UNIVERSITY OF MISKOLC

MIKOVINY SÁMUEL DOCTORAL SCHOOL OF SCIENCES

Head of the Doctoral School:

Prof. Dr. Peter Szücs

Estimation of Hydrogeological and Geotechnical Soil Parameters Using Artificial Neural Networks

PhD Thesis

Author:

Hasan Eteraf

Geotechnical Engineer

Supervisors:

Dr. Balázs Kovács

Dr. Viktória Mikita

Miskolc, 2024

Table of Contents

1- Introduction

1.1. Background

The complex characteristics of the ground's soil layers are fundamental to the safety and success of many geotechnical engineering projects (Das, 2017). To do this, precise measurements of the properties of the soil that govern the behavior of the ground beneath our feet are crucial (Ameratunga, 2016). In the field of geotechnical practice, soil permeability—which is commonly evaluated as hydraulic conductivity—and strength characteristics are key components that have a major influence on the stability and operation of structures (Han, 2020).

Critical to many engineering applications, hydraulic conductivity is especially important for hydrogeology, geotechnical engineering, and groundwater management. Engineers need to understand hydraulic conductivity in order to evaluate the flow of water through soil, which is necessary for managing groundwater resources, building efficient drainage systems, and assessing the possibility of pollutant transfer (Gao, 2024). For instance, accurate measurement of hydraulic conductivity aids in assessing the viability and effectiveness of different remediation methods, such as groundwater extraction and treatment, in groundwater remediation projects. Hydraulic conductivity is also useful in irrigation management for streamlining water distribution networks and guaranteeing correct irrigation of crops while reducing water waste (Gupta, 2024).

Apart from that, a key soil characteristic that controls the safety and stability of civil engineering structures, such as retaining walls, slopes, and foundations, is shear strength (Hu, 2020). Shear strength is a crucial factor in the design of geotechnical engineering since it indicates the soil's resistance to deformation and failure under applied loads (Vanapalli, 2009). For example, in foundation engineering, understanding the parameters of shear strength is crucial to ascertaining the soil's carrying capability, which in turn affects the design of both shallow and deep foundations. Similar to this, in slope stability analysis, shear strength knowledge aids engineers in determining whether naturally occurring and artificially constructed slopes are stable, reducing the possibility of landslides and other slope collapses that could jeopardize infrastructure and human life.

1.2. Problem statement

Several direct methods exist for measuring hydraulic conductivity and shear strength parameters of soil. For hydraulic conductivity, these include the constant head permeability test, falling head permeability test, packer test, slug test, and pumping test. Also, for shear strength, the methods include the direct shear test, triaxial compression test, vane shear test, and unconfined compression test. However, due to the significant dependence on in situ and laboratory testing techniques, conventional methods for measuring these parameters frequently prove to be laborious and resource-intensive (Hicher, 1996). The installation of monitoring wells for hydraulic conductivity testing and the gathering of undisturbed soil samples for shear strength testing are two common fieldwork requirements for these technologies (Craig, 2004). While laboratory and in-situ measurements provide high safety and accuracy, they can be time-consuming and resource-intensive. Finding a way to estimate these parameters with an acceptable safety index and reduced time requirements would be advantageous, as it would help mitigate delays and lower project costs. Furthermore, it can be difficult to get representative samples due to the inherent regional heterogeneity in soil properties, which can result in inaccuracies in the measured data (Dołęgowska, 2016). Because of this, engineers frequently struggle to precisely describe the behavior of soil and forecast how it will react to various loading scenarios. The challenges in obtaining accurate measurements are made more difficult by the dependence on sophisticated equipment and specialist knowledge. Accurate performance and interpretation of test findings are crucial for in situ methods like cone penetration test and borehole permeability test, which call for specialized equipment and experts with advanced training (Elhakim, 2016).

Similar to this, in order to guarantee accurate results, laboratory tests for determining shear strength—such as direct shear tests and triaxial compression tests—require specific equipment and skilled personnel. The expense and complexity of testing are increased by the requirement for specialized tools and knowledge, which also restricts the methods' applicability in distant or resource-constrained places where these resources can be hard to come by or unavailable (Nam, 2011). Additionally, the time-consuming nature of traditional testing methods poses significant challenges in project planning and execution. Delays in obtaining test results can impede decision-making processes, leading to uncertainties in design parameters and construction schedules. In fast-paced construction environments, where timely decisions are crucial to project success, the prolonged testing procedures associated with traditional methods can hinder progress and increase project risks (Viana da Fonseca, 2015). As a result, there is a growing demand for alternative approaches that offer faster, more cost-effective, and reliable solutions for estimating soil parameters, such as hydraulic conductivity and shear strength.

It is noteworthy to mention that getting soil samples for lab testing might be especially difficult in some situations, such as historical sites or places with restricted access. Strict preservation guidelines are frequently in place for historical sites, limiting the amount of infrastructure and soil disturbance (Prieto-Taboada, 2014). In these situations, typical sample techniques cannot be practical, forcing engineers to use non-intrusive approaches to indirectly infer soil parameters, including geophysical surveys or remote sensing technology (Cozzolino, 2018). Although these techniques provide insightful information on subsurface conditions, they might not be as accurate or comprehensive as direct sampling and testing. Moreover, in the preliminary design phase of engineering projects, there is often a need for quick and reliable data to inform decision-making and design optimization. Laboratory testing methods, which involve sample collection, transportation, preparation, and analysis, can be time-consuming and may not align with the fast-paced nature of preliminary design processes (Sharma, 2021). Engineers face the challenge of balancing the need for comprehensive soil characterization with the time constraints of project schedules. Consequently, there is a growing demand for innovative techniques that can rapidly estimate soil parameters with minimal time and resource requirements, allowing for more efficient and informed decision-making during the early stages of project development.

1.3. Research aims and significance

To address the limitations of conventional approaches for measuring soil parameters, several scientists and researchers have attempted to create new strategies that use easily obtainable soil properties for estimation. Grain size distribution is one such feature that is frequently easy to find in any engineering project. Grain size distribution is a key indicator of soil's hydraulic conductivity and shear strength, offering important details about the composition and structure of the material (Belkhatir, 2013). Because of this, scientists have looked into a number of methods for estimating these characteristics using information on grain size.

For estimating soil characteristics from grain size distribution, empirical approaches have been frequently used. Based on trends seen in empirical data, these methods create connections between hydraulic conductivity or shear strength and grain size measurements (Meskini-Vishkaee, 2018). Although empirical methods are straightforward and simple to apply, they might not be as robust and broadly applicable as models built on larger datasets.

In addition to empirical approaches, researchers have employed statistical methods such as regression analysis to estimate soil parameters from grain size data. Regression models are developed by fitting mathematical equations to empirical data, allowing for the quantification of relationships between independent variables and dependent variables (Nemes, 2004). By analyzing large datasets of soil samples with known parameters, regression models can identify statistically significant correlations and derive predictive equations for parameter estimation. However, the accuracy of regression-based models depends

heavily on the quality and representativeness of the training data, as well as the appropriateness of the chosen model structure (Klein, 1999).

The introduction of artificial intelligence (AI) has transformed the estimation of soil parameters based on grain size distribution, providing sophisticated methodologies that meet the issues associated with older methods (Khalili-Maleki, 2022). Among the many AI techniques, artificial neural networks (ANNs) have received a lot of attention due to their capacity to manage the complex and nonlinear correlations that exist in soil data. ANNs are computational models based on the structure and function of the human brain, with interconnected nodes (neurons) structured in layers (Citakoglu, 2017). These networks excel in extracting patterns and correlations from huge datasets, making them ideal for predicting soil characteristics based on grain size distributions. One of the primary advantages of utilizing ANNs for soil parameter estimation is their ability to capture complex correlations between input variables and output parameters. This versatility allows ANNs to account for complex soil behavior and environmental influences that simpler models may not fully capture. Moreover, ANNs demonstrate resilience and flexibility in the face of noisy or imperfect datasets, which are frequent problems in geotechnical engineering research (Park, 2011). ANNs are able to produce reliable predictions even in the face of uncertainty by adapting over time and learning from past events, which allows them to overcome data constraints (Prieto, 2016). Additionally, ANNs are scalable, which makes it possible to include a variety of input factors and ambient variables in the modeling process. This adaptability improves soil parameter estimations' accuracy and comprehensiveness, enabling researchers and engineers to make well-informed decisions in a range of geotechnical applications (Bolón-Canedo, 2011). It is noteworthy to mention that the faster estimates provided by ANNs also enhance risk management by allowing for earlier identification of potential issues, enabling more proactive and effective mitigation strategies, and ultimately supporting project safety and success.

In this study, I attempted to estimate the hydraulic conductivity of soil and shear strength parameters based on grain size distribution using artificial neural networks. Recognizing the challenges associated with traditional methods and the potential of ANNs to overcome them, a comprehensive dataset comprising laboratory-tested soil samples was collected. Subsequently, the collected data were utilized to train and validate ANN models for predicting hydraulic conductivity and shear strength parameters. Also, the performance of the ANN models was compared to other indirect methods. By evaluating the accuracy and reliability of ANN-based predictions compared to conventional methods, this study aims to highlight the effectiveness of artificial neural networks in soil parameter estimation. Additionally, it seeks to demonstrate how ANNs can advance geotechnical engineering practices by enhancing safety management and reducing risk levels.

2- Methodology

2.1. Data Collection and Preparation

The soil samples and materials used in this study were collected from the Hejőpapi area shown in Figure 6, situated in Borsod-Abaúj-Zemplén County, Hungary. The geographical coordinates of the collection site are 47°54'00" North latitude and 20°55'00" East longitude. In the laboratory, the collected soil samples were systematically separated into their constituent particles: gravel, sand, silt, and clay. This separation process ensured that each type of particle could be individually analyzed and combined in precise proportions to create new soil samples for experimentation.

The new soil samples were prepared by carefully combining varying quantities of each type of particle. The goal of this procedure was to reproduce different soil compositions and investigate their characteristics in a controlled environment. To ensure that the experimental results were free of biases or mistakes, great care was taken to preserve consistency across all samples, including maintaining uniform water content and unit weight. The precise ratios of gravel, sand, silt, and clay were adjusted to see how different soil compositions affected the characteristics under study. It is noteworthy that for the scope of this study, which focused on specific soil parameters, only small-sized gravel particles were used. All prepared samples were subjected to identical environmental conditions and handling procedures to ensure uniformity. This standardization was crucial for minimizing variability and enhancing the reliability of the experimental results. By maintaining these stringent preparation protocols, including consistent water content and unit weight, the study aimed to produce accurate and reproducible findings that could contribute valuable insights into soil behavior and characteristics.

