

Article

Novel Insights in Soil Mechanics: Integrating Experimental Investigation with Machine Learning for Unconfined Compression Parameter Prediction of Expansive Soil

Ammar Alnmr ^{1,*} , Haidar Hosamo Hosamo ² , Chuangxin Lyu ², Richard Paul Ray ¹ 
and Mounzer Omran Alzawi ³

¹ Department of Structural and Geotechnical Engineering, Széchenyi István University, 9026 Győr, Hungary; ray@sze.hu

² Department of Built Environment, OsloMet—Oslo Metropolitan University, 0176 Oslo, Norway; haidarho@oslomet.no (H.H.H.); chuangxin.lyu@oslomet.no (C.L.)

³ Department of Geotechnical Engineering, Tishreen University, Latakia P.O. Box 2230, Syria; momran1@scs-net.org

* Correspondence: alnmr.ammar@hallgato.sze.hu

Abstract: This paper presents a novel application of machine learning models to clarify the intricate behaviors of expansive soils, focusing on the impact of sand content, saturation level, and dry density. Departing from conventional methods, this research utilizes a data-centric approach, employing a suite of sophisticated machine learning models to predict soil properties with remarkable precision. The inclusion of a 30% sand mixture is identified as a critical threshold for optimizing soil strength and stiffness, a finding that underscores the transformative potential of sand amendment in soil engineering. In a significant advancement, the study benchmarks the predictive power of several models including extreme gradient boosting (XGBoost), gradient boosting regression (GBR), random forest regression (RFR), decision tree regression (DTR), support vector regression (SVR), symbolic regression (SR), and artificial neural networks (ANNs and proposed ANN-GMDH). Symbolic regression equations have been developed to predict the elasticity modulus and unconfined compressive strength of the investigated expansive soil. Despite the complex behaviors of expansive soil, the trained models allow for optimally predicting the values of unconfined compressive parameters. As a result, this paper provides for the first time a reliable and simply applicable approach for estimating the unconfined compressive parameters of expansive soils. The proposed ANN-GMDH model emerges as the pre-eminent model, demonstrating exceptional accuracy with the best metrics. These results not only highlight the ANN's superior performance but also mark this study as a groundbreaking endeavor in the application of machine learning to soil behavior prediction, setting a new benchmark in the field.

Keywords: expansive soils; sand content; machine learning innovation; predictive modeling; soil property optimization; artificial neural networks



Citation: Alnmr, A.; Hosamo, H.H.; Lyu, C.; Ray, R.P.; Alzawi, M.O. Novel Insights in Soil Mechanics: Integrating Experimental Investigation with Machine Learning for Unconfined Compression Parameter Prediction of Expansive Soil. *Appl. Sci.* **2024**, *14*, 4819. <https://doi.org/10.3390/app14114819>

Academic Editor: Wei Gao

Received: 30 April 2024

Revised: 30 May 2024

Accepted: 31 May 2024

Published: 2 June 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Understanding unconfined compression parameters (q_u and E_t) is critical in geotechnical engineering. These parameters aid in determining soil cohesion (c), which is required for calculating cohesive soil bearing capacity. Furthermore, the elasticity modulus (E_t) is important in predicting settlements. To perform the unconfined compression test, a soil sample must be collected from the field, which requires advanced skills and extensive experience. However, performing these tests in a laboratory is known to be both tedious and time-consuming [1]. To overcome these challenges, researchers investigated the relationships between unconfined compression parameters and physical soil properties. This correlation is useful for reducing the number of tests and the overall investigation costs.

Several studies have been conducted over the last decade to establish relationships between soil's mechanical and physical properties. This approach, aimed at saving time and money during investigations, is based on previous research in soil mechanics and foundation engineering [1,2]. Preliminary investigations and studies for any project can be conducted using predictive models and correlations in conjunction with relevant information, making the process of obtaining design parameters simpler. Machine learning beats traditional lab tests and basic regression models in unconfined compression testing because of its automation, speed, and cost-effectiveness. Machine learning reduces the need for tedious manual testing by rapidly analyzing soil data and predicting compression parameters, leading to faster results and reduced expenses.

Expansive soils, which cover large areas of land worldwide, present a significant challenge in the construction of structures and roads on them due to the high cost associated with repairing damage caused by expansive soil shrink–swell cycles, including building and road cracks and collapse, costing billions of dollars annually [3–6]. Nelson and Miller [7] assert that expansive soils may result in greater financial losses than earthquakes or floods. In addressing this challenge, the behavior of expansive soils has been studied in numerous studies by adding different materials to improve their characteristics [8–15]. Although these studies have shown that soil characteristics can be improved via the addition of materials, many unresolved questions remain. Therefore, the main purpose of this study is to investigate the effects of sand content on the unconfined compression test characteristics of expansive soil. This study evaluates the impact of varied percentages of sand, initial saturation levels, and dry unit weights on expansive soil behavior. Additionally, it aims to build a dataset for predicting unconfined compression test parameters.

1.1. Sand as an Additive Material and Its Impact on Soil Parameters

Sand is a granular material of natural origin with high strength and is often utilized as a filler material in cohesive soil mixtures to alter their plasticity, compaction, and strength properties [9,13,16–18]. The sand fraction significantly influences the cracking behavior of expansive soil. As the sand content increases, it can moderate the volume fluctuations caused by soil swelling and shrinking, minimizing the amount of cracking. However, an excessive sand component may modify the soil's mechanical properties, thereby influencing its overall behavior. Balancing the sand fraction is critical in engineering applications because it improves soil performance and reduces the negative impacts of cracking in expansive soil [19]. The addition of sand to expansive soils has been extensively investigated by many researchers to determine its effect on the soil parameters [10,20–26]. Previous studies have confirmed that increasing the percentage of sand increases the unconfined compressive strength (UCS) of expansive soils. Khan et al. [27] performed unconfined compressive experiments on expansive soils. They found that the compressive strength increases with increasing dry density. Increased water content, on the other hand, shows the opposite trend. It is worth noting, however, that their study only looked at two levels of added sand (20% and 40%). Nagaraj [10] performed unconfined compression tests to investigate the impact of sand on the unconfined compressive strength (UCS) of expansive soil, using a static compactor to prepare samples at moisture levels determined by the proctor test only. Their findings showed that, up to a certain point, the unconfined compression strength increased with the percentage of sand. But as the sand percentage increased further, the strength decreased. Mallikarjuna Rao and Subba Rao [28] also investigated the effect of sand on the unconfined compressive strength (UCS) of expansive soil. Nonetheless, their research was limited to samples prepared with the maximum dry density and optimum moisture content with only three percentages of added sand (25%, 40%, and 70%). Sometimes, disparities in results emerged across studies, as can be seen in Nagaraj [10] versus Mallikarjuna Rao and Subba Rao [28]. The primary cause of this variation was significant differences in the percentages of added sand used in each study. Nagaraj [10], for example, used percentages ranging from 10% to 80%, whereas Mallikarjuna Rao and Subba Rao [28] used 25%, 40%, and 70%. Furthermore, the existing

sand content in the expansive soil was not taken into consideration in some studies. It is best to remove all pre-existing sand content from the expansive soil to ensure consistency before conducting experiments and adding different percentages of sand.

It is worth noting that saturated soil has all its voids entirely filled with water. The soil is referred to as partially saturated or unsaturated soil when the voids are partially filled with water. The Soil–Water Characteristic Curve (SWCC) is a fundamental relationship between the content of water and soil suction [29] that effectively describes the behavior of unsaturated soils. The retention behavior of soil water can be influenced by various factors, including dry unit weight, degree of saturation, stress conditions, and soil amendments [26,30,31]. The degree of saturation, moisture content, and volumetric water content are terms commonly used to describe the amount of water in the soil. As moisture content in the soil changes from full saturation to drought, the wet area of contact between grains or agglomerates decreases as suction stresses in the soil increase. Put differently, there is a correlation between the SWCC and the shear strength of unsaturated soils [29,32,33]. However, obtaining this curve requires special equipment, effort, and time, which can lead to inconsistent results. As a simpler and more precise alternative, this study focuses on the degree of saturation to predict the elasticity modulus and strength of expansive soil, without considering the characteristic curve.

Prior experimental investigations into the impact of sand on unconfined compression tests were constrained by a lack of comprehensive assessments that considered all factors at the same time. Sand percentage, initial dry density, and initial degree of saturation were frequently studied separately. Some studies varied sand proportions, whereas others focused solely on the Proctor test's maximum dry density and optimal moisture content. This research seeks to overcome these constraints to gain a more comprehensive understanding of unconfined compression experiments and the potential for accurate predictive modeling.

1.2. Artificial Intelligence as a Predictive Tool in Geotechnical Engineering

The flexible and precise data-driven approach of artificial intelligence (AI) has gained increasing attention as computer technology advances. Many scholars have tried to integrate machine learning (ML) and geotechnical reliability analysis because of the complexity of expansive soils to improve computational precision and efficacy. This endeavor has resulted in a number of successful applications [34–38]. Within geotechnical reliability research, the major goal of machine learning (ML) is to recreate complicated, high-dimensional implicit performance functions by exploiting insights from carefully selected data. These datasets are primarily made up of input factors such as plastic index, liquid limit, degree of saturation, sand content, and unit weight. Similarly, relevant parameters of interest, such as unconfined compressive strength and elasticity modulus, are included in the datasets. Using this methodology to implement the ML analysis model provides a powerful method for accurately predicting outcomes in the geotechnical sector. Via training and validation processes, the model's accuracy and computational efficiency are validated, confirming its accuracy and computational efficiency in meeting particular performance requirements.

Effective applications of various machine learning (ML) models have been found in geotechnical analysis. These models cover a broad range, including support vector machines (SVMs) [39], multivariate adaptive regression splines (MARS) [40,41], relevant vector machines (RVMs) [42], decision tree regression (DTR) [43], gradient boosting regression (GBR) [43], K nearest neighbor regression (KNN) [43], particle swarm optimization (PSO) [44], random forest regression (RFR) [43], extreme gradient boosting (XGBoost) [38,45–47], extreme learning machines (ELMs) [40], symbolic regression (SR) [48,49], and artificial neural networks (ANNs) [50–54]. Table 1 summarizes several studies that employed machine learning in geotechnical engineering.

Table 1. Overview of machine learning models used in the geotechnical sector.

Reference	Model	Description
Díaz 2023 [55]	GBR, RFR, DTR, ET, Linear R, LASSO, EN, and others	The current study predicts the UCS of jet grouting columns using supervised machine learning algorithms. The study concludes that DTR, GBR, RFR, and extremely randomized trees (ET) are the top algorithms in terms of performance.
Kumar 2019 [40]	MARS and ELM	The study focused on predicting pile-bearing capacity in sandy soil using machine learning algorithms (ELM and MARS). In terms of performance measures, the MARS model outperformed the ELM model significantly.
Kardani 2022 [43]	GBR, DTR, KNR, RFR, ST-ENSM, BG-ENSM, and others	The study evaluated machine learning models for estimating the resilient modulus (M_R) of compacted subgrade soils. Bagging ensemble (BG-ENSM) ranked first, while RFR ranked last. Stacking ensemble (ST-ENSM) and DTR also ranked fairly well.
Wu 2023 [56]	RFR, SVR, and ANN	In this study, SVR, RFR, and ANN models were trained on coral reef limestone (CRL) samples using X-ray CT and lab data. On test data, the ANN model performed admirably, predicting UCS with 92% accuracy.
Zhang 2022 [45]	XGBoost and GBR	The study presented an effective use of XGBoost and GBR in time-varying reliability analysis. The probabilities of a landslide failing at particular times were computed by these models effectively and precisely. While both models accurately predicted outcomes, XGBoost exhibited marginally superior performance.
Chang 2023 [57]	RFR and ANN	The multi-scale segmentation method was used in this study to identify slope units in Chongyi County, China. The various characteristics of these units were used to construct various slope-machine learning models. In terms of predictive accuracy, RF models outperform ANN models.
Wang 2021 [36]	LR, SVM, RFR, GBR, and ANN	This study provided a new machine learning method for identifying natural-terrain landslides. It revealed that ANN excelled because of its superior feature extraction and data processing capabilities. Boosting algorithms rank second in accuracy after ANN, followed by RFR, LR, and SVM.

The symbolic regression (SR) algorithm is used effectively in many areas of civil engineering. It adeptly models the Colebrook equation governing hydraulic flow friction in hydraulics, significantly improving accuracy in turbulent scenarios [58]. Symbolic regression is used in structural engineering to estimate seismic peak drift ratio, penetration depth into concrete blocks, shear capacity of concrete beams reinforced with steel fibers, fire response of concrete structures, seismic fragility analysis, remaining fatigue life, and shear resistance in bolted connections [59–64]. Notably, in some cases [59,63,64], symbolic regression equations outperformed traditional formulas.

Geotechnical engineers are becoming interested in ANNs because of their exceptional capacity to represent nonlinear problems with many different parameters. ANNs can create non-linear relationships between variables and make relatively accurate predictions [65]. Several geotechnical investigations have demonstrated that ANN models outperform multiple linear regression (MLR) models [51,65–67].

Table 2 summarizes several studies that used machine learning to predict the behavior of expansive soils, including some of the most used regression algorithms such as GBR, SVR, and DT. According to Tables 1 and 2, investigating the best behavior prediction remains a difficult task because there is no general agreement on the best algorithm for predicting the behavior of expansive soil. The following machine learning models that have demonstrated the best performance according to Table 2 will be used in this paper: artificial neural networks (ANNs and the proposed ANN-GMDH), extreme gradient boosting (XGBoost), gradient boosting regression (GBR), random forest regression (RFR), decision tree regression (DTR), and support vector regression (SVR). The outcomes of those models

will be compared, and the most accurate one will be used for prediction. Additionally, symbolic regression (SR) will be used to formulate the prediction equations for E_t and q_u .

Table 2. Overview of machine learning models used in the literature to predict expansive soil behavior.

Reference	Model	Description
Najjar 1996; Najjar 1998 [68,69]	ANN	These papers focus on the development of statistical models and neural network-based models to predict swelling potential. It has been demonstrated that neural models outperform statistical models in terms of prediction accuracy.
Doris 2008 [70]	ANN	This study employs artificial neural networks (ANNs) to forecast soil surface movements caused by changes in moisture caused by climate variations. Based on climatic changes, ANN predictions match field measurements of soil movements caused by moisture changes as well as a physics-based method based on soil water content measurements.
Merouane 2018; Ikizler 2010; Ashayeri 2009; Erzin 2011 [50,66,71,72]	ANN	ANNs were used to predict the swelling pressure and amplitude of unsaturated clay
Dutta 2019 [51]	ANN	ANNs were used to predict the free swell index (<i>FSI</i>) of the expansive soil. The ANN model predicted the <i>FSI</i> more accurately than the multiple linear regression model.
Ikeagwuani 2021 [73]	MARS, RFR, and GBR	Three distinct machine learning models were employed in this study to predict the California Bearing Ratio (<i>CBR</i>) of an expansive soil subgrade treated with regular Portland cement, quarry dust, and sawdust ash. The findings of the study revealed that the random forest model surpassed both the gradient boosting machine and MARS models in terms of predictive accuracy.
Eyo 2022 [74]	ANN, SVR, LR, RFR, and others	This work used a broad range of machine learning techniques to evaluate and estimate the swelling of soils with different plastic characteristics under expanding behavior during inundation. According to the findings, the support vector machine (SVR) outperformed other techniques, including random decision forests (RFR), logistic regressors (LRs), and artificial neural networks (ANNs).
Benbouras 2021 [75]	SVR, RFR, ANN, DNN, LASSO, ELM, and others	This study used many advanced algorithms to predict the swelling index of expansive soil. The results show that the random forest method yielded more effective and accurate results than the other methods in modeling geotechnical phenomena.
Chen 2022 [76]	RFR, XGBoost, SG, ANN, SVR, and MARS	This work developed a prediction model of the uniaxial tensile strength using three ensemble machine learning approaches and three classical methods. Based on all metrics, the results indicate that the stacked generalization (SG) and XGBoost perform best, with the ANN, RF, MARS, and SVM following closely after.

1.3. Research Scope

A series of tests on expansive soils, including the Atterberg limits test, unconfined compressive test, and Proctor test, were carried out in this study, resulting in a dataset of 225 records with 5 input features and 2 labels. To develop individual predictive models for the unconfined compressive parameters, eight ML approaches were considered, including three classical ones (the proposed ANN-GMDH, ANN, and SVR), four ensemble ones (DTR, RFR, GBR, and XGBoost), and symbolic modeling (SR). The proposed ANN-GMDH is intended to improve the ANN's prediction accuracy by incorporating new relationships between input features. The suggested ML models' performances were carefully tested and compared.

This pioneering research marks a significant advance in the field of soil mechanics. To date, no research effort has systematically applied a comprehensive suite of advanced

machine learning models to analyze both q_u and E_t properties of expansive soil while accounting for the influence of sand. Our work fills this crucial gap by integrating advanced machine learning techniques with experimental data on expansive soil properties. The study's innovative approach provides a unique combination of classical and ensemble algorithms to analyze and predict the unconfined compression parameters of expansive soils with varying sand content, saturation levels, and dry densities.

