# Computation of Permeability of Soil using Artificial Intelligence Approaches

**Jitendra Khatti, Kamaldeep Singh Grover**

*Abstract: The Gaussian Process Regression (GPR), Decision Tree (DT), Relevance Vector Machine (RVM), and Artificial Neural Network (ANN) AI approaches are constructed in MATLAB R2020a with different hyperparameters namely, kernel function, leaf size, backpropagation algorithms, number of neurons and hidden layers to compute the permeability of soil. The present study is carried out using 158 datasets of soil. The soil dataset consists of fine content (FC), sand content (SC), liquid limit (LL), specific gravity (SG), plasticity index (PI), maximum dry density (MDD) and optimum moisture content (OMC), permeability (K). Excluding the permeability of soil, rest of properties of soil is used as input parameters of the AI models. The best architectural and optimum performance models are identified by comparing the performance of the models. Based on the performance of the AI models, the NISEK_K_GPR, 10LF_K_DT, Poly_K_RVM, and GDANN_K_10H5 models have been identified as the best architectural AI models. The comparison of performance of the best architectural models, it is observed that the NISEK_K_GPR model outperformed the other best architectural AI models. In this study, it is also observed that GPR model is outperformed ANN models because of small dataset. The performance of NISEK_K_GPR model is compared with models available in literature and it is concluded that the GPR model has better performance and least prediction error than models available in literature study.*

*Keywords: Artificial Neural Network, Gaussian Process Regression, Optimum Performance, Relevance Vector Machine*

## I. INTRODUCTION

The soil consists of different shapes and size of particles, and these particles creates persistent or connected gaps. The characteristics of soil that allow the flow of water are called the permeability of soil [5]. The soil is classified as pervious, impervious, and semi-pervious if the permeability is more significant than $10^{-3}$ mm/sec, less than $10^{-5}$ mm/sec, and between $10^{-5}$ to $10^{-3}$ mm/sec, respectively as per USBR. The constant-head permeability test and variable head permeability test methods are laboratory methods and used to determine the permeability of soil.

The determination of permeability of soil using laboratory methods is tedious and cumbersome task because it has lengthy procedures. Therefore, new methods and methodologies are developed by various researchers. These methods and methodologies include statistical methods and artificial intelligence (AI) approaches. The covSE, covRQ, covMatern 5, covMatern 3, and covExp kernel function-based GPR models has the capabilities to predict the permeability of soil [16],[2]. The random forest or tree AI approach predicts the permeability of soil with better performance [2], [15]. The sand, fly ash, time, head, specific gravity plays an important role in predicting the permeability of soil [15]. The performance of random forest or tree AI model can be improved by optimizing the model [10]. The support vector machine can also predict the permeability of soil with high accuracy and performance [2]. The artificial neural network is the most popular AI approaches and applied by numerous researchers to predict the permeability of soil. The artificial neural network models can predict the permeability of soil [16],[15],[17]. The permeability of fine soil can be predicted using artificial neural network AI approach [1] The comparison of performance of RF (simple & GA optimized), SVM, RA, GPR, ANN, ANFIS was carried out by different researchers and the following statements were made –

- The ANN models outperformed the GPR models having performance of 0.86 in predicting the permeability of soil. [16].

- The support vector machine AI models outperformed the RF, GP, RA models in predicting the permeability of soil. [2].

- The random tree models can predict the permeability of soil with the performance of 0.9125 which is better than ANNs, ANFIS AI approaches. [15]

- The optimized random forest can predict the permeability of soil with high performance i.e., 0.9171 and least prediction error and outperformed simple random forest model. [10]

- The artificial neural network outperformed the published models of Najjar and Basheer (1996) [1]

- The artificial neural network model outperformed statistical model and ANN model can predict the permeability of soil with performance of 0.901. [17]

The literature study shows that the artificial intelligence approaches have the potential to predict the permeability of soil. Most of the researchers applied the artificial neural network AI approach to compute the permeability of soil. The published study was carried out for different datasets.

**Jitendra Khatti\*,** PhD Fellow, Department of Civil Engineering, Rajasthan Technical University, Kota, (Rajasthan), India. E-mail. Id Jitendrakhatti197@gmail.com
**Kamaldeep Singh Grover,** Professor, Department of Civil Engineering, Rajasthan Technical University, Kota, (Rajasthan), India. E-mail id ksgrover@rtu.ac.in

From the literature study, it has also been observed that the hybrid learning AI approach has not been used to predict the permeability of soil and no comparison was made with machine learning and deep learning AI approaches. Based on the literature study, the present study aims to –

- Develop the Gaussian Process Regression (GPR), Decision Tree (DT), Relevance Vector Machine (RVM), and Artificial neural network AI models to compute the permeability of soil.
- Draw the comparison the between the performance of GPR, DT, RVM and ANN models to identify the best architectural and optimum performance models.
- Map the comparison between the performance of optimum performance model and published AI models in the literature study.

## II. METHOD & METHODOLOGY

Four AI approaches have been used to predict the permeability of soil in this study. Two approaches, namely Gaussian Process Regression, Decision Tree are associated with machine learning. The relevance vector machine and artificial neural network is associated with hybrid learning and deep learning, respectively. The description of the adopted AI techniques is given below.

### 2.1 Details of Gaussian Process Regression

The Gaussian Process (GP) solves the regression and classification problems under the supervised learning method. A Gaussian process is a stochastic process of statistics and probability theories. In the stochastic process, the dataset of random variables arranged by time or space having a multivariate normal distribution (MND). The Gaussian process (GP) distribution is joint distribution [4]. A process that handles an infinite-dimensional generalization of MND is known as Gaussian Process. Carl Friedrich Gauss introduced the concept of GP. The same data points should have similar target values. The concept of the Kernel for prior on the GP distribution is

$$y \sim GP(m(x), k(x, x')) \tag{1}$$

where, $k(x, x')$ kernel function of models the covariance between each pain in $x$, $m(x)$ Mean function.

**Radial Basis Function (RBF) Kernel –** The RBF is a stationary kernel known as a squared exponential kernel. The RBF kernel is –

$$k(x_a, x_b) = exp\left(-\frac{\|x_a - x_b\|^2}{2l^2}\right) \tag{2}$$

Length scale parameter $l > 0$ can be an isotropic variant and anisotropic variant of the Kernel.

**Rational Quadratic Kernel –** It's a scale mixture of gaussian kernels with different characteristics length scales. The length scale parameter is $l > 0$, and a scale mixture parameter $\beta > 0$,

$$k(x_a, x_b) = \sigma^2 exp\left(-\frac{\|x_a - x_b\|^2}{2l^2}\right)^{-\beta} \tag{3}$$

where $l$ is the length scale of the Kernel, d is the Euclidean distance, and $\beta$ is scale mixture parameters

In the present work, for the development of a rational quadratic kernel-based Gaussian process regression model, the $x_a$ = [FG, S, SG, LL, PI, OMC, MDD], and $x_b$ = k selected.

