Machine learning-based intelligent modeling of hydraulic conductivity of sandy soil considering a wide range of grain sizes

Zia ur Rehman

Corresponding Author

ziaur.rehman@uettaxila.edu.pk
lei-m16@tsinghua.org.cn
Department of Civil Engineering, University of Engineering and Technology, Taxila 47080, Pakistan
State Key Laboratory of Hydroscience and Engineering, Tsinghua University Beijing, 100400, P.R. China.

Usama Khalid

National Institute of Transportation (NIT), National University of Sciences and Technology (NUST), Risalpur, 23200, Pakistan

Nauman Ijaz

Corresponding Author

nauman_ijaz99@tongji.edu.cn
nauman_ijaz99@hotmail.com
Key Laboratory of Geotechnical and Underground Engineering of Ministry of Education, College of Civil Engineering, Tongji University, Shanghai, 200092, China.

Hassan Mujtaba

Department of Civil Engineering, University of Engineering & Technology, Lahore, 54890, Pakistan

Abbas Haider

NUST Institute of Civil Engineering (NICE), School of Civil and Environmental Engineering (SCEE), National University of Sciences and Technology (NUST), Islamabad, 44000, Pakistan

Khalid Farooq

Department of Civil Engineering, University of Engineering & Technology, Lahore, 54890, Pakistan

Zain Ijaz

Key Laboratory of Geotechnical and Underground Engineering of Ministry of Education, College of Civil Engineering, Tongji University, Shanghai, 200092, China.

1. Introduction

The response of sandy soils under various hydraulic conditions is important for the stability of numerous engineering projects e.g., pumping of underground water, water resource management and filtration systems, water retaining structures, and waste disposal. The ability of a porous medium to transmit water through its voids is known as hydraulic conductivity (*k*), and it is one of the most important geological parameters of such geomaterials to analyze many natural geological phenomena and engineering projects, such as the management of water resources, drinking water supply, stability of waste repositories, basin-scale hydrogeologic circulation, slope stability and landslide, seepage in different geological features, deep-water pressure management, percolation mechanisms in porous media, foundation settlement, water regimes in stratified geological deposits, pollution migration from waste disposal, movement of hydrate and gas, storage of CO_2 and nuclear waste and many other issues in subsurface hydrology and geological engineering (Di Maio et al., 2021; Morbidelli et al., 2014; Ren et al., 2016; Zeng et al., 2020; Zhai et al., 2021). The *k*-value of sandy soil can be directly estimated through in-situ methods i.e., pumping, tracer, or slug tests; on the other hand, due to the specialized nature of these in-situ methods, a laboratory method i.e., constant head is also in practice as an alternate direct estimation method. However, direct estimation methods of the *k*-value are expensive and time-consuming, owing to which different researchers have established predictive models for the *k*-value based on gradation, texture, and porosity for the specific geological media, which are quick and inexpensive to determine (Deng et al., 2015; Ren et al., 2016).

According to the literature, the *k*-value for a singular fluid flow can be indirectly estimated using capillary models, empirical relationships, hydraulic radius theories, and statistical models (Chapuis, 2012). Most of these models rely on a small number of geotechnical characteristics of soil and predict the *k*-value of soil with certain limitations. Theoretically, the *k*-value of

sandy soil is dependent on the size, distribution, and interconnection of the voids within a soil matrix, which are challenging to determine (Feng et al., 2019). Alternatively, gradation parameters are considered best suitable for the prediction of the *k*-value, which defines the solid distribution within a soil matrix (Chapuis, 2004; Ren and Santamarina, 2018). A variety of predictive equations has been developed for sandy soils and a list of grain sizes (quantified by *D*-values) at different passing percentages i.e., D_{10}, D_5, D_{17}, D_{15}, and smallest equivalent size (D_{eq}) were regarded as strongly correlated parameters (Chapuis, 2012; Mujtaba et al., 2021). All of these parameters relate to the small-sized grains, which are considered to control the voids within a soil matrix. On the other hand, soil packing, and large and medium-sized grains also contribute to the distribution and interconnection of voids that govern the permeability of soil (de Bono and McDowell, 2020; Khokonov and Khokonov, 2021; McDowell and de Bono, 2021; Mujtaba et al., 2021; Zhai et al., 2018). Moreover, the use of small *D*-values as predictors also limits the applicability of these models to a bounded range of gradation (Schaap and Lebron, 2001). Further, the *k*-value is reported to have high variability (having a coefficient of variability up to 240%) within a similar soil deposit rendering that any singular class of *D*-values may not be able to predict such variability (Elhakim, 2016). Therefore, past models to predict the *k*-value of sandy soil are criticized on account of simplicity, over/under-estimation, limited ability to account for the output variability, and applicability to a limited gradation range (Chapuis, 2012; Kashani et al., 2020). Additionally, literature manifests that difference between predicted and experimental *k*-values could stretch more than one order of magnitude due to diversity in grain size other than used as the predictors (Ren and Santamarina, 2018). To counter these factors, intelligent predictive modeling of the *k*-value of sandy soil is required involving all representative *D*-values, gradation parameters and density for a large gradation spectrum. However, such modeling is challenging using traditional statistical and theoretical

89 modeling schemes, which have been used hitherto to model the *k*-value of sandy soil in the
90 literature (Chapuis, 2012).

91 Machine learning-based (ML) techniques are gaining popularity for intelligent modeling due
92 to their superior predictive capacity in comparison to the traditional empirical and statistical
93 methods (Bardhan et al., 2021; Diaz et al., 2021; Leong et al., 2015; Zhang et al., 2020).
94 Further, to correlate a large number of variables, mechanistic learning is required to develop
95 an intelligent structure, which is missing in conventional statistical or theoretical modeling
96 approaches. Considering the foregoing discussion, ML-based modeling of the *k*-value of sandy
97 soil with all representative *D*-values is envisaged to yield more accuracy (McDowell and de
98 Bono, 2021). Moreover, in the last few decades, several ML techniques have been established
99 using the white box, black box, and gray box correlation techniques. These techniques are
100 classified based on the correlation mechanism between inputs and outputs. Physical laws,
101 regressive data-driven systems, and logical systems are used to define these correlation
102 mechanisms in the white box, black box, and gray-box techniques, respectively (Naghadehi et
103 al., 2018). On the other hand, all of these ML techniques have merits and demerits of their own.
104 Recently, different ML techniques have been employed to model different geotechnical
105 responses of the soil and to gain more insight, comparative studies on various ML techniques
106 are still being carried out (Jong et al., 2021; Leong et al., 2018; Naghadehi et al., 2018; P.
107 Zhang et al., 2021; W. Zhang et al., 2021). However, there is a scarcity of literature on ML-
108 based modeling techniques to predict the *k*-value of sandy soil especially using novel genetic
109 and evolutionary ML algorithms.