By tapping into the vast potential of these diverse models, the research provides a nuanced understanding of how different variables influence soil behavior, a task that traditional methods have found challenging.

The study's novelty stems from the direct application of sophisticated machine learning techniques to experimental data, resulting in remarkably precise predictions of previously unattainable soil parameters. Unlike traditional methods, which rely on theoretical assumptions or simplified models, this novel methodology uses advanced methods to analyze raw experimental data. Through this, the study achieves remarkably precise predictions of soil parameters, outperforming previous methods. One important feature of this novel approach is its consideration of the percentage of sand in the soil, which has frequently been overlooked in previous research efforts. By incorporating this critical factor into the analysis, the predictive models developed in this study provide a more complete picture of soil behavior. The inclusion of sand percentage allows for more accurate predictions, filling a significant gap in the literature and increasing the findings' applicability to real-world scenarios. Furthermore, this study presents predictive models and equations designed to estimate the elasticity modulus (E_t) of expansive soil. Unlike previous studies, which may have focused solely on parameters such as unconfined compressive strength (q_u) or investigated different soil types, our study fills a gap by developing models specifically for expansive soils. Considering the unique properties and behaviors of expansive soils, which require tailored predictive models for precise assessment of elastic modulus and unconfined compressive strength, this study aims to contribute to the field of geotechnical engineering by addressing identified gaps in the literature and proposing innovative methodologies. The use of advanced machine learning techniques, combined with the consideration of previously overlooked factors such as sand percentage and the creation of specialized models for expansive soils, improves prediction accuracy and reliability. Finally, these advancements improve our understanding of soil behavior and have practical engineering applications, laying the groundwork for future field study.

2. Materials and Methods

2.1. Methodology Overview

Figure 1 depicts the overall workflow, which includes laboratory and machine learning predictions for unconfined compressive parameters (q_u and E_t). The process followed standard machine learning procedures, beginning with the collection of experimental data from 225 unconfined compression test samples in the laboratory. Two cross-validation techniques were used: the Train/Validation/Test Split and (k-fold) methods. For the (k-fold) technique, data were randomly divided into 80% for model training and 20% for testing, with $k = 5$. Meanwhile, the Train/Validation/Test Split approach randomly divided the data into three sets: 70% for model training, 10% for validation, and 20% for testing. Preprocessing entailed dimensionless normalization of input to speed up model convergence. Following model training, the mean absolute error (MAE), root-mean-square error (RMSE), and coefficient of determination (R^2) were used to assess machine learning model performance.

2.2. Laboratory Tests

The materials utilized in this study included expansive clayey soil sourced from Demsarkhu, Lattakia, Syria and fine sea sand. Figure 2 displays the case study location. For further details on the characteristics of these materials, refer to Alnmr and Ray [13]. The sand and clay were mixed in varying proportions, ranging from 10% to 50% relative

to the dry weight. All the laboratory tests were conducted according to ASTM. Granular gradient curves were plotted and presented in Figure 3. To ensure that the materials were free of impurities, the clay and sand samples were cleaned using sieve N200. The chemical composition of the expansive clay was determined and is presented in Table 3.

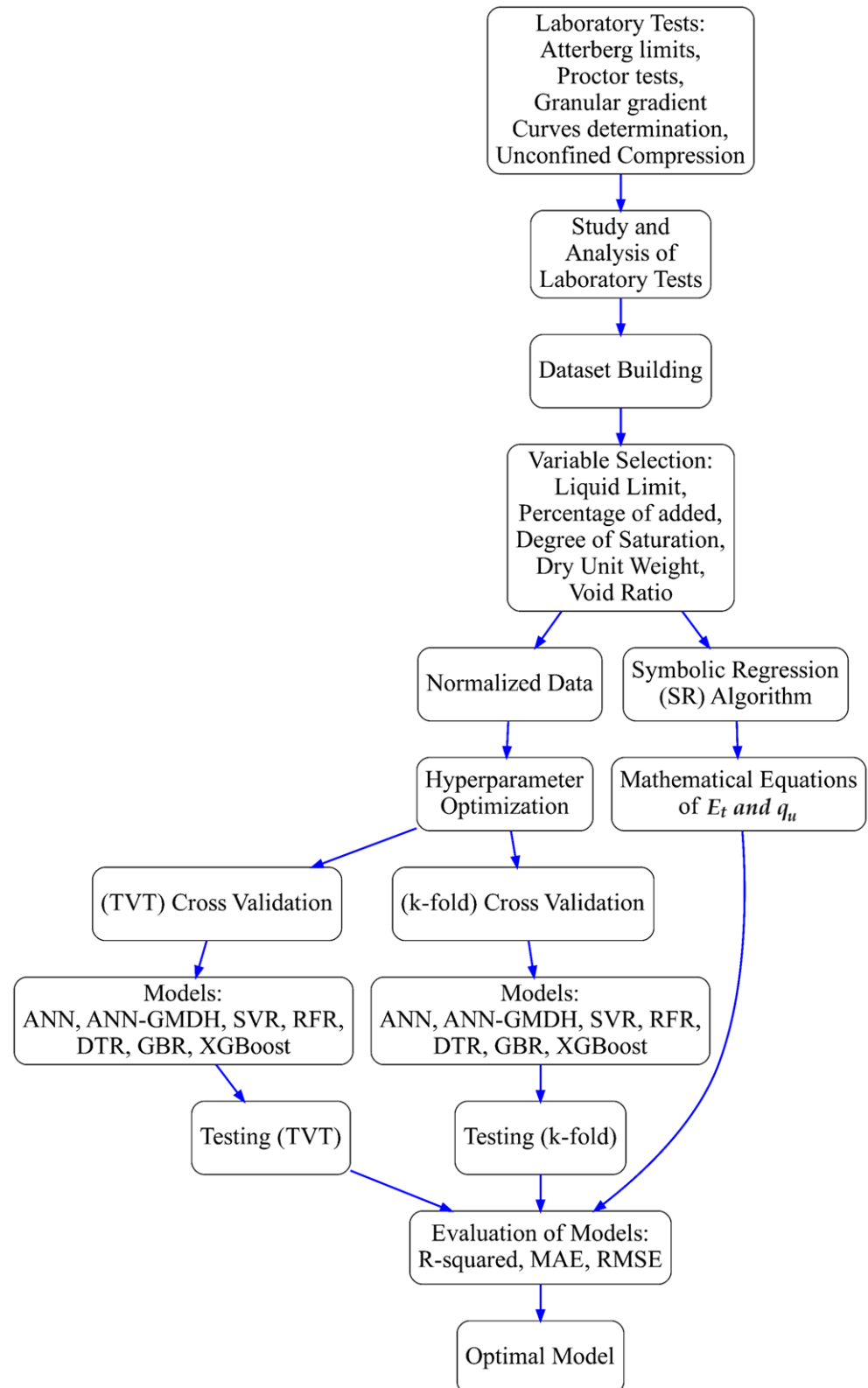


Figure 1. Machine learning model implementation workflow.



Figure 2. Case study location: Coordinates: 35°33′12.2″ N, 35°46′35.9″ E.

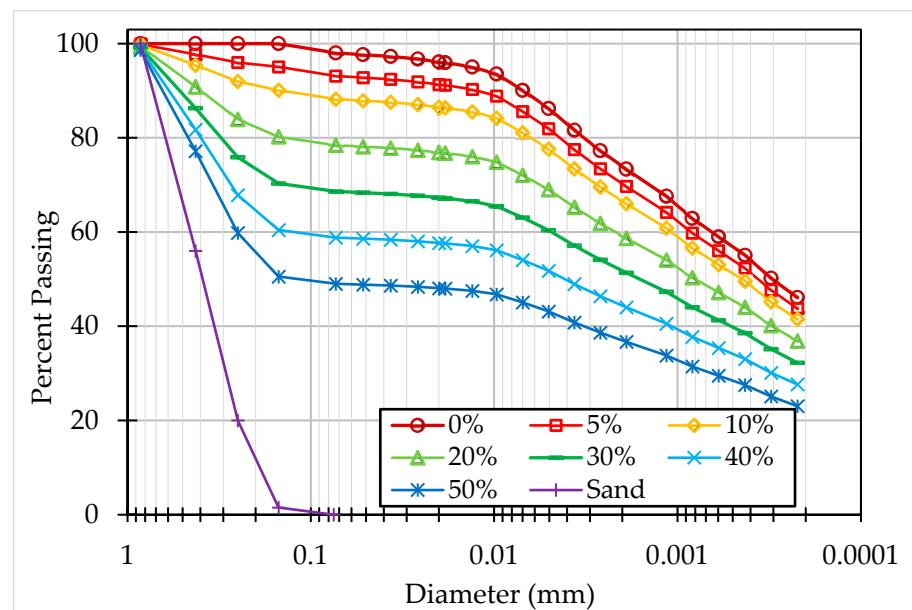


Figure 3. Variation in granular gradients across tested mixtures.

Table 3. Chemical composition of the expansive clay.

Chemical Composition	Alumina (Al_2O_3)	Silica (SiO_2)	Magnesium (MgO)	Ferric (Fe_2O_3)	Potassium (K_2O)	Calcium (CaO)	(Sodium (Na_2O))	Loss of Ignition (LoI)
%	11.51	49.79	2.41	5.49	0.37	12	1.2	17.23

According to the Unified Soil Classification System [77] and the AASHTO classification system [78], the expansive clay employed in this investigation was classified as CH and A-7-5, respectively.

Prior research [79,80] established a relationship between the Atterberg limits plot position on a plasticity chart and the soil's clay mineralogy. Plotting the liquid limit (LL) and plastic index (PI) for each percentage of sand (F_s) resulted in a plasticity chart (Figure 4) with ranges for kaolinite, montmorillonite, illite, and chlorite. The results indicated that the samples in this study with varying amounts of added sand are composed of illite.

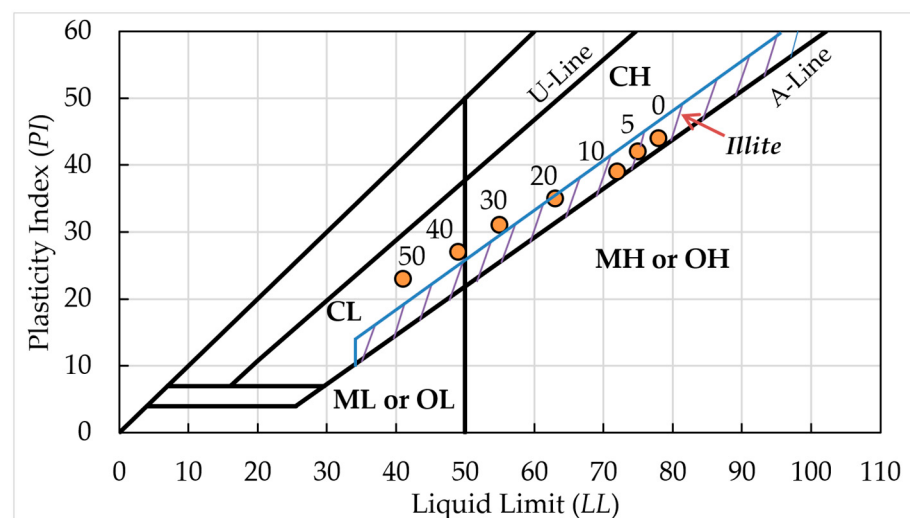


Figure 4. Effect of sand content on the Atterberg limits.

The standard proctor test was used to determine the optimal moisture content and maximum dry density of the sand–expansive clay mixtures (Figure 5). Refer to Alnmr and Ray [13] for detailed information on Proctor, Atterberg, and consolidation tests.

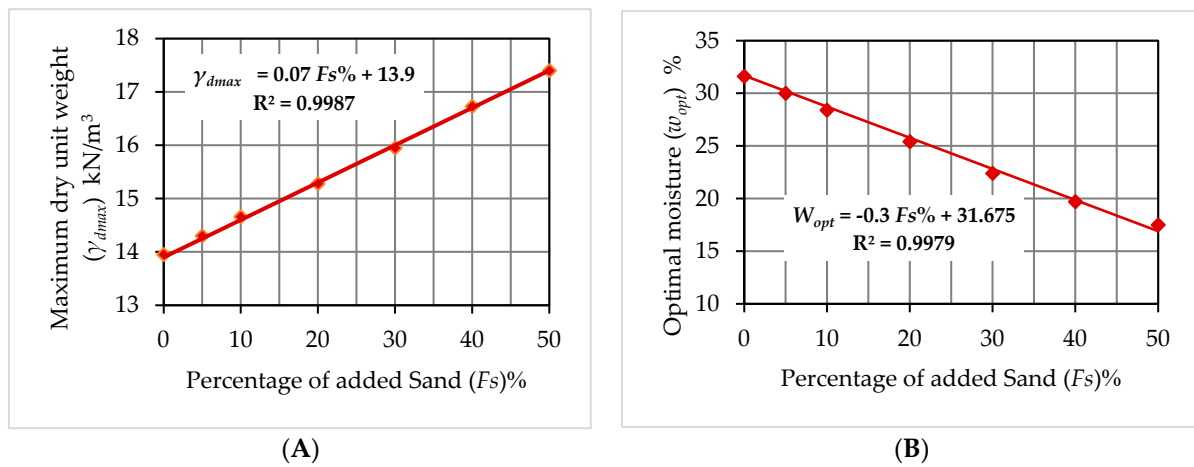


Figure 5. The maximum dry density and optimal moisture values as a function of the sand content; (A) the change in γ_{dmax} as a function of F_s , (B) the change in w_{opt} as a function of F_s .

In the current study, unconfined compression tests (UCTs) were performed on cylindrical samples that were twice their diameter in length. The tests were conducted at various densities and saturations for each sand percentage added, with a deformation rate of 2 mm/min.

Figure 6 depicts the step-by-step procedure used in the modified sample preparation at the laboratory. Initially, sand and clay soils were thoroughly mixed to create a homogeneous mixture (A). Following that, the blend was sealed in a plastic bag and left isolated for 24 h to ensure consistent moisture distribution (B). After that, the mixture was put into an unconfined compression test mold. The specimen was compressed to the required unit weight using a hydraulic jack and static pressure (C). After being prepared, the specimen was placed in the unconfined compression apparatus (D). The apparatus applied a vertical load until the specimen failed (E). Figure 7 shows the relationship between strain and vertical stress to clarify how E_t and q_u are defined.

A dataset comprising 225 unconfined compression tests was compiled from experimental outcomes for utilization in machine learning models. The input data consist of specific gravity (G), plastic index (PI), liquid limit (LL), dry unit weight (γ_d), sand content (F_s), degree of saturation (SR), optimum degree of saturation (SR_{opt}), maximum dry unit weight (γ_{dmax}), and void ratio (e_0). The dataset is used to predict the elasticity modulus (E_t) and unconfined compressive strength (q_u). Density plots and distribution histograms of the dataset are presented in Figure 8.

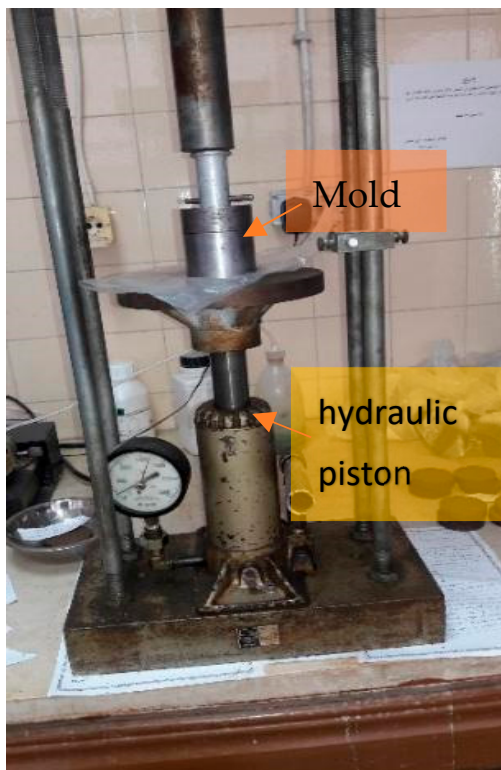


(A)

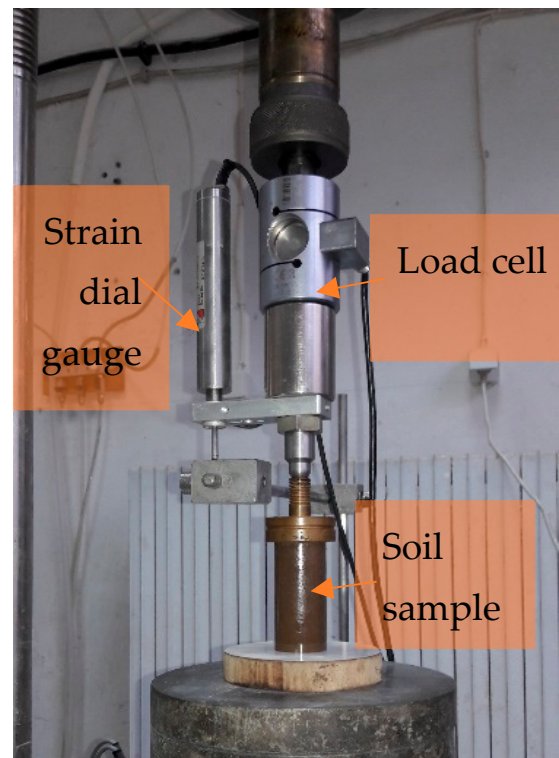


(B)

Figure 6. Cont.



