$\hat{y}_1 = \alpha f_1(x_1) = \hat{y}_0 + \alpha f_1(x_1) \tag{15}$$

$$\hat{y}_2 = \alpha \sum_{j=1}^{2} f_j(x_j) = \hat{y}_1 + \alpha f_2(x_2) \tag{16}$$

$$\ldots.$$

$$\hat{y}_T = \alpha \sum_{j=1}^{T} f_j(x_j) = \hat{y}_{T-1} + \alpha f_T(x_T) \tag{17}$$

where, T is number of regression trees in model, α is the learning rate ($0<\alpha<1$), $\hat{y}_j$ represents the prediction of output using first j^{th} regression trees, and $f_j(\bullet)$ represents the output of j^{th} regression tree. The initial prediction is generally assumed to be zero or average of all the outputs. The objective function of XGBoost includes an additional regularization term along with conventional loss function which is as follows:

$$Obj=\sum_{i=1}^{L} l\left(y_i,\hat{y}_i\right)+\sum_t \Omega\left(f_t\right) \qquad (18)$$

$$\Omega\left(f_t\right)=\gamma L+\frac{1}{2}\lambda\left\|w\right\|^2 \qquad (19)$$

Here $l\left(y_i,\hat{y}_i\right)=\left(y_i-\hat{y}_i\right)^2$ represents the square loss function, $\Omega\left(f_t\right)$ represents the regularization term that penalizes the complexity of model, *γ* is the complexity cost of introducing additional leaves; *L* is the number of leaf node in t^{th} tree; *λ* is *L2* regularization term on leaf scores; and $\left\|w\right\|^2$ is the weight of the k^{th} leaf node. The objective function for t^{th} regression tree at i^{th} leaf node can be written as:

$$Obj^{(t)}=\sum_{i=1}^{L} l\left(y_i,\hat{y}_i^{t-1}+\alpha f_t\left(x_i\right)\right)+\sum_t \Omega\left(f_t\right) \qquad (20)$$

Using Taylor second order approximation,

$$Obj^{(t)}=\sum_{i=1}^{L} l\left(y_i,\hat{y}_i^{(t-1)}+\alpha g_i f_t\left(x_i\right)+\frac{1}{2}\alpha^2 h_i f_t^2\left(x_i\right)\right)+\sum_t \Omega\left(f_t\right) \qquad (21)$$

$g_i=\partial_{\hat{y}_i^{(t-1)}} l\left(y_i,\hat{y}_i^{(t-1)}\right)$ and $h_i=\partial^2_{\hat{y}_i^{(t-1)}} l\left(y_i,\hat{y}_i^{(t-1)}\right)$ are the first and second order derivative of loss function with respect to $\hat{y}_i^{(t-1)}$. After differentiating Eqn. 21 with respect to $f_t\left(\bullet\right)$ output of t^{th} regression tree and equating it to zero, the optimal value of output for a t^{th} regression tree and corresponding objective function can be given as:

$$f^t=-\frac{\left(\sum_{i=1}^{L} g_i\right)}{\left(\sum_{i=1}^{L} h_i+\lambda\right)} \qquad (22) \qquad\qquad Obj^{(t)}=-\frac{1}{2}\frac{\left(\sum_{i=1}^{L} g_i\right)}{\left(\sum_{i=1}^{L} h_i+\lambda\right)}+\gamma L \qquad (23)$$

The training of model is finished once the objective function for all tree is determined and Eqn. 17 can be used to perform a prediction.

3 Data Generation

The dataset used for training and testing the ML models has been generated using MATLAB code meant for estimation of filter dimension using Schaffernak’s method. The ratio of crest width to height of the dam (*B*/*H*$_d$) has been varied from 0.2 to 1.5 with an interval of 0.1. The range has been selected based on the dam configurations encountered by Nakamura and Yamazaki (1988) during the investigation of damaged and undamaged earthen dams for irrigation during the 1983 Nihonkai-Chubu earthquake. The ratio of upstream water level to height of the dam (*H*/*H*$_d$) has been varied from 0.1 to 1.0 with an interval of 0.1. The range has been selected as a result of visualization of water present on the upstream side of the dam. In the present study, the length of dam is considered larger as compare to cross sectional dimensions which confines the problem to a plane-strain case. Thus, the cross-sectional dimension plays an eminent role in shaping seepage lines through dam. The two parameters defining the cross-section of the dams are the angles of upstream and downstream slopes which are acute angle for a typical dam configuration. In order to have a better perspective of the variation in the outputs, these angles have been chosen in the range of 15º-75º at a regular interval of 10º. In total, 49 different combinations of upstream and downstream angles have been investigated. The variation of (B/H_d) results in 14 dam configurations for each set of upstream and downstream angles, giving a total of 686 dam configurations. Further, the variation of (H/H_d) in above specified interval and imposing this condition over the other inputs gives a total of 6860 dam models.