**Exponential Quadratic Kernel –** The radial basis function (RBF) kernel or square exponential Kernel is known as the exponential quadratic Kernel. The exponential quadratic Kernel is –

$$k(x_a, x_b) = \sigma^2 exp\left(-\frac{\|x_a - x_b\|^2}{2l^2}\right) \tag{4}$$

where, $\sigma^2$ is overall variance, $l$ is lengthscale

**Matern Kernel –** The Matern kernel is a generalization of the RBF kernel, and it is the stationary Kernel. An additional parameter, i.e., $\upsilon$, controls the smoothness of the resulting function. The Matern kernel is –

$$k(x_a, x_b) = \frac{1}{\Gamma(\upsilon)2^{\upsilon-1}}\left(\frac{\sqrt{2\upsilon}}{l}d(x_a, x_b)\right)^\upsilon K_\upsilon\left(\frac{\sqrt{2\upsilon}}{l}d(x_a, x_b)\right) \tag{5}$$

where, d(.,.) is Euclidean distance, $K_\upsilon(.)$ is a modified Bessel function, $\Gamma(.)$ is a gamma function.

The converge of Matern kernel to RBF kernel as $\upsilon = \infty$. If $\upsilon = 1/2$, the Kernel turns to the absolute exponential Kernel.

$$k(x_a, x_b) = exp\left(-\frac{1}{l}d(x_a, x_b)\right) \tag{6}$$

If $\upsilon = 3/2$;

$$k(x_a, x_b) = \left(1 + \frac{\sqrt{3}}{l}d(x_a, x_b)\right)exp\left(-\frac{\sqrt{3}}{l}d(x_a, x_b)\right) \tag{7}$$

If $\upsilon = 5/2$;

$$k(x_a, x_b) = \left(1 + \frac{\sqrt{5}}{l}d(x_a, x_b) + \frac{\sqrt{5}}{3l}d(x_a, x_b)^2\right)exp\left(-\frac{\sqrt{5}}{l}d(x_a, x_b)\right) \tag{8}$$

Ten models of the Gaussian process regression approach have been developed using different kernels in the present work. The architecture of GPR models with kernel functions is given in Table I.
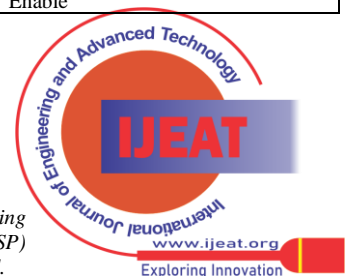
**Table- I: Architecture of the GPR Models**

| Kernel Function | Notation | Model Architecture |
|---|---|---|
| Non Isotropic Rational Quadratic | NIRQ | NIRQK_K_GPR |
| Isotropic Rational Quadratic | IRQ | IRQK_K_GPR |
| Non Isotropic Squared Exponential | NISE | NISEK_K_GPR |
| Isotropic Squared Exponential | ISE | ISEK_K_GPR |
| Non Isotropic Matern 5/2 | NIM5/2 | NIM5/2K_K_GPR |
| Isotropic Matern 5/2 | IM5/2 | IM5/2K_K_GPR |
| Non Isotropic Matern 3/2 | NIM3/2 | NIM3/2K_K_GPR |
| Isotropic Matern 3/2 | IM3/2 | IM3/2K_K_GPR |
| Non Isotropic Exponential | NIE | NIEK_K_GPR |
| Isotropic Exponential | IE | IEK_K_GPR |

The proposed GPR models have been developed using different hyperparameters. The hyperparameters and optimizer options of GPR models are given in Table II.

**Table- II: Hyperparameters of the GPR Models**

| Optimize Hyperparameters | Value |
|---|---|
| Basic Function | Auto (Constant, Linear, Zero) |
| Kernel Function | NIRQ, IRQ, NISE, ISE, NIM5/2, IM5/2, NIM3/2, IM3/2, NIE, IE |
| Kernel Scale | Auto |
| Signal Standard Deviation | Auto |
| Sigma | Auto |
| Standardize Data | Enable |

| Optimize Numeric Parameters | Enable |
|---|---|
| **Optimizer Options** | |
| Optimizer | Bayesian Optimization |
| Acquisitions Function | Expected Improvement per second plus |
| Iterations | 30 |
| Max. Training Time in Seconds | 300 |
| Number of Grid Division | 10 |

## 2.2    Details of Decision Tree

Decision tree learning predictive approaches introduced in statistics, data mining, and machine learning in the early centuries. It is a model consisting of branches and leaves. In this model, the observations go to conclusions through branches to leaves to get the target values. Tree models where the target variable can take a discrete set of values are called classification trees. The leaves represent class labels, and branches represent conjunctions of features that lead to those class labels [21],[14].

In 1984, a term umbrella term was used by Breiman to define the classification and regression decision tree procedure. The method of constructing more than one decision tree is known as the ensemble method. The ensemble methods/ techniques are –

**Boosted trees –** These trees are incremental tree which learns from the previous mistakes while building an ensemble by training. The best example of boosted trees is AdaBoost. The AdaBoost can solve the regression and classification problems [7], [9].

**Bootstrap aggregated or Bagged –** It an early ensemble method. It is based on the development of multiple trees by resampling training data with replacement and voting the trees for a consensus prediction [3]. The single-sided decision tree has one leaf node and one internal node as a child.

The Iterative Dichotomiser 3 algorithm was the first algorithm of the decision tree developed by Ross Quinlan. The ID3 algorithm starts with the original dataset D as the root node. It iterates through every unutilized attribute of the dataset D and calculates the entropy H(D) or the information gain IG(D) of that attribute. It then selects the attribute which has the most significant information gain. That attribute prepares the subsets of data by splitting the dataset D. The Iterative Dichotomiser 3 metrics are –

## 2.3    Entropy –

$$H(D) = \sum_{x \in X} -p(x) log_2 p(x) \qquad (9)$$

where D is the current dataset, X is the set of classes in D, $p(x)$ is part of the element in class $x$ to number of elements in set D

Information gain – It measures the difference in entropy from initial to end the set D.

$$IG(D, A) = H(D) - \sum_{t \in T} p(t) H(t) = H(D) - H(D|A) \quad (10)$$

where H(D) is the entropy of set D, T is the subsets created from splitting dataset D, $p(t)$ is the portion of the number of elements in t, $H(t)$ is the entropy of subset t

Various investigators and mathematicians developed the advanced edition of the ID3 algorithm. The advanced edition of ID3 algorithms is the C4.5 algorithm, classification and regression tree algorithm (CART), Chi-square automatic interaction detection (CHAID) algorithm, and MARS. In the present work, ten models of decision trees have been developed using one to ten leaves. The

architecture of DT models is **nLF_K_DT**, where n is the number of leaves. The proposed DT models have been developed using different hyperparameters. The hyperparameters and optimizer options of DT models are given in Table 3.