(C)



(D)



(E)

Figure 6. Visualization of sample preparation process: (A) sand and expansive clay mixing, (B) sample encasement in plastic bags for moisture retention, (C) mold placement and application of static pressure via hydraulic piston, (D) final prepared specimen in unconfined compression apparatus, and (E) post-failure specimen.

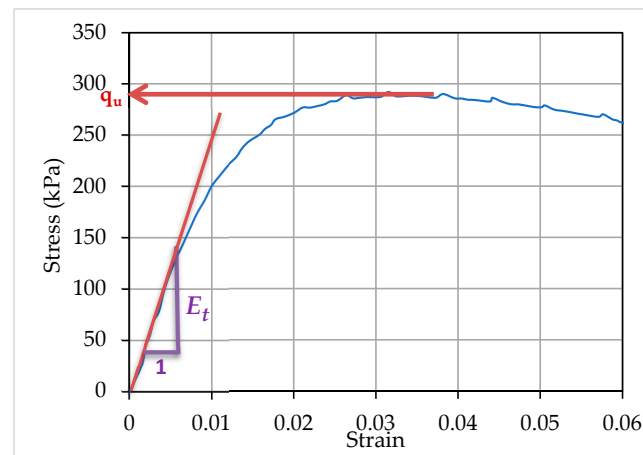


Figure 7. Relationship between the strain and vertical stress ($G = 2.69$; $e_0 = 0.758$; $F_s = 20\%$; $\gamma_d = \gamma_{dmax} = 15.3 \text{ kN/m}^3$; $SR = SR_{opt} = 88.7\%$; $LL = 63.3\%$; $PI = 35.6\%$).

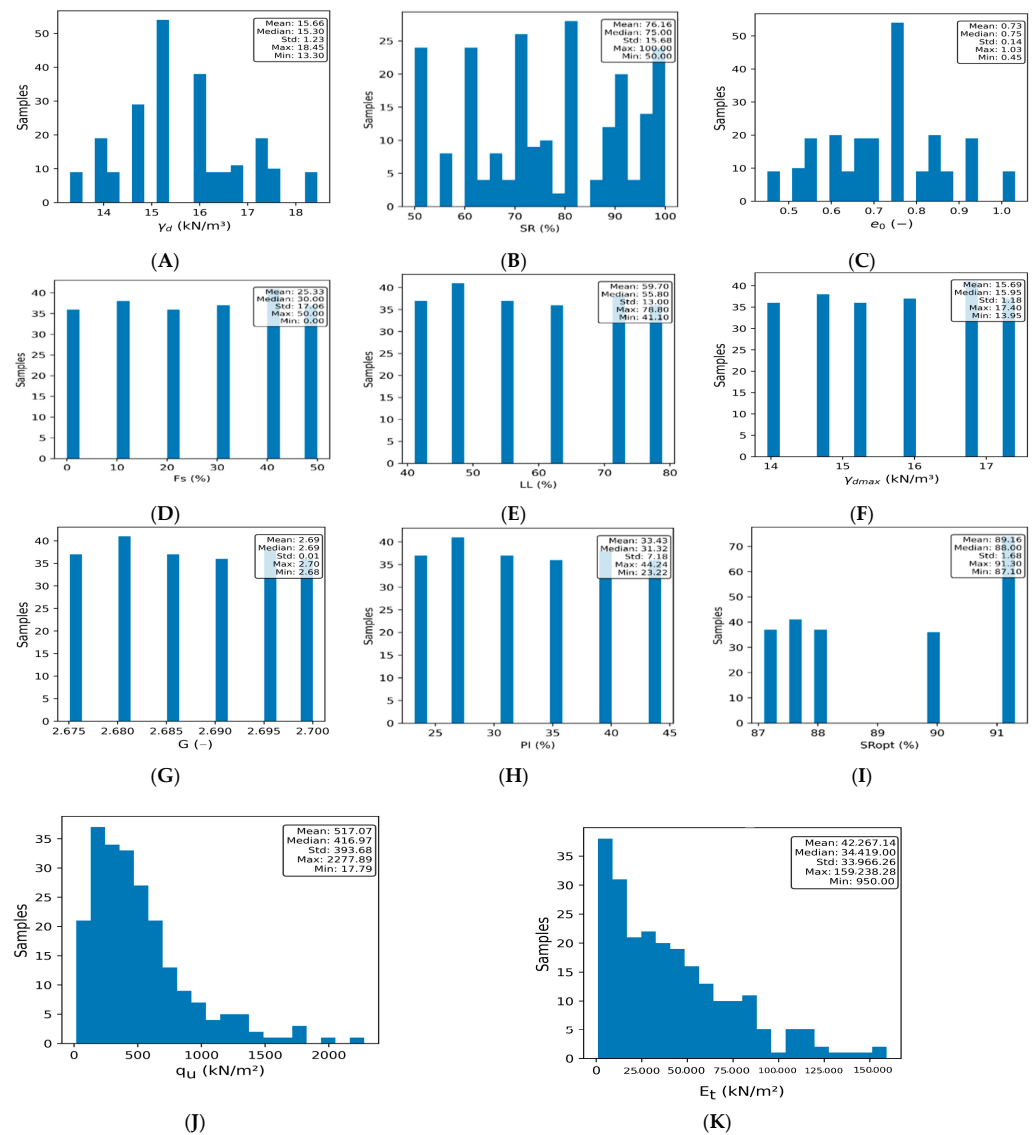


Figure 8. Density plots and distribution histograms of the dataset used in this study: (A) γ_d , (B) SR , (C) e_0 , (D) F_s , (E) LL , (F) γ_{dmax} , (G) G , (H) PI , (I) SR_{opt} , (J) q_u , (K) E_t .

2.3. Selection of Variables

Correlation analysis plays a vital role in exploring the connections among different inputs in datasets. However, given the diverse nature of datasets, the assumptions associated with each correlation method often vary, influencing their suitability and reliability.

The primary criterion for selection involves eliminating variables that are challenging to obtain and opting for readily available data. In this study, Spearman, Kendall, and Pearson correlation coefficients are calculated among different parameters to examine the correlation between input and output parameters, as illustrated in Figure 9. The strength of the correlation between factors is depicted via numerical values ranging from 0 to 1, where a value of 1 indicates a robust positive correlation, while a value of -1 signifies a strong negative correlation. When comparing the three methods (Spearman, Kendall, and Pearson) across different datasets with different inputs, the variables SR , γ_d , and e_0 consistently show the highest correlation coefficients. Regardless of the dataset's composition or the underlying presumptions of each correlation technique, this steady trend points to a solid and robust relationship between SR , γ_d , and e_0 and E_t and q_u . Consequently, LL , Fs , and these three parameters (SR , γ_d , and e_0) were selected as inputs. The high correlation coefficients observed for degree of saturation (SR), dry unit weight, and void ratio (e_0), with unconfined compressive strength and the elasticity modulus, can be attributed to their fundamental roles in soil mechanics. The degree of saturation (SR) reflects the soil's moisture content relative to its maximum water-holding capacity, directly impacting soil strength and stiffness via its influence on pore water pressure. Dry unit weight and void ratio directly influence soil density, porosity, compressibility, and shear strength, explaining their strong correlations with unconfined compressive strength and elasticity modulus. Conversely, the lower correlation coefficients observed for sand content (Fs) and liquid limit (LL) suggest their relatively weaker direct influence on soil mechanics. Fs primarily reflects the composition of soil particles, which may affect soil strength and stiffness to a lesser extent than density and porosity represented by dry unit weight and void ratio. LL , indicating soil plasticity, may indirectly affect mechanical properties via its influence on soil structure and behavior under different moisture conditions. However, despite achieving low correlation coefficients, LL and Fs were selected as inputs due to their importance in soil classification and the ease of determination via simple tests. This decision aids machine learning models in comprehending the complex relationships within multidimensional data, ensuring a more comprehensive analysis of soil behavior. The statistical characteristics of the chosen variables are presented in Table 4.

2.4. Machine Learning Algorithms Used in the Study

In this section, a comprehensive comparison is conducted among the algorithms employed in this study, including decision tree regression (DTR), random forest regression (RFR), gradient boosting regression (GBR), support vector regression (SVR), and artificial neural networks (ANNs). These algorithms were chosen for their common use in previous studies.

Table 4. The statistical attributes of the selected variables.

	LL	Fs	e_0	SR	γ_d	E_t	q_u
count	225	225	225	225	225	225	225
mean	59.7	25.3	0.73	76.2	15.66	42,267	517.1
std	13	17.06	0.14	15.68	1.228	33,966	393.7
min	41.1	0	0.45	50	13.3	950	17.79
25%	48.6	10	0.63	60	14.66	14,717	238.3
50%	55.8	30	0.74	74.25	15.3	34,418	416.97
75%	72.1	40	0.858	90	16.6	61,706	665.11
max	78.8	50	1.032	100	18.451	159,238	2277.9

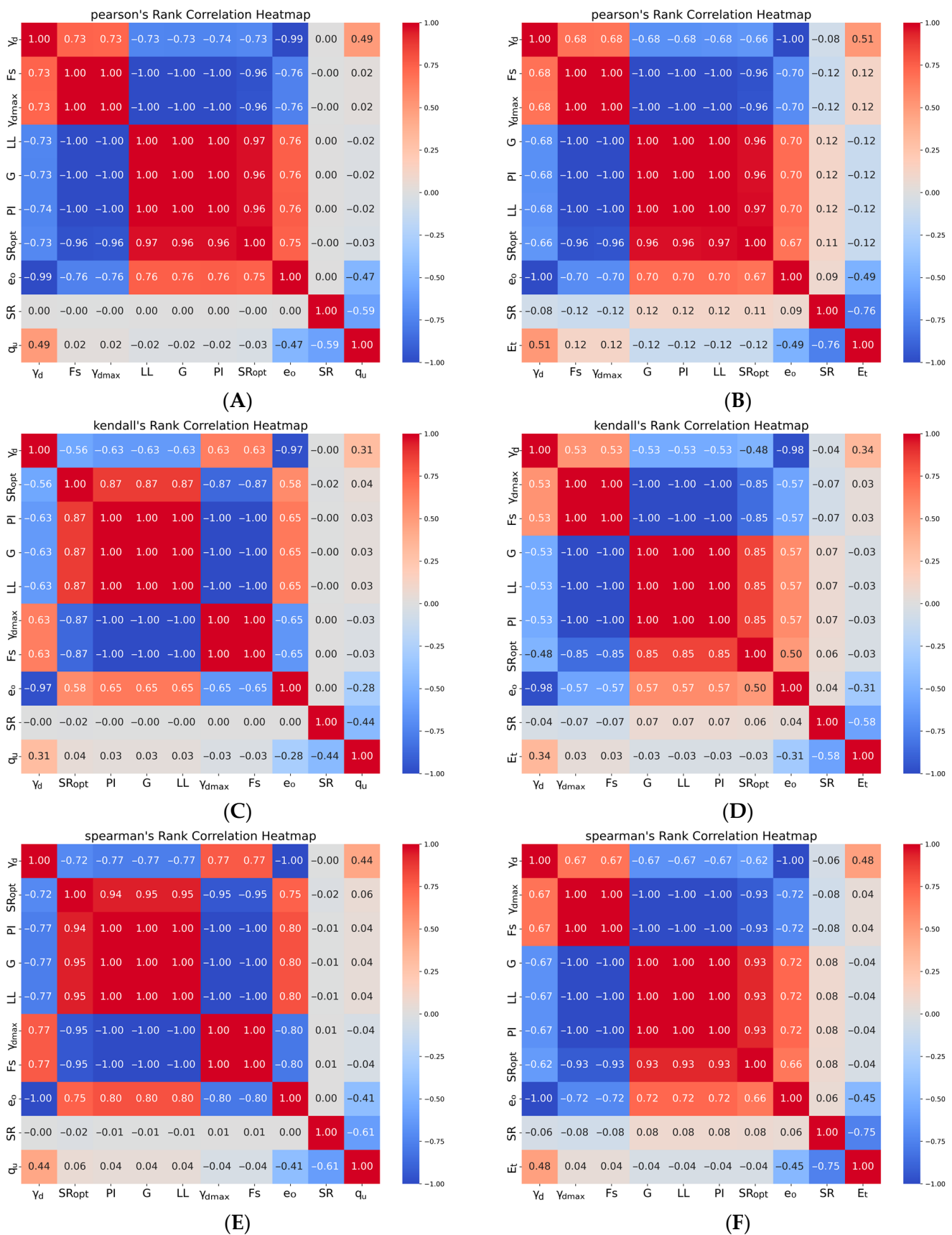


Figure 9. Assessing the impact of different input parameters for q_u and E_t via correlation coefficients: (A) Pearson for q_u , (B) Pearson for E_t , (C) Kendall for q_u , (D) Kendall for E_t , (E) Spearman for q_u , and (F) Spearman for E_t .

The dataset's complexity influenced the selection of machine learning (ML) models. ML models that are adept at handling such complexities may be necessary because traditional models may not be able to capture these subtleties. Aiming primarily for higher predictive accuracy for variables such as q_u and E_t in expansive soils, machine learning models were chosen due to their capacity to efficiently navigate intricate relationships.

This study employs a variety of machine learning models, including extreme gradient boosting (XGBoost), gradient boosting regression (GBR), artificial neural networks (ANN-GMDH, and ANNs), support vector regression (SVR), decision tree regression (DTR), random forest regression (RFR), and symbolic regression (SR) in the Python software 3.11 environment. These models were chosen based on their widespread use in previous studies. For more information on the machine learning models used in this study, including their most important hyperparameters, see [81].

When evaluating efficacy of the ML models, the first step is to analyze their performance on the training data. Following this, to thoroughly evaluate the models' ability to generalize, two cross-validation techniques are utilized. This entails testing the models on datasets distinct from those utilized during training. In this study, each predictive model was developed using a total of 225 datasets. Among these, 80% were assigned to the training phase, while the remaining 20% were set aside for critical testing purposes.

The quality of the predictive models were evaluated using statistical indices, including MAE (mean absolute error), RMSE (root-mean-square error), and R^2 (coefficient of determination), which are defined in Equations (1)–(3), respectively [82–84].

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^n (y'_i - y_i)^2}{n}} \quad (1)$$

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |(y'_i - y_i)| \quad (2)$$

$$R^2 = \left(\frac{\sum_{i=1}^n (y_i - \bar{y})(y'_i - \bar{y}')}{\left(\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2} \right) \left(\sqrt{\sum_{i=1}^n (y'_i - \bar{y}')^2} \right)} \right)^2 \quad (3)$$

where “ y ” denotes the actual values, “ y' ” stands for the predicted values, “ $\bar{y}$ ” represents the mean of the set “ y ”, and “ n ” signifies the total number of data points.

2.4.1. Decision Tree Regression

Decision tree regression (DTR) is a structured hierarchical data representation characterized by variable branches and nodes. In this framework, certain nodes extend outward with connecting lines, while others, termed ‘leaves’, do not. The data points intended for regression or categorization are systematically divided into distinct categories by utilizing specific internal nodes. In the training phase, the algorithm compares the values of input variables to designated functions. Strategically, the algorithm seeks to construct the most effective decision trees by iteratively minimizing the fitness function. The dataset is partitioned at multiple points within each set of independent variables. At these points, the algorithm calculates prediction errors, reflecting differences between expected and actual values based on the fitness function. Identifying the optimal split point involves selecting the variable that yields the smallest fitness function value, considering split point errors across all variables [85].

2.4.2. Random Forest Regression

The random forest regression (RFR) algorithm, which relies on ensemble learning, is extensively applied in both academic and industrial settings for tasks involving classification and regression due to its efficacy. By employing bootstrap aggregation, it coordinates

an ensemble of varied, randomly formed, and unpruned decision trees. An RFR is built by assembling a collection of decision trees that have been systematically altered. Random feature selection emerges as a critical component of the training process. The careful selection of different attributes is required to ensure a rich diversity of decision trees. As a result, each decision tree contributes a unique evaluative perspective to the collective forest, culminating in the outcome being determined by the aggregated votes of these trees. The random forest approach is explained in detail in earlier research [73,86].