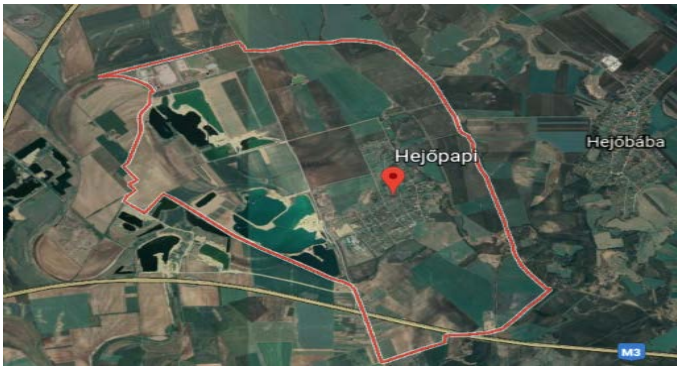


Figure 2-1. Collection Site of Soil Sample (Source: My Own Edit)

In general, 205 soil compositions were reconstructed in the laboratory. Depending on the type of parameter under investigation, all or some of these samples were investigated. The percentages of each component were variable and included 0, 10, 15, 20, 25, 30, 40, 50, 75 and 100 percent (Annex 1). It should be noted that the naming order of the samples in this research was determined after obtaining the results and performing the final analysis, which will be discussed in detail in the next chapter. The desired tests were performed on each reconstructed sample according to the standard to obtain the investigated parameters. It should be mentioned that the laboratory work for this study was carried out across two institutions. The initial phase of the experiments was conducted in the soil mechanics laboratory at the University of Miskolc. The subsequent phase of the laboratory work was carried out at the Budapest University of Technology.

2.2. Experiments

2.2.1. Grain size distribution

Following the creation of the desired soil combinations, the grain size distribution for each new soil sample was determined. To achieve this, both the sieve analysis test and the hydrometer test were conducted in

accordance with the Eurocode 7 standard. The grain size distribution analysis was crucial for providing a comprehensive understanding of the sample characteristics before conducting further experiments. It allowed for the categorization of samples, ensuring that each one was appropriate for specific experimental procedures. Additionally, it facilitated the separation of samples into coarse and fine grains, which is essential for tailoring experiments to different soil types. The analysis also enabled the determination of key parameters such as D_{10}, D_{30} and D_{60}. These parameters are critical for various aspects of my investigation, including the assessment of soil permeability and the mechanical behavior of the soil samples.

2.2.2. Hydraulic conductivity

Hydraulic conductivity, a crucial parameter for understanding water movement in soil, was measured for 205 samples based on Eurocode 7 criteria in this study. It was tried to have a similar initial condition for all specimens before conducting the tests. Following the concepts outlined in Darcy's law in previous chapter, hydraulic conductivity, also known as coefficient of permeability or permeability, refers to the ease with which water permeates rock or soil. To capture soil type heterogeneity, two different laboratory procedures were used: the constant head permeability test for coarse-grained soils and the falling head permeability test for fine-grained soils. The constant head permeability test indicates soils with high permeability, such as sand. It involves measuring the discharge flow rate (Q) through the soil sample under a constant hydraulic head (Δh) during a set duration (t). Conversely, the falling head permeability test, suitable for soils with low to intermediate permeability like clays and silts, monitored the gradual decrease in water level (h_1 to h_2) within a standpipe installed on top of the soil sample. By recording the change in water level over time (t) and considering sample dimensions, hydraulic conductivity was computed using the pertinent equations.

2.2.3. Shear strength parameters

The soil's shear strength parameters were determined using direct shear tests on a total of 95 samples, in accordance with the Eurocode 7 criteria. These tests were critical in determining the soil's resistance to internal friction and cohesion, especially for coarse and fine-grained soils. Following Eurocode 7 protocols enhanced scientific accuracy and complying with international standards, hence increasing the reliability and credibility of the findings. Each direct shear test was methodically carried out, with special attention given to sample preparation, loading conditions, and testing techniques. To account for variability and assure the correctness of the data, each test was performed three times, as is standard practice in geotechnical engineering experiments.

The direct shear test involves exposing soil samples to controlled shear stress along a designated plane to determine cohesion (c) and internal friction angle (ϕ). These characteristics were determined after a careful study of the test findings, which included calculations based on the measured shear stress and normal stress applied to the samples. The cohesion (c) was calculated using the intercept of the shear stress-normal stress plot, while the internal friction angle (ϕ) was found using the slope of the linear component of the curve. The complete investigation included differences in soil composition, grain size distribution, and other relevant parameters, resulting in a comprehensive understanding of the soil's shear strength behavior.

2.3. Multiple linear regression

In this study, Multiple Linear Regression (MLR) is initially utilized to estimate hydraulic conductivity based on the percentages of clay, silt, sand, and gravel in soil samples. Based on regression analysis, a linear relationship is established between the “response variable” and one or more “explanatory variables.” In multiple linear regression, the parameters of a linear model are estimated using an objective function and the values of the variables. Thus, if there are *n* observations from the *P* dimensional independent variable *X* and it is required to establish a relationship with the response variable *Y*, the multiple linear regression model can be used as shown in equation 1 (Uyanık, 2013).

$$Yi = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon \quad (1)$$

Where Y_i is dependent variable, x_i is explanatory variables, β_0 is y-intercept (constant term), βp is the slope coefficients for each explanatory, and variable ϵ is the model's error term (also known as the residuals)

The coefficients in an MLR model are estimated using the method of least squares. This method aims to minimize the sum of the squared differences between the observed values and the values predicted by the model. The performance of the Multiple Linear Regression (MLR) model in my work was evaluated using several key metrics. One important metric was R-squared (R^2), which represents the proportion of variance in the dependent variable explained by the independent variables. It is calculated as the ratio of the sum of squares due to regression (SSR) to the total sum of squares (SST). The coefficient of determination, R^2, is interpreted as the proportion of the variability in the dependent variable accounted for by the regression model. This value is frequently expressed as a percentage, indicating the additional explanation of variability provided by the model among the total variability (Nathans, 2012). In accordance with the classification proposed by (Chin, 1998) the range of R-squared values serves as a meaningful indicator of the explanatory power of a regression model. When R-squared is greater than or equal to 0.67, it is considered substantial, indicating that the independent variables explain a significant portion of the variability in the dependent variable. R-squared values between 0.33 and 0.67 are categorized as moderate, suggesting a moderate level of explanatory power. Lastly, R-squared values of 0.19 or higher but less than 0.33 are classified as weak, indicating a relatively lower level of explanatory power. This classification provides valuable insight into the strength of the relationship between the independent and dependent variables captured by the regression model. Additionally, I examine the p-values of the coefficients to determine whether each independent variable significantly contributes to the model. Variables with low p-values (typically less than 0.05) are considered significant predictors, indicating their strong influence on the dependent variable. Furthermore, I utilize the Mean Squared Error (MSE) to assess the average of the squared differences between observed and predicted values. MSE is a measure of the model's accuracy; lower MSE values indicate better fit, signifying that the predicted values are close to the actual values. By evaluating these metrics collectively—R-squared, adjusted R-squared, p-values of the coefficients, and MSE— it is possible to validate the robustness and reliability of the MLR model for estimating hydraulic conductivity based on grain size distribution. To ensure I developed a reliable model, I divided my data into two parts. I used 80% of the observations to train the Multiple Linear Regression (MLR) model and kept the remaining 20% for validation. This method helps us rigorously assess how well the model performs and how well it can generalize to new data. By training the model on the larger portion of the data, I make sure it effectively learns the underlying patterns and relationships between the independent variables (percentages of clay, silt, sand, and gravel) and the dependent variable (hydraulic conductivity). This way, I can be confident that my model isn't just fitting the specific data I have, but can also predict outcomes accurately for new, unseen data.

2.4. Artificial Neural Networks (ANNs)

Artificial neural networks contain artificial neurons known as units. These units are organized into a succession of layers that collectively form the entire artificial neural network in a system. A layer can have a dozen or millions of units, depending on how complicated neural networks are necessary to understand the dataset's underlying patterns. An artificial neural network typically consists of three layers: input, output, and hidden. The input layer collects inputs from the outside world that the neural network must interpret or learn about. The data is then passed through one or more hidden layers, which turns it into useful data for the output layer. Lastly, the artificial neural networks' reaction to the supplied input data is presented as an output by the output layer. Units are connected from one layer to another in most neural networks. The weights assigned to each of these relationships indicate how much effect one unit has upon the others (Abdolrasol, 2021). The neural network gains more and more knowledge about the data as it

moves from one unit to the next, ultimately producing an output from the output layer. Figure 2-2 shows a simple neural network architecture.

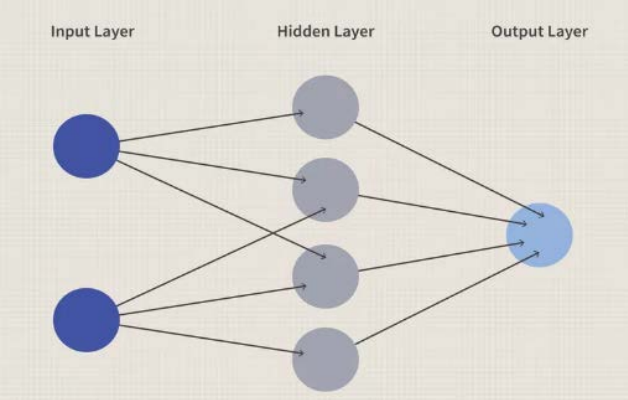


Figure 2-2. A Simple Neural Network Structure (Source: www.mediasoft.ir).

An artificial neural network's first layer, known as the input layer, transfers data from outside sources to the second layer, known as the hidden layer. Each neuron in the hidden layer takes in information from the neurons in the layer above, calculates the weighted sum, and then relays it to the neurons in the layer below. Because these connections are weighted, the effects of the inputs from the preceding layer are essentially maximized by giving each input a unique weight, which is then modified during training to improve model performance. These systems are designed to recognize patterns, process data, and learn from experience. An ANN consists of interconnected groups of artificial neurons (nodes), which work collaboratively to solve specific problems. Each neuron processes inputs and generates an output that is transmitted to other neurons in the network. Figure 2-3 shows a biological neuron and a conventional mathematical model of neurons. The main components of an artificial neural network are neurons, layers, weights (W), biases (b), and activation functions. Neurons (N) are the basic units of an ANN, analogous to biological neurons, receiving inputs (x), processing them, and producing outputs (y) using activation functions (f). A biological neuron consists of a cell body or soma for processing impulses, dendrites for receiving them, and an axon for transmitting them to other neurons. Artificial neural networks' input nodes receive input signals; the hidden layer nodes compute these input signals; and the output layer nodes compute the final output by processing the hidden layer's results with activation functions. Synapses connect biological neurons and allow impulses to be transmitted from dendrites to the cell body. Synapses are the weights that connect one-layer nodes to next-layer nodes in artificial neurons. Weight determines the strength of the linkages. Learning in biological neurons occurs in the cell body nucleus, also known as the soma, which contains a nucleus that aids in impulse processing. If the impulses are strong enough to cross the threshold, an action potential is generated and propagates via the axons. This is made possible by synaptic plasticity, which is the ability of synapses to strengthen or weaken over time in response to changes in their activity. Backpropagation is a learning approach in artificial neural networks that modifies node weights based on errors or disparities between expected and actual outcomes. When an impulse is strong enough to cross the threshold and cause a neuron to fire, this is known as activation in biological neurons. An activation function is a mathematical function that maps input to output and performs activations in artificial neural networks (Zou, 2009).