**Table- III: Hyperparameters of the GPR Models**

| Optimize Hyperparameters | Value |
|---|---|
| Minimum Leaf Size | Manual (1-10) |
| Surrogate Decision Splits | Off |
| Maximum Surrogate Per Node | 10 |
| **Optimizer Options** | |
| Optimizer | Bayesian Optimization |
| Acquisitions Function | Expected Improvement per second plus |
| Iterations | 30 |
| Max. Training Time in Seconds | 300 |
| Number of Grid Division | 10 |

## 2.4    Details of Relevance Vector Machine

The relevance vector machine (RVM) is a machine learning technique. It uses Bayesian inference (optimizer) to obtain promising results for regression and classification [18]. The relevance vector machine is an advanced version of the support vector machine. There is no much difference between the Gaussian process model and the RVM model with covariance function.

$$k(x, x') = \sum_{j=1}^{N} \frac{1}{\beta_j} \varphi(x, x_j) \varphi(x', x_j) \qquad (11)$$

The RVM is a Bayesian formulation, and it avoids the set of free parameters of the SVM. Usually, SVM requires cross-validation based post-optimization. The RVM uses an expectation-maximization learning method and therefore at risk of local minima. The mathematical expression of relevance vector machine kernels is –

Gaussian Kernel

$$K(x, x') = exp\left(\frac{\|x-x'\|^2}{2\sigma^2}\right) \qquad (12)$$

Linear Kernel

$$K(x) = C + \left(ai * (x, x_i)\right) (12) \qquad (13)$$

Laplacian Kernel

$$K(x, y) = exp\left(\frac{\|x-y\|}{\sigma}\right) \qquad (14)$$

Polynomial Kernel

$$K(x, y) = (x^T y + c)^d \qquad (15)$$

where d is the degree of polynomials, x and y are elements,

The Gaussian, Linear, Laplacian, and Polynomial kernels have been used for developing the relevance vector machine models. The architecture of the RVM models with kernel functions is shown in Table IV.

**Table- IV: Architecture of the RVM Models**

| Kernel Function | Notation | Model Architecture |
|---|---|---|
| Gaussian Kernel | GaussK | GaussK_K_RVM |
| Linear Kernel | LinK | LinK_K_RVM |
| Laplacian Kernel | LapK | LapK_K_RVM |
| Polynomial Kernel | PolyK | PolyK_K_RVM |

The RVM models have been optimized by the Genetic Algorithm (GA). Genetic algorithms are a family of heuristic algorithms [11]. It is used to determine the optimum or near-optimum of any functions, called the objective basis. It is necessary to make any particular hypothesis as decent gradient algorithms regarding their derivability. The population of individuals of constant size is manipulated using genetic algorithms. This population of constant size is targeted to competition between individuals. Each individual is fed as a single string of characters called a chromosome. Based on the natural phenomena of selection, crossing, and mutation operators, individuals over generations are evolved by genetic algorithms. The genetic algorithm is used to generate the high-quality solution to optimization. The proposed Alan Turing machine learning is based on parallel the principle of evaluation. The flow chart of the genetic algorithm is shown in Fig. 1.
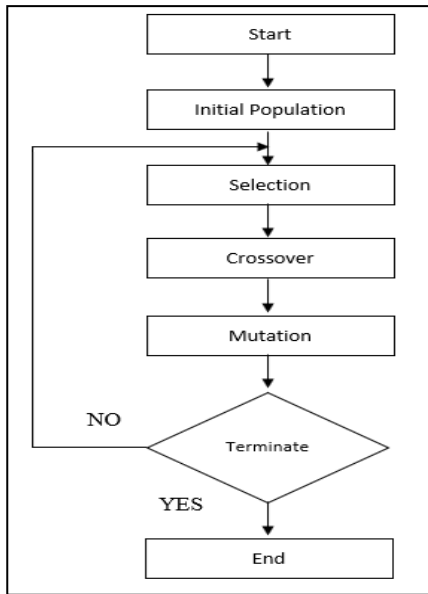


**Fig. 1.Flow chart of genetic algorithm**

The proposed RVM models have been developed using different hyperparameters. The hyperparameters and optimizer options of RVM models are given in Table V.

## 2.5 Details of Artificial Neural Networks

The artificial neural network is one of the deep learning technologies. ANN is the artificial intelligence system that understands and draws "how the human brain analyses and processes information?". The artificial neural network is a network of layers consists of different or the same hyperparameters. These hyperparameters are the number of hidden layers, neurons, activation functions, backpropagation algorithm, etc. The linear and non-linear activation functions are used at the hidden layer(s) and output layer.

**Table- V: Hyperparameters of the RVM Models**

| Optimize Hyperparameters | Value |
|---|---|
| Kernel Type | Gaussian, Linear, Polynomial, Laplacian |
| Kernel Degree | Polynomial (2) |
| Cost | 0.1 |
| Freebasis | Enable |
| MaxIter | 1000 |
| **Optimizer Options** | |

| Optimizer | Genetic Algorithm |
|---|---|
| Structure | Kernel |
| lb | $2^{-6}$ |
| ub | $2^6$ |
| Kfold | 5 |
| MaxIter | 100 |

Each neural network has a forward and backpropagation process to compute the results. In this forward process, the information travels from the input layer to the output layer. In the backpropagation process, the information travels from the output layer to the input layer. The LM, BFG, SCG, GDM, GD, and GDM algorithms are used for the backpropagation process. The algorithms are nothing, but it is a mathematical formula. The mathematical formulation of LM, BFG, SCG, GDM, GD, and GDM algorithms are –

Levenberg Marquardt (LM) Algorithm

$$[J^T J + \lambda diag(J^T J)]\, \delta = J^T[y - f(\beta)] \tag{16}$$

BFGs Quasi-Newton Method

$$B_{k+1}^{-1} = B_k^{-1} + \frac{(s_k^T y_k + y_k^T B_k^{-1} y_k)(s_k s_k^T)}{(s_k^T y_k)^2} - \frac{B_k^{-1} y_k s_k^T + s_k y_k^T B_k^{-1}}{s_k^T y_k} \tag{17}$$

Scaled Conjugate Gradient (SCG) Algorithm

$$f(x_{k+1}) = f(x_k) - \left[\frac{(a_k + (b_k + 2\alpha_k \varepsilon_k))^2}{2(b_k + 2\alpha_k \varepsilon_k)} - \rho\alpha_k\right] \le f(x_k) \tag{18}$$

Gradient Descent with Momentum (GDM) Algorithm

$$w_i'(t) = c_1 e^{\lambda i,1 t} + c_2 e^{\lambda i,2 t} \tag{19}$$

Gradient Descent (DG) Algorithm

$$w := w - \eta \nabla Q_i(w) \tag{20}$$

Gradient Descent with Adaptive Learning (GDA)

$$w_j := w_j \frac{\eta}{\sqrt{G_{i,j}}} g_j \tag{21}$$

The MLP class-based neural network has been developed in the present study to predict the permeability of the soil. The structure of the MLP class-based artificial neural network models is shown in Fig. 2.
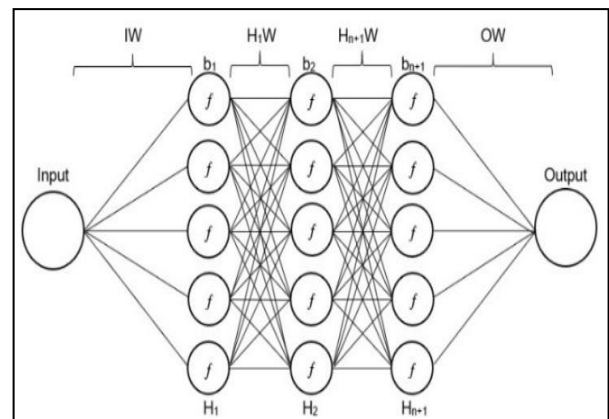


**Fig. 2.Structure of the MLP class-based ANN model**

The different hyperparameters have been used to develop the MLP class-based ANN model, as shown in Table VI.