2.4.3. Gradient Boosting Regression

Gradient boosting regression (GBR) refers to an ensemble-trained supervised machine learning model. Boosting is a technique that involves combining many simple models into a unified composite model. Boosting is frequently described as an additive model because basic models are sequentially added while the trees of the model remain unchanged. Combining more fundamental models results in more accurate predictions. Gradient boosting aims to reduce losses via gradient descent, a first-order iterative optimization procedure. Employing a squared error loss function, decision trees act as the weak learners within the framework of gradient boosting. GBR proceeds by training a weak model to map features to the predicted residuals of the weak model. These residuals are incorporated into the input of the current model, guiding it toward the intended outcome. When this process is repeated on a regular basis, the model's overall predicted accuracy steadily improves [87,88].

2.4.4. Extreme Gradient Boosting

Extreme gradient boosting (XGBoost), a powerful implementation of gradient boosting machines (GBM), is regarded as one of the best performers in the field of supervised learning. Its capability spans across both regression and classification problems, making it a versatile tool in data analysis. Data scientists frequently favor XGBoost because of its faster execution speed, particularly when dealing with out-of-core computations. The performance of the algorithm is driven by a combination of advanced processes such as gradient boosting, tree pruning, and regularization. These mechanisms collaborate to improve its prediction powers. Notably, XGBoost is adaptable in dealing with missing data instances and provides regularization procedures, validating its appropriateness for predicting applications. To completely unlock the algorithm's latent potential, hyperparameters must be carefully fine-tuned. This tuning method entails tweaking crucial parameters such as learning rate and maximum depth to achieve maximum efficiency [89].

2.4.5. Support Vector Regression

Support vector machine (SVM) finds extensive application across diverse prediction scenarios, encompassing tasks related to classification and regression. In the domain of regression, it is commonly referred to as support vector regression (SVR) [90]. SVR operates by transforming the model's inputs into a higher-dimensional space, a fundamental step in its training process based on the principles of structural risk minimization (SRM) [91]. Facilitating this transformation is the kernel function, tasked with mapping inputs into this expanded space to enable complex processing.

2.4.6. Artificial Neural Networks

Artificial neural networks (ANNs) represent a computational architecture comprising artificial neurons, engineered to mimic the information processing mechanisms observed in the human brain, thereby enabling knowledge acquisition. These networks construct intricate input–output models capable of identifying intricate correlations within multi-dimensional datasets. Their ubiquitous use is found in a variety of engineering specialties [92,93]. ANNs support a wide range of network architectures, including single and multiple layer networks. One instance is the feedforward back propagation neural network

(FFBP), consisting of an input layer for receiving data, an intermediate hidden layer, and an output layer that provides outcome-specific insights in reaction to input data.

This paper proposes a comprehensive integration between the GMDH features and the ANN network model. The group method of data handling (GMDH) algorithm finds applications in diverse research domains. For instance, the authors of [94,95] employed GMDH for pore pressure analysis and permeability modeling, resulting in accurate permeability predictions. Additionally, GMDH has been utilized in rock deformation prediction, as demonstrated by [96]. In GMDH, the Kolmogorov–Gabor polynomial, referenced in [97], serves as a mathematical framework for understanding the relationship between input and output variables. For instance, when predicting E_t using input features like LL , Fs , and SR , the mathematical Equation (4) is formulated as follows:

$$E_t = a_0LL + a_1Fs + a_2SR + a_3LL^2 + a_4Fs^2 + a_5SR^2 + a_6(LL \times FS) + a_7(LL \times SR) + a_8(Fs \times SR) \quad (4)$$

where LL , Fs , and SR denote the input features representing liquid limit, sand percentage, and degree of saturation, respectively; $a_0, a_1, a_2, \dots, a_8$ are the coefficients; and E_t symbolizes the output, specifically the elasticity modulus.

In the conventional ANN, five features (SR , Fs , e_0 , LL , and γ_d) are utilized. In contrast, the proposed GMDH incorporates 20 features derived from the original five features to capture interaction effects. These features include both individual and interaction terms: SR , Fs , e_0 , LL , γ_d , SR^2 , Fs^2 , e_0^2 , LL^2 , γ_d^2 , $SR \times Fs$, $SR \times e_0$, $SR \times LL$, $SR \times \gamma_d$, $Fs \times e_0$, $Fs \times LL$, $Fs \times \gamma_d$, $e_0 \times LL$, $e_0 \times \gamma_d$, and $LL \times \gamma_d$. The architecture of the suggested ANN-GMDH model, depicting the input characteristics, is illustrated in Figure 10B, while the conventional ANN model is represented in Figure 10A.

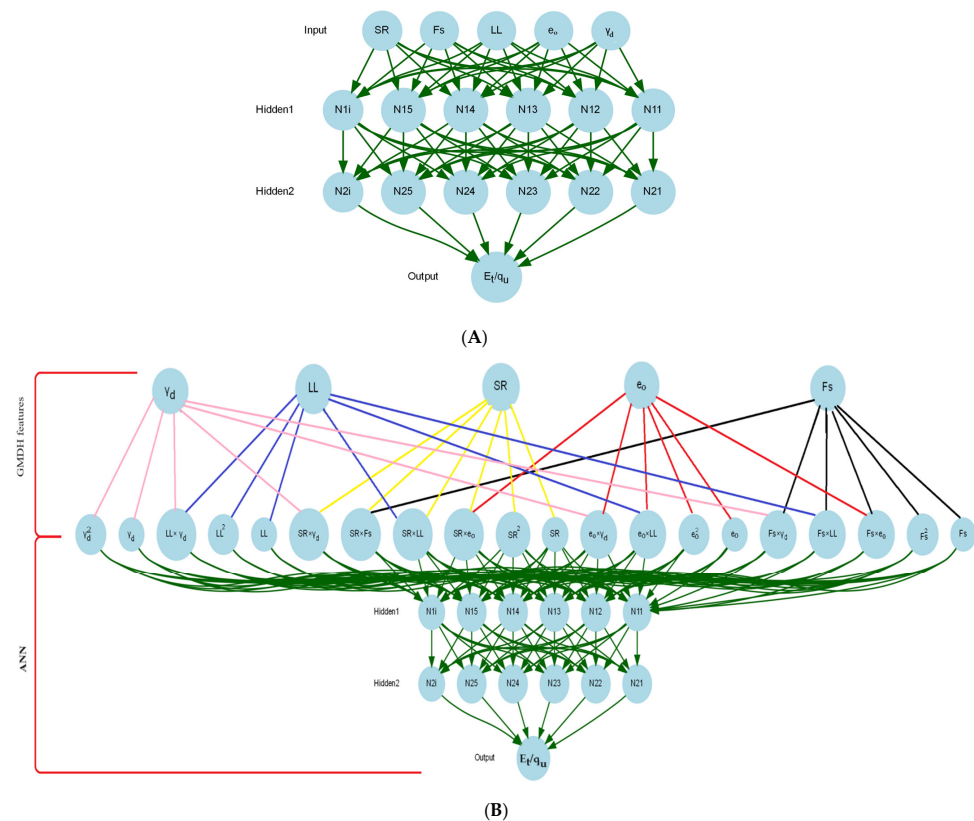


Figure 10. Illustration of a graphical representation of the neural network models: (A) conventional ANN and (B) proposed ANN-GMDH.

2.4.7. Symbolic Regression

Symbolic regression (SR) is a machine learning algorithm that searches for mathematical equations or formulas that best describe and fit data relationships. It allows for the modeling of complex relationships and the creation of predictive equations without making any assumptions about the underlying mathematical form. In capturing complex nonlinear relationships, symbolic regression outperforms multiple linear regression (MLR), which is limited to linear relationships between variables [98]. The symbolic regression algorithm is used in this study to determine the equation of q_u and E_t . In this study, EUREQA 1.24.0 software is used for this purpose. EUREQA begins the symbolic regression process by configuring algorithm parameters and establishing the equation space, as shown in Figure 11. It begins with a small set of equations. These equations are subjected to a fitness evaluation, which determines how well they correspond to the given dataset. The algorithm grows and refines this set by including and excluding equations dynamically based on their fitness and relevance. Equations with high fitness are subject to adaptive evolutionary mechanisms, allowing the algorithm to iteratively explore and refine equations. EUREQA adjusts its equation population, concentrating computational resources on promising equations. Iterations of this dynamic evolution continue until convergence or the fulfillment of predetermined termination criteria. In the end, EUREQA chooses the best-fit equation or equations to analyze and validate their correctness in capturing the underlying patterns in the dataset.

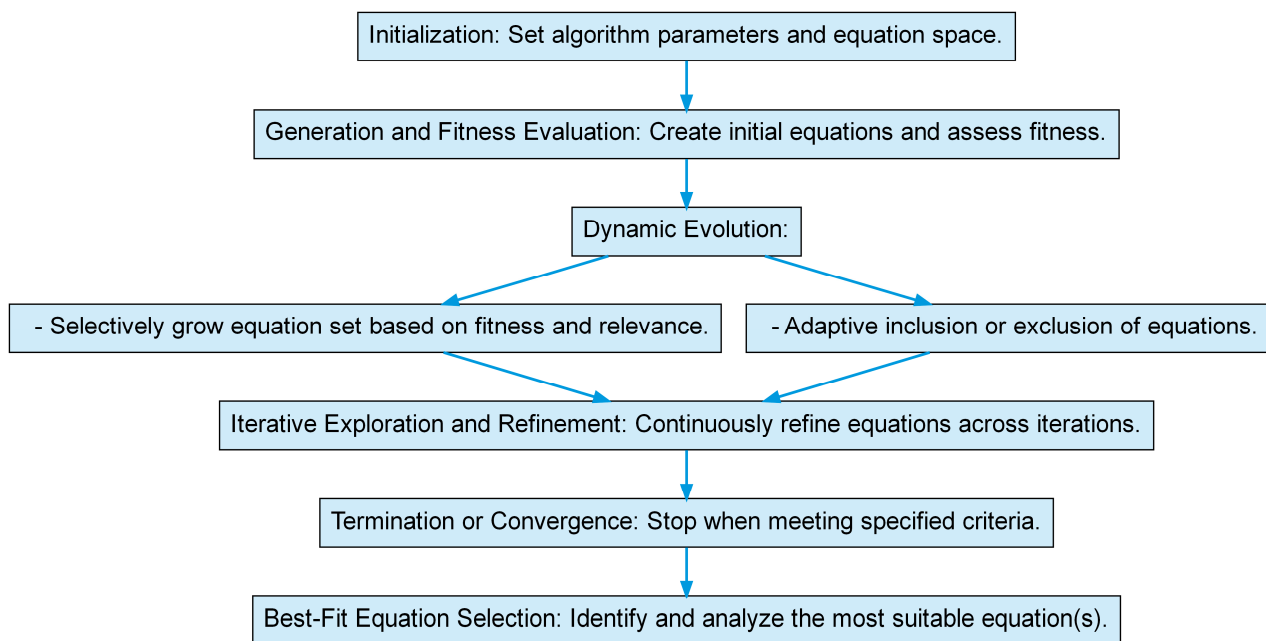


Figure 11. Flowchart of symbolic regression process.

2.5. Hyperparameters Optimization

In this study, a systematic strategy is employed for hyperparameter tuning, which focuses on refining machine learning models via the GridSearchCV technique. Specifically, GridSearchCV is utilized across GBR, XGBoost, SVR, ANN, ANN-GMDH RFR and DTR, models. This approach involves systematically exploring a predefined grid of hyperparameter values, coupled with cross-validation for each combination, to identify the most effective set that enhances model performance. Considering the more compact and manageable hyperparameter spaces of these models, a comprehensive exploration proves both feasible and advantageous [99–101]. While acknowledging the larger and intricate hyperparameter spaces of the GBR and ANN models, requiring additional optimization time, the study opts for GridSearchCV for consistency. This strategic choice guarantees

a systematic and exhaustive examination of the hyperparameter space, aligning with the study's methodology and obviating the necessity for Randomized Search CV [101].

Artificial neural networks (ANNs) powered by TensorFlow are instrumental in unraveling complex patterns within datasets. Specifically designed for regression tasks, ANNs utilize a structure of interconnected nodes organized into layers to establish correlations between input features and continuous output values. In TensorFlow-based ANNs, hyperparameter optimization involves thorough exploration of the parameter space, encompassing architecture parameters like activation functions, neuron count per layer, and layer count, as well as training process parameters such as number of training epochs, batch size, and learning rate. This systematic approach aims to identify the optimal combination of hyperparameters that enhances predictive accuracy, exploring various configurations and leveraging TensorFlow's capabilities for efficient evaluation and training.

This process entails a thorough exploration of these models' parameter spaces to identify optimal configurations. The optimal settings for each model regarding q_u and E_t , respectively, are displayed in Tables 5 and 6.

Table 5. Optimal hyperparameters for models predicting ' q_u '.

Model	Hyperparameter	Value
DTR	max_depth	10
	min_samples_leaf	1
	min_samples_split	1
RFR	max_depth	10
	min_samples_leaf	1
	min_samples_split	1
	n_estimators	130
GBR	subsample	0.8
	n_estimators	100
	min_samples_split	12
	min_samples_leaf	2
	max_depth	3
	learning_rate	0.6
XGBoost	colsample_bytree	0.8
	learning_rate	0.4
	max_depth	3
	n_estimators	200
	subsample	0.6
SVR	C	100,000
	epsilon	10
	kernel	rbf
ANN	number of neurons in layer1	10
	number of neurons in layer2	5
	learning_rate	0.05
	batch size	10
	hidden layers function	ReLU
	linkage between the hidden layer and the ultimate output layer function	ReLU
ANN-GMDH	number of neurons in layer1	20
	number of neurons in layer2	30
	learning_rate	0.09
	batch size	14
	hidden layers function	ReLU
	linkage between the hidden layer and the ultimate output layer function	ReLU

Table 6. Optimal hyperparameters for models predicting ' E_t '.

Model	Hyperparameter	Value
DTR	max_depth	10
	min_samples_leaf	1
	min_samples_split	1
RFR	max_depth	12
	min_samples_leaf	1
	min_samples_split	1
	n_estimators	70
GBR	subsample	0.8
	n_estimators	100
	min_samples_split	12
	min_samples_leaf	2
	max_depth	3
	learning_rate	0.6
XGBoost	colsample_bytree	0.6
	learning_rate	0.2
	max_depth	3
	n_estimators	200
	subsample	0.6
SVR	C	100,000
	epsilon	0.1
	kernel	rbf
ANN	number of neurons in layer1	18
	number of neurons in layer2	18
	learning_rate	0.07
	batch size	8
	hidden layers function	ReLU
	linkage between the hidden layer and the ultimate output layer function	ReLU
ANN-GMDH	number of neurons in layer1	20
	number of neurons in layer2	16
	learning_rate	0.085
	batch size	41
	hidden layers function	ReLU
	linkage between the hidden layer and the ultimate output layer function	ReLU

3. Results and Discussion

3.1. Effect of Initial Dry Unit Weight

The study investigated the effect of initial dry unit weight on the elasticity modulus (E_t) and unconfined compressive strength (q_u) of sand–clay mixtures with varying percentages of added sand. The initial saturation level was set at 75% for all samples, while the densities were varied for each sample. Figure 12 shows the change in q_u and E_t with the initial dry unit weight for each percentage of sand.

The results showed an increase in both q_u and E_t with an increase in the dry unit weight for each sand-added percentage. On the other hand, both q_u and E_t decreased with an increase in the percentage of sand for all added sand percentages. The reason for this trend is that as the percentage of sand increases for the same dry density, the size of the voids in the soil increases, while the amount of clay minerals that have high water affinity decreases, leading to a reduction in suction forces within the pores of the soil.