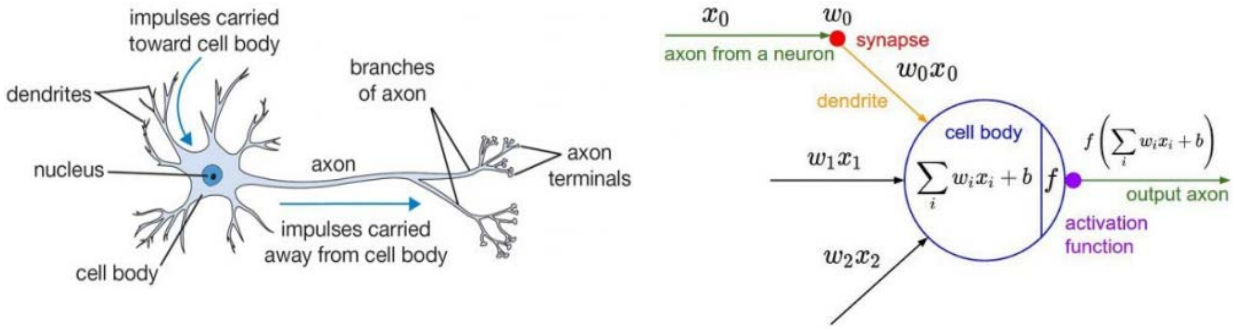


Figure 2-3. A Biological Neuron (www. hamruyesh.com) and A Conventional Mathematical Model of Neurons (Subhashini, 2020)

The output layer generates the final set of computations, estimates, or classifications based on the input data and outcomes processed by the hidden layers. Mathematically, the output of a neuron can be represented as

$$Y = \Sigma\,(weights \times input) + bias \qquad (2)$$

Weights determine the signal (or strength of the link) between two neurons. In other words, the weight determines how much influence the input has on the output. Constant biases are an additional input into the following layer with a value of 1. Bias units are unaffected by the previous layer (there are no incoming connections), but they do have outbound connections with their own weights. The bias unit ensures that even if all inputs are zeros, the neuron will still be activated.

Activation functions are applied to the weighted sum of inputs to determine the output of a neuron, allowing the network to handle non-linear transformations and complex patterns in data. Choosing the activation function for the hidden and output layers is an important decision when developing a neural network. It means that the activation functions play an important part in determining whether a neuron should be activated or not. Neural networks cannot function without activation functions; otherwise, the model's output would just be a linear function of the input. Stated differently, it would not be able to manage substantial amounts of intricate data. In every forward propagation layer, activation functions are an extra yet important step. Even if the network had numerous layers, neurons, or nodes, problems between layers could not be analyzed without activation functions. By introducing nonlinearity via activation functions, neural networks may mimic more complex functions within each node, allowing the neural network to learn more efficiently. Activation functions can be mainly classed into three types: binary step, linear, and non-linear, with several subcategories, derivatives, variants, and other calculations being employed in neural networks. The simplest sort of activation function is the binary step, which produces a binary output based on whether the input is greater than or less than a specific threshold. Linear functions are also generally simple, with the output proportionate to the input. Non-linear functions, like sigmoid and tanh functions, are more complex and bring nonlinearity into the model. In each situation, the activation function is chosen based on the individual problem and challenge to be solved (Rasamoelina, 2020). It's not always clear which one data scientists and machine learning engineers should employ, so trial and error is sometimes necessary. However, that is always the beginning point for selecting the appropriate activation function for a neural network or any other complex algorithmic-based model that requires activation functions.

The activation function used in this study was hyperbolic Tangent Function. The hyperbolic tangent function (tanh) is a widely used activation function. It converts input numbers to a range of -1 to 1. The tanh activation function can be beneficial in artificial neural networks because it is zero-centered, which helps to mitigate the vanishing gradient problem. Also, the values are more easily transferred to a scale of extremely negative, neutral, or positive. Adding to a neuron's output introduces nonlinearity, allowing the network to acquire complicated representations. However, it is important to note that Tanh suffers from the same saturation problem as the sigmoid function, in which gradients become extremely small for large input values. Despite this drawback, tanh is still a good choice for hidden layers in neural networks because of its balanced behavior near zero. Additionally, tanh is sigmoidal (s-shaped), as shown in figure 2-4. It is worth noting that feed-forward networks, which will be discussed later, use both the tanh and logistic sigmoid activation functions (Shakiba, 2020).

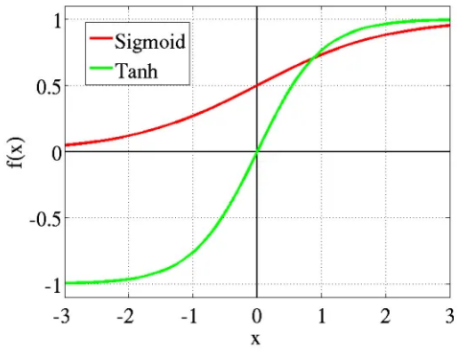


Figure2-4. Hyperbolic Tangent Function (Source: www.ashutoshtripathi.com)

The goal of this study was to use artificial neural networks (ANNs) to estimate hydraulic conductivity based on soil component weight percentages. A total of 205 soil samples were collected, with each sample's hydraulic conductivity and weight percentages of various soil components (such as sand, silt, clay, and gravel) determined. The dataset was divided into two subsets: 70% (144 samples) was utilized to train the neural network, and the remaining 30% (61 samples) was set aside for testing and validation. The data was divided at random to achieve a representative distribution of samples across both subsets. This split ratio seeks to offer enough data for training while keeping enough samples to assess the network's performance.

A feedforward neural network was selected because of its simple architecture and its efficacy in related applications. FFNs' flexibility enables the integration of many input variables as well as model architecture optimization to obtain greater predictive performance when compared to classic regression techniques. This neural network is best suited for tabular data, as each input feature (soil component) helps estimate a continuous output (hydraulic conductivity). FFNs are also well-suited to this task because they can capture complex nonlinear relationships between input variables (such as soil components) and output variables (such as hydraulic conductivity) without requiring prior assumptions about the underlying data distribution. Furthermore, FFNs are very versatile and can handle big datasets with a variety of input features, making them ideal for dealing with the multidimensional nature of soil properties and interactions. It also was taken into account that; previously conducted studies have shown that FFNs have been widely used in soil science and hydrology research because of their capacity to accurately predict complicated and nonlinear interactions between soil parameters.

The network's design consists of three layers: input, hidden, and output. MATLAB software, notably the Neural Network Toolbox, was utilized in this study to code and build the neural network. It offers a user-friendly environment for developing, training, and evaluating neural networks, as well as support for

function approximation and nonlinear regression problems. The input layer contained nodes representing the number of soil components considered (e.g., sand, silt, clay, gravel). The output layer contained a single node indicating the estimated hydraulic conductivity. The neural network was trained using the Levenberg-Marquardt backpropagation algorithm which is a popular method for training FFNNs by MATLAB because of its efficiency and speed in convergence to a solution. This method combines the benefits of gradient descent and Gauss-Newton methods, making it ideal for training networks on my dataset. After choosing the feedforward method for simulating soil hydraulic conductivity based on soil components, the investigation progressed to optimizing the neural network architecture by testing with different numbers of layers and neurons. This strategy is based on the observation that neural network performance can be highly sensitive to the architecture used, and determining the ideal configuration is critical for making correct predictions. Each architecture's performance was evaluated iteratively using criteria like prediction accuracy. The goal of carefully evaluating a variety of designs was to find the configuration that produced the optimal balance of model complexity and predictive performance.

The same approach was used in other neural network models to predict the soil's shear strength characteristics, namely cohesion and internal friction angle. For this, values from experimental procedures were applied to a dataset consisting of 95 soil samples for each parameter. An input layer representing several soil components (sand, silt, clay, and gravel) and a single output node corresponding to the expected Cohesion or internal friction angle were used in the creation of the neural networks. To ensure a representative distribution across both subsets, the datasets for each parameter were divided into two groups: 70% for training and 30% for testing and validation. The feedforward neural network (FFN) technique, which had previously been validated for estimating hydraulic conductivity, was used separately to predict cohesion and internal friction angle. Each FFN's architecture was adjusted using iterative testing with varying numbers of layers and neurons, with the goal of balancing model complexity and prediction accuracy. The construction, training, and evaluation procedures were facilitated by MATLAB's Neural Network Toolbox, which utilized the Levenberg-Marquardt backpropagation algorithm for its convergence efficiency.

3- Data analysis and results

I used several approaches, such as indirect methods, the multiple linear regression (MLR) method, and the artificial neural networks method, to create a comprehensive comparison. Empirical calculations based on easily measured soil parameters are used in indirect approaches. The MLR approach entails creating a statistical model that links several soil properties to hydraulic conductivity, improving forecast accuracy by taking into account the combined impact of numerous variables. The primary objective was to identify the optimal ANN model architecture that provides the most accurate and reliable predictions. The ANNs method is a complex strategy that may increase prediction precision. It models complex, nonlinear relationships among soil parameters and hydraulic conductivity by utilizing machine learning methods. This section compares the ways in which these approaches work in order to assess how well they predict hydraulic conductivity. The task of selecting the optimal ANN architecture for predicting hydraulic conductivity is inherently challenging due to the numerous possible configurations and the need to balance model complexity with generalization ability. Below is a detailed comparison and analysis of the results obtained from each method

3.1. Estimation of hydraulic conductivity by indirect method

As it was discussed in second chapter, Numerous empirical formulas have been proposed by scientists to predict hydraulic conductivity for both fine-grained and coarse-grained soils. These formulas typically utilize soil properties such as grain size distribution, porosity, void ratio, and plasticity characteristics. For coarse-grained soils, formulas often rely on parameters like effective grain size and uniformity coefficient, while for fine-grained soils, plasticity index and liquid limit are more commonly used. The diversity in

empirical approaches reflects the complex nature of soil properties and their impact on hydraulic conductivity. To check the validation of the empirical formula, I used the Carrier and Beckman method to predict hydraulic conductivity for a range of fine-grained soil samples with different compositions. This method was chosen due to its close alignment with the initial properties of my samples, ensuring a relevant comparison. This method is particularly suitable for remolded or disturbed samples, making it highly applicable in practical scenarios where obtaining undisturbed samples is challenging. Its reliance on easily measurable soil properties like plasticity index and liquid limit makes it versatile for various types of clays, providing a reliable estimate of hydraulic conductivity under different conditions. The Carrier and Beckman method, developed in 1984, offers an empirical formula specifically tailored for fine-grained soils, particularly clays. The formula is:

$$K = 0.174 \times \frac{[e - 0.027\,(PL - 0.242PI)\,/(PI)]^{4.29}}{(1+e)}\ (m/s) \qquad (3)$$

where k is the hydraulic conductivity, (PL) is the liquid limit, (PI) is the plasticity index, and e is the void ratio.