**Table- VI: Hyperparameters of the ANN Models**

| Hyperparameters | Value |
|---|---|
| Hidden layer(s) | One, two, three, four, five |
| Neurons | Five, ten, fifteen |
| Backpropagation algorithm(s) | LM, BFG, SCG, GDM, GD, GDM |
| Activation function(s) | Sigmoid at hidden, linear at output layer |
| Train: Validation ratio | 70: 30 |
| Epochs | 1000 |
| Network type | Feed-forward backpropagation |
| Network class | Multilayer perceptron class |
| Mu | 0.001 |
| Max fail | 6 |
| Min gradient | 10e-7 |

The architecture of the artificial neural network with different algorithms is shown in Table VII.

**Table- VII: Architecture of the ANN Models**

| Backpropagation Algorithm | Notation | Model Architecture |
|---|---|---|
| Levenberg Marquardt | LM | LMNN_K_XHY |
| BFGS Quasi-Newton | BFG | BFGNN_K_XHY |
| Scaled Conjugate Gradient | SCG | SCGNN_K_XHY |
| Gradient Descent with Momentum | GDM | GDMNN_K_XHY |
| Gradient Descent | GD | GDNN_K_XHY |
| Gradient Descent with Adaptive Learning | GDA | GDANN_K_XHY |

### III. DATA ANALYSIS

A detailed description of soil datasets has been given in this section. This section also consists of data sources, data cleaning, normalization of data, etc.

### 3.1 Data Source

The datasets of soil have been collected from the following sources to carry out the present research work –

- Najjar, Y.M., Basheer, I.A., 1996, "Utilizing Computational Neural Networks for Evaluating the Permeability of Compacted Clay Liners." J. Geotech. Geol. Eng. 14, 193–212.
- Sinha, S.K., Wang, M.C. (2008), "Artificial Neural Network Prediction Models for Soil Compaction and Permeability," Geotech Geol Eng, 26,47–64.
- Wang MC, Huang CC (1984), "Soil Compaction and Permeability Prediction Models," J Environ Eng, ASCE 110(6):1063–1083.

### 3.2 Data Cleaning

The process of removing or fixing incorrect, duplicate, incomplete, corrupted, and inaccurate formatted data is known as data cleaning. A total of 159 datasets of permeability have been collected from different sources. The data cleaning process has been carried out because the datasets have been collected from various sources and may have missing or duplicate data. The outliers have been removed from the datasets to improve the quality of datasets in data cleaning. A total of 158 datasets of soil permeability have obtained after removing the outliers and redundant data.

### 3.3 Normalization of Dataset

A process of adjusting values measure on different scales to a standard scale is known as normalization. The non-dimensional ratio of errors, standard deviations, mean, and residuals are the different normalizations in statistics. The input and output variables have been normalized by min-max and log normalizing function in this research work.

### 3.4 Data Statistics

The soil properties such as fine-grain, sand content, specific gravity, liquid limit, plasticity index, optimum moisture content, and maximum dry density of 158 datasets have been used to predict soil permeability. Furthermore, the statistical parameters minima, maxima, mean, mode, median, standard deviation, and confidence level (95%) of soil properties have been calculated using the Data Analysis tool of Microsoft 2016. The statistics of the properties of soil are given in Table VIII.

**Table- VIII: Statistical parameters of soil properties**

| Parameters | Soil properties | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | FG | S | SG | LL | PI | OMC | MDD | K |
| Min | 0.00 | 0.00 | 2.52 | 22.19 | 6.78 | 7.61 | 1.50 | 0.05 |
| Max | 100.00 | 100.00 | 2.83 | 59.48 | 30.90 | 25.00 | 2.01 | 13.38 |
| Mean | 58.03 | 34.17 | 2.68 | 36.10 | 15.70 | 15.89 | 1.76 | 3.22 |
| Mode | 68.00 | 30.00 | 2.70 | 49.20 | 16.70 | 16.50 | 1.67 | 0.23 |
| Median | 66.50 | 28.25 | 2.70 | 34.60 | 15.67 | 16.06 | 1.73 | 1.58 |
| St.Dev | 30.30 | 26.27 | 0.06 | 8.48 | 4.51 | 4.36 | 0.13 | 3.26 |
| CL (95%) | 4.76 | 4.13 | 0.01 | 1.33 | 0.71 | 0.68 | 0.02 | 0.51 |

It was suggested that the value of the coefficient of correlation more than 0.8, between 0.8 & 0.2, and less than 0.2 shows the strong, good, and weak correlation for the elements and labels [8]. The coefficient of correlation for permeability of soil with different soil properties has been determined and shown in Fig. 3.
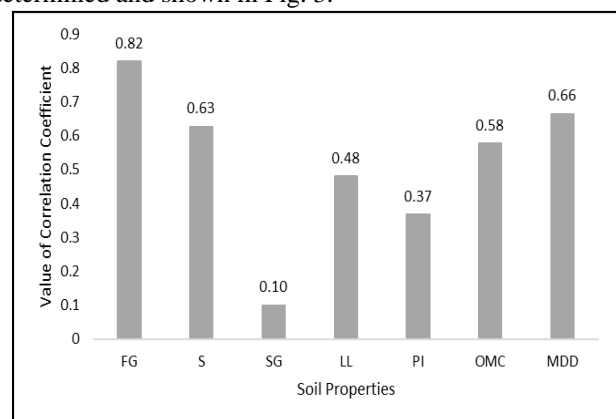


**Fig. 3.Coefficient of correlation for permeability of soil**

The permeability of soil is affected by the shape and size of the particles. Figure 3 shows the strong correlation between permeability and fine-grains.

The correlation coefficient of S, LL, PI, OMC, and MDD is 0.63, 0.48, 0.37, 0.58, and 0.66, respectively. The correlation coefficient of SG is 0.10, which is relatively poor. The Pearson correlation matrix of input and output parameters is given in Table IX.

**Table- IX: Pearson matrix of soil properties**

| (r) | FG | S | SG | LL | PI | OMC | MDD | K |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| FG | 1 | 0.84 | 0.20 | 0.61 | 0.39 | 0.71 | 0.79 | 0.81 |
| S | - | 1 | 0.15 | 0.77 | 0.62 | 0.86 | 0.85 | 0.62 |
| SG | - | - | 1 | 0.17 | 0.15 | 0.11 | 0.14 | 0.09 |
| LL | - | - | - | 1 | 0.84 | 0.82 | 0.77 | 0.48 |
| PI | - | - | - | - | 1 | 0.63 | 0.57 | 0.36 |
| OMC | - | - | - | - | - | 1 | 0.92 | 0.57 |
| MDD | - | - | - | - | - | - | 1 | 0.66 |
| K | - | - | - | - | - | - | - | 1 |

### 3.5 Training, Validation and Testing Datasets

A total of 158 datasets have been used to train, validate, and test AI models in the present research work.

The AI models, Gaussian process regression, decision tree, and relevance vector machine have been trained using 105 datasets. In the case of ANNs, the 105 datasets have sub-divided by 70% and 30%.

The MLP class-based neural network models have been trained and validated by 73 and 32 datasets. Finally, the proposed Gaussian process, decision tree, relevance vector machine, and MLP neural network models have been tested on 53 datasets.

## IV. RESULTS AND DISCUSSION

The literature survey shows that artificial intelligence approaches have the potential to predict the permeability of the soil.

The Gaussian process regression, decision tree, relevance vector machine, and artificial neural network models have been developed with different hyperparameters in MATLAB R2020a to predict soil permeability.