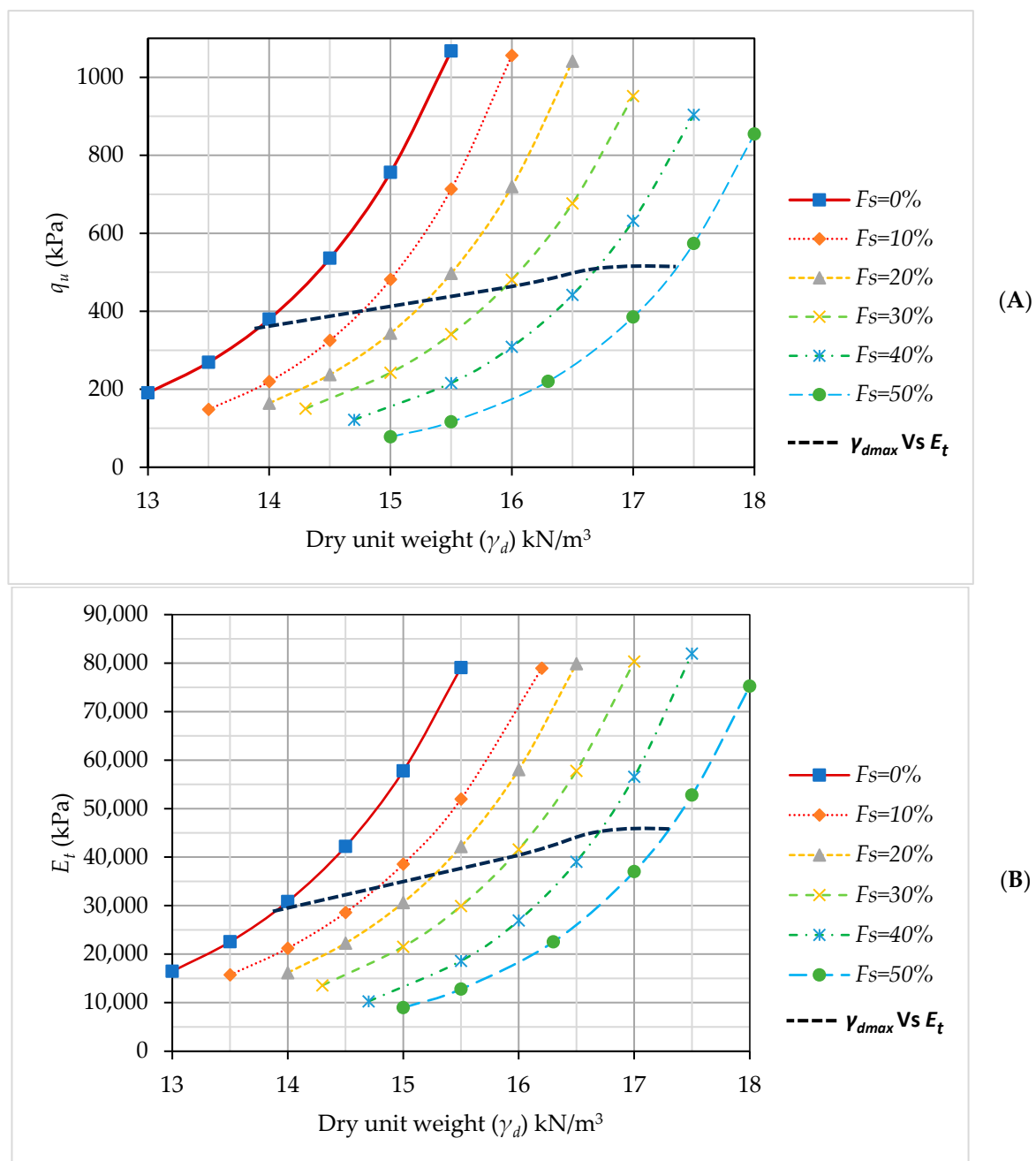


Figure 12. The change in q_u and E_t with γ_d according to F_s ; (A) the change in q_u with γ_d according to F_s ; (B) the change in E_t with γ_d according to F_s .

Figure 12A: This graph is a scatter plot with ascending lines reflecting the trend for a certain percentage of added sand (F_s) inside the soil mixture. The x-axis shows the dry unit weight (γ_d kN/m³), which ranges from 13 kN/m³ to 18% kN/m³, and the y-axis indicates the unconfined compressive strength (q_u), which ranges from 0 to approximately 1100 kPa. The general trend indicates that as dry unit weight increases, so does unconfined compressive strength, especially after exceeding the maximum dry unit weight of the standard proctor test for each sand content.

Figure 12B: This graph, like Figure 12A, is a scatter plot with ascending lines, but it illustrates the change in elasticity modulus (E_t). The y-axis has a significantly broader scale, extending from 0 to 90,000 kPa, representing the higher values typical of elasticity modulus measurements. The lines use the same color and style codes as in Figure 12A and

demonstrate the similar pattern of increasing elasticity modulus with increasing dry unit weight. The sharp rise in modulus indicates that the soil becomes much stiffer and less deformable as it becomes denser.

The trend in Figure 12A shows a clear direct relationship between γ_d and q_u . As γ_d increases, q_u increases across all sand content percentages (F_s), indicating that density compromises soil strength. The results show a significant increase in strength after exceeding the maximum dry unit weight of a standard proctor. The various lines for different F_s percentages indicate that mixtures with more added sand tend to retain lower strength at the same dry unit weight than those with less sand. This behavior implies that soils with lower sand content are more susceptible to strength gain with increased dry unit weight.

In Figure 12B, the elasticity modulus (E_t) shows a similar upward trend with increasing dry unit weight, indicating an increase in soil stiffness. The lines plotted for different F_s percentages all slope upward, indicating that as the soil becomes denser, its ability to elastically deform and return to its original shape improves.

Overall, the results of this study suggest that the addition of sand to clayey soils can enhance their strength and stiffness properties, but the percentage of sand should be carefully considered to avoid reducing these properties due to an increase in voids size and a decrease in water affinity of clay minerals.

3.2. Effect of Initial Degree of Saturation

The impact of the initial degree of saturation on the mechanical properties of sand-clay mixtures with varying sand content was investigated. The maximum dry density was calculated for each percentage of sand using the standard Proctor test, and it was found that the maximum dry density increased with an increasing percentage of sand, as depicted in Figure 5. The relationship between q_u and E_t with the degree of saturation (SR) was studied separately for each sand percentage, and the results are presented in Figure 13. A sharp decrease in both the unconfined compressive strength and elasticity modulus was observed as SR increased, and the decrease became less severe as SR increased due to the filling of soil pores with water and the corresponding decrease in suction forces.

Figure 13A: This graph is a scatter plot with a series of descending lines, each line representing the trend for a specific percentage of sand (F_s) within the soil mixture. The x-axis represents $SR\%$ ranging from about 30% to 100%, and the y-axis represents q_u measured in kilopascals (kPa), which ranges from approximately 0 to 1800 kPa. The general trend shows that as saturation increases, the unconfined compressive strength decreases sharply initially and then levels off, indicating a reduction in strength gain as saturation approaches 100%.

Figure 13B: This graph is also a scatter plot with descending lines, similar to Figure 13A, but it depicts the change in elasticity modulus (E_t). The y-axis has a much larger scale, ranging from 0 to 160,000 kPa, reflecting the higher values typical of elasticity modulus measurements. The lines follow the same color and style codes as Figure 13A and show the same trend of reduction in the elasticity modulus with an increase in soil saturation. The sharp decline in modulus suggests that the material becomes significantly less stiff and more deformable as it becomes wetter, which is a critical consideration for soil mechanics and engineering.

The trend in Figure 13A demonstrates a clear inverse relationship between SR and q_u of the soil. As the saturation increases, q_u decreases across all sand content percentages (F_s), indicating that soil strength is compromised by the presence of water. The results show a steep decline in strength at lower saturation levels, which then tapers as saturation approaches full capacity. The various lines for different F_s percentages suggest that the mixtures with more added sand tend to retain higher strength at the same saturation level compared to those with less sand. This behavior implies that soils with lower sand content are more susceptible to strength loss with increased water content.

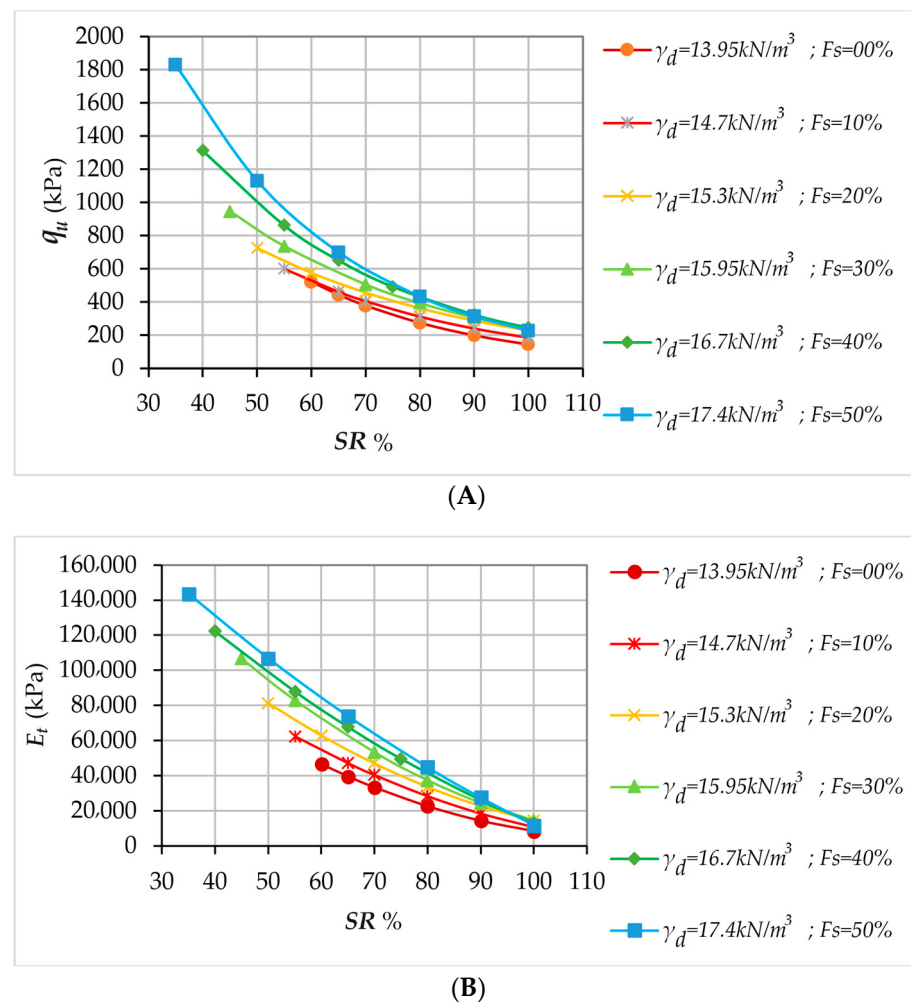


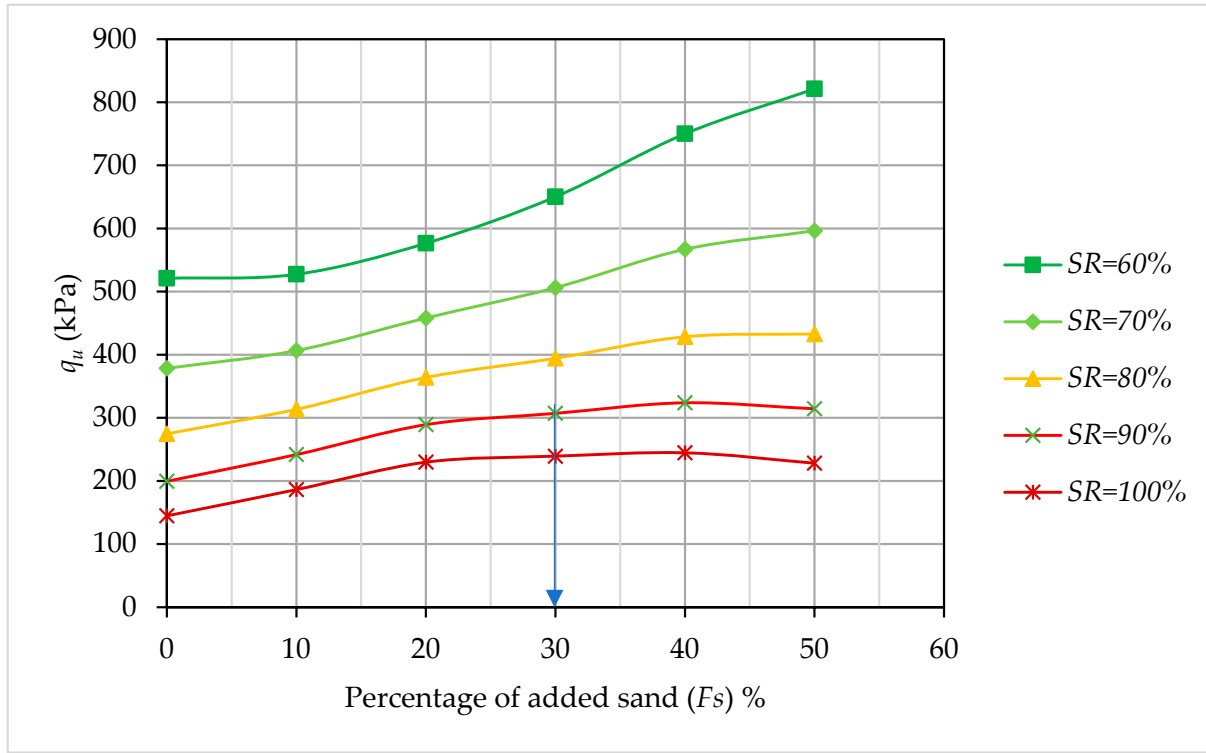
Figure 13. The change in q_u and E_t with SR according to F_s ; (A) the changes in q_u with SR according to F_s ; (B) the changes in E_t with SR according to F_s .

In Figure 13B, the elasticity modulus (E_t) exhibits a similar downward trend with increasing soil saturation, highlighting a decrease in soil stiffness. The lines plotted for different F_s percentages all slope downward, indicating that as the soil becomes more saturated, its ability to elastically deform and return to its original shape is diminished. The sharpness of the slope varies with the percentage of added sand, with higher sand content mixtures showing a slightly less severe reduction in E_t . This reduction in elasticity with increased saturation is crucial for civil engineering applications, as it affects how soil will behave under structural loads, with potential implications for the design and stability of foundations and other soil–structure interactions.

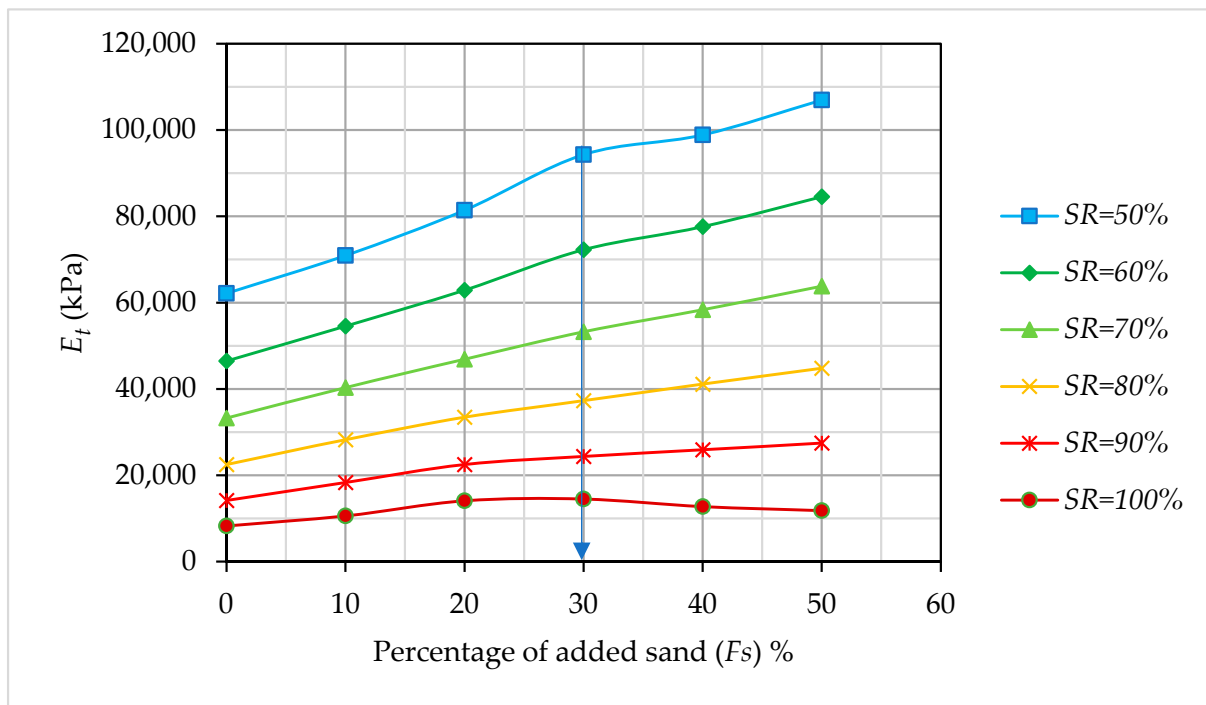
3.3. Effect of Sand Content

The effect of varying percentages of added sand (F_s) on the elasticity modulus and unconfined compressive strength is investigated by setting the maximum dry density for each percentage of added sand. Figure 14A,B are scatter plots with a series of lines, with each line representing the trend for a specific degree of saturation (SR) within the soil mixture. Figure 14 displays the change in unconfined compressive strength and elasticity modulus as a function of the percentage of sand for each degree of saturation. The results indicate that an increase in the percentage of sand leads to an increase in both elasticity modulus and unconfined compressive strength up to a certain percentage related to the degree of saturation. However, this tendency of increase becomes less significant with the increase in the degree of saturation. Moreover, after reaching a specific percentage of added

sand, the strength decreases. This percentage is about 30% for high saturation degrees, where the percentage of increase in E_t and q_u is about 75% and 65%, respectively. Notably, this critical percentage increases with a decrease in the degree of saturation.