Figure 3-1 presents a comparison between the measured hydraulic conductivity data and the values predicted by the Carrier and Beckman method. The plot shows the measured data on the x-axis and the predicted data on the y-axis, with a trendline indicating the correlation between the two datasets. The coefficient of determination (R^2) is 0.5187, suggesting a moderate correlation. While an R^2 value of 0.51 indicates some level of predictive capability, it is not particularly strong, suggesting that the empirical formula may not fully capture the variability in the hydraulic conductivity of my soil samples. This discrepancy highlights the need for further refinement of empirical models or the use of complementary methods to improve the accuracy of hydraulic conductivity predictions for fine-grained soils. Furthermore, a significant portion of the data points, as shown in Figure 3-1, fell below the bisector line, demonstrating a persistent underestimation of the values predicted by the Carrier and Beckman approach in comparison to the measured data. So, the empirical formula might not be completely reliable for some factors impacting hydraulic conductivity in my samples, as suggested by this consistent underestimating.

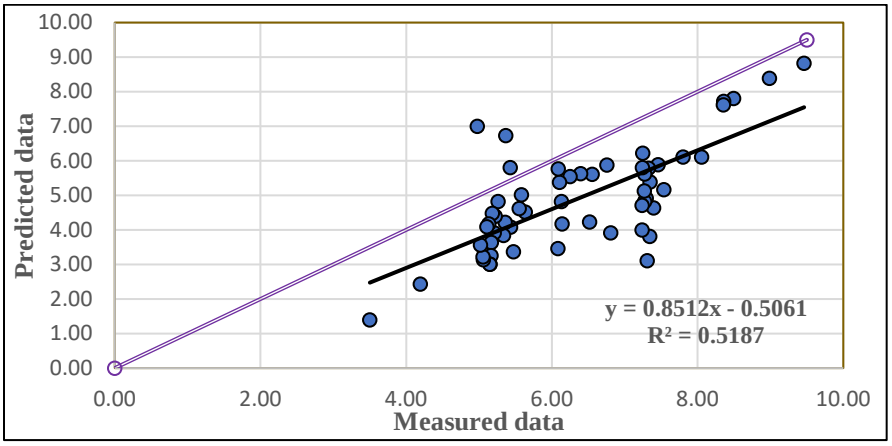


Figure 3-1. Comparison of (−LogK) Values Obtained by Empirical Formulae and Laboratory Tests (Source: My Own Edit)

3.2. Estimation of hydraulic conductivity by Multiple linear regression analysis

To conduct the Multiple linear regression analysis, I used SPSS software. Initially descriptive statistics obtained to gain insight into the central tendency and dispersion of the variables, providing a comprehensive overview of the data's characteristics. According to Table2, the standard deviation for normalized hydraulic

conductivity is approximately 1.212. After performing the multiple regression, the desired results were extracted, which I will continue to interpret. According to Table 4, since the correlation coefficient (R = 0.906) and determination coefficient (R^2= 0.822) as well as adjusted determination coefficient (Adjusted R^2 = 0.819) were calculated, it seems that the regression model is appropriate. The closer these values are to 1, the more the model expresses the relationship between the dependent and independent variables. In other words, the regression model was able to cover or express a greater percentage of changes in the dependent variable.

Table 3-1. MLR Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	DW	MSE
1	0.906	0.822	0.819	0.515	1.797	0.266

In Table 4-2, the variance analysis for the regression model has been done. Considering the size of F and the value of Sig=<0.0001, I conclude that the regression model will be appropriate. Because most of the changes in the dependent variable have been seen in the regression model. This means that the contribution of the model (Regression) in the total changes that can be seen in the last row (Total) of the column (Sum of Squares) is much higher than the contribution of error or residuals.

Table 3-2. Analysis of variance.

Source	DF	Sum of squares	Mean squares	F	Sig	p-values signification codes
Model	3.000	246.050	82.017	308.691	<0.0001	***
Error	201.000	53.404	0.266			
Corrected Total	204.000	299.454				
Computed against model Y=Mean(Y)						
*Signification codes: 0 < *** < 0.001 < ** < 0.01 < * < 0.05 <. < 0.1 < ° < 1*						

Finally, the scatter plot depicted in figure 3-2 derived from the multiple linear regression (MLR) analysis which represents the relationship between predicted and real data. The coefficient of determination (R^2) is 0.81, indicating that 81% of the variability in the real data is explained by the model. This value suggests that the MLR model has a relatively good fit and effectively captures the relationship between the independent variables and the dependent variable.

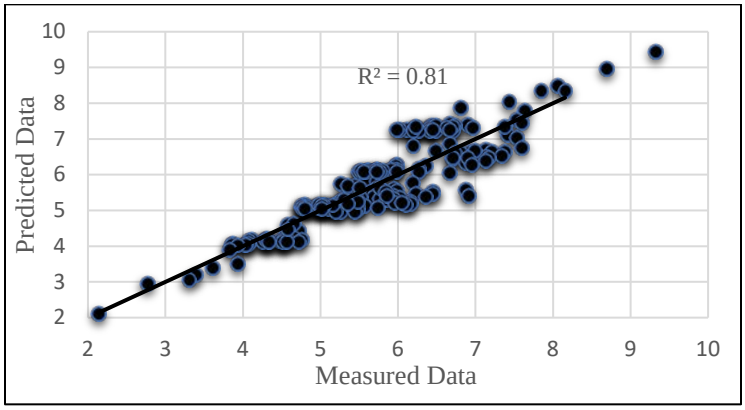


Figure 3-2. Multiple Linear Regression Analysis Result (Source: My Own Edit)

3.3. Artificial Neural Network model for hydraulic conductivity parameter

In this section, I present the results of my analysis for estimating hydraulic conductivity based on the weight percentages of soil components (clay, silt, sand, and gravel) using artificial neural networks. The primary goal was to identify the optimal ANN model architecture that provides the most accurate and reliable predictions. The task of selecting the optimal ANN architecture for predicting hydraulic conductivity is inherently challenging due to the numerous possible configurations and the need to balance model complexity with generalization ability. I evaluated multiple models with varying configurations of hidden layers and neurons.

I began my exploration with two neurons in a single layer. Subsequently, I expanded my investigation by varying the number of neurons within this initial layer. Following this, I introduced an additional layer, increasing the depth of my network to two hidden layers. I conducted several iterations, adjusting the number of neurons in each layer to discern their impact on performance. In pursuit of further insights and more confident outcomes, I conducted a test by introducing a third hidden layer. Despite altering the number of neurons across these layers, my efforts did not yield any noticeable improvements in performance.

All in all, I tested a total of ten different ANN models, each with a unique architecture. The architectures varied in terms of the number of hidden layers and the number of neurons within those layers. The selection of the optimal artificial neural network structure was guided by key performance metrics including the coefficient of determination (R^2), Root Mean Square Error (RMSE), and Mean Squared Error (MSE). These metrics were employed as indicators of the model's predictive accuracy and goodness of fit. R^2, also known as the coefficient of determination, measures the proportion of the variance in the dependent variable that is predictable from the independent variables, thus assessing the model's explanatory power. RMSE represents the square root of the average squared differences between predicted and observed values, providing a measure of the model's prediction error. Lastly, MSE calculates the average of the squared differences between predicted and observed values, offering insight into the variance of the prediction errors. In the context of ANN modeling, these metrics serve as crucial evaluation tools to iteratively refine and optimize the network architecture, ensuring robust and reliable predictions. The equations 4, 5, and 6 correspond to these metrics.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - x_i)^2}{n}} \quad (4), \quad MSE = \frac{\sum_{i=1}^{n}(y_i - x_i)^2}{n} (5), \quad R^2 = \frac{\sum_{i=1}^{n}(y_i - x_i)^2}{\sum_{i=1}^{n}(y_i - \hat{y})^2} (6)$$

Which, n represents the number of measurements, y_i and x_i are the observed and predicted values of the dependent variable, respectively, and $\bar{y}$ is the mean of the observed values of the dependent variable.
Table 3-3 shows the Regression analysis for all ten models across training, validation, testing, and overall datasets. Model number 4, highlighted in the table, performed the best across all phases, with highest R values. Figure 3-3 shows regression plots for the best-performing model (Model 4), with separate plots for training, validation, testing, and all data combined.

Table 3-3. ANN Models for Hydraulic Conductivity Prediction.

				R Values			
Model no	**output**	**Number of hidden layers**	**Number of neurons**	**Training**	**Validation**	**Testing**	**All**
1	-Log K	1	2	0.67245	0.7122	0.69542	0.6829
2	-Log K	1	4	0.7157	0.72214	0.6519	0.70154
3	-Log K	2	2,2	0.82521	0.8021	0.78025	0.8102
4	-Log K	2	1,3	0.94939	0.96013	0.9373	0.95471
5	-Log K	2	4,1	0.77591	0.7928	0.84146	0.79013
6	-Log K	2	3,4	0.86813	0.84292	0.803	0.8562
7	-Log K	3	2,5,1	0.88235	0.7452	0.80187	0.84094
8	-Log K	3	4,1,6	0.8125	0.72491	0.8232	0.80191
9	-Log K	3	3,2,2	0.87513	0.81793	0.7029	0.85517
10	-Log K	3	1,4,1	0.88922	0.9032	0.87961	0.89183

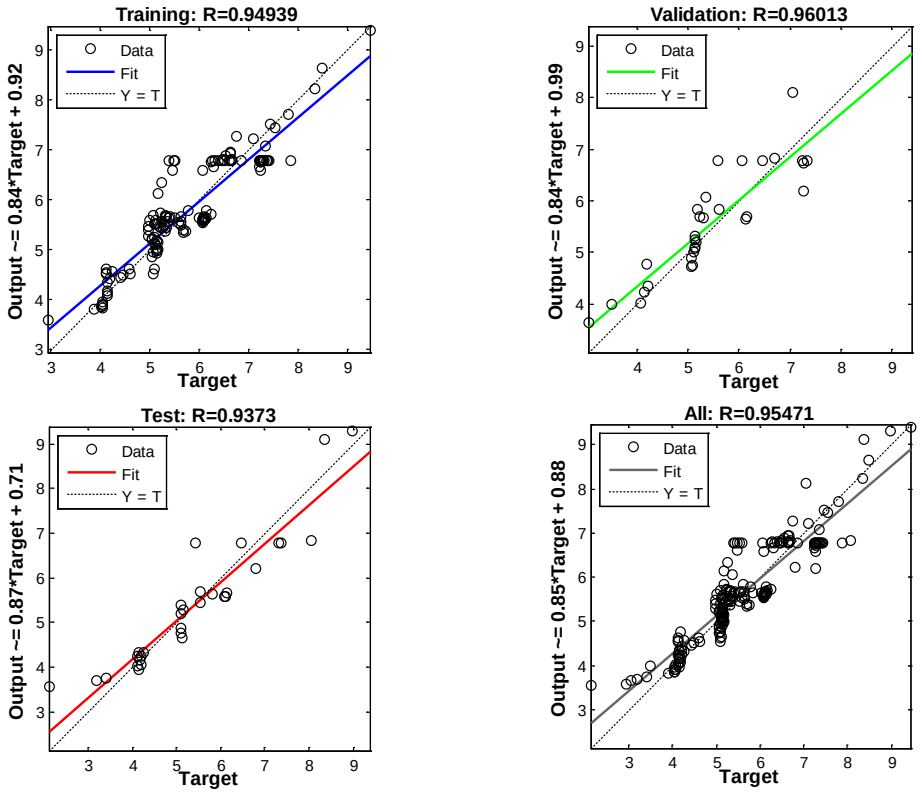


Figure 3-3. Regression Plots of the Best-Performing Model for (-Log K) Prediction (Source: My Own Edit)

Thus, the analysis shows that ANN model number 4 with two hidden layers correspondence number of neurons is the most effective for predicting hydraulic conductivity based on soil component weight percentages.