In this section, the performance and results of the proposed AI models have been analyzed and discussed. The performance of the AI models has been determined using RMSE, R, and MAE. The formula of RMSE, R, and MAE is –

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(T_i - P_i)^2} \tag{22}$$

$$R = \frac{\sum_{i=1}^{N}(T_i-\overline{T})(P_i-\overline{P})}{\sqrt{\sum_{i=1}^{N}(T_i-\overline{T})^2 \sum_{i=1}^{N}(P_i-\overline{P})^2}} \tag{23}$$

$$MAE = \frac{1}{N}\left(\sum_{i=1}^{N} abs(T_i - P)_i\right) \tag{24}$$

Based on the performance of AI models, the best architecture AI model has been identified from each AI technique.

The four best architecture AI models have been identified in the present research work. The performance of the best architecture AI models has been compared to recognize the optimum performance AI model.

The performance and results of AI models have analysed and discussed below –

### 4.1 Results of GPR

The proposed models of GPR have been developed using NIRQ, IRQ, NISE, ISE, NIM5/2, IM5/2, NIM3/2, IM3/2, NIE, and IE kernels to predict the permeability of the soil. The training period of NIRQ, IRQ, NISE, ISE, NIM5/2, IM5/2, NIM3/2, IM3/2, NIE, and IE kernel-based models has been 168.44, 102.73, 184.81, 94.61, 155.50, 102.05, 122.91, 82.04, 151.07 and 143.43 recorded in seconds, respectively.

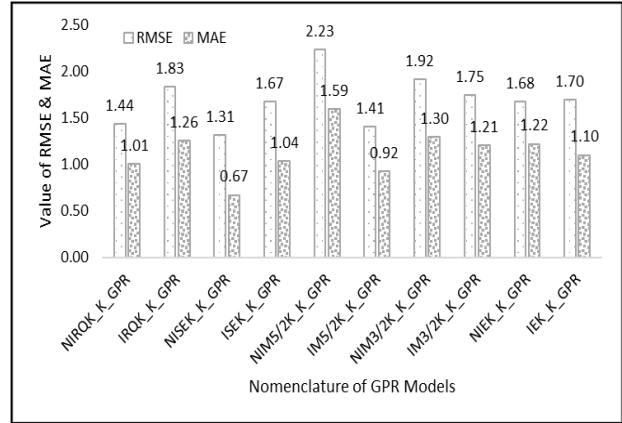The performance RMSE and MAE of the GPR models are determined and shown in Figure 4.



**Fig. 4. Performance RMSE and MAE of GPR Models**

The squared exponential kernel is very smooth (Eric P. Xing et al., 2015). The squared exponential kernel has the property of infinitely differentiable with probability one, and it also has a clear connection with a Gaussian density. From figure 4, it has been observed that the NISEK_K_GPR model has predicted the permeability of soil with 1.31 RMSE and 0.67 MAE that is significantly less to other GPR models. The performance (R) of the GPR models has also been determined and shown in Figure 5.
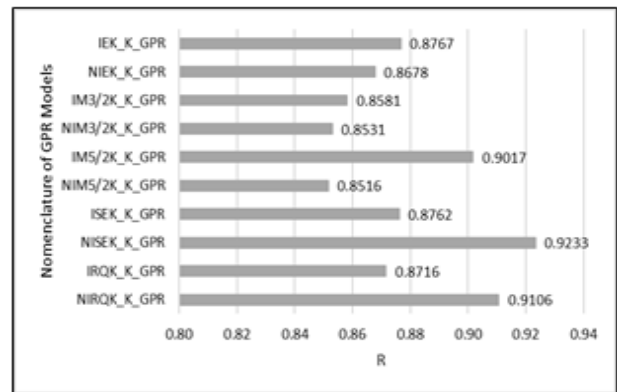


**Fig. 5. Performance R of GPR Models**

Figure 5 shows that the performance of the NISEK_K_GPR model is 0.9233, which is comparatively higher than other GPR models. The NISEK_K_GPR (Non-isotropic squared exponential Kernel) model has been identified as the best architecture GPR model based on the performance of the GPR models.

## 4.2    Results of Decision Tree

The proposed DT models have been developed using one to ten leaves to predict the permeability of the soil. The training period of 1LF, 2LF, 3LF, 4LF, 5LF, 6LF, 7LF, 8LF, 9LF, and 10LF has been 3.31, 0.60, 0.55, 0.57, 0.38, 0.42, 0.37, 0.38, 0.42, and 0.35 recorded in seconds, respectively.

The training period shows that the computation time decreases approximately exponentially with increasing leaf size [13]. The performance RMSE and MAE of the DT models have been determined and shown in Figure 6.
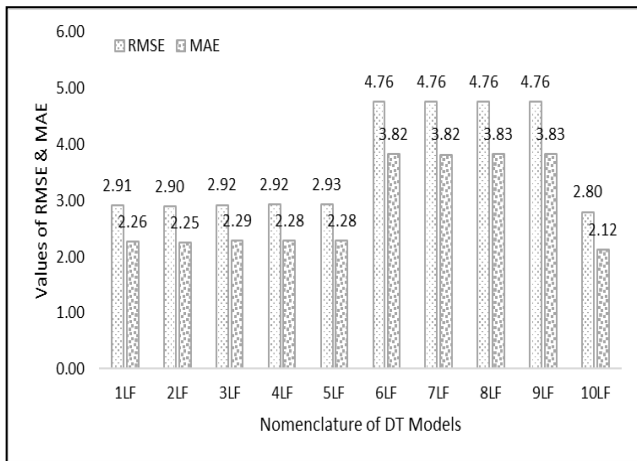


**Fig. 6.Performance RMSE and MAE of DT Models**

From fig. 6, it has been observed that the 10LF_K_DT model has predicted the permeability of soil with 2.80 RMSE and 2.12 MAE that is significantly less to other DT models. The performance (R) of the DT models has also been determined and shown in Fig. 7.
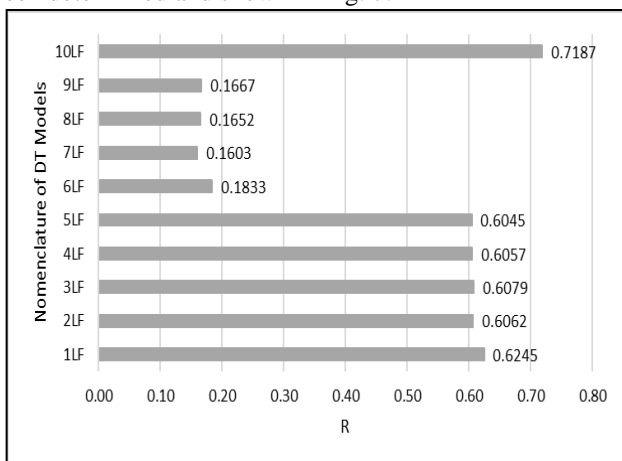


**Fig. 7.Performance R of DT Models**

Fig. 7 shows that the performance of the 10LF_K_DT model is 0.7187, which is comparatively higher than other DT models. The 10LF_K_DT model of DT has been identified as the best architecture DT model based on the model's performance.