110 This study, for the first time, aimed to bridge the aforementioned research gaps by employing
111 various ML-based modeling techniques, i.e., artificial neural network (ANN), multi-expression
112 programming (MEP), and genetic expression programming (GEP) to model the *k*-value of
113 sandy soil using a large set of variables representing a wide range of *D*-values, gradation

parameters, and density. This study provides the scientific basis for engaging a wide range of gradation parameters to cover output variability in the prediction of *k*-value for a wide spectrum of gradation and put forward a robust mathematical model for the quick determination of *k*-value with a considerable degree of accuracy. Overall, the current study addresses the shortcoming of the past models as identified by the recent literature (de Bono and McDowell, 2020; Khokonov and Khokonov, 2021; McDowell and de Bono, 2021; Mujtaba et al., 2021; Ren and Santamarina, 2018; Zhai et al., 2018) related to the output variability due to dependence on a singular class of grain size against a variety of grain size distributions by employing ML algorithms (including neural and novel genetic programming).

Modeling data

2.1. Materials and Testing

The schematic illustration of the testing and modeling program adopted in the current study is presented in Figure 1. In this study, sandy soils having varying grain size distributions (GSDs) were collected for testing from various parts of Pakistan regarded as pit and riverbed sand. The locations for the sample collection were on the Indus plain, which is drained by the Indus and its tributaries, i.e., the Jhelum, Chenab, Ravi, Beas, and Sutlej rivers, creating a sophisticated system of interfluves having sandy soils with varying deposition characteristics. The major identified sites for the sand collection were at different localities of Dera Ghazi Khan (Chenab sand), Dina (Jehlum sand), Khusab (pit sand), Lahore (Ravi sand), Lawrencepur, Mianwali (Indus sand), Qadirabad, Sakhi Sarwar and Wazirabad. These sites are located on different terrains ensuring varying grain sizes in the collected samples. To enrich the database several natural sandy soil samples collected from the aforementioned sites were also mixed in varying proportions. These composite samples were a mix of two or more aforementioned natural soil specimens. A comprehensive testing plan was devised to characterize all the natural and composite soil specimens. For basic geotechnical characterization, GSD analysis (ASTM D-

422 and D7928) and Atterberg limit tests (ASTM D-4318) tests were performed. To evaluate the permeability of the soil, the samples were compacted as per ASTM D1557-07 and then a constant head permeability test (ASTM D-2434) was performed on the compacted sample.

GSD analysis was used to determine *D*-values, with US standard sieves of 19.0, 4.75, 2.00, 0.425, 0.15, and 0.075 mm sizes. To extend GSD curves beyond 0.075 mm hydrometer analysis was also performed for specimens having more than 5 % of fine-grained material (<0.075 mm). Further, for the soil specimen having greater than 5% fines Atterberg's limit test i.e., liquid limit and plastic limit tests were performed. Compaction tests were performed to determine optimum moisture content (w_{opt}) and maximum dry density (γ_{dmax}). Constant head permeability tests were performed to determine the *k*-value, for which samples were compacted at w_{opt} and γ_{dmax} determined earlier in a mold with a height of 12.7 cm and internal diameter of 10.16 cm. A constant head was maintained over-saturated sample and discharge of water (q) through the saturated sample was estimated at an outflow pipe; afterward, the *k*-value was estimated as per the designated procedure by the ASTM D-2434 using the following Equation:

$$k = \frac{qL}{Ah} \tag{1}$$

where L, A, and h are the length of the specimen, area of the specimen, and head causing flow, respectively.

2.2. Test results

A dataset of 247 soil samples was prepared by employing the aforementioned testing scheme; a summary of test results is given in Table 1. Sieve analysis results showed that the soil sample tested for this study was sandy soil (i.e., 4.75-0.075 mm) predominantly having medium to small-sized sand grains in a range of 90 to 100%. A small fraction of gravelly soil (i.e., >4.75 mm) was also observed in the specimen in a range of 0-9%. Similarly, fine particles (i.e., <0.075 mm silt and clay) were observed to be in a range of 0-10%. Based on GSD curve analysis, different gradation parameters were computed which correspond to the sizes and

distribution of soil grains within the soil matrix (Fig. 2). *D*-value for 90, 60, 50, 30, 17, 15, 10, and 5% passing (D_{90}, D_{60}, D_{50}, D_{30}, D_{17}, D_{15}, D_{10}, and D_5, respectively) and D_{eq} were determined, which correspond to the large, medium and small-sized grains (Table 1) (Fig. 2). Further, coefficient of uniformity (C_u) and coefficient of curvature (C_c) were determined to evaluate the gradation of *D* from GSD curves, which were found to be in a range of 1.65-8.0 and 0.22-6.73, respectively. All the samples were classified as per ASTM D-2487 majorly into sand-class (S). The samples were further classified into 73% poorly graded sand (SP), 21% poorly graded sand with silt (SP-SM), 4% well-graded sand (SW), and 2% well-graded sand with silt (SW-SM). The w_{opt} and γ_{dmax} were found to be in a range of 14.5-16% and 1.47-2.02 kN/m^3, respectively. Further, all of the samples were observed to have the *k*-value ranging from 0.00014 to 0.0077 cm/sec, with an average of 0.00264 cm/sec. Out of total sample approximately 83% of samples had a *k*-value between 0.0011 and 0.005 cm/sec, 10% had a value between 0.0051-0.01 cm/sec, and 7% had a value less than 0.0011 cm/sec.