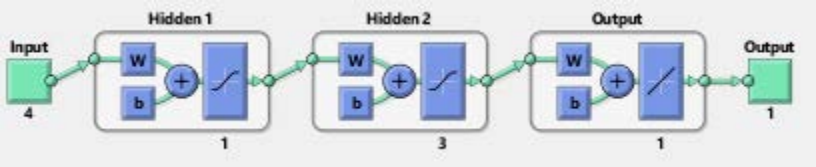


Figure 3-4. Neural Network Structure for (-Log K) Prediction (Source: My Own Edit)

Figure 3-5 compares the actual (-log K) values with the neural network's predictions for the testing dataset. To better visualize the model's performance, the data was sorted in descending order. This sorting allows for a clearer comparison and helps highlight the extent of the fitting between the predicted and actual values. According to the figure3-5, except for a few outliers, the neural network's predictions are very close to the actual values. The close alignment of the predicted values (red line) with the actual values (blue line) demonstrates the model's ability to accurately predict hydraulic conductivity based on soil composition. It indicates that the model has learned the complex relationships between soil components (clay, silt, sand, and gravel) and hydraulic conductivity. This high level of agreement between predicted and actual values confirms the accuracy of the ANN model.

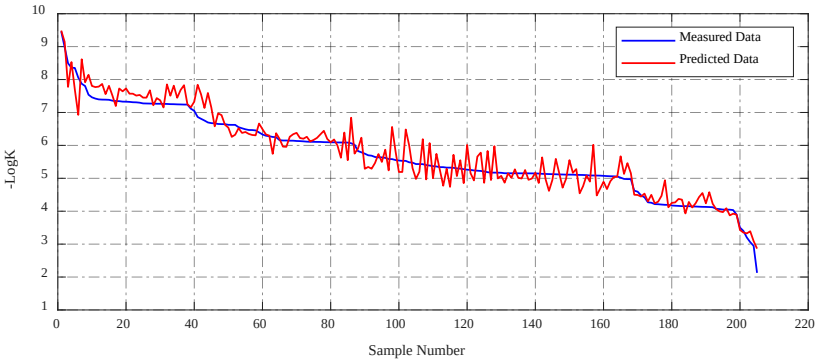


Figure 3-5. Comparison of (- Log K) Values Obtained by ANN and Experiments (Source: My Own Edit)

As a next step of coding, I evaluated the performance of my neural network by calculating the Root Mean Squared Error (RMSE). I computed the 'RMSE' value by taking the square root of the mean of the squared differences between the predicted values and the measured values. It is noteworthy to mention that the RMSE value obtained by the model was 0.0768, indicating a high level of accuracy in the model's predictions. This low root mean square error suggests that the differences between the predicted values and the actual values are minimal, thereby validating the model's effectiveness in capturing the underlying patterns of the dataset. Another important point obtained from the model was the importance of each parameter in estimation. It was found that the importance of the clay and gravel components is evident from the importance weight assigned to each variable. As mentioned earlier, the data was sorted in descending order to help understand the performance of the model; this means that as the number of tests increases, the values of -Log K drop. According to figures 3-6, there is a visible trend between these parameters and the output parameter, and they have less deviation of data in comparison to silt and sand. So, the results indicate the weight percentages of clay and gravel have a significant and meaningful effect on hydraulic conductivity, more so than silt and sand. It means that they are critical factors influencing the model's

predictions. The high importance weight of these variables underscores their significant role in the underlying processes modeled by the ANN, demonstrating that accurate measurement and inclusion of these components are essential for reliable model performance.

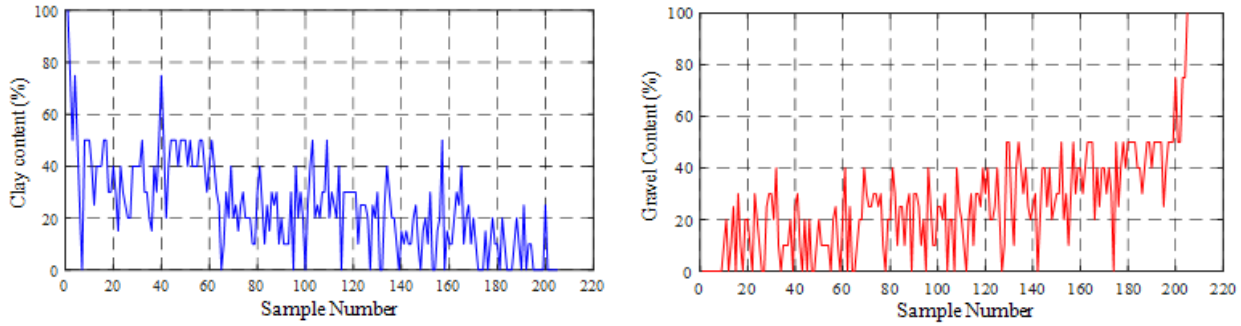


Figure 3-6. Relationship Between Weight Percentage of clay and Gravel and Hydraulic Conductivity Value (Source: My Own Edit)

In summary, I presented an analysis of estimating hydraulic conductivity based on the weight percentages of soil components (clay, silt, sand, and gravel) using artificial neural networks. I aimed to identify the optimal ANN model architecture that provides the most accurate and reliable predictions. I evaluated multiple models and found out the most effective, achieving R^2 values of 0.92. This high level of accuracy, with an RMSE value of 0.0768, underscores the model's effectiveness in capturing the underlying patterns in the data. Notably, I could not find any previous studies using the same input data to predict the K value, though various other modeling approaches exist. Comparison of my results with those from similar studies showed that my model achieved superior predictive accuracy with R^2 values consistently above 0.90, positioning it among the best-performing models. The analysis also highlighted the significant impact of clay and gravel on hydraulic conductivity, more so than silt and sand highlighting their essential role in the model's predictions. Regarding complexity, the MLR method is the least complex and easy to interpret, but it is limited in handling non-linear relationships. The empirical formula is simple to use but offers limited accuracy and flexibility. On the other hand, the ANN model is the most complex and computationally intensive, but it is highly flexible and capable of capturing intricate patterns in the data. According to figure 3-7, the ANN approach has the best accuracy R^2 value of 0.92, indicating better predictive accuracy. With an R^2 value of 0.81, the MLR approach demonstrated moderate accuracy; this is a better fit but less precise than the ANN model. With an R^2 value of 0.52 and the lowest accuracy of all investigated methods, the empirical formula proposed by Carrier and Beckman clearly needed improvement.

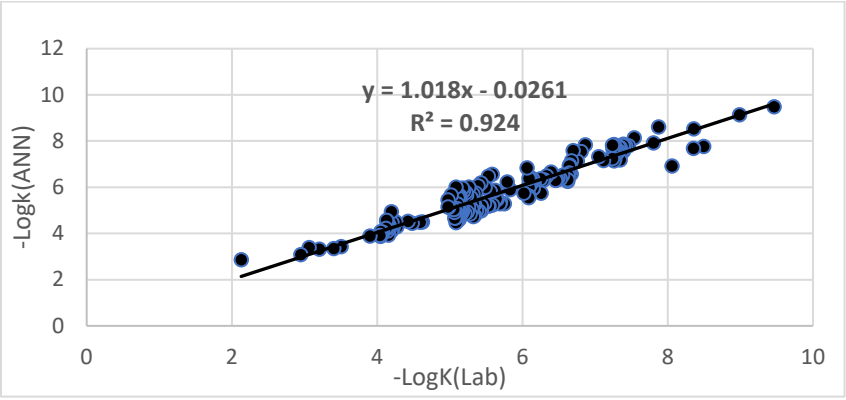


Figure 3-7. Coefficient of Determination Value for the Best Model of ANN (Source: My Own Edit)

3.4. Artificial Neural Network model for cohesion parameter

In this section, I extend the application of artificial neural networks to predict the shear strength characteristics of soil, specifically cohesion and internal friction angle. Following the methodology used in previous neural network models for hydraulic conductivity, I employed a dataset comprising 95 soil samples for each shear strength parameter. Each sample included the weight percentages of soil components (sand, silt, clay, and gravel) as input features. The datasets for both cohesion and internal friction angle were divided into training (70%) and testing/validation (30%) groups to ensure a representative distribution across both phases.

The Feedforward Neural Network (FFN) technique, validated in the hydraulic conductivity study, was applied to predict cohesion and internal friction angle independently. Iterative testing with varying architectures was conducted to identify the optimal balance between model complexity and prediction accuracy. The construction, training, and evaluation of the neural networks were performed using MATLAB's Neural Network Toolbox, utilizing the Levenberg-Marquardt backpropagation algorithm for its efficient convergence properties. The neural network architecture for estimating soil cohesion is illustrated in figure 3-8. It comprises an input layer with four nodes corresponding to the weight percentages of clay, silt, sand, and gravel. This is followed by three hidden layers, each with 3, 1 and 4 neurons respectively, and a single-node output layer that estimates cohesion. Through iterative testing, the mentioned structure was identified as optimal, achieving a balance between underfitting and overfitting. The model's performance was evaluated using key metrics. According to figure 3-9, the correlation coefficient (R) for cohesion was approximately 0.90 which indicates a relatively good correlation between the predicted and actual values. Additionally, the Root Mean Square Error (RMSE) value for cohesion was found to be 0.53, presenting the model's predictive accuracy.