Similarly, for validating the model the same code has been used for generation of validation dataset. These dataset helps in understanding the prediction accuracy of ML models for arbitrary dam configuration and condition. The ratio of crest width to height of the dam (B/H_d) has been assigned random 500 values within range of (0.357, 0.826). The ratio of upstream water level to height of the dam (H/H_d) has been assigned random 500 values within range of (0.456, 0.789)**,** the downstream and upstream slopes has been assigned random 500 values within range of 15º to 75º. Thus, a total of 500 synthetic dataset has been generated for assessing the efficacy of ML models. Further elaborations on the chosen ranges of various parameters can be found in preceding literature [Anand, 2012; Anand and Dey, 2012]

4 Result and Discussion

The 6860 data-points generated from the Schaffernak's solution has been split into training and testing dataset such that the training dataset comprises of 80% (5488) of total data and rest 20% (1372) assigned as test data. The validation of ML algorithms has been done using synthetic data (500). Based on XGBoost and RF models, figures 4-7 shows the typical comparison of output value predicted by ML model with the actual value obtained from Schaffernak's solution for both test as well as synthetic data. Similar plots have been obtained from other ML algorithms (MLP, RR and SVR) as well, however, they are not presented here for the sake of brevity.

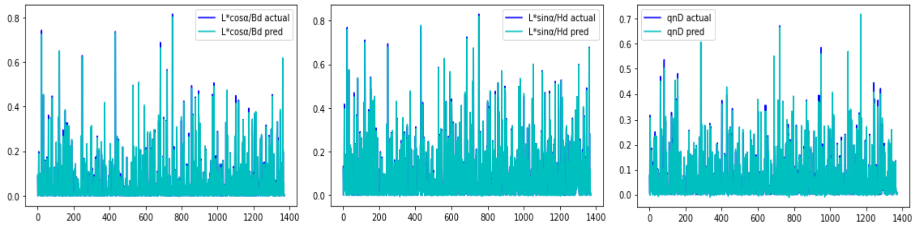

Fig. 4. Comparison between actual and predicted output by XGBoost for test data

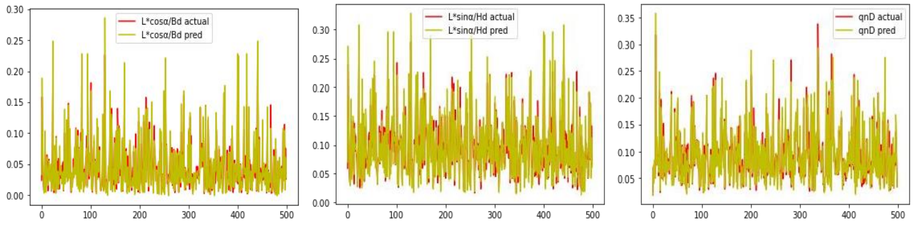

Fig. 5. Comparison between actual and predicted output by XGBoost for synthetic data