2.3. Input analysis

The number of data points improves the ML-based model training and consequently enhances the model performance. Therefore, large experimental data of the *k*-value of sand for a wide gradation range was used in this study for the development of ML-based models. The dataset included a variety of *D*-values and distribution parameters, compaction characteristics, and *k*-value of sandy soils (Table 1). However, all the parameters given in Table 1 were not selected as the input parameter since the selection of appropriate input parameters is an important dimension for model development. Input parameters in this study were selected based on multiple criteria encountering the physical and statistical significance of these parameters. For instance, in the current study, a wide range of *D*-values i.e., large, medium, and small, gradation parameters and density of soil matrix were considered to be responsible for controlling the *k*-value of sandy soil. For this purpose, the theoretical framework of the selection of input

parameters based on the aforementioned consideration is presented in Figure 2. To capture large, medium, and small *D*-values the range of *D* i.e., D_{90}, D_{60}, D_{50}, D_{30}, D_{17}, D_{15}, D_{10}, D_5, and D_{eq} was considered as representative input parameters (Fig. 2). It is pertinent to mention that physically significant and well-known *D*-values in conventional soil mechanics practice were used in this study in the domain of large, medium and small sizes. To involve the gradation pattern of these *D*-values within the soil matrix, C_u and C_c were considered, and to engage deposition, γ_d was considered as an input parameter.

Further, since the selection of input parameters and the number of data points synergistically affect the model performance; therefore, statistical considerations were further taken into account. Literature suggests that selection of input parameters must not be too extensive and limited for better model performance. A ratio of >5 is considered to be reasonable between total data points and a number of input parameters. In this study, 12 input parameters were selected with a sample size to input parameter ratio around 19, which is significantly higher than the threshold value.

Further, the probability distribution of input and output parameters was also assessed to observe the spectrum of input and output variables (Fig. 3). The statistical analysis of the input and output data is presented in Table 2. The frequency histogram of *D*-values and parameters showed that the database had a greater range of GSD data (Fig. 3). Further, the *k*-value was also observed to have a wide spectrum in accordance with GSD. Statistical analysis showed that the data distribution of the input and output parameters chosen for modeling was skewed on the right-hand side owing to having the positive skewness value (Table 2). The skewness trend of all the considered parameters was observed to be in line with each other, thus underlying a satisfying trend in terms of the selection of parameters for the current study. Furthermore, excess kurtosis values showed that the data distribution of most of the parameters was almost the same, i.e. leptokurtic except for D_{17}, D_{15}, and C_c which showed platykurtic

distribution. Overall, almost an identical distribution of data in terms of skewness and tailedness showed that input and output data were in good agreement with each other. The preceding analysis demonstrated that the present database had a broad range of *D*-values and gradation parameters, and input parameters could play an important role in predicting the output parameter since they had a similar data distribution trend.

Moreover, correlation analysis was carried out to determine the interrelationship strength between input and output parameters. A full-scale Pearson correlation matrix in terms of correlation coefficient (*r*) is presented in Figure 4; all the input and output parameters were considered in this analysis. It was observed that not a single parameter could be taken as a lone predictor owing to the non-unity of *r*-value of input parameters in relation to *k*-value. Further, a reasonable *r*-value was observed for the particle size and *k*-value, which was observed to be increased as the particle size decreased from D_{90} to D_{eq} (Fig. 4). Further, C_u and C_c and γ_d yielded low *r*-value against *k*-value; however, they had significant physical meaning in defining the gradation and packing of soil grains, therefore, they were also considered as inputs for stepwise and iterative ML modeling. In addition, the *r*-value between the input parameters was ≤ 0.85 indicating that the correlation between input parameters is within a prescribed range of ≤ 0.9 as per (Yoo et al., 2014) to avoid multicollinearity issues. Thus, 12 input parameters were initially selected to predict the *k*-value by incorporating the ML algorithm. It is important to note the significantly influencing parameters were further scrutinized for inclusion in the model framework to predict *k*-value by iterative ML algorithms.

2. Modeling methods

In this study, ML-based modeling of the *k*-value of sandy soil was carried out by considering a variety of *D*-values in a sandy specimen i.e., large, medium, and small sizes. For this purpose, different representative *D*-value corresponding to 90 to 5% (i.e., D_{90}, D_{60}, D_{50}, D_{30}, D_{17}, D_{15}, D_{10}, $D_{5,}$ and D_{eq}) passing were considered. Further, for rationalization of the models, different

other relevant parameters i.e., C_u and C_c which represent gradation patterns i.e., well or poor gradation of soil specimen were also considered for the prediction of the *k*-value. In addition, density also plays a vital role in the permeability of soil; therefore, it was also included as a parameter in the model. Thus, a variety of parameters as discussed above were used to predict the *k*-value in the current study as follows.

$$k = f(D_{90}, D_{60}, D_{50}, D_{30}, D_{17}, D_{15}, D_{10}, D_5, D_{eq}, C_u, C_c, \gamma_d) \quad (2)$$

Three different ML-based modeling approaches were adopted in the current study which involves two major modeling paradigms i.e., grey box and black box modeling schemes. GEP, MEP, and ANN were employed and compared to predict the *k*-value using the aforementioned input parameters. Further, MATLAB is used to run ANN and commercially available software packages (MEPx and GeneXpro) were used to run MEP and GEP algorithms. These modeling paradigms have been employed for various applications; however, their use to predict the *k*-value of sandy soil is state-of-the-art. For ML modeling, the data set is initially analyzed and screened based on statistical and data interpretation techniques. The final dataset was then divided into training and testing classes using standard 70 and 30% division as per (Shahin et al., 2004). ML-based algorithms were trained and subsequently tested in an iterative manner to achieve the best possible programs to predict the *k*-value. A number of models were developed using GEP, MEP, and ANN, individually; however, models having the highest statistical health are presented in this paper. The description of the ML algorithm used in the current study is as follows.