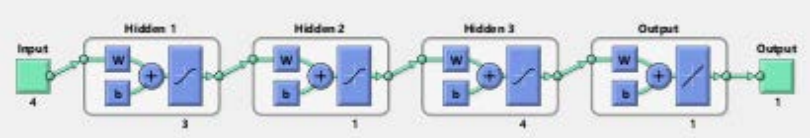


Figure3-8. Neural Network Structure for cohesion Prediction (Source: My Own Edit)

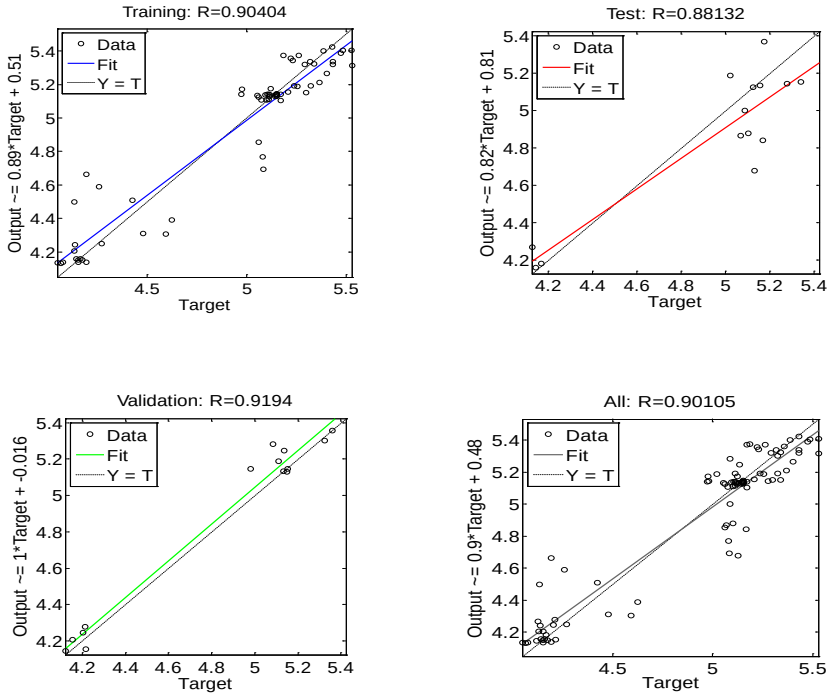


Figure 3-9. Regression Plots of the Best-Performing Model for Cohesion Prediction (Source: My Own Edit)

3.5. Artificial Neural Network model for internal friction angle parameter

The same procedure was followed to estimate the other shear strength parameter, the internal friction angle, using the ANN model. The neural network architecture for predicting the internal friction angle is depicted in figure 3-10. It consists of an input layer with four nodes representing the weight percentages of clay, silt, sand, and gravel. This is followed by two hidden layers, each with different numbers of neurons, and a single-node output layer that estimates the internal friction angle. The optimal network architecture was determined to have [3 2] neurons, which provided a suitable balance between underfitting and overfitting for the data. The model's effectiveness in predicting the internal friction angle was similarly assessed using performance metrics. The correlation coefficient (R) for the internal friction angle was approximately 0.91 that shows a good correlation between the predicted and actual values as shown in figure 3-11. Furthermore, the obtained RMSE value for the internal friction angle was 0.39, demonstrating the model's accuracy in prediction.

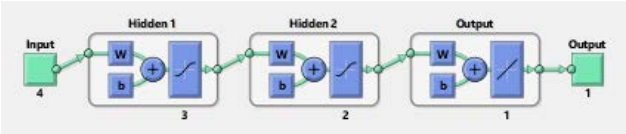


Figure 3-10. Neural Network Structure for Prediction of Internal Friction Angle (Source: My Own Edit)

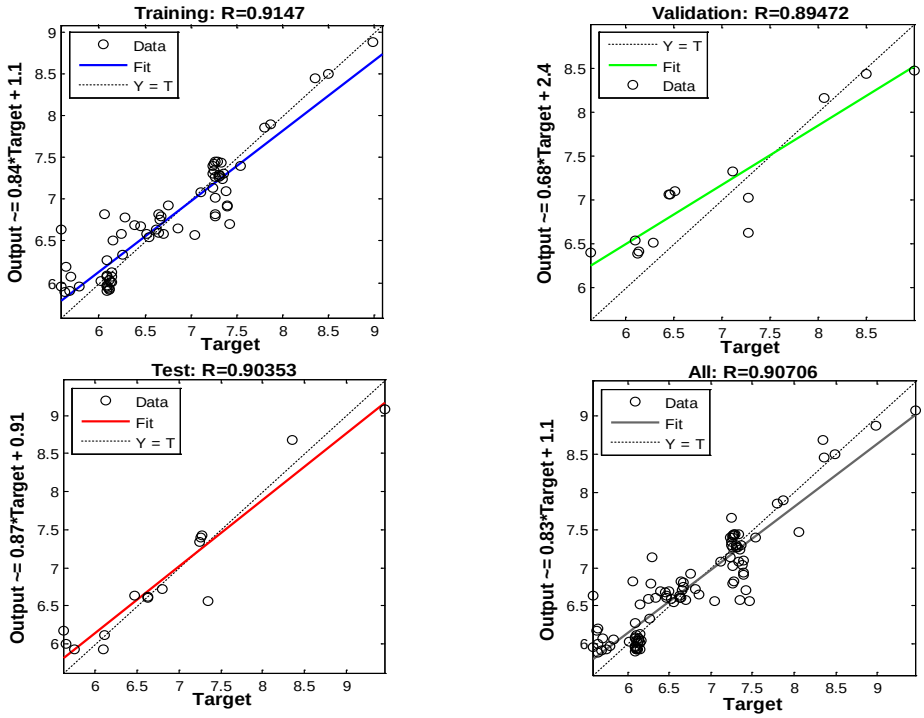


Figure 3-11. Regression Plots of the Best-Performing Model for Prediction of Internal friction Angle (Source: My Own Edit)

3.6. Simultaneous modeling of shear strength parameters

One powerful application of ANNs is the creation of joint models. Joint models in artificial neural networks represent a sophisticated approach to predictive modeling by simultaneously predicting multiple output variables from a shared set of input features. Unlike traditional models that handle each output independently, joint models capture interdependencies and correlations between outputs, enhancing

predictive accuracy and computational efficiency. This approach is particularly advantageous in fields where outputs are inherently related, such as geotechnical engineering, where predicting both cohesion and friction angle together improves the understanding of soil behavior. The design of a joint model involves configuring the neural network architecture with multiple output neurons corresponding to each variable of interest and utilizing a multi-output loss function during training to optimize predictions across all outputs simultaneously.

The decision to opt for a joint model was primarily driven by the understanding that cohesion and friction angle exhibit correlated behavior in soil mechanics. As friction angle increases, cohesion typically decreases, and vice versa. By modeling cohesion and friction angle together, rather than separately, the jointly model can capture synergistic effects and dependencies that affect both parameters simultaneously. This approach leads to more accurate predictions of soil behavior across a wide range of scenarios. Engineers and researchers can optimize geotechnical designs and analyses more effectively when they have a comprehensive understanding of both these parameters. This includes designing foundations, slopes, excavations, and other structures that rely on accurate soil parameter predictions.

In traditional neural networks, each output corresponds to a single neuron in the output layer, predicting a single target variable. However, to model the parameters simultaneously, the output layer was configured to have multiple neurons, each representing a different output (cohesion and friction angle). By having multiple neurons in the output layer, the neural network learns to optimize weights and biases across all outputs simultaneously. The shared input features (Weight percentage of clay, silt, sand and gravel) are processed through hidden layers that extract relevant patterns and relationships. These hidden layers act as shared representations, contributing to the prediction of both shear parameters and leveraging correlations between them. The network architecture comprises hidden layers with sizes [3, 2, 4], utilizing hyperbolic tangent sigmoid ('tansig') functions for activation in the hidden layers and a linear ('purelin') function for the output layer.

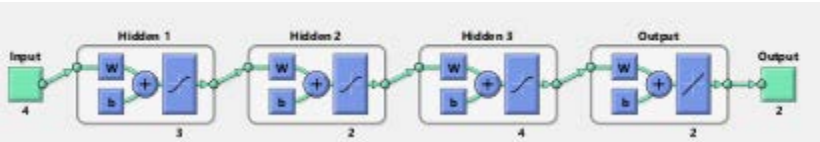


Figure 3-12. Neural Network Structure for Simultaneous Model of shear strength parameters (Source: My Own Edit)

Upon implementation, the joint neural network achieved significant improvements in predictive performance compared to previous independent modeling efforts. The new Root Mean Squared Error (RMSE) values were notably reduced: RMSE for cohesion was 0.41 and for friction angle was 0.25, indicating enhanced accuracy in predicting these soil parameters based on the input features. Moreover, the new R^2 values exhibited substantial increases, with R^2 value of 0.97 for cohesion and 0.90 for friction angle. These R^2 values signify a marked improvement over the previous independent models, where R^2 for cohesion was 0.81 and for friction angle was 0.82.

To visualize the model's performance, the regression curves were plotted for both cohesion and friction angle by comparing actual values against predicted values. This is done using scatter plots, where each point represents an actual-predicted pair. Additionally, I incorporate trendlines to better illustrate the correlation between actual and predicted values. Specifically, for each subplot, I use the ‘polyfit’ function to compute the coefficients of a linear fit for the data points, and the ‘polyval’ function to evaluate this linear fit across the range of actual values.

The reduction in RMSE and increase in R^2 values directly contribute to the reduction in the standard deviation of the errors. RMSE quantifies the average magnitude of prediction errors, where lower values indicate that predictions are closer to actual values, thus reducing the variability in prediction errors. Similarly, R^2 measures the proportion of variance in the dependent variable that is predictable from the independent variables; higher R^2 values indicate a better fit of the model to the data, thereby reducing the overall spread of errors around the fitted line. By achieving lower RMSE and higher R^2 values, the joint neural network effectively minimizes the variability and standard deviation of prediction errors. This enhancement underscores the model's capability to leverage shared representations and correlations between outputs, thereby refining predictions and supporting more accurate and reliable decision-making in geotechnical engineering applications. Consequently, this joint neural network model stands out as the preferred choice for predicting both cohesion and friction angle due to its superior performance metrics and comprehensive approach to capturing soil behavior characteristics.