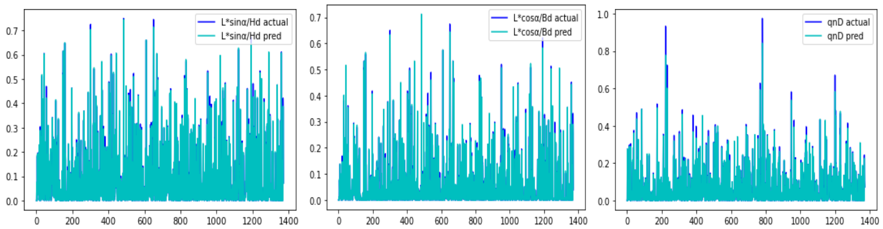

Fig. 6. Comparison between actual and predicted output by RF for test data

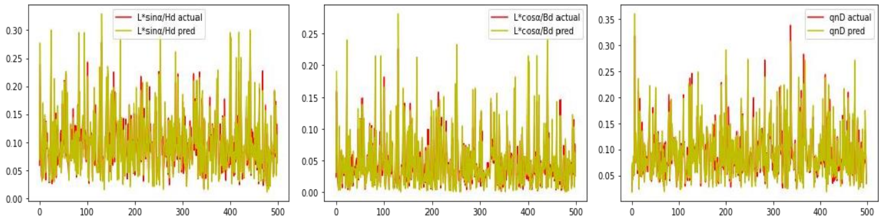

Fig. 7. Comparison between actual and predicted output by RF for synthetic data

Table 2. illustrates the performance index of ML algorithms in terms of different metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and R^2 score that were used to measure the predictive performance of each algorithm. MAE measures the difference between the actual and predicted value in absolute term, RMSE measures the square root of average of squared difference between actual and predicted output and R^2 (also known as coefficient of determination) measures the amount of variance of prediction. For a residual, $r_i = y_i - \bar{y}_i \forall\, i \in (0, n)$, R^2 can be defined as follows:

$$R^2 = 1 - \frac{\sum r_i^2}{\sum_i \left(y_i - \bar{y}\right)^2} \qquad (24)$$

where, $\bar{y}$ is average computed over all sample and $\bar{y}_i$ is average of sample on each side of threshold (splitting node). R^2 varies from 0 to 1, value closer to 1 indicate almost perfect regression fit and value closer to 0 (or < 0) indicates a poor fit. It is evident from Table 2 that XGBoost shows best fitting for test data while Random Forest shows best fitting for synthetic data. For the three output entities considered, the R^2 score for XGBoost model is 1, 1, 1 for test data and 0.91, 0.89,0.90 for synthetic data; while Random Forest has R^2 score of 1, 1, 0.99 for test data and 0.92, 0.90,0.91 for synthetic data. For SVR model, R^2 score is 0.36, 0.05, 0.20 for test data and -0.37, -0.5, -0.20 for synthetic data which is minimum among all model, thereby indicating its poor fitting (Fig. 8). The Root Mean Square Error (RMSE) for XGBoost for all three-output parameter is 1.513x10^{-05}, 2.029x10^{-05}, 2.471x10^{-05} for test data and 2.452x10^{-05}, 1.710×10^{-04}, 2.433×10^{-04} for Random Forest for synthetic data, which is minimum among all the ML models (Fig. 9). This observation further justifies that XGBoost performs best for test data and Random Forest perform best for synthetic data. On the other hand, for SVR, the obtained RMSE is 9.92×10^{-03}, 9.02×10^{-03}, 9.55×10^{-03} for test data and 3.904×10^{-03}, 6.165×10^{-03}, 3.114×10^{-03} for synthetic data, which is maximum among all the predictive models, thereby indicating its poor prediction capability. In terms of mean absolute error, MAE (%), the trend is little bit different from what illustrated by the other two performance indices (Fig. 10). The MAE (%) for Random Forest model is 0.225, 0.199, 0.243 for test data and 1.158, 0.826, 1.194 for synthetic data, which is minimum among all models; and for SVR, the obtained MAE (%) is 8.425, 8.487, 7.727 for test data and 5.839, 7.058, 4.954 for synthetic data, which is maximum among all models. The MAE (%) for XGBoost is 0.233, 0.251, 0.273 for test data and 0.12, 0.861, 1.174 for synthetic data which comes next to those obtained from Random Forest. It is interesting to note that both SVR and RR algorithms shows better predictive performance for synthetic data as the RMSE and MAE (%) is lesser for both algorithms for synthetic data. On the basis of above performance indices, it can be clearly drawn out that XGBoost and Random Forest models are the best model for prediction of filter dimension and exit discharge. The predictive performance of XGBoost and RF has been further expressed in terms of plot for coefficient of determination (*R^2*) in Figures 11-13. The prediction made by XGBoost and RF model is represented by red dotted line while the blue dots represent the scatter plot of actual output values.