3.1. Model descriptions

3.1.1. Gene expression programming (GEP)

Genetic modeling (GM) is based on the genetic evaluation process and works on the principle of regression and neural techniques. This is a computer-based technique to solve intricate problems by employing Darwin's principle. Initially, simple regression-based functions are

defined, for which GM is capable of removing and adding parameters based on suitability to yield proper outcomes. GEP is an improved and distinct variant of GM with encoded linear fixed chromosomes and parses tree-like structures. Owing to its multigene activity, GEP employs basic conditions to build genetic variation and solve complicated tasks. Expression trees (ETs) are a type of parse trees that expresses the varied shapes and sizes of non-linear items. The GEP procedure initiates with the creation of a single individual's fixed-length chromosome. ETs then embody the chromosomes, and thereby their health is assessed. Finally, the reproduction procedure commences, and fitness functions are used to assess the results. The schematic diagram of the GE algorithm adopted in this study is shown in Figure 5.

3.1.2. Multiple expression programming (MEP)

In GM, fixed binary length strings can also be used instead of nonlinear parse trees, which is a significant variance from GEP. This kind of GM is regarded as MEP, in which simulation is a linear string instruction, where strings are a collection of mathematical functions and variables. The MEP is capable of encoding multiple computer programs in singular chromosomes which makes it unique. The schematic diagram of the MEP algorithm used in this study is shown in Figure 6. The random generation of the chromosomal population is the first step in MEP evolution. After that, the most suited chromosome is chosen based on the best fitness values, and the final solution is created. The binary environment is used to choose two parents, which are then recombined to produce two distinct offspring. The offspring are mutated, and the procedure is repeated until the best program/solution is found before the halting requirements are met (Fig. 6). Similar to GEP functions set, crossover probability, code length, subpopulation size, and the number of subpopulations are the major governing factors for MEP.

3.1.3. Artificial neural networks (ANN)

The artificial neuron is the primary building block of ANNs, which are algorithms that approximately duplicate and imitate the microstructures of the biological nervous system. It is

primarily based on a black box correlation paradigm, in which complicated and nonlinear functions could be readily represented by the network with various parameters or variables. These ANN are trained in such a manner that the simulation output matches the experimental output. Moreover, each network has three types of layers i.e., input, hidden, and output layers. Input and output layers are mathematically linked together with a broad network of in-between hidden layers. A feed-forward ANN paradigm was used in this study as shown in Figure 7. It is important to mention that a large quantity of data is necessary for training ANN models to duplicate the output from previously unknown inputs. ANN has been widely used to predict different geotechnical characteristics of soils based on multiple parameters, thus its implementation for the prediction of the *k*-value of soil is important. Further, a comparison of all these ML-based algorithms is also imperative based on the *k*-value.

3.2. Model performance evaluation criteria

The performance of the ML-based model was evaluated based on various key performance indices (KPIs), which include error indices (i.e., root mean squared error (RMSE), relative root mean square error value (RRMSE), and mean absolute error (MAE) and correlation indices (i.e., Nash sufficient model efficiency coefficient (NSE), correlation coefficient (R^2) and adjusted correlation coefficient (*adj* R^2)) as follows.

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(E_i - M_i)^2}{n}} \quad (3)$$

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n}|E_i - M_i|}{n} \quad (4)$$

$$\mathrm{RSE} = \frac{\sum_{i=1}^{n}(M_i - E_i)^2}{\sum_{i=1}^{n}(\bar{E} - E_i)^2} \quad (5)$$

$$\mathrm{NSE} = 1 - \frac{\sum_{i=1}^{n}(E_i - M_i)^2}{\sum_{i=1}^{n}(M_i - \overline{M_i})^2} \quad (6)$$

$$R^2 = \frac{\sum_{i=1}^{n}(E_i - \bar{E}_i)(M_i - \bar{M}_i)}{\sqrt{\sum_{i=1}^{n}(E_i - \bar{E}_i)^2 \sum_{i=1}^{n}(M_i - \bar{M}_i)^2}} \quad (7)$$

$$Adj\ R^2 = 1 - \frac{n-1}{n-p}(1-R^2) \quad (8)$$

where M_i and E_i are the ith predicted and experimental outputs, respectively, and $\bar{M}_i$ and $\bar{E}_i$ are their average values; n is sample quantity. A low value of error indices and a value close to unity are required of the correlation indices for a good model. Further, variance analysis (ANOVA) was also conducted for the developed models to check their statistical health.

To further validate the proposed model an independent dataset of 47 samples given by (Mujtaba et al., 2021) was also employed to check the prediction ability of the proposed model and compared it with the existing model. Different external validation parameters such as R, R_m, and slope of regression lines (k and k') and corresponding indices (R_0 and R_0') were also evaluated as follows:

$$R_m = R^2 \times \left(1 - \sqrt{\left|R^2 - R_0^2\right|}\right) \quad (9)$$

$$k = \frac{\sum_i^n (o_i \times p_i)}{\sum_i^n {o_i}^2} \quad (10)$$

$$k = \frac{\sum_i^n (o_i \times p_i)}{\sum_i^n {p_i}^2} \quad (11)$$

$$R_0^2 = 1 - \frac{\sum_i^n (o - p_i^0)^2}{\sum_i^n (o_i - \bar{p}_i^0)^2},\ \mathrm{o}_i^0 = k \cdot p_i \quad (12)$$

$$R_0'^2 = 1 - \frac{\sum_i^n (o_i - p_i^0)^2}{\sum_i^n (o_i - \bar{o}_i^0)^2},\ \mathrm{p}_i^0 = k' \cdot o_i \quad (13)$$

where o_i and p_i are the experimental and predicted values at the ith interval.

3. Results and discussion

4.1. GEP-based model

A comprehensive GEP modeling was engaged to establish a model to predict the *k*-value based on the aforementioned input parameters. The optimal setting was employed for genetic operators, constants, and general parameters in GEP as presented in Table 3. A stepwise GEP modeling approach was employed by running a simple algorithm to complex for better performance based on multiple iterations. Prior to running the GEP algorithm, the basic function sets i.e., addition, subtraction, and division were engaged along with the structural relationship of chromosomes. After performing a series of GEP algorithms starting with the lowest head size with a single gene chromosome, the GEP models for *k*-value were developed. The simplest mathematical models were obtained by using fundamental operators, the smallest head size, and the smallest number of chromosomes. However, the statistical health of such a mathematical model could be inferior, especially when the model involves large inputs; thus, more complex mathematical operators and large head sizes were selected to enhance the performance of models (Table 3). However, as the complexity in terms of head size and mathematical operator increases beyond a certain point, concerns with overfitting, model complexity, and slowness of algorithm run of the model may develop. To avoid this, a randomly distributed dataset was used for training and testing and the complexity of GEP was enhanced iteratively and in an optimized manner (Table 3). The GEP approach was initiated with a population of the most feasible options. Following that, an iterative process from one generation to the next was used to find the best possible solution. By employing this process, the best model was generated having R^2 around 0.965 and 0.945 for training and testing datasets, respectively. Figure 8 presents the performance of the GEP model for the *k*-value of sand for training and testing datasets. It can be observed that a small variation occurred in the predicted and experimental *k*-value of sand from equity (45°-line) showing reasonable statistical health of the developed model. Thus, a reasonable GEP-based model was obtained.