## 4.3    Results of RVM

The relevance vector machine is an advanced machine learning technique of a support vector machine. The Kernel is the set of the mathematical function used in the support vector machine and proves the window for manipulating the data. The RVM is based on the sparse Bayesian learning

method, and the SVM is based on the rule of structure risk minimum. The RVM technique has been selected to predict the permeability of soil in the present work. The Linear, Gaussian, Laplacian, and Polynomial kernels have been used to develop the RVM models. The training period of LinK_K_RVM, GaussK_K_RVM, LapK_K_RVM, and PolyK_K_RVM has been 0.0794, 0.1511, 0.7494, and 1.0779 recorded in seconds, respectively. The performance RMSE and MAE of the RVM models have been determined and shown in Fig. 8.
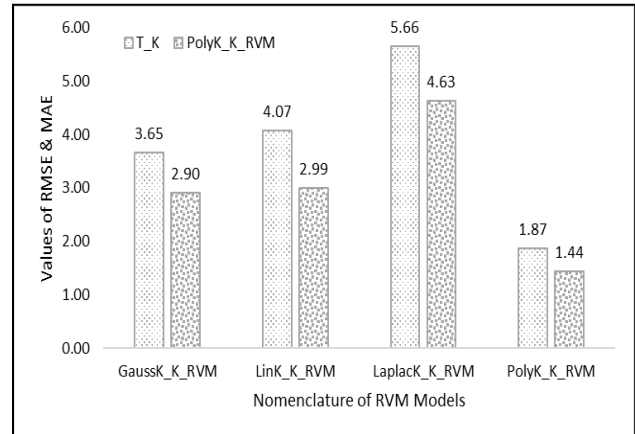


**Fig. 8.Performance RMSE and MAE of RVM Models**

From fig. 8, it has been observed that the PolyK_K_RVM model has predicted the permeability of soil with 1.87 RMSE and 1.44 MAE that is significantly less to other RVM models. The performance (R) of the RVM models has been determined and shown in Fig. 9.
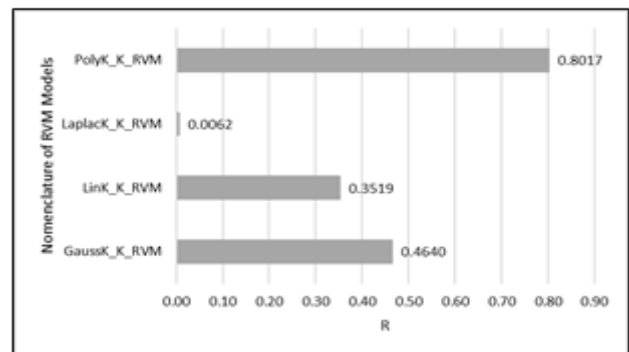


**Fig. 9.Performance R of RVM Models**

Figure 9 shows that the performance of the PolyK_K_RVM model is 0.8071, which is comparatively higher than other RVM models. The PolyK_K_RVM model of RVM has been identified as the best architecture RVM model based on the model's performance.

## 4.4    Results of ANNs

The artificial neural network is a network of layers. The LMNN_K_XHY, BFGNN_K_XHY, SCGNN_K_XHY, GDMNN_K_XHY, GDNN_K_XHY, and GDA_K_XHY models have been developed in the present research work to predict the permeability of the soil.

RMSE, R, and MAE have been used to determine the performance of the ANN models. The best architecture ANN models have been determined based on the performance of the ANN models. The RMSE and MAE performance of all ANN models is shown in Fig. 10. From fig. 10, it has been observed that the GDANN_K_10H5 NN model has predicted the permeability of soil with 1.84 RMSE and 1.54 MAE that is significantly less to other ANN models. The performance (R) of the ANN models has been determined and shown in Fig 11.
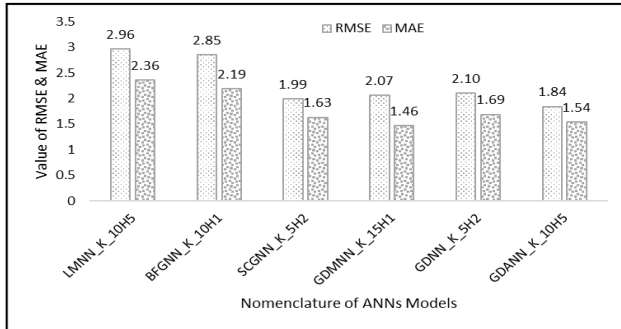


**Fig. 10. Performance RMSE and MAE of ANN Models**
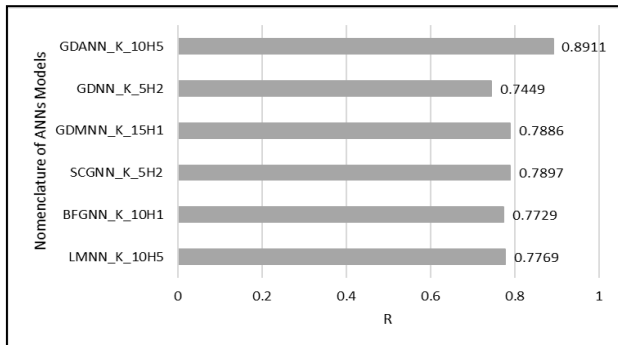


**Fig. 11. Performance R of ANN Models**

Fig. 11 shows that the performance of the GDANN_K_10H5 model is 0.8911, which is comparatively higher than other ANN models. The GDANN_K_10H5 model has been identified as the best architecture NN model to predict the permeability of the soil.

**4.5 Optimum Performance Model**

The optimum performance AI model has been recognized by comparing the performance of the best architecture neural network. The comparison of performance RMSE and MAE of the best architecture NISEK_K_GPR, 10LF_K_DT, PolyK_K_RVM, and GDANN_K_10H5 AI models have been drawn and shown in Fig. 12.
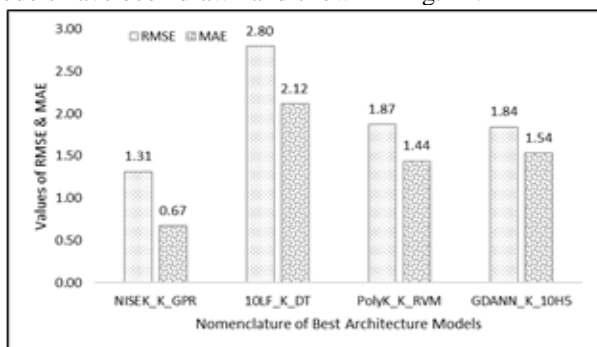


**Fig. 12. Comparison of performance RMSE and MAE of the best architectural AI models**

From fig. 12, it has been observed that the NISEK_K_GPR model has predicted the permeability of soil with 1.31 RMSE and 0.67 MAE that is significantly less to the other best architecture AI models. The comparison of performance (R) of the NISEK_K_GPR, 10LF_K_DT, PolyK_K_RVM, and GDANN_K_10H5 AI models has also been drawn and shown in Fig. 13.
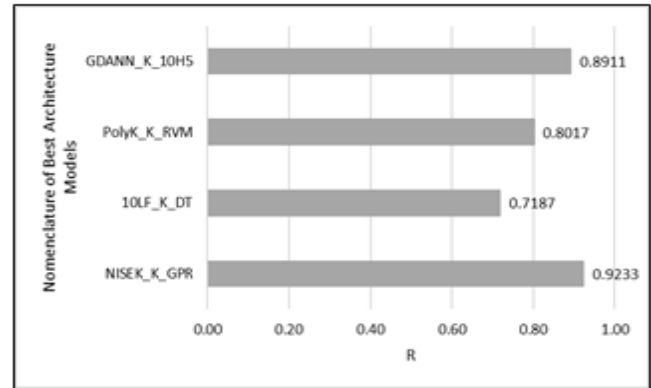