Table 2. Assessment of model performance on the basis of different metrics

Output	Test Data			Synthetic Data			Model
Parameter	R^2	RMSE	MAE (%)	R^2	RMSE	MAE (%)	
$L*(\sin\alpha)/H_d$	1.00	1.513×10^{-05}	0.233	0.91	2.639×10^{-04}	1.200	XGBoost
	1.00	2.896×10^{-05}	0.225	0.92	2.452×10^{-04}	1.158	RF
	0.86	2.385×10^{-03}	3.755	0.45	1.594×10^{-03}	3.551	MLP
	0.67	6.038×10^{-03}	5.592	0.67	9.339×10^{-04}	2.839	RR
	0.36	9.920×10^{-03}	8.425	-0.37	3.904×10^{-03}	5.839	SVR
$L*(\cos\alpha)/B_d$	1.00	2.029×10^{-05}	0.251	0.89	1.891×10^{-04}	0.861	XGBoost
	1.00	2.587×10^{-05}	0.199	0.90	1.710×10^{-04}	0.826	RF
	0.76	2.438×10^{-03}	3.737	0.26	7.729×10^{-04}	2.170	MLP
	0.54	5.478×10^{-03}	4.916	0.302	1.162×10^{-03}	3.281	RR
	0.05	9.022×10^{-03}	8.487	-0.5	6.165×10^{-03}	7.058	SVR
q_{nD}	1.00	2.471×10^{-05}	0.273	0.90	2.571×10^{-04}	1.174	XGBoost
	0.99	8.206×10^{-05}	0.243	0.91	2.433×10^{-04}	1.194	RF
	0.62	3.091×10^{-03}	3.581	0.21	2.602×10^{-03}	4.387	MLP
	0.37	6.394×10^{-03}	4.824	0.58	1.102×10^{-03}	2.214	RR
	0.20	9.553×10^{-03}	7.727	-0.20	3.114×10^{-03}	4.954	SVR

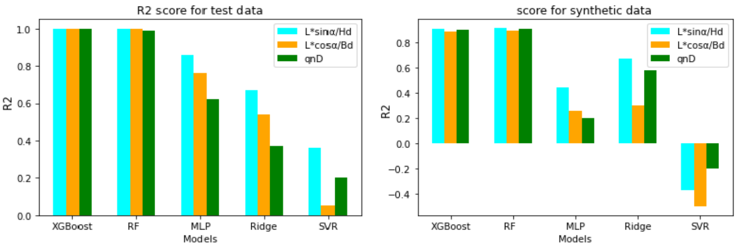

Fig. 8. Coefficient of determination (R^2) for all models for both test as well as synthetic data