4.2. MEP-based model

In order to construct a comprehensive and robust model, MEP requires certain algorithm input parameters to be determined before model creation. The algorithm input parameters were chosen with care, taking into account the aforementioned recommendations and the trial-and-error method. For the initial stages, a subpopulation size of 10 and generations of 100 were explored, along with fundamental mathematical parameters. The number of subpopulations and generations were increased to 700 and 1000, respectively, and more complex mathematical operators were included stepwise to achieve an optimal setting for the most productive model (Table 4). The MEP algorithm was initiated by building a population of all potential solutions. The algorithm's approach was iterative; each generation brought the prediction closer to the experimental value. Different KPIs i.e., MSE and R and R^2 were used to assess the performance of each generation of the model. The generation evolved stepwise in an optimal manner to achieve the best possible model having R^2 around 0.946 and 0.880 for training and testing datasets. The predicted k-values by the developed model are presented in Figure 9 and compared with experimental values. It can be seen that the predicted k-value of sandy soil was close to that of experimental values rendering the effectiveness of the developed MEP model.

4.3. ANN-based model

A feed-forward propagation algorithm was employed to train the ANN-based model to predict the k-value of the sandy soil based on the selected input parameters. The number of neurons in the hidden layer governs the performance of the ANN model; a small number of neuron tends to yield non-reasonable solutions and a high number of neurons results in overfitting and lengthier run time. Therefore, an optimization routine was employed in which neurons were changed simultaneously from hidden layers to achieve an optimum number of neurons. A reasonable ANN-based model having R^2 of 0.944 and 0.820 for training and testing datasets, respectively, was achieved by several trials. Moreover, a strong correlation was observed

between the predicted *k*-value from the ANN model and experimental values (Figure 10); thus, this model is reasonable for the prediction of the *k*-value.

4.4. Model comparison

Different KPIs were determined for each developed model and the performance was compared (Table 5). A general rule of thumb and Pellinen criteria describe a value of correlation indices (i.e, R^2, adj R^2 and NSE) between 0-0.3, 0.3-0.5, 0.5-0.7, 0.7-0.9, and 0.9-1.00 to show negligible, low, moderately high and very high correlation strength, respectively (Mujtaba et al., 2021). Generally, in terms of correlation indices (i.e R^2, adj R^2 and NSE), all the developed models showed a very high correlation strength on the training and testing dataset. Meanwhile, for the training dataset, the GEP-based model performed slightly better than MEP and ANN as shown in Figure 11. On the other hand, R^2, adj $R^{2,}$ and NSE of MEP and ANN-based models were almost similar for the training dataset. Further, for all the models, these correlation indices were observed to be decreased as the dataset changed from training to testing. GEP and ANN-based models showed the lowest and highest decrease in the correlation indices for the testing dataset, respectively (Fig. 12).

Further, in terms of error computed based on (RMSE and MAE) all the developed models showed a marginal value close to zero for both training and testing datasets (a general rule of thumb manifests that a value >0.5 reflects the poor quality of the model to accurately predict the data). In comparison, the GEP model showed the least error for both testing and training datasets among all the developed models. Similar to correlation indices, error indices indicated a decrease in model performance as the dataset changed from training to testing. GEP and ANN-based models showed the least and highest decrease in model performance in terms of error indices. The highest decrease in the ANN-based model performance in comparison to GEP and MEP from the training to testing phase indicated a typical response associated with the black box model. In comparison to gray-box models (GEP and MEP), ANN showed

overfitting in the training phase and limitation in establishing the possible relationships between variables which resulted in compromised prediction. Nevertheless, the performance of all the models was satisfactory as a high *F*-value (a value close to 1 implies an insignificant model having the same model and residual variances) and least Prob>*F* (a value less than 0.05 indicates model terms are significant) was observed in ANOVA. Further, the comparison, showed that GEP based model performed better than MEP and ANN-based models based on all KPIs; thus it could be preferred over the other models. This can be attributed to the fact that by astutely utilizing the specialized chromosome, ETs, and internal mechanism of establishing explicit mathematical relationships, GEP overcomes the limitations of other ML algorithms. For instance, ANN is based on a weight matrix with various biases that are difficult to explicate; in GEP, this problem is overcome by explicit mathematical relations. Meanwhile, in comparison to MEP, the operating mechanism of GEP is more sophisticated. Further, in GEP, the proposed solution is encoded by the multigenic chromosomes (having head and tail) in the form of multiple sub-programs which are subsequently translated into ETs. The partition of the GEP chromosome into head and tail, each of which includes specific symbols, allows for a more innovative and effective technique of encoding syntactically sound computer programs, making it better in performance than MEP. In general, the complicated gene/tree system of GEP akin to the DNA/protein life system on Earth, not only examines all of the niches and routes of the solution space but also can investigate complex degrees of interrelationship to efficiently program an output.