(A)



(B)

Figure 14. The change in q_u and E_t with F_s according to SR: (A) change in q_u with F_s according to SR' (B) change in E_t with F_s according to SR.

This behavior can be explained by the suction forces within the soil. The increase in density increases suction forces, while an increase in saturation and sand percentage decreases suction forces. For instance, in Figure 14A,B, for 90% or greater saturation, both strength and stiffness increase with the percentage of sand due to the increase in maximum dry density. However, after about 30% of the added sand, there is a greater decrease in suction forces due to the increase in the percentage of sand, which overcomes the increase in suction forces caused by the increase in dry density. The study found that there is a substantial rise in suction forces when the saturation degree decreases, resulting in a continuous increase in unconfined compressive strength for larger percentages of the added sand.

Unconfined compression tests on mixtures corresponding to the optimum moisture and maximum dry density according to standard Proctor's tests revealed that the strength and elasticity modulus increase with the percentage of sand. The reason for the continued increase in strength and elasticity modulus is that the saturation degree corresponding to the optimum moisture is less than 90% for added sand percentages greater than 30%. Additionally, it is worth noting that there is an inflection point between 20% and 30% of the added sand, after which the increase becomes less severe, as shown in Figure 15. As a result, a sand percentage of 30% can be considered the ideal percentage in clay soil. The percentage of increase in the unconfined compressive strength and elasticity modulus when the sand percentage increases from 0 to 30% is approximately 68% and 81%, respectively.

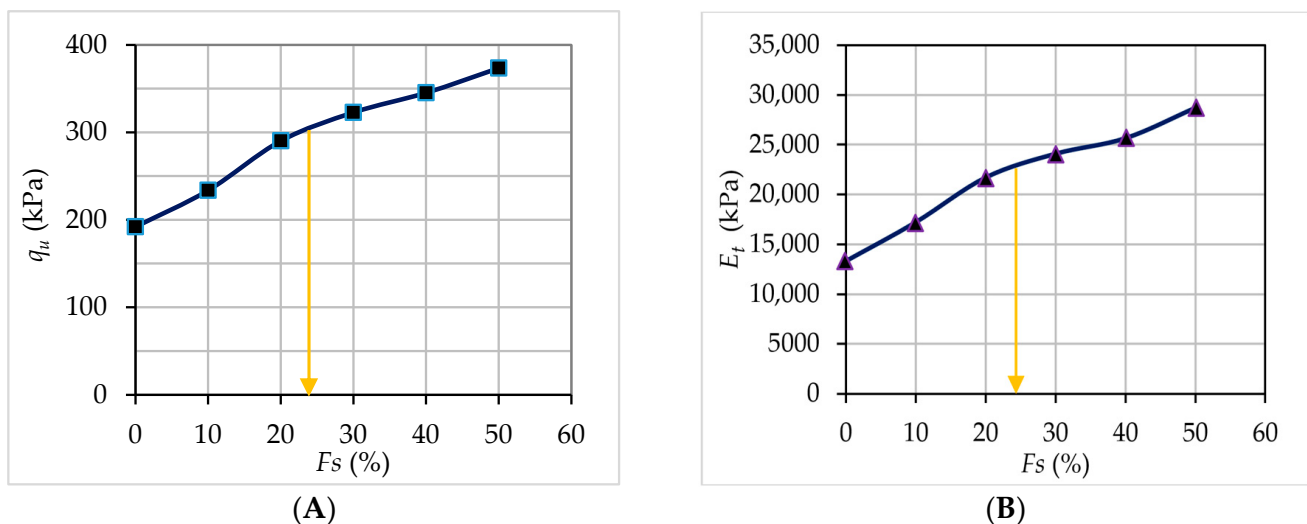


Figure 15. Unconfined compression tests on mixtures corresponding to the optimal moisture and maximum dry density according to standard Proctor's tests: (A) the change in q_u as a function of F_s , (B) the change in E_t as a function of F_s .

3.4. Machine Learning to Predict q_u and E_t

Initially, several established explicit models such as linear regression, Lasso, Ridge, and GMDH were utilized. The corresponding equations and associated metrics are summarized in Table 7. Notably, the GMDH model yielded the most favorable metrics, indicating its potential for generating reliable results. However, due to the lengthy equations involved, we opted to integrate the GMDH features with the most effective implicit model, namely ANN, as elaborated in the subsequent section. This combined model is denoted as ANN-GMDH. Lastly, a user-friendly program will be developed to facilitate straightforward predictive tasks.

Table 7. Equation and metrics of the explicit models.

Model	Equation	Metrics	Training	Testing
Linear regression	$q_u = 20.43LL - 0.19Fs + 2282.73e_0 - 14.69SR + 581.26\gamma_d - 10337.54$	RMSE MAE R ²	156.58 113.37 0.8502	152.37 111.70 0.8302
	$E_t = 102277.05e_0 - 3650.79LL + 33535.90\gamma_d - 3766.92Fs - 1467.53SR - 129466.22$	RMSE MAE R ²	9750.54 7673.28 0.9196	10,376.38 8035.28 0.8960
Lasso	$q_u = 21.89LL - 0.08Fs + 152.78e_0 - 14.67SR + 352.01\gamma_d - 5290.28$	RMSE MAE R ²	159.24 115.22 0.8450	151.95 116.70 0.8312
	$E_t = 59441.79e_0 - 2414.77LL + 28635.07\gamma_d - 2835.58Fs - 1466.10SR - 119044.50$	RMSE MAE R ²	9777.38 7667.83 0.9192	10,362.64 7998.31 0.8963
Ridge	$q_u = 20.69 * LL - 0.61Fs + 973.61e_0 - 14.68SR + 440.47\gamma_d - 7186.86$	RMSE MAE R ²	157.59 114.34 0.8482	151.61 114.57 0.8319
	$E_t = -881.03e_0 - 3028.96LL + 21526.18\gamma_d - 3307.39Fs - 1463.22SR + 84560.49$	RMSE MAE R ²	9850.17 7680.33 0.9180	10,303.14 8002.29 0.8974
GMDH	$q_u = 883490.10 - 7004.26LL - 5180.73Fs - 346890.29e_0 + 412.66SR - 48116.59\gamma_d + 31.89LL^2 + 47.51LLFs + 701.47LLe_0 - 1.10LLSR + 99.11LL\gamma_d + 17.92Fs^2 + 538.04FSeq - 0.51FsSR + 65.26Fs\gamma_d + 61461.40e_0^2 - 101.58e_0SR + 6983.98e_0\gamma_d + 0.20SR^2 - 19.49SR\gamma_d + 836.99\gamma_d^2$	RMSE MAE R ²	49.52 34.68 0.9850	56.56 38.45 0.9766
	$E_t = 46991712.22 - 11093891.97e_0 - 1106276.20LL - 1064926.86\gamma_d - 839473.26Fs - 3565.98SR + 2067379.01e_0^2 - 10309.94e_0LL + 1069107.43e_0\gamma_d - 16039.98e_0Fs - 1160.06e_0SR + 7192.25LL^2 - 2621.05LL\gamma_d + 11280.37LLFs + 94.18LLSR + 45439.18\gamma_d^2 - 3445.81\gamma_dFs - 461.22\gamma_dSR + 4426.94Fs^2 + 79.94FsSR + 17.17SR^2$	RMSE MAE R ²	2544.52 1932.47 0.9945	2889.17 2248.35 0.9919

3.4.1. Comparison of Machine Learning Models

This section presents a comparison of the models used in this study. The metrics derived from k-fold and TVT cross-validation techniques are shown in Tables 8 and 9, respectively. Tables 8 and 9 show that ANN-GMDH, ANN and SVR have superior performance metrics, with SVR closely matching ANN's performance. As a result, ANN-GMDH, ANN and SVR outperform XGBoost, GBR, RFR, and DTR in predicting E_t and q_u for the expansive soil under consideration.

Table 8. The study's metrics for the various models using the Train/Validation/Test Split cross-validation technique.

Model	Metrics	E_t				q_u			
		Training	Validation	Testing	r _d (%)	Training	Validation	Testing	r _d (%)
DTR	R ²	0.9975	0.8243	0.8372	16.07	0.99971	0.8132	0.9065	9.32
	RMSE	102.41	808.88	787.88	669.34	6.88	140.64	113.09	1543.75
	MAE	22.73	634.1	622.84	2640.17	1.4963	91.589	75.847	4968.97
RFR	R ²	0.9857	0.9088	0.898	8.90	0.982	0.83	0.883	10.08
	RMSE	244.33	582.83	623.55	155.21	54.21	134.144	126.724	133.76
	MAE	184.3	483.757	484.45	162.86	36.84	89.92	77.85	111.32
GBR	R ²	0.9991	0.985	0.9764	2.27	0.99886	0.9133	0.9495	4.94
	RMSE	59.39	236.44	300	405.14	13.637	95.806	83.11	509.44
	MAE	48	164.36	224.48	367.67	10.462	55.105	49.745	375.48
XGBoost	R ²	0.99824	0.988	0.9853	1.30	0.99902	0.9568	0.9557	4.34
	RMSE	85.81	211.14	237.2	176.42	12.664	67.644	77.818	514.48
	MAE	66.1	148.44	169.82	156.91	9.155	44.433	51.099	458.15
SVR	R ²	0.996	0.9947	0.9909	0.51	0.9972	0.99655	0.98873	0.85
	RMSE	131.02	140.61	186.27	42.17	21.5785	19.124	39.265	81.96
	MAE	62.83	110.55	128.37	104.31	10.99	13.27	22.374	103.59
ANN	R ²	0.9951	0.9955	0.9924	0.27	0.995	0.9963	0.9944	0.06
	RMSE	145.4	129.15	170.03	16.94	28.54	19.8	27.63	3.19
	MAE	109.116	108.792	115.408	5.77	19.471	15.83	18.9	2.93
ANN-GMDH	R ²	0.9963	0.9956	0.9935	0.28	0.996	0.9968	0.9949	0.11
	RMSE	125.67	128.24	157.31	25.18	25.54	18.43	26.37	3.25
	MAE	95.66	96.658	120.86	26.34	17.73	14.24	18.45	4.06

Table 9. The study’s metrics for the various models using (k-fold) cross-validation technique.

Model	Metrics	E_t			q_u		
		Training	Testing	r_d (%)	Training	Testing	r_d (%)
DTR	R^2	0.99492	0.8766	11.89	0.9835	0.9042	8.06
	RMSE	145.52	681.11	368.05	51.083	114.465	124.08
	MAE	66.272	546.266	724.28	27.73	80.778	191.30
RFR	R^2	0.9854	0.9388	4.73	0.9828	0.9007	8.35
	RMSE	241.62	521.71	115.92	52.17	116.51	123.33
	MAE	189.75	408.33	115.19	33.3	76.91	130.96
GBR	R^2	0.999694	0.9861	1.36	0.99755	0.9591	3.85
	RMSE	35.72	228.93	540.90	19.694	74.827	279.95
	MAE	29.35	167.56	470.90	14.83	48.52	227.17
XGBoost	R^2	0.99803	0.9906	0.74	0.99995	0.9652	3.48
	RMSE	90.51	187.53	107.19	2.7669	69.011	2394.16
	MAE	69.5	148.97	114.35	2.08	40.92	1867.31
SVR	R^2	0.99394	0.99323	0.07	0.9958	0.98926	0.66
	RMSE	158.865	159.557	0.44	26.203	35.051	33.77
	MAE	97.77	121.494	24.27	13.409	23.053	71.92
ANN	R^2	0.99487	0.9901	0.48	0.9914	0.9903	0.11
	RMSE	147.19	187.71	27.53	36.971	36.0973	2.36
	MAE	107.69	145.744	35.34	23.58	24.35	3.27

A comprehensive assessment of machine learning models was undertaken to evaluate their effectiveness in predicting E_t and q_u , with results presented across mean absolute error (MAE), root-mean-square error (RMSE), and R-squared (R^2) metrics in Tables 8 and 9. Decision tree regression (DTR) demonstrates strong performance on the training set but struggles with overfitting on validation and testing sets, limiting its extrapolative capability. Random forest regression (RFR) consistently produces robust results but shows slightly higher error metrics, indicating some constraints in its predictive capacity. Gradient boosting regression (GBR) maintains remarkable performance with acceptable generalization ability, while XGBoost outshines others with exceptional performance and robust generalization. Artificial neural networks (ANNs) and support vector regression (SVR) exhibit commendable performance, with ANN-GMDH emerging as a top performer, boasting high R^2 values and low error metrics.

It should be noted that an ML model’s capacity for generalization is assessed by its performance on unseen (testing) data. Metric scores during testing typically indicate a slight decrease in performance across all models compared to the training stage due to the unfamiliar data. To quantify this decline in performance, the degradation rate, or r_d , is defined in Equation (5) [76]:

$$r_d = \left| \frac{m_{train} - m_{test}}{m_{train}} \right| \times 100\% \quad (5)$$

where m_{test} and m_{train} stand for a particular measure’s values during the testing and training stages, respectively.

Notably, the K-fold technique produces the least amount of performance decay and outperforms the TVT technique, except for ANN, where the TVT technique outperforms the ANN technique in terms of consistency for E_t prediction.

All performance metrics show that the ANN-GMDH and ANN produce the least amount of performance decay among all models, with only the r_d of RMSE exceeding 10% for E_t predictions. The SVR follows. The performance of the other models significantly declined, particularly when it came to the MAE and RMSE for DTR using TVT technique for q_u predictions, where the r_d values were 1544% and 4969%, respectively. Throughout both

training and testing phases, the ANN-GMDH and ANN models demonstrated remarkable consistency in their performance, with ANN-GMDH, ANN and SVR delivering the most accurate predictions. However, because of the widespread use of ANN in geotechnical engineering [50–52,54,66,68–72] and consistently superior performance in achieving better metrics. ANNs build complex input–output models that can learn intricate relationships within multidimensional data. Their widespread application spans a wide range of engineering disciplines [92,93]. This paper will primarily focus on ANN-GMDH and ANN predictions using the TVT technique because it provides more consistent results than the K-fold for ANN predictions.

ANN Models to Predict Elasticity Modulus (E_t)

Figure 16 casts light on the regression cross plots of E_t model, which harnesses the power of artificial neural networks (ANNs) for elasticity modulus prediction. The figure delineates the congruence and disparities between true and predicted values across the training, validation, and test sets. Figure 16A–H highlights the precise predictions of ANN E_t models, displaying a favorable alignment with the $x = y$ reference line and best fit trend line (ideal line) across a range of elasticity moduli [54]. In the training set, the E_t models achieve an R-squared of 0.9951 for ANN and 0.9963 for ANN-GMDH. the mean absolute error is 109.116 for ANN and 95.66 for ANN-GMDH, while the root-mean-square error (RMSE) is 145.4 for ANN and 125.67 for ANN-GMDH. This evidences the models' capability to apprehend intricate patterns within the training data and produce predictions that exhibit a commendable level of accuracy. In the validation set, the models maintain their predictive strength, achieving an R-squared of 0.9955 for ANN and 0.9956 for ANN-GMDH. The mean absolute error (MAE) is 108.792 for ANN and 96.658 for ANN-GMDH, while the root-mean-square error (RMSE) is 129.15 for ANN and 128.24 for ANN-GMDH. This highlights the models' prowess in extrapolating predictions to new data, with minimal deviations from the actual values. In the test set, consistent performance is observed, with an R-squared of 0.9924 for ANN and 0.9935 for ANN-GMDH. The mean absolute error (MAE) is 115.408 for ANN and 120.86 for ANN-GMDH, while the root-mean-square error (RMSE) is 170.03 for ANN and 157.31 for ANN-GMDH. This performance reiterates the models' proficiency in generating dependable predictions on independent data instances. When juxtaposed with SR model, it becomes apparent that ANN models occupy a favorable position. The ANN-GMDH model consistently outperforms the ANN model across all evaluation metrics, including R^2 , MAE, and RMSE. However, it is important to note that the r_d value of the ANN-GMDH model is slightly worse than that of the ANN model. Overall, the ANN-GMDH model demonstrates better predictive accuracy and error minimization compared to the ANN model. The models aptly capture underlying relationships within the data and translates them into reliable predictions. This performance is poised to deliver valuable insights for elasticity prediction applications.