**Fig. 13. Comparison of performance R of the best architectural AI models**

Fig. 13 shows that the performance of the NISEK_K_GPR model is 0.9233, which is comparatively higher than the other best architecture AI models. Finally, it has been observed that the NISEK_K_GPR model is the optimum performance model based on the optimal performance. The predicted permeability of the soil using the NISEK_K_GPR model has compared with laboratory test permeability (T_K) and shown in Fig 14.
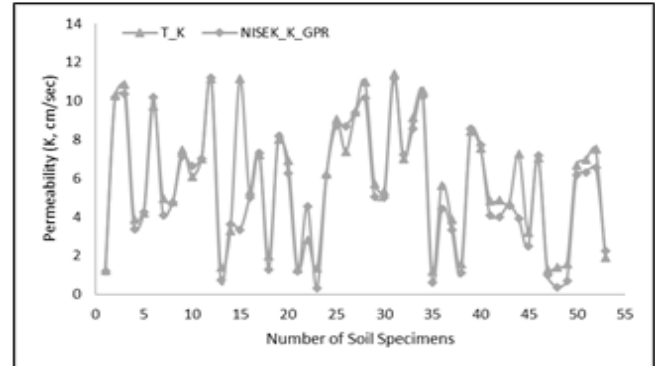


**Fig. 14. Comparison of laboratory test and predicted permeability of soil using NISEK_K_GPR model**

The residuals have also been calculated for predicted permeability of soil using the NISEK_K_GPR optimum performance model. The residuals of predicted permeability of soil are shown in Fig. 15.

Fig. 15 shows the minimal residuals in the prediction of permeability of the soil, which is equally or randomly spaced around the horizontal axis. Fig. 15 also shows that the NISEK_K_GPR model is the best quality model. The coefficient of determination has been calculated for test (T_K) and predicted permeability of soil as shown in Fig. 16.
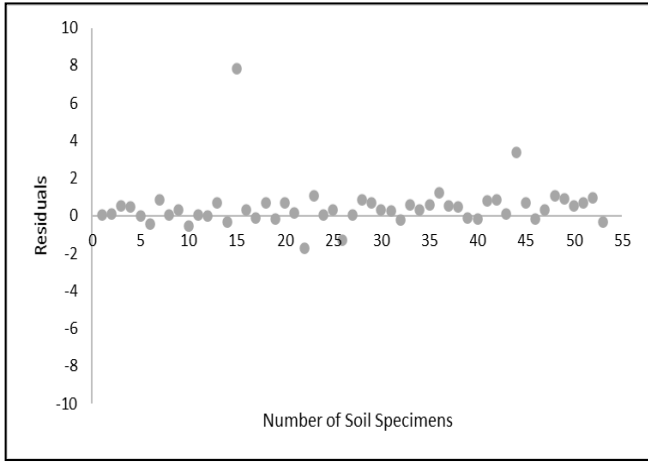
264

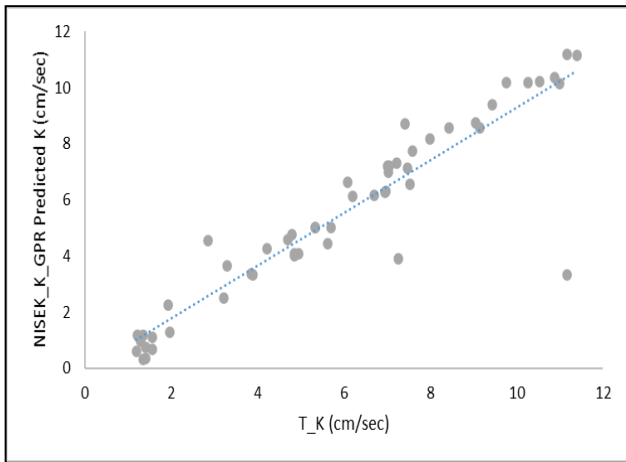**Fig. 15. Residuals in predicted permeability of soil**



**Fig. 16. Test vs predicted plot of permeability of soil**

Fig. 16 shows the performance of the optimum performance NISEK_KGPR AI model is 0.9233. The value of 0.9233 indicates that the independent variables predict 92.33% of the dependent variables. Thus, the confidence interval and prediction interval have been determined for the NISEK_K_GPR model. The confidence and prediction interval of the NISEK_K_GPR model is drawn and shown in Fig. 17 and Fig. 18.
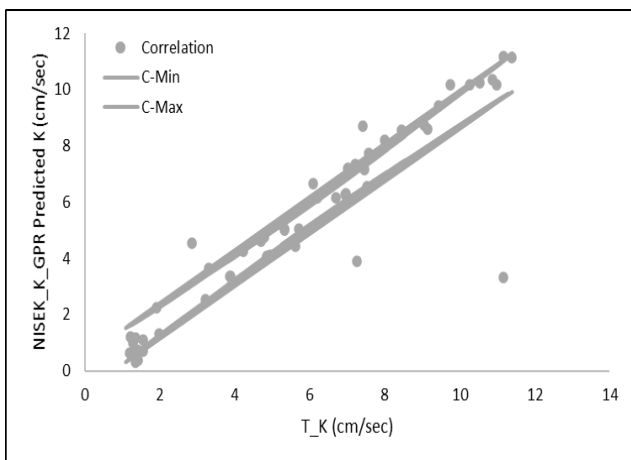


**Fig. 17. Confidence interval of predicted permeability of soil**

Fig. 17 shows that permeability of the soil has been predicted near the upper & lower confidence level. The

difference between upper and lower confidence levels is ±1.237, which shows the stable estimate of the permeability. On the other hand, Fig. 18 shows that permeability is predicted between upper and lower prediction levels. Hence, it may be stated that the optimum performance NISEK_K_GPR model has the potential to predict the permeability of the soil.
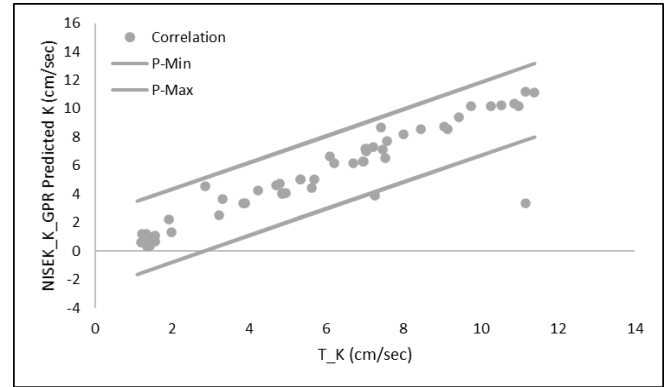


**Fig. 18. Prediction interval of predicted permeability of soil**

## V. COMPARISON WITH LITERATURE STUDY

The NISEK_K_GPR AI model has been identified as the optimum performance model of GPR. The NISEK_K_GPR AI model has been predicted the permeability of soil with 0.9233 performance. The performance of the NISEK_K_GPR model has been compared with previously proposed models in the literature survey and shown in Table X.

**Table- IX: Comparison of performance of NISEK_K_GPR with literature study**

| Authors | Performance | AI Approach |
|---|---|---|
| Jitendra Khatti et al., Present Case | 0.9233 | GPR |
| Junjui Wang et al., 2020 | 0.9171 | GA_RF |
| Praveen Sihag et al., 2018 | 0.9125 | RF |
| Soloman Asante et al. 2018 | 0.8600 | ANN |
| Sunil K Sinha et al., 2008 | 0.9010 | ANN |

From Table IX, it has been observed that the proposed NISEK_K_GPR model has outperformed the presently available AI models.