4.5. Proposed model formulation

Based on the aforementioned results and analysis, a GEP-based model was found to show maximum performance; therefore this model is proposed for the prediction of the *k*-value of sandy soil in the current study. This model can be derived by decoding the ETs as shown in

Figure 13. The final expression of the proposed model to predict the *k*-value of the sandy soil using *D*-values, gradation and density parameters is as follows.

$$k = -2.15 - 10^{\left(\left(\left(\frac{C_c - 6.47}{2}\right) + \left(\frac{0.94}{D_{60}}\right)\right) + \left(\frac{1}{C_c}\right)\right) - 6.47} + 0.44 \times 10^{10^{\left(e^{D_{60}} + (D_{eq} - C_u)\right) - ((3.13 - D_{60}) + D_{90})}}$$

$$+ \left(\left(e^{D_{60}} \cdot \left(\frac{D_{10} + D_5}{2}\right)\right) + (1.5D_{10} + 0.22)\right) \cdot e^{-4.95} + \left(e^{\left((2.34 - D_{50}) \cdot (D_{10} + \gamma_d)\right) \cdot \left(\frac{(D_{30} - D_{60}) - 4.08}{2}\right)} - 0.70\right)^{\frac{1}{3}} \quad (14)$$

4.6. Model validation

To further assess the validity of the proposed model, an independent dataset of 47 samples was obtained from the documented literature (Mujtaba et al., 2021). The performance of the proposed model was assessed based on the normalized Taylor diagram (Zhou et al., 2021) by comparing Pearson correlation and standard deviation of the predicted *k*-value in accordance with the experimental values (Fig. 14). In addition, the performance of models from literature (Chapuis, 2012; Hazen, 1911; Mujtaba et al., 2021; Shahabi et al., 1984) was also assessed using the same dataset and compared with the current model. Further, a simple regression model (having framework: $k = a \cdot D_{10} + b$) using D_{10} as a sole parameter considering the traditional modeling approach based on the current dataset is also compared with the proposed model to assess the suitability of the proposed model to cover output variability. The simple regression model showed a R^2 value of around 0.86. In a normalized Taylor diagram, the distance between the predicted and experimental/reference points describes the accuracy of the predictive models (Zhou et al., 2021). It can be observed that the GEP-based model yielded the *k*-values of sandy soil closest to that of experimental values as compared to other models (Fig. 14). The GEP-based model developed in the current study appeared to be more accurate in the prediction of *k*-value as compared to the existing predictive models. Also, the GEP-based

model showed more accuracy as compared to the simple regression model based on the current dataset referring that GEP-based model resolves the limited output variability issue of the traditional modeling approaches.

External validation of the proposed model and existing models was also performed based on different indices as given in Table 6. The slope of the *k* and *k'* regression line was observed to be close to 1 for the GEP-based model, whereas for other models, it was found to be higher or lower than 1 which exceeded the desirable limit and showed low predictive performance (Table 6). Further, R, R_m, $R_{0,}$ and R_0' were observed to be in a desirable range for the GEP-based model, simple regression model and model proposed by Mujtaba et al. (2021) and Hazen (1911), with GEP based model showing the highest values of these indices (Table 6). The external validation results also indicated that the GEP-based model proposed in the current study performed efficiently in predicting the *k*-value of sandy soil. It is important to note that the past predictive models involved just D_{10} to be the main predictor, which compromised the prediction ability of these models in terms of variability of the *k*-value due to changes in other *D*-values. A similar trend is observed for simple regression based model developed using the current dataset. Meanwhile, the proposed GEP-based model included the entire set of representative *D*-values as the predictors, which enhanced its ability to tackle the variability of the *k*-value and yielded results with high accuracy.

4.7. Parametric and sensitivity analyses

The parametric and sensitivity analyses were conducted to determine the influence of different input parameters on the model prediction. ML-based models often perform well on training and testing datasets; however, their performance is uncertain on another independent dataset. In this regard, the optimized response of different model inputs is important to comprehend the prediction framework of a model for better application to independent datasets. Albeit the proposed GEP-based model performed reasonably on an independent dataset, its modeling

framework is analyzed by performing parametric and sensitivity analysis. In the parametric analysis, the influence of each input parameter was examined by varying a singular input parameter based on its distribution within the dataset and by keeping other parameters to be constant at their average value (Fig. 15). The parametric analysis showed that the overall *k*-value increased as the *D*-values increased at different passing percentages and decreased with an increase in the γ_d (Fig. 15). This is owing to the fact that with an increase in the larger *D*-values of soil within a soil matrix and a decrease in the γ_d, soil packing becomes loose which leads to an increase in the interconnected voids and thereby permits more water to pass through it. Further, the *k*-value of sandy soil related linearly with representative small *D*-values (i.e., D_{10}, D_5, D_{eq}) and hyperbolically with large and medium grain rain sizes (i.e., D_{60}, D_{50}, D_{30}). Past literature showed a similar relationship (Chapuis, 2012); thus, the model framework is reasonably incorporating different input parameters.

Further, sensitivity analysis was performed to determine the contribution of input parameters in the model framework by employing the method as per (Iooss and Saltelli, 2017). The following Equations are used to determine the sensitivity (S_i) of the input parameter:

$$\left.\begin{aligned} S_i &= \frac{N_i}{\sum_{j=1}^{n} N_j} \times 100\ ; \\ N_i &= f_{\max}(x_i) - f_{\min}(x_i) \end{aligned}\right\} \quad (15)$$

where f_{max} and f_{min} are the maximum and minimum predicted values for a singular input parameter while keeping other parameters to be constant at an average value within the dataset. The sensitivity analysis showed that D_{10} is the most sensitive parameter in the model framework followed by D_{50}, D_{60}, D_5, D_{30}, γ_d, and C_u subsequently (Fig. 16). Further, the contribution of D_{90}, D_{15}, D_{17}, and C_c in the model framework was found to be insignificant. D_{10} and compaction characteristics are considered to be the most useful parameters to predict the *k*-value of sandy soil in the past literature; thus current model framework is in line with

past models (Fig. 16). However, the current model involved small, medium, and large *D*-values, gradation parameter (i.e., C_u) and γ_d altogether in its framework which enhanced its prediction ability. Moreover, large to small *D*-values, gradation, and loose dense packing altogether physically control the permeability of soil as advocated by (de Bono and McDowell, 2020; McDowell and de Bono, 2021); thus, the current model holds an astute, reasonable and comprehensive framework.