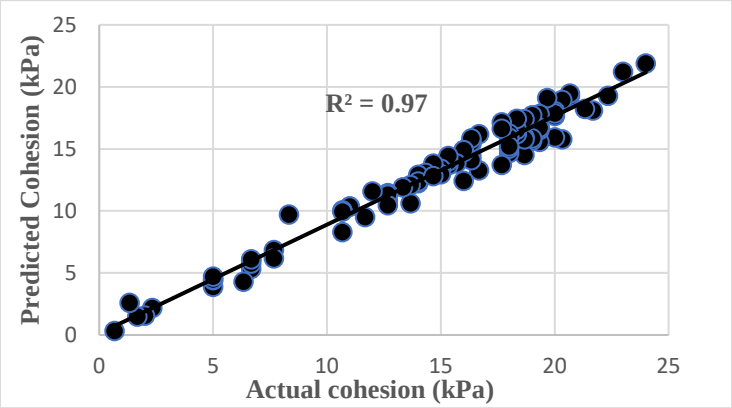


Figure 3-13. Comparison of Cohesion Values Obtained by ANN and Laboratory Tests

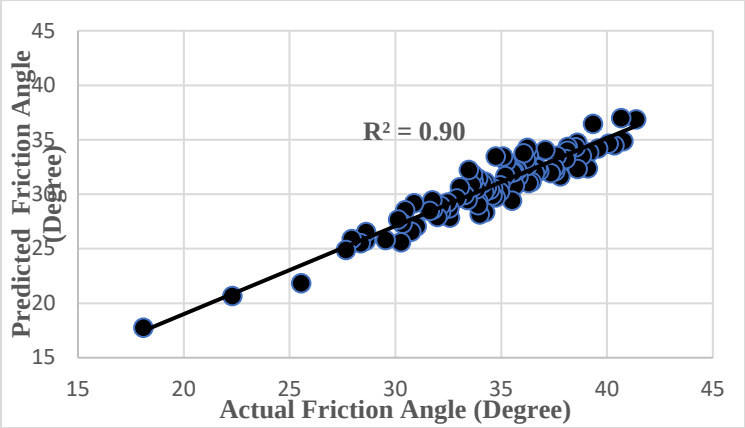


Figure 3-14. Comparison of Friction Angle Values Obtained by ANN and Laboratory Tests (Source: My Own Edit)

4- Theses

In this chapter, the main achievements of this research have been highlighted and explained in a general context. To provide a clearer understanding, some of the visualized graphs and plots presented in previous chapters as well as some new ones are revisited here.

Thesis 1

It is proven that artificial neural network (ANN) models can estimate the most important hydraulic and geotechnical parameters, hydraulic conductivity, cohesion and angle of internal friction, with an accuracy worth for preliminary phases of planning based on weight percentages of soil particles (clay, silt, sand and gravel).

Figures 3-7, 3-13 and 3-14 show the accuracy of my models. My ANN models which employed the feedforward method predict values with higher accuracy than the other known and published methods. Based on my analysis, the ANN model for hydraulic conductivity, which has two hidden layers with 1 and 3 neurons respectively, and the joint model for cohesion and internal friction angle, which has three hidden layers with 3, 2, and 4 neurons respectively as well as an output layer with 2 neurons, demonstrated the highest accuracy.

My innovative programs utilize artificial neural networks (ANNs) to predict soil parameters, employing a specific function for neuron activation. The designed Programs for predicting both hydraulic conductivity and shear strength parameters follow the below fundamental mathematical concepts.

$$y_k = tanh\left(\sum_{j=1}^{m} w_{kj} a_j + b_k\right)$$

$$a_j = tanh\left(\sum_{i=1}^{n} w_{ji} x_i + b_j\right)$$

Where:

- y_k is the output of the k-th neuron in the output layer,
- a_j is the output of the j-th neuron in the hidden layer, computed using the tanh activation function,
- w_{ji} is the weight connecting the i-th input to the j-th neuron,
- x_i is the i-th input to the network,
- b_j is the bias of the j-th neuron,
- w_{kj} is the weight connecting the j-th neuron in the hidden layer to the k-th neuron in the output layer,
- and b_k is the bias of the k-th neuron in the output layer.

Thesis 2

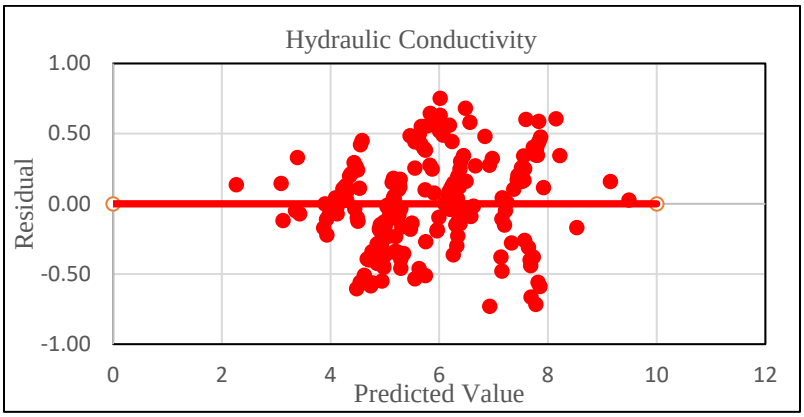
The accuracy of the ANN model I developed and suggested for use is independent of grain size, unlike most other models and developed methods that are typically suitable only for fine-grain or coarse-grain soils.

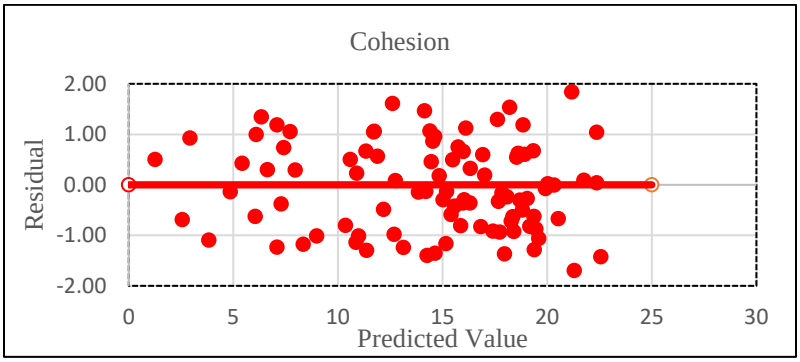
My ANN model has been designed to effectively predict soil parameters regardless of whether the soil is predominantly fine-grained or coarse-grained. This capability ensures that my model provides reliable results across a wide range of soil types. Importantly, my approach simplifies the initial assessment of site soils for preliminary design by focusing solely on essential soil components. This streamlined methodology not only enhances efficiency but also ensures that my model can swiftly provide reliable insights into soil behavior across diverse environmental conditions. Thus, my ANN-based approach represents a practical and effective tool for engineers and researchers seeking rapid and insightful soil assessments during early project phases.

Input Data	Output Data
Weight Percentage of Clay :	Hydraulic conductivity:
Weight Percentage of Silt:	Cohesion:
Weight Percentage of Sand :	Internal friction Angle:
Weight Percentage of Gravel :	

Figure 4-1. Simplified Input and Output box of my Program

The scatter plots for hydraulic conductivity, cohesion, and internal friction angle display residuals that are randomly distributed around the horizontal axis. This random distribution indicates that the residuals do not show any systematic pattern or trend. In other words, the errors in the predictions do not vary with the magnitude of the predicted values. This lack of systematic bias across the predicted values suggests that the model's performance is consistent and unbiased across different samples. The absence of such patterns confirms that the ANN model is robust, and its predictions are not dependent on the grain size of the soil.





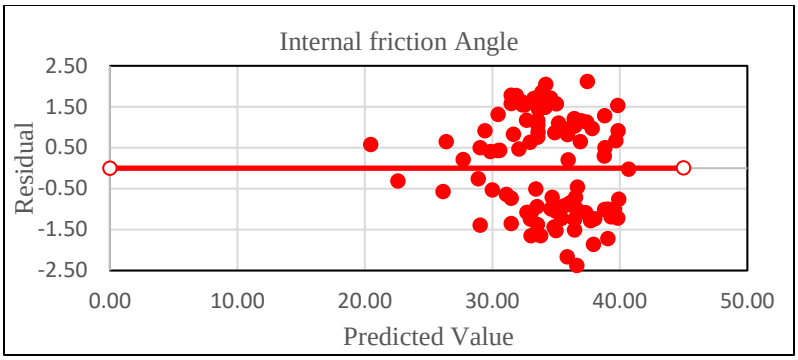


Figure 4-2. Plot of Residuals vs. Predicted Values (Source: My Own Edit).

Thesis 3

It was proven that artificial neural network method provides more reliable and accurate predicted hydraulic conductivity values independently form grainsize distribution than multiple linear regression or any known empirical correlation or pedotransfer function.

Considering figures 3-1, 3-2 and 3-7 and the results showed that artificial neural network model is more reliable than the others as it presented higher accuracy. I utilized laboratory measurements to evaluate the performance of various methods and models in predicting hydraulic conductivity. The indirect method, specifically the Carrier and Beckman approach, showed a moderate correlation with an R^2 value of 0.5187. However, this method consistently underestimated hydraulic conductivity values, highlighting the need for refinement or complementary methods. Multiple linear regression analysis demonstrated a better fit with an R^2 value of 0.81, indicating that 82% of the variability in the real data was explained by the model. My artificial neural network (ANN) models achieved superior accuracy with an R^2 value of 0.92 and an RMSE of 0.0768. This high level of accuracy underscores the effectiveness of ANNs in capturing the underlying patterns in the data.

Thesis 4

It is statistically proven by using the artificial neural network model that the extremities in grainsize (gravel and clay content) indicate a higher effect on the hydraulic conductivity of the material than weight percentages of silt and sand.

By examining the effect of each parameter in the compositions and the amount of hydraulic conductivity changes, as shown in figure 3-6, it becomes evident that clay and gravel exhibit a considerable slope as well as less deviation, indicating their substantial influence. In contrast, silt and sand do not display a noticeable slope, suggesting their lesser impact on the hydraulic gradient parameter. This highlights the critical importance of gravel and clay, where even minor variations significantly affect the hydraulic conductivity predictions. To plot the figures, the data was sorted in descending order to help understand the performance of the model; this means that as the number of tests increases, the values of -Log K drop.