ANN Model to Predict Unconfined Compressive Strength (q_u)

Figure 17 casts light on the regression cross plots of q_u model, which harnesses the power of artificial neural networks (ANNs) for unconfined compression strength prediction. The figure delineates the congruence and disparities between true and predicted values across the training, validation, and test sets. In Figure 17A–H, the ANN q_u models display precise predictions, aligning favorably with the $x = y$ reference line and best-fit trend line (ideal line) across a range of unconfined compression strengths [54]. In the training set, the ANN model achieves an R^2 of 0.995 with an MAE of 19.471 and an RMSE of 28.54, while the ANN-GMDH model achieves an R^2 of 0.996 with an MAE of 17.73 and an RMSE of 25.54. This evidences the models' capability to apprehend intricate patterns within the training data and produce predictions that exhibit a commendable level of accuracy. In the test set, both the ANN and ANN-GMDH models exhibit consistent performance, with the ANN model achieving an R-squared of 0.9944, an MAE of 18.9, and an RMSE of 27.63, while the ANN-GMDH model achieves an R-squared of 0.9949, an MAE of 18.45, and an RMSE of 26.37. This performance

reiterates the models' proficiency in generating dependable predictions on independent data instances. When juxtaposed with SR model, it becomes apparent that ANN-GMDH and ANN models occupy a favorable position. The models aptly capture underlying relationships within the data and translates them into reliable predictions. This performance is poised to deliver valuable insights for compressive strength prediction applications. While both the ANN-GMDH and ANN models achieve almost the same metrics, the ANN-GMDH model achieves better accuracy overall. However, it is important to note that its r_d value is slightly worse than that of the ANN model.

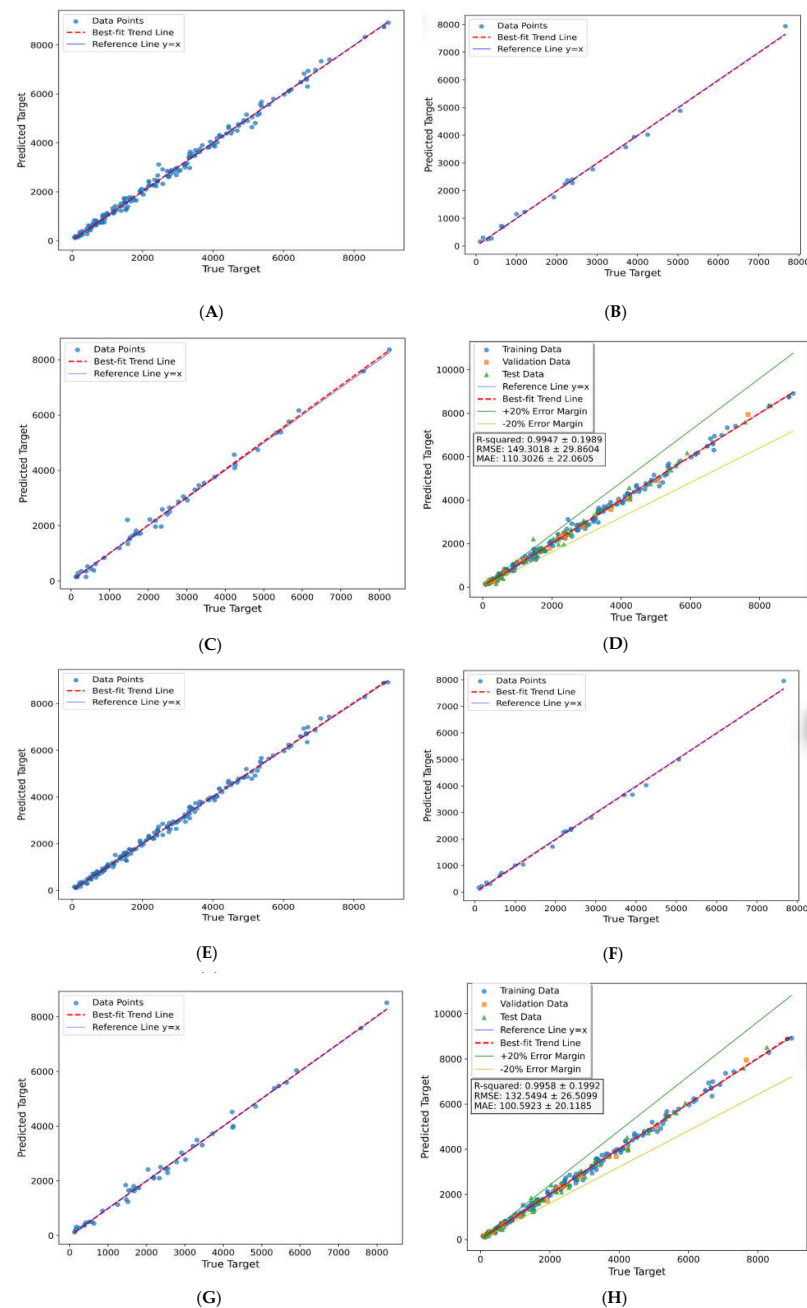


Figure 16. The cross plots for actual and predicted $\frac{E_t}{\gamma_d}$ targets of training, validation, test, and combined data for ANN (E_t) model and ANN-GMDH (E_t) model: (A) ANN (E_t) model for training data, (B) ANN (E_t) model for validation data, (C) ANN (E_t) model for test data, (D) ANN (E_t) model for combined data, (E) ANN-GMDH (E_t) model for training data, (F) ANN-GMDH (E_t) model for validation data, (G) ANN-GMDH (E_t) model for test data, and (H) ANN-GMDH (E_t) model for combined data.

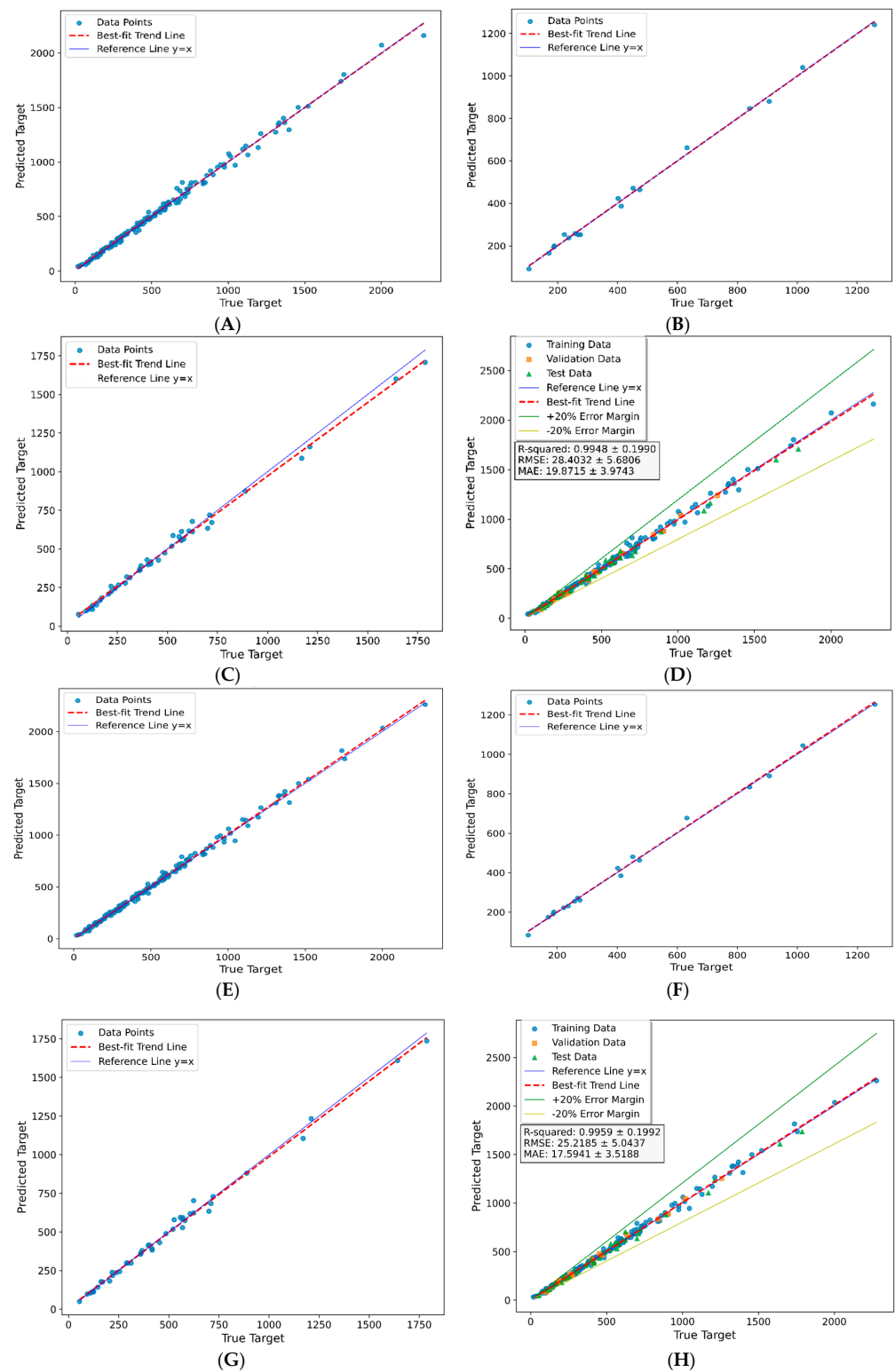


Figure 17. Comparison of true and predicted q_u targets in training, validation, test, and combined data for ANN (q_u) model and ANN-GMDH (q_u) model: (A) ANN (q_u) model for training data, (B) ANN (q_u) model for validation data, (C) ANN (q_u) model for test data, (D) ANN (q_u) model for combined data, (E) ANN-GMDH (q_u) model for training data, (F) ANN-GMDH (q_u) model for validation data, (G) ANN-GMDH (q_u) model for test data, and (H) ANN-GMDH (q_u) model for combined data.

Moreover, establishing an acceptable error margin is pivotal in assessing the sufficiency of the training dataset. Figures 16 and 17 depict a 20% error margin on each cross plot to illustrate this concept. The performance of both the E_t and q_u models was commendable, as the majority of predictions fell within the error margin lines across the training, validation, and testing phases. This underscores the adequacy of the training dataset in supporting the predictive models for E_t and q_u using the proposed ANN model in this study.

The precision observed in predicted accuracy is the result of a strategy based on statistical rigor and practical relevance. The models were trained to detect subtle patterns in the data by carefully selecting variables and exhaustively assessing their statistical features. Additionally, a diverse array of machine learning methods including ANN, GBR, XGBoost, SVR, DTR, and RFR were employed to ensure comprehensive exploration of predictive patterns. Furthermore, the intensive evaluation approach, which included cross-validation techniques on 225 datasets, enabled a thorough assessment of the models' generalization capabilities. Hyperparameter tuning fine-tuned the models' performance, ensuring they were well-suited to the dataset's subtleties. Recognizing the inherent randomness and possibility for human error, the methodology promoted openness and accountability by employing statistical indicators such as RMSE, MAE, and R^2 . This comprehensive approach emphasizes the methodology's rigor and the predictive models' reliability.

3.4.2. Streamlined Interface for ANN Model Predictions

The incorporation of a user-friendly interface enhances this approach by providing a streamlined, one-click prediction process. This interface simplifies the utilization of the ANN model, ensuring accessibility for users of varying technical expertise. For instance, to utilize the ANN q_u model, users can easily open the "ANN q_u model.ipynb" (Supplementary Materials) Python file using the Jupyter Notebook app. By simply pressing shift + enter, the interface promptly appears, facilitating a user-friendly experience, as depicted in Figure 18. This straightforward process remains consistent for the ANN E_t model, ensuring uniform and accessible user interaction.

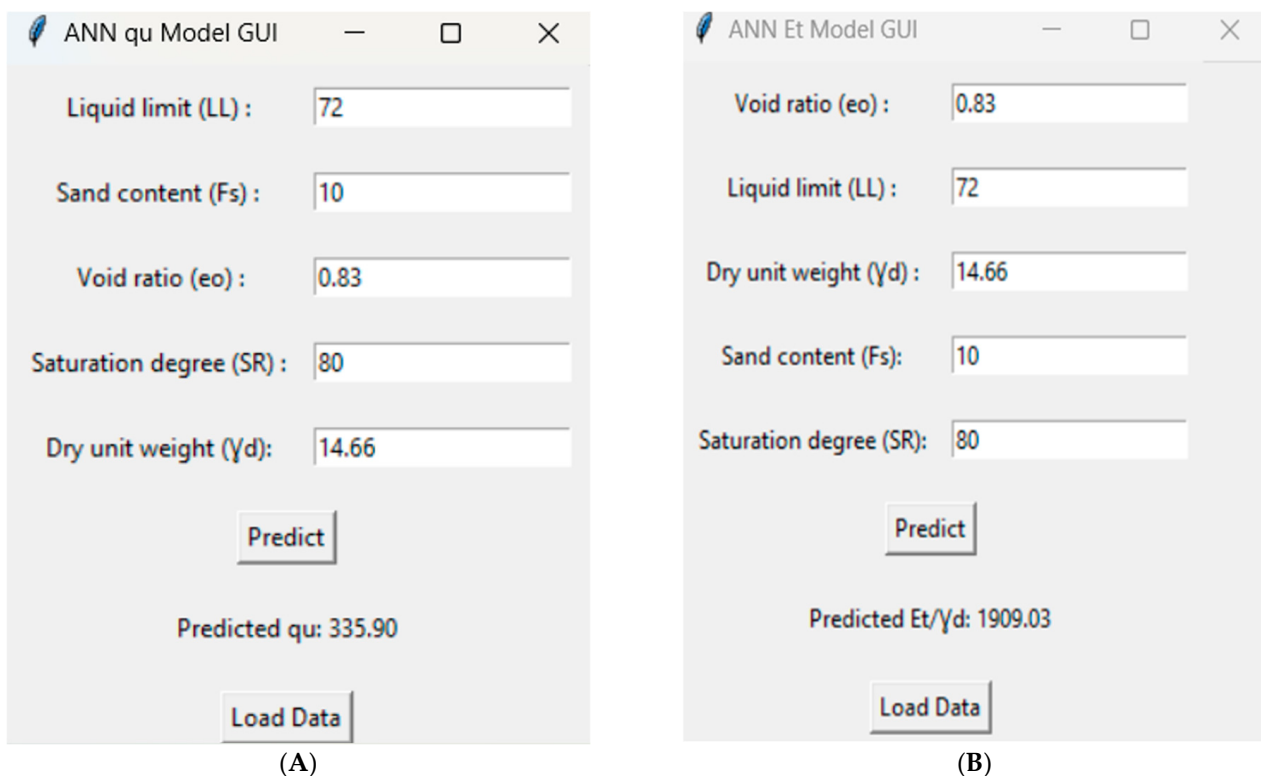


Figure 18. Illustration of user-friendly interface: (A) ANN q_u model, (B) ANN E_t model.

3.4.3. Symbolic Regression Equations

The EUREQA software is used in this study to uncover latent patterns and relationships in the dataset and generate equations for predicting the unconfined compressive strength (q_u) and elasticity modulus (E_t) [98].

Proposed Equations of Elasticity Modulus (E_t)

Equations (6)–(8) are formulated via the utilization of the EUREQA software. It is crucial to acknowledge that as the metric is enhanced, the resulting equations tend to become more intricate, necessitating additional time to achieve optimal performance.

$$E_t = \frac{\gamma_d * LL * SR^2}{e^{\frac{e_0 + SR}{\gamma_d}}} \quad (R^2 = 0.90816; \text{RMSE} = 615.6; \text{MAE} = 495.03) \quad (6)$$

$$E_t = \gamma_d * \sqrt{\frac{e^{e_0 + \gamma_d + \frac{LL}{\gamma_d}}}{e^{\frac{SR}{\gamma_d}}}} \quad (R^2 = 0.93394; \text{RMSE} = 522.13; \text{MAE} = 412.17) \quad (7)$$

$$E_t = \gamma_d * \frac{LL * SR^2 + FS * \gamma_d^2}{SR^{\frac{1}{\gamma_d} + \frac{SR * \ln(e_0)}{\gamma_d^2}} e^{\frac{SR}{\gamma_d} + \frac{e_0 * \gamma_d * FS * SR^2}{e_0 + FS * LL^3}}} \quad (R^2 = 0.9676; \text{RMSE} = 365.54; \text{MAE} = 283.45) \quad (8)$$

where the following are defined:

E_t —Elasticity modulus;

e_0 —Initial void ratio: $e_0 = \frac{\gamma_s}{\gamma_d} - 1$;

γ_d —Dry unit weight kN/m³;

γ_s —Unit weight of solids kN/m³;

SR —Initial degree of saturation %;

FS —Percentage of added sand %;

LL —Liquid limit %.

Figure 19 illustrates the distinction between true and predicted (E_t/γ_d) data points. The data points in the comparative visual representation align more closely with the best-fit line (ideal line) for Equations (6)–(8), respectively, as the metrics R^2 , RMSE, and MAE improve. It should be noted that most data points are located between the 20% error margin lines. However, caution must be exercised when ($E_t/\gamma_d \leq 1000$ m) where there are notable bounds out of the +20% error margin line, so it is preferable to reduce it by 20% to be on the safe side in this case.