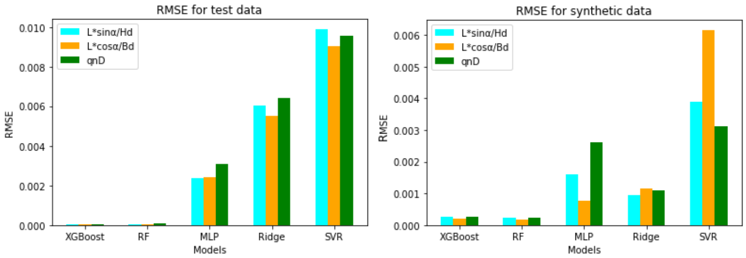

Fig. 9. RMSE for all models for both test as well as synthetic data

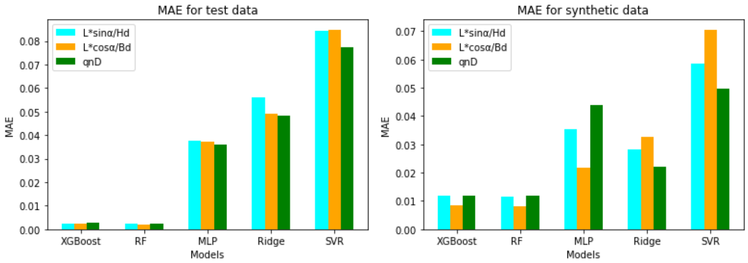

Fig. 10. MAE for all models for both test as well as synthetic data

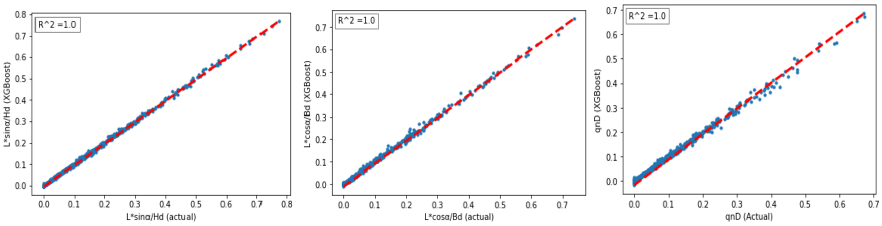

Fig. 11. R^2 between actual and predicted (XGBoost) for test data

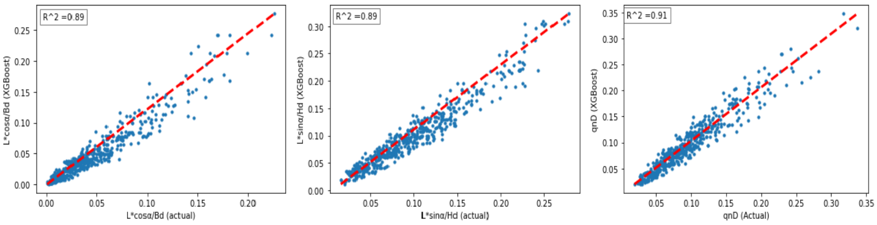

Fig. 12. R^2 between actual and predicted (XGBoost) for synthetic data

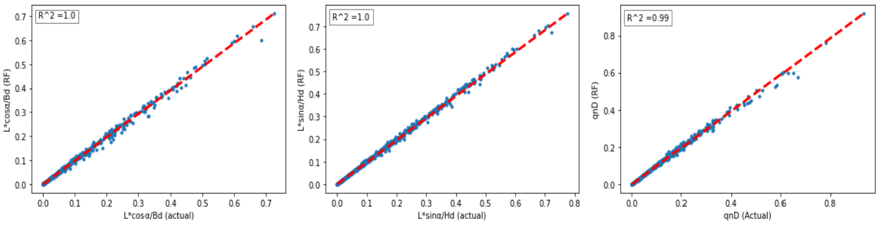

Fig. 13. R^2 between actual and predicted (RF) for test data

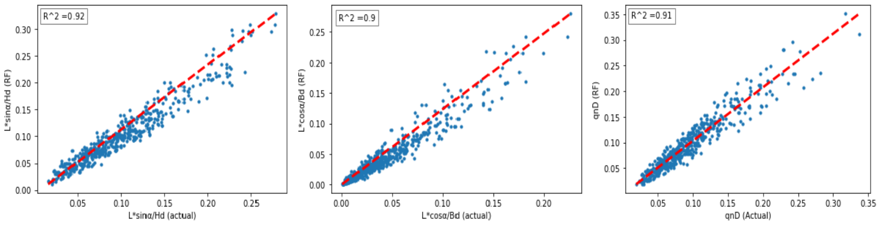

Fig. 22. R^2 between actual and predicted (RF) for synthetic data

5 Conclusion

In the present study, Machine Learning (ML) models have been proposed for the estimation of filter dimension and exit discharge based on Schaffernak's analysis. The parameters affecting the filter dimension has been identified out and has been varied in all possible range to include all dam configuration and condition. A MATLAB code based on Schaffernak seepage analysis theory has been developed for the estimation of filter dimension and exit discharge for the above generated dam configuration and conditions. The filter dimension and exit discharge act as output and all the governing parameters act as input for ML models. Several machine learning algorithms such as Multilayer perceptron (MLP), Random Forest (RF), Support Vector Regression (SVR), Ridge Regression (RR) and Xtreme Gradient Boosting (XGBoost) has been trained and tested using the generated dataset and their predictive performance has been further analysed using synthetic dataset. XGBoost and RF gives the maximum R^2 score for test and synthetic dataset respectively, thus making them best options to be used as ML-based predictive models for the assessment of filter dimensions.

References

1. Anand, V.: Seepage through Earthen Dams: Estimation of Filter Dimensions. Summer Training Report, Department of Civil Engineering, IIT Guwahati, (July 2012)
2. Anand, V., Dey, A.: Estimation of Filter Dimensions for a Homogeneous Earth Dam resting on Impervious Foundation based on Basic Seepage Analyses. International Congress on Computational Mechanics and Simulation (ICCMS), IIT Hyderabad, (December 2012)
3. Chen, T., Jia, X.: A comparison of information value and logistic regression models in landslide susceptibility mapping by using GIS. Environ Earth Sci, Heidelberg vol. 75, Iss. 10, (May 2016)
4. Mitu, M., Nayan, K.: Implementation of Machine Learning Algorithms in Spectral Analysis of Surface Waves (SASW) Inversion. Appl. Sci., vol. 11, Iss. 2557, (2021)
5. Nakamura, M., Yamazaki, A.: An investigation of damaged or undamaged small earth dams for irrigation during the 1983 Nihonkai-Chubu earthquake. Proceedings of the 9th World Conference on Earthquake Engineering, Tokyo-Kyoto vol. 7 (1988)
6. Schaffernak, F.: Über die Standicherheit durchlaessiger geschuetteter Dämme, Allgem. Bauzeitung, (1917)
7. Zhou, J., Shi, X.: Feasibility of Random-Forest Approach for Prediction of Ground Settlements Induced by the Construction of a Shield-Driven Tunnel. International Journal of Geomechanics, vol. 17, Iss. 6, (June 2017)