## VI. CONCLUSIONS

In the present work, the permeability of soil was predicted using machine learning and deep learning techniques. The Gaussian process regression, decision tree, and relevance vector machine are the approaches or techniques of machine learning. The machine learning and deep learning approaches were compared based on the performance in this research work. The artificial neural network is a deep learning technique. Based on the performance of AI models, the NISEK_K_GPR, 10LF_K_DT, PolyK_K_RVM, and GDANN_K_10H5 AI models were identified as the best architecture AI models.

The performance of the NISEK_K_GPR, 10LF_K_DT, PolyK_K_RVM, and GDANN_K_10H5 AI model was 0.9233, 0.7187, 0.8017, and 0.8181, respectively. The Gaussian process regression model outperformed the artificial neural network model. The reason for achieving less performance by artificial neural network models is the inadequacy of the datasets [9]. Still, in the present study, a total number of 158 datasets were used to predict the permeability of the soil. The performance of the RVM model was comparatively less than the GPR model because the RVM approach is based on the Bayesian inference, and it contains a minimum basis function. Therefore, the RVM performs better with large datasets [19]. Comparing the performance of the best architecture AI models, the NISEK_K_GPR AI model was determined as the optimum performance AI model. Furthermore, the performance of the NISEK_K_GPR AI model was compared with previously proposed models in the literature survey. The comparison showed that the NISEK_K_GPR AI model has better performance than the proposed models in the literature survey. Hence, the NISEK_K_GPR may be used to predict the permeability of the soil.

## REFERENCES

1. A. Boroumand, M. H. Bazaar (2005), Determination of compacted clay permeability by artificial neural networks, Ninth International Water Technology Conference, VTC9, 2005, Sharm El-Sheikh, Egypt.
2. Balraj Singh, Praveen Sihag, Siraj Muhammed Pandhiani, Sourav Debnath, Saurabh Gautam (2019), Estimation of permeability of soil using easy measured soil parameters: assessing the artificial intelligence-based models, ISH Journal of Hydraulic Engineering, pp. 1-11. https://doi.org/10.1080/09715010.2019.1574615
3. Breiman, L. (1996), Bagging Predictors, Machine Learning, 24 (2), pp. 123–140.
4. Carl Eduard Rasmussen, Christopher K.I. Williams (2006), Gaussian Processes for Machine Learning, MIT Press 2006. ISBN 026218253X
5. Dr. K. R. Arora (2004), Soil mechanics and foundation engineering, Standard Publisher Distributors, New Delhi, India.
6. Ehsan Momeni, Mohammad Bagher Dowlatshahi, Fereydoon Omidinasab, Harnedi Maizir, Danial Jahed
7. Friedman, J. H. (1999), Stochastic gradient boosting, Stanford University.
8. G. N. Smith (1986), Probability and statistics in civil engineering – An introduction, Collins, London.
9. Hastie, T.; Tibshirani, R.; Friedman, J. H. (2009), The Elements of Statistical Learning: Data Mining, Inference, and prediction, New York, Springer Verlag.
10. Junhui Wang, Wanzi, Zhijun, Wan, Yi Wang, Jiakun Lv, Aiping Zhou (2020), Prediction of permeability using random forest and genetic algorithm model, Computer Modeling in Engineering and Sciences, 125(3), pp. 1135-1157. doi:10.32604/cmes.2020.014313
11. Mohammed el Amin Bourouis; Abdeldjalil Zadjaoui; Abdelkader Djedid; Abderrahmen Bensenouci (2017), Design of neural networks by using genetic algorithm for the prediction of immersed CBR index, Long-Term Behaviour and Environmentally Friendly Rehabilitation Technologies of Dams, Proceedings of the 4th International Conference, Teharan, Iran. doi:10.3217/978-3-85125-564-5-046
12. Najjar, Y.M., Basheer, I.A., 1996, "Utilizing Computational Neural Networks for Evaluating the Permeability of Compacted Clay Liners." J. Geotech. Geol. Eng. 14, 193–212.
13. Philipp Probst, Marvin Wright, Anne-Laure Boulesteix (2019), Hyperparameters and tuning strategies for random forest, arXiv:1804.03515v2, pp. 1-19. doi: 10.1002/widm.1301
14. Piryonesi S. Madeh, El-Diraby Tamer E. (2020), Data Analytics in Asset Management: Cost-Effective Prediction of the Pavement Condition Index, Journal of Infrastructure Systems, 26 (1), pp. 04019036. DOI: 10.1061/(ASCE)IS.1943-555X.0000512
15. Praveen Sihag, Siraj Muhammed Pandhiani, Varun Singh, Sourav Debnath (2018), Prediction of permeability of soil using Random tree, Journal of Indian Water Resource Society, 38(3), pp. 12-18.
16. Soloman Asante-Okyere, Chuanbo Shen, Yao Yevenyo Ziggah, Mercy Moses Rulegeya, Xiangfeng Zhu (2018), Investigating the predictive performance of Gaussian process regression in evaluating reservoir porosity and permeability, 11(12), pp. 1-13. doi:10.3390/en11123261
17. Sunil K. Sinha, Mian C. Wang (2008), Artificial neural network prediction model for soil compaction and permeability, Geotechnical and Geological Engineering, 26, pp. 47-67. doi: 10.1007/s10706-007-9146-3
18. Tipping, Michael E. (2001), Sparse Bayesian Learning and the Relevance Vector Machine, Journal of Machine Learning Research, 1, pp. 211–244.
19. Vina Ayumi, Mohamad I. Fanany (2015), A comparison of SVM and RVM for Human Action Recognition, International Conference on Industrial Internet of Things, Samosir Island, North Sumatra, Indonesia, pp. 1-5.
20. Wang MC, Huang CC (1984), "Soil Compaction and Permeability Prediction Models," J Environ Eng, ASCE 110(6):1063–1083.
21. Wu, Xindong; Kumar, Vipin; Ross Quinlan, J.; Ghosh, Joydeep; Yang, Qiang; Motoda, Hiroshi; McLachlan, Geoffrey J.; Ng, Angus; Liu, Bing; Yu, Philip S.; Zhou, Zhi-Hua (2008), Top 10 algorithms in data mining, Knowledge and Information Systems, 14 (1), pp. 1–37. doi; 10.1007/s10115-007-0114-2

## AUTHORS PROFILE

**Mr. Jitendra Khatti** is pursuing PhD at Department of Civil Engineering, Rajasthan Technical University, Kota, Rajasthan. He completed his Master's from Rajasthan Technical University, Kota. He published many research articles in International and National Journals/ Conferences. The research areas of Mr. Jitendra Khatti are Applications of Artificial Intelligence in Civil Engineering, Stabilization of Soil, Pavement Engineering, Use of Waste in Concrete and Tall Structures.

**Dr. Kamaldeep Singh Grover** is Professor at Department of Civil Engineering, Rajasthan Technical University, Kota, Rajasthan. Dr. Kamaldeep Singh Grover is specialized in Geotechnical Engineering and Structural Engineering. The research areas of Dr. Kamaldeep Singh Grover are Artificial Intelligence, Concrete Technology, Tall Structures, Pavement Engineering and Stabilization of Soil.