5. Field implications

The outcome of this study provides a comprehensive predictive model that provides a quick and accurate estimation of the *k*-value without performing tedious and expensive testing procedures for various geotechnical and geological engineering applications by employing ML algorithms (including neural and novel genetic programming). The modeling framework of this study addresses the core shortcomings of the existing empirical predictive methods i.e., dependence on a singular grain type, low input variables and non-inclusion of robust modeling methods. Such an accurate model of *k*-value as demonstrated in the current study is important for the meticulous stability analysis of various geological phenomena and geotechnical structures, i.e., slopes, landslides, foundations, engineered barriers for waste disposals and water repositories, etc., (Zhai et al., 2021, 2018). Also, for the precise preliminary feasibility assessments of various geological utilities e.g., systems for water pumping, water management, and liquid and gas extraction through stratified layers, such an accurate model based on simplified parameters could be critical, in order to boost the confidence level of practitioners during the design phase. Apart from the aforementioned applications, there could be other geotechnical and geological applications of the specific modeling strategy adopted in the current study. For example, as per literature, the *k*-value is important in numerically simulating the liquefaction phenomenon and liquefiable soil could have varying depositions rendering wavering grain size distributions. Thereby, for the known and different grain size distributions

of the sandy soils, *k*-value can be quickly and accurately acquired using the proposed model for liquefaction simulations. Further, the constitutive response of the clogging could also be simulated considering the modeling strategy adopted in the current study. In general, the parametric study manifested that the *k*-value decreases as the dominance of finer/small grain size increases within a soil matrix which could be a base line to quantify clogging phenomenon (Bai et al., 2021). Further in the future, the current modeling method could be extended to other soil types e.g., silty/sandy/clayey soils and peats to yield more reasonable predictive models (Ong et al., 2022). Thus, the model developed in this study has a vast field implication pertaining to engineering geology and it could serve as a reference for future studies.

6. Summary and conclusions

This study presents novel intelligent modeling of the *k*-value of sandy soil by employing ML-based algorithms i.e., ANN, MEP, and GEP on a large dataset. For this purpose, an extensive testing program was carried out to evaluate the *k*-value, gradation, and γ_d of a large gradation spectrum of the sandy soil. A broader range of input parameters comprising of the large, medium, and small *D*-values, gradation parameters, and γ_d, which can be quickly determined through simple GSD and γ_d determination tests, was used to predict the *k*-value of sandy soil. A number of possible models were generated by algorithm-guided iteration and varying ML algorithm inputs; thereof best models were scrutinized. The performance of ANN, MEP, and GEP-based models was compared based on various KPIs, i.e., error indices, correlation indices, and ANOVA; thereof the best model was proposed for the prediction of the *k*-value. The performance of the proposed model was validated and compared with existing models in the literature based on an independent dataset based on the Taylor diagram analysis and validation indices. Finally, the parametric and sensitivity analyses were carried out to understand the prediction framework of the proposed model for applicability to the independent datasets. The following are the main outcomes of the current study.

- All the ML-based models exhibited excellent correlation indices close to one, minimum error indices, and reasonable best fit ANOVA results on training and testing datasets. However, comparatively GEP-based model performed better than the ANN and MEP-based models. Therefore GEP-based model (Eq. 13) was proposed for the prediction of the *k*-value of the sandy soils. Further, on the training dataset, ANN and MEP-based models performed almost similarly, while the GEP-based model showed slightly better performance than both of these models. On the other hand, the performance of all the models decreased as the dataset changed from training to testing, whereas GEP- and ANN-based showed the lowest and highest decrease in the performance among all the subjected ML-based models, respectively. This shows the limited ability and complexity of ANN-based models in comparison to MEP or GEP-based models. Further, the ability of GEP in utilizing multigenic chromosomes, ETs, and mechanisms to establish empirical relations promoted it over MEP and ANN.
- The proposed model showed a reasonable prediction ability of the *k*-value in accordance with the independent validation dataset on a scale of Pearson correlation and standard deviation on Taylor's diagram. Further, in comparison to the existing models and simplified regression model based on the current dataset, the proposed model yielded the prediction results closest to the experimental data. This referred that the GEP-based model with large numbers of input parameters (i.e. large, medium, and small *D*-values, gradation, and compaction parameters) performed better than the conventional modeling approaches. Thus, the rigorous modeling approach of ML and involvement of a wide range of grain sizes in a modeling framework to predict the *k*-value overcome the output variability issue to a reasonable degree.
- The sensitivity analysis showed that D_{10} is the most sensitive parameter in the model framework followed by D_{50}, D_{60}, D_5, D_{30}, γ_d, and C_u subsequently. Further, the

contribution of D_{90}, D_{15}, D_{17}, and C_c in the model framework was found to be insignificant. D_{10} and γ_d are considered the most useful parameters to predict the *k*-value of sandy soil in the past literature. The sensitivity analysis findings were found to be quite consistent with the trend in the literature, supporting the validity of the GEP-based model. As a result, such modeling approaches are advised because of their advantages of decent monotonicity and precise formulation.

Acknowledgment

UET Lahore is acknowledged for providing technical data and support during the experimental phase of this study.

References

Bai, X.-D., Cheng, W.-C., Ong, D.E.L., Li, G., 2021. Evaluation of geological conditions and clogging of tunneling using machine learning. Geomech. Eng. 25, 59–73.

Bardhan, A., Gokceoglu, C., Burman, A., Samui, P., Asteris, P.G., 2021. Efficient computational techniques for predicting the California bearing ratio of soil in soaked conditions. Eng. Geol. 291, 106239.

Chapuis, R.P., 2012. Predicting the saturated hydraulic conductivity of soils: a review. Bull. Eng. Geol. Environ. 71, 401–434.

Chapuis, R.P., 2004. Predicting the saturated hydraulic conductivity of sand and gravel using effective diameter and void ratio. Can. Geotech. J. 41, 787–795.

de Bono, J.P., McDowell, G.R., 2020. On the packing and crushing of granular materials. Int. J. Solids Struct. 187, 133–140.

Deng, Y., Yue, X., Liu, S., Chen, Y., Zhang, D., 2015. Hydraulic conductivity of cement-stabilized marine clay with metakaolin and its correlation with pore size distribution. Eng. Geol. 193, 146–152.

Di Maio, C., De Rosa, J., Vassallo, R., 2021. Pore water pressures and hydraulic conductivity in the slip zone of a clayey earthflow: Experimentation and modelling. Eng. Geol. 292, 106263.