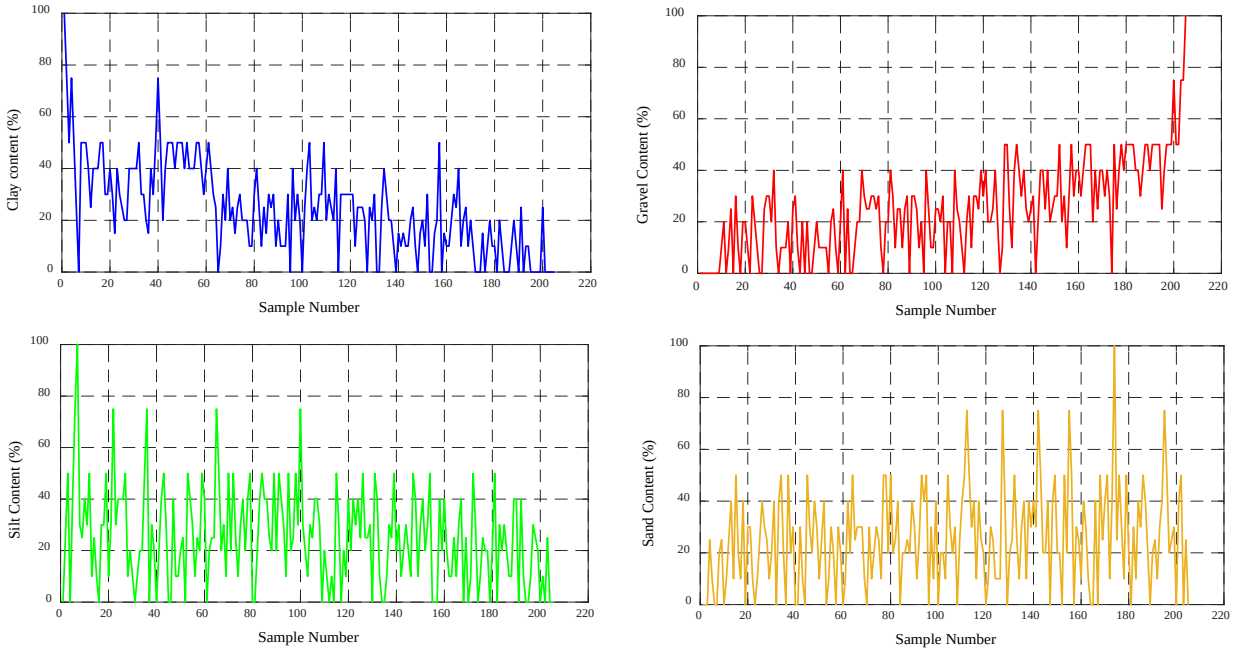


Figure 4-3. Effect of Each Component on Output (Source: My Own Edit).

Thesis 5

The simultaneous ANN model of cohesion and internal friction proved that contrary to separate ANN models, it achieved statistically significant improvements in predictive performance and accuracy as shown in figures 3-13 and 3-14.

The joint approach not only delivered higher accuracy but also demonstrated greater consistency in predictions. This model effectively captures the complex interactions between cohesion and internal friction. Unlike traditional models that handle each output independently, my joint model captures interdependencies and correlations between outputs, enhancing predictive accuracy and computational efficiency. The design of a joint model involves configuring the neural network architecture with multiple output neurons corresponding to each variable of interest and utilizing a multi-output loss function during training to optimize predictions across all outputs simultaneously. The decision to opt for a joint model was primarily driven by the understanding that cohesion and friction angle exhibit correlated behavior in soil mechanics. By having multiple neurons in the output layer, the neural network learned to optimize weights and biases across both outputs simultaneously. The shared input features (Weight percentage of clay, silt, sand and gravel) are processed through hidden layers that extract relevant patterns and relationships. These hidden layers act as shared representations, contributing to the prediction of both shear parameters and leveraging correlations between them. The network architecture comprises hidden layers with sizes [3, 2, 4], utilizing hyperbolic tangent sigmoid ('tansig') functions for activation in the hidden layers and a linear ('purelin') function for the output layer. Comparison of Simultaneous Model and Separate Models statistically demonstrated higher accuracy as mentioned in table5-1.

Table 4-1. Comparison of Simultaneous Model and Separate Models

Model	Year	Method	RMSE-c	RMSE- φ	R^2-c	R^2 - φ
Simultaneous Model	2024	ANN	0.41	0.25	0.97	0.90
Separate Model- φ	2024	ANN	-	0.39		0.82
Separate Model- c	2024	ANN	0.53	-	0.81	

List of research related to the dissertation

Eteraf, H., Kovács, B., & Mikita, V. (2024). Advanced Approach Integrating Methods to Estimate Hydraulic Conductivity of Different Soil Types with Support from a Machine Learning Model, Open Geoscience Journal. De Gruyter, ISSN: 2391-5447. (submitted)

Eteraf, H., Kovács, B., & Mikita, V. (2024). Enhanced Methodology for Estimating Soil Shear Strength Parameters Across Varied Soil Types Using Machine Learning Integration, Open Geoscience Journal. De Gruyter, ISSN: 2391-5447. (Submitted)

Eteraf, H., Kovacs, B., Mikita, V., & Delshad, Z. (2023). Effect of particle size distribution on shear strength of soil. In MATEC Web of Conferences (Vol. 385, p. 01043). EDP Sciences. DOI: 10.1051/matecconf/202338501043 (Published)

Eteraf, H., Mosallaei, A., Kovács, B., & Mikita, V. (2020). Estimation of hydraulic conductivity of soil from grain size distribution: A Review. Muszaki Foldtudomanyi Kozlemenyek, 89(2), 316-326. (Published)

Eteraf, H., Mosallaei, A., Kovács, B., & Mikita, V. (2022). Effect of Water Content on Consolidation Parameters. Research Developments in Geotechnics, Geo-Informatics and Remote Sensing, 161–163. https://doi.org/10.1007/978-3-030-72896-0_35 (Published)

Eteraf, H., Kovács, B., & Mikita, V. (2019). Effect of the compaction energy on pre-consolidation stress and consolidation parameters of clayey soil-experimental study. In Szigyártó, I.-L., & Szikszai, A. (Eds.), XV. Kárpát-medencei Környezettudományi Konferencia (pp. 199-210). Cluj-Napoca, Romania: Ábel Kiadó. (Kárpát-Medencei Környezettudományi Konferencia, ISSN 1842-9815). Cluj-Napoca: Sapientia Hungarian University of Transylvania. (Published)

Eteraf, H., Madarász, T., Mosallaei, A., Kovács, B., & Mikita, V. (2022). A Review of the Role of Geosynthetic Clay Liners in Contaminated Site Remediation. Geosciences and Engineering, 10(15), 80–95. https://doi.org/10.33030/geosciences.2022.15.080 (Published)

Mikita, V., Kovács, B., Czinkota, I., **Eteraf, H.**, & Pinjung, Z. (2022). The Investigation of the Stress Field Evolution in Soil Samples During Vertical Loads with Computer Tomography. Research Developments in Geotechnics, Geo-Informatics and Remote Sensing, 123–126. https://doi.org/10.1007/978-3-030-72896-0_27 (Published)

References

A. Prieto, B. P. (2016). Neural Networks: An Overview of Early Research, Current Frameworks and New Challenges. Neurocomputing, 242-268. doi:https://doi.org/10.1016/j.neucom.2016.06.014

A. Nemes, W. J. (2004). Soil Texture and Particle-Size Distribution as Input to Estimate Soil Hydraulic Properties. In Developments in Soil Science (Vol. 30, pp. 47-70).

Abdolrasol, M. G. (2021). Artificial neural networks based optimization techniques: A review. Electronics, 143. doi:https://doi.org/10.3390/electronics10212689

Ameratunga, N. S. (2016). Correlations of Soil and Rock Properties in Geotechnical Engineering. doi:https://doi.org/10.1007/978-81-322-2629-1

B. D. Klein, D. R. (1999). Data Quality in Linear Regression Models: Effect of Errors in Test Data and Errors in Training Data on Predictive Accuracy. Informing Science, 33. doi:http://dx.doi.org/10.28945/599

B. M. Das, N. S. (2017). Fundamentals of Geotechnical Engineering. Cengage Learning.

Citakoglu, H. (2017). Comparison of Artificial Intelligence Techniques for Prediction of Soil Temperatures in Turkey. Theoretical and Applied Climatology, 545-556. doi:https://link.springer.com/article/10.1007/s00704-016-1914-7

Craig, R. F. (2004). Craig's Soil Mechanics. CRC Press. doi:https://doi.org/10.4324/9780203494103

F. Meskini-Vishkaee, N. D. (2018). Evaluation of Different Predictor Models for Detailed Soil Particle-Size Distribution. Pedosphere, 28(1), 157-164.

Nathans, L. L. (2012). Interpreting multiple linear regression: A guidebook of variable importance. Practical Assessment, Research & Evaluation.

P. Y. Hicher, A. M. (1996). Identifying soil parameters by means of laboratory and in situ testing. Computers and Geotechnics, 153-170. doi:https://doi.org/10.1016/0266-352X(96)00040-7

Q. Gao, M. H. (2024). Geophysical Estimation of 2D Hydraulic Conductivity for Groundwater Assessment in Hard Rock. Acta Geophysica, 1-12. doi:https://doi.org/10.1007/s11600-024-01310-w

S. Dołęgowska, A. G. (2016). Heterogeneous Areas—Identification of Outliers and Calculation of Soil Sampling Uncertainty Using the Modified RANOVA Method. Environmental Monitoring and Assessment, 188, 1-11. doi:https://doi.org/10.1007%2Fs10661-016-5584-9

S. K. Gupta, M. K. (2024). A Comprehensive Review of the Impacts of Chemicals, Soil Salinity, Organic Matter, and Land Use. IOP Conference Series: Earth and Environmental Science (p. 012032). IOP Publishing.

S. Nam, M. G. (2011). Determination of the Shear Strength of Unsaturated Soils Using the Multistage Direct Shear Test. Engineering Geology, 122(3-4), 272-280. doi:https://doi.org/10.1016/j.enggeo.2011.06.003

S. Sharma, S. A. (2021). A Survey on Applications of Artificial Intelligence for Pre-Parametric Project Cost and Soil Shear-Strength Estimation in Construction and Geotechnical Engineering. Sensors, 21(2), 463. doi:https://doi.org/10.3390/s21020463

Shakiba, F. M. (2020). Novel analog implementation of a hyperbolic tangent neuron in artificial neural networks. IEEE Transactions on Industrial Electronics, 10856-10867.

Uyanık, G. K. (2013). A study on multiple linear regression analysis. Procedia - Social and Behavioral Sciences, 234-240.

V. Bolón-Canedo, D. P.-B.-B.-B.-M. (2011). Scalability Analysis of ANN Training Algorithms with Feature Selection. Conference of the Spanish Association for Artificial Intelligence (pp. 84-93). Berlin, Heidelberg: Springer Berlin Heidelberg.

Vanapalli, S. K. (November 2009). Shear Strength of Unsaturated Soils and Its Applications in Geotechnical Engineering Practice. 4th Asia-Pacific Conference on Unsaturated Soils, (pp. 579-598). New Castle, Australia.

Weidong Hu, K. L. (2020). Active Earth Pressure against Rigid Retaining Walls for Finite Soils in Sloping Condition considering Shear Stress and Soil Arching. Advances in Civil Engineering. doi:https://doi.org/10.1155/2020/6791301

Z. Han, J. L. (2020). Determining the Shear Strength and Permeability of Soils for Engineering of New Paddy Field Construction in a Hilly Mountainous Region of Southwestern China. International Journal of Environmental Research and Public Health, 17(5), 1555.

Zou, J. H. (2009). Overview of artificial neural networks. In Artificial Neural Networks: Methods and Applications. 14-22. doi:https://doi.org/10.1007/978-1-60327-101-1_2