Diaz, E., Pastor, J.L., Rabat, Á., Tomas, R., 2021. Machine learning techniques for relating liquid limit obtained by Casagrande cup and fall cone test in low-medium plasticity fine grained soils. Eng. Geol. 294, 106381.

Elhakim, A.F., 2016. Estimation of soil permeability. Alexandria Eng. J. 55, 2631–2638.

Feng, S., Vardanega, P.J., Ibraim, E., Widyatmoko, I., Ojum, C., 2019. Permeability assessment of some granular mixtures. Géotechnique 69, 646–654.

Hazen, A., 1911. Discussion of dams on sand formations by A.C. Koenig. Am. Soc. Civ. Eng. 73, 199–203.

Iooss, B., Saltelli, A., 2017. Introduction to sensitivity analysis. Handb. Uncertain. Quantif. 1103–1122.

608 Jong, S.C., Ong, D.E.L., Oh, E., 2021. State-of-the-art review of geotechnical-driven artificial
609 intelligence techniques in underground soil-structure interaction. Tunn. Undergr. Sp.
610 Technol. 113, 103946.
611 Kashani, M.H., Ghorbani, M.A., Shahabi, M., Naganna, S.R., Diop, L., 2020. Multiple AI
612 model integration strategy—application to saturated hydraulic conductivity prediction
613 from easily available soil properties. Soil Tillage Res. 196, 104449.
614 Khokonov, M.K., Khokonov, A.K., 2021. Cluster size distribution in a system of randomly
615 spaced particles. J. Stat. Phys. 182, 1–20.
616 Leong, H.Y., Ong, D.E.L., Sanjayan, J.G., Nazari, A., 2015. A genetic programming predictive
617 model for parametric study of factors affecting strength of geopolymers. RSC Adv. 5,
618 85630–85639.
619 Leong, H.Y., Ong, D.E.L., Sanjayan, J.G., Nazari, A., Kueh, S.M., 2018. Effects of significant
620 variables on compressive strength of soil-fly ash geopolymer: variable analytical
621 approach based on neural networks and genetic programming. J. Mater. Civ. Eng. 30,
622 4018129.
623 McDowell, G., de Bono, J., 2021. Relating hydraulic conductivity to particle size using DEM.
624 Int. J. Geomech. 21, 6020034.
625 Morbidelli, R., Saltalippi, C., Flammini, A., Rossi, E., Corradini, C., 2014. Soil water content
626 vertical profiles under natural conditions: Matching of experiments and simulations by a
627 conceptual model. Hydrol. Process. 28, 4732–4742.
628 Mujtaba, H., Shimobe, S., Farooq, K., Khalid, U., 2021. Relating gradational parameters with
629 hydraulic conductivity of sandy soils: a renewed attempt. Arab. J. Geosci. 14, 1–17.
630 Naghadehi, M.Z., Samaei, M., Ranjbarnia, M., Nourani, V., 2018. State-of-the-art predictive
631 modeling of TBM performance in changing geological conditions through gene
632 expression programming. Measurement 126, 46–57.
633 Ong, D.E.L., Jong, S.C., Cheng, W.C., 2022. Ground and Groundwater Responses Due to Shaft
634 Excavation in Organic Soils 2. J. Geotech. Geoenvironmental Eng. 148, 5022003.
635 Ren, X., Zhao, Y., Deng, Q., Kang, J., Li, D., Wang, D., 2016. A relation of hydraulic
636 conductivity—void ratio for soils based on Kozeny-Carman equation. Eng. Geol. 213,
637 89–97.
638 Ren, X.W., Santamarina, J.C., 2018. The hydraulic conductivity of sediments: A pore size
639 perspective. Eng. Geol. 233, 48–54.
640 Schaap, M.G., Lebron, I., 2001. Using microscope observations of thin sections to estimate
641 soil permeability with the Kozeny–Carman equation. J. Hydrol. 251, 186–201.
642 Shahabi, A.A., Das, B.M., Tarquin, A.J., 1984. An empirical relation for coefficient of
643 permeability of sand.
644 Shahin, M.A., Maier, H.R., Jaksa, M.B., 2004. Data division for developing neural networks
645 applied to geotechnical engineering. J. Comput. Civ. Eng. 18, 105–114.
646 Yoo, W., Mayberry, R., Bae, S., Singh, K., He, Q.P., Lillard Jr, J.W., 2014. A study of effects
647 of multicollinearity in the multivariable analysis. Int. J. Appl. Sci. Technol. 4, 9.
648 Zeng, Z., Cui, Y.-J., Talandier, J., 2020. Evaluating the influence of soil plasticity on hydraulic
649 conductivity based on a general capillary model. Eng. Geol. 278, 105826.
650 Zhai, Q., Rahardjo, H., Satyanaga, A., 2018. A pore-size distribution function based method
651 for estimation of hydraulic properties of sandy soils. Eng. Geol. 246, 288–292.
652 Zhai, Q., Rahardjo, H., Satyanaga, A., Zhu, Y., Dai, G., Zhao, X., 2021. Estimation of wetting
653 hydraulic conductivity function for unsaturated sandy soil. Eng. Geol. 285, 106034.
654 Zhang, P., Yin, Z.-Y., Jin, Y.-F., 2021. Machine learning-based modelling of soil properties
655 for geotechnical design: review, tool development and comparison. Arch. Comput.
656 Methods Eng. 1–17.
657 Zhang, W., Li, H., Li, Y., Liu, H., Chen, Y., Ding, X., 2021. Application of deep learning

algorithms in geotechnical engineering: a short critical review. Artif. Intell. Rev. 54, 5633–5673.
Zhang, W., Zhang, R., Wu, C., Goh, A.T.C., Lacasse, S., Liu, Z., Liu, H., 2020. State-of-the-art review of soft computing applications in underground excavations. Geosci. Front. 11, 1095–1106.
Zhou, Q., Chen, D., Hu, Z., Chen, X., 2021. Decompositions of Taylor diagram and DISO performance criteria. Int. J. Climatol. 41, 5726–5732.

List of Figures

724

725

726

727

728

729

730

731

732

733

734

735
736
737

List of Tables