Proposed Equations of Unconfined Compression Strength q_u

For the prediction of unconfined compression strength (q_u), Equations (9)–(11) are formulated using the EUREQA software.

$$q_u = \frac{LL^2 * \gamma_d^2}{e_0^3 * SR^2} + \frac{\gamma_d - FS}{e_0} \quad (R^2 = 0.96; \text{RMSE} = 77; \text{MAE} = 51.6) \quad (9)$$

$$q_u = \frac{e_0 * \gamma_d * \gamma_d^{\ln(LL)}}{e_0^3 * SR^2 - SR} + LL + \gamma_d - SR \quad (R^2 = 0.971; \text{RMSE} = 65.66; \text{MAE} = 35.7) \quad (10)$$

$$q_u = \frac{e_0^2 * \gamma_d * \gamma_d^{\ln(LL)}}{e_0^4 * SR^2 + LL - \gamma_d - SR} + LL + \gamma_d - SR \quad (R^2 = 0.983; \text{RMSE} = 50.24; \text{MAE} = 29.9) \quad (11)$$

where the following are defined:

q_u —unconfined compressive strength.

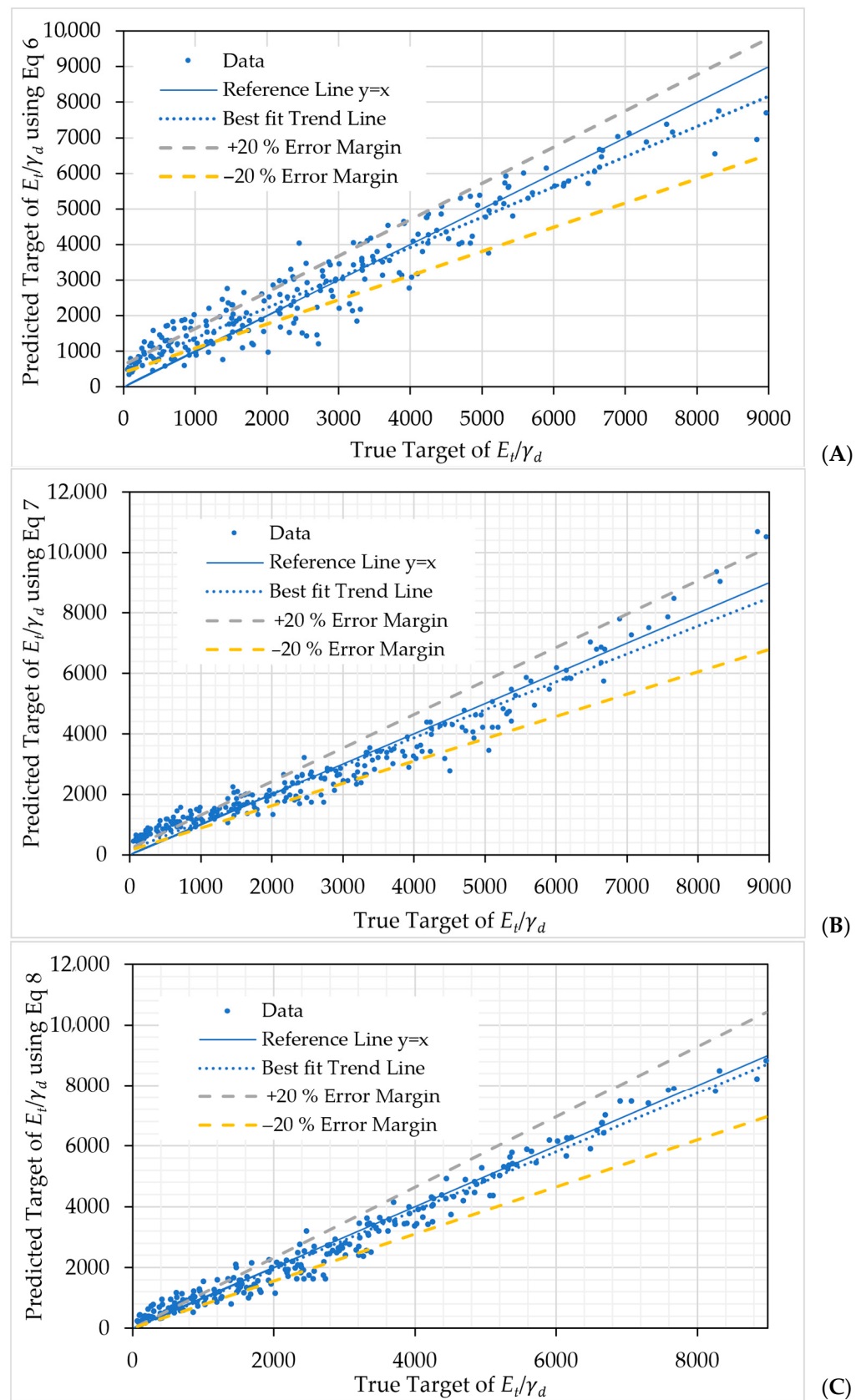


Figure 19. The cross plots of actual and predicted targets for (A) Equation (6), (B) Equation (7), and (C) Equation (8).

Figure 20 shows improved alignment of true and predicted data points with the reference line $y = x$ and best-fit trend line (ideal line) for Equations (9)–(11) as metrics (R^2 , RMSE, and MAE) improve. Equation 11 has the highest metrics, while Equation (9) has the lowest. It should be noted that most data points are located between the 20% error margin lines. This demonstrates that q_u equations are capable of accurately predicting the q_u .

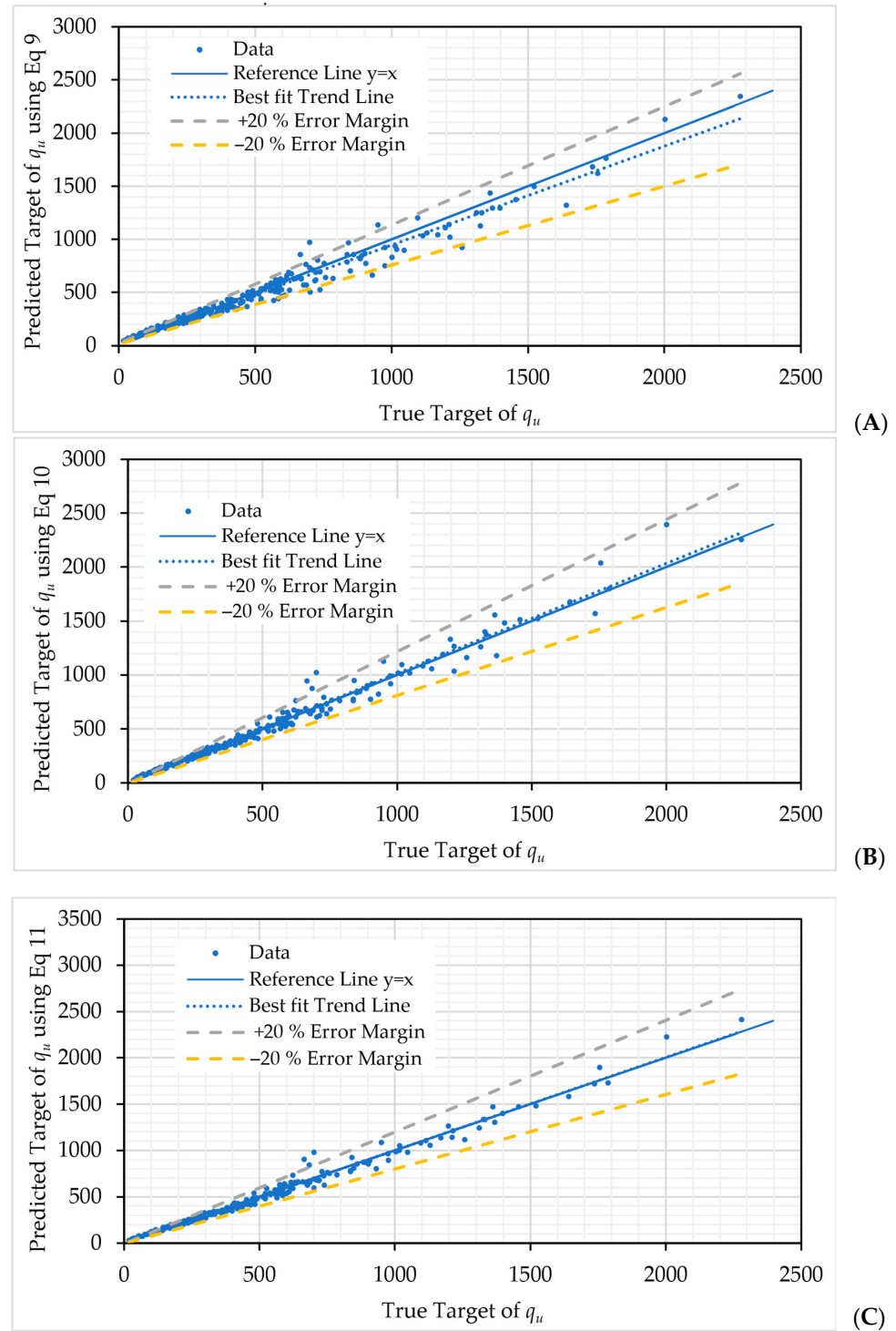


Figure 20. The cross plots of actual and predicted targets for (A) Equation (9), (B) Equation (10), and (C) Equation (11).

It is crucial to highlight that as metrics are improved, the generated equations become more complex, necessitating more time to achieve optimal performance. EUREQA can generate many equations. Equations (6)–(11) are chosen for their practicality and ability to produce favorable metrics.

To ensure the validity of the Equations (6)–(11), the following conditions must be met: sand content below 50%, The Atterberg limits ought to be situated slightly above and in proximity to line A on the Casagrande chart, initial degree of saturation exceeding 50%, and a liquid limit within the range of 40–80%.

When the metrics of ANN and SR are compared (Figure 21), as well as the visual examination of true and predicted targets in Figure 22, artificial neural networks outperform symbolic regression. This observation highlights the superior performance of artificial neural networks in comprehending complex patterns within data, outperforming symbolic regression.

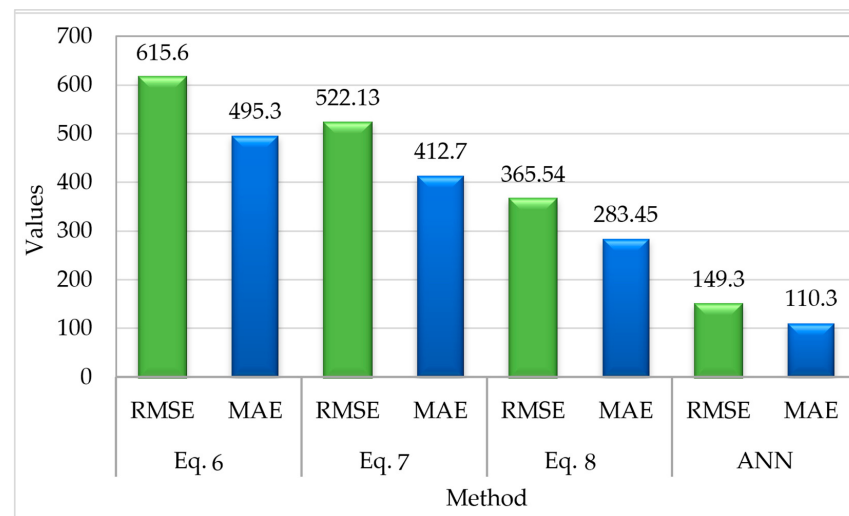


Figure 21. RMSE and MAE metrics ANN (E_t) vs. symbolic regression E_t equations.

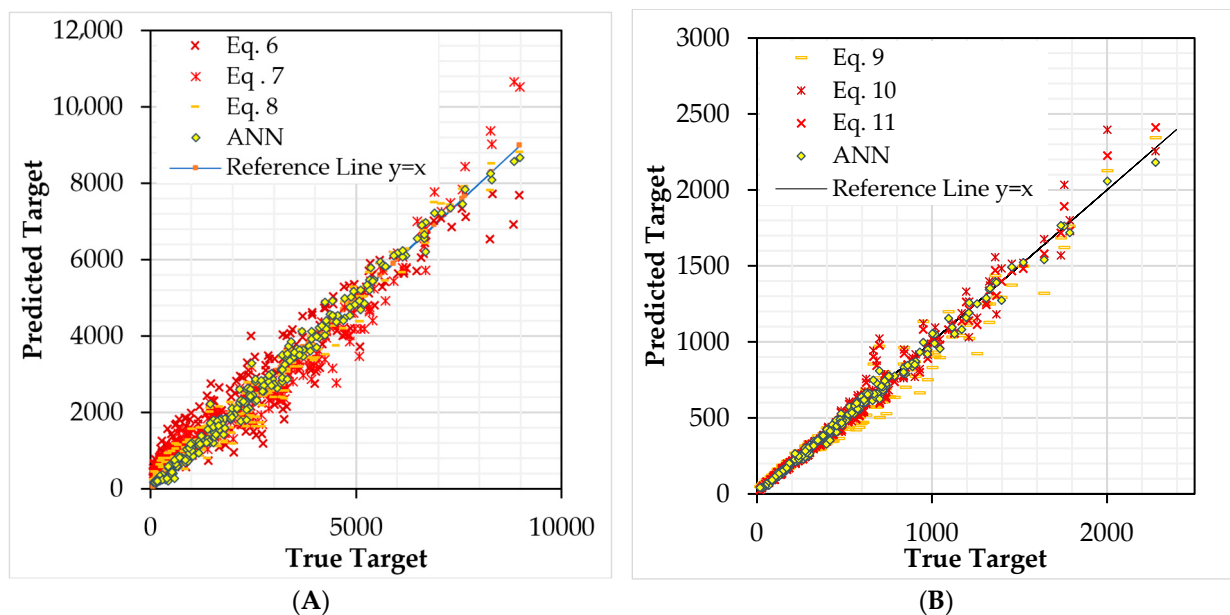


Figure 22. Comparison of true and predicted targets for Symbolic regression and ANN models: (A) E_t / γ_d models, (B) q_u models.

Prior research using single or multiple regression techniques focused on only one or two inputs, including (w_n , LL , and/or γ_d) [1,2]. The highest achieved R-squared values were 0.89 in work by Khalid et al. (2015) and 0.606 in work by Senoon and Hussein, (2018), confirming symbolic regression equations' superior performance over single and multiple regression models. Upon applying the equations from Senoon and Hussein (2018) and Khalid et al. (2015)'s works to our dataset, a large number of data points were predicted to have negative values. This happened as a result of the models in Senoon and Hussein (2018) and Khalid et al. (2015)'s works depending only on one or two inputs. Furthermore, their datasets were not sufficiently explored in terms of expansive soil properties. Senoon and Hussein (2018) concentrated on MH classified soils and specifically avoided expansive soil behavior, whereas Khalid et al. (2015) only addressed a small portion of this soil classification. The power of the equations in our work comes from the ability to accurately predict q_u and E_t in expansive soils that are partially or fully saturated.

4. Feature Importance

Two methods were used to evaluate the importance of the selected variables: the SHapley Additive ExPlanations (SHAP) values [102] and the algorithm's inherent computation of traditional feature importance. In the latter approach, following algorithm training, relative relevance scores for each input characteristic can be derived [103]. A higher score indicates a greater impact on prediction. The analysis results are shown in Table 10, indicating that across all models, saturation degree holds the highest influence on q_u and E_t , followed by void ratio and dry unit weight, with liquid limit and sand percentage having the least influence. Notably, XGBoost places saturation degree on top, followed by void ratio and liquid limit, with dry unit weight and sand percentage at the bottom.

Table 10. Analysis of the importance of variables.

Model	DTR		RFR		GBR		XGBoost		Correlation According to SHAP	Correlation According to Previous Research
	q_u	E_t	q_u	E_t	q_u	E_t	q_u	E_t		
SR	0.406	0.677	0.429	0.67	0.394	0.662	0.291	0.577	Negative	Negative [104,105]
e_0	0.163	0.127	0.263	0.147	0.325	0.162	0.332	0.122	Negative	Negative [106]
γ_d	0.294	0.125	0.204	0.11	0.154	0.101	0.097	0.135	Positive	Positive [106,107]
LL	0.089	0.052	0.052	0.033	0.039	0.03	0.213	0.118	Positive	Positive [106,107]
Fs	0.048	0.019	0.052	0.04	0.088	0.045	0.067	0.048	Negative	Negative [108]

Additionally, the SHAP method was employed, providing a consistent approach to explaining the output of any tree-based model and offering distinct advantages over other methods. Figures 23 and 24 illustrates the combination of feature importance with feature effects. Each point in the graph represents a SHAP value for a feature and prediction, with the y-axis displaying features in descending order of importance. SHAP values are plotted along the x-axis, where a value of zero indicates no contribution to the prediction, with increasing contributions as the SHAP value moves away from zero. Furthermore, color gradation from low (blue) to high (red) feature values provides insight into the directional impact of features. The distribution of blue and red dots offers an overall understanding of the directionality impact of features [55].

According to the random forest regression (RFR) model, the degree of saturation emerges as the most influential variable impacting both q_u and E_t , showing a negative correlation, consistent with findings from [104,105]. Following closely is the void ratio, exhibiting a smaller yet still significant negative correlation, as observed in the study by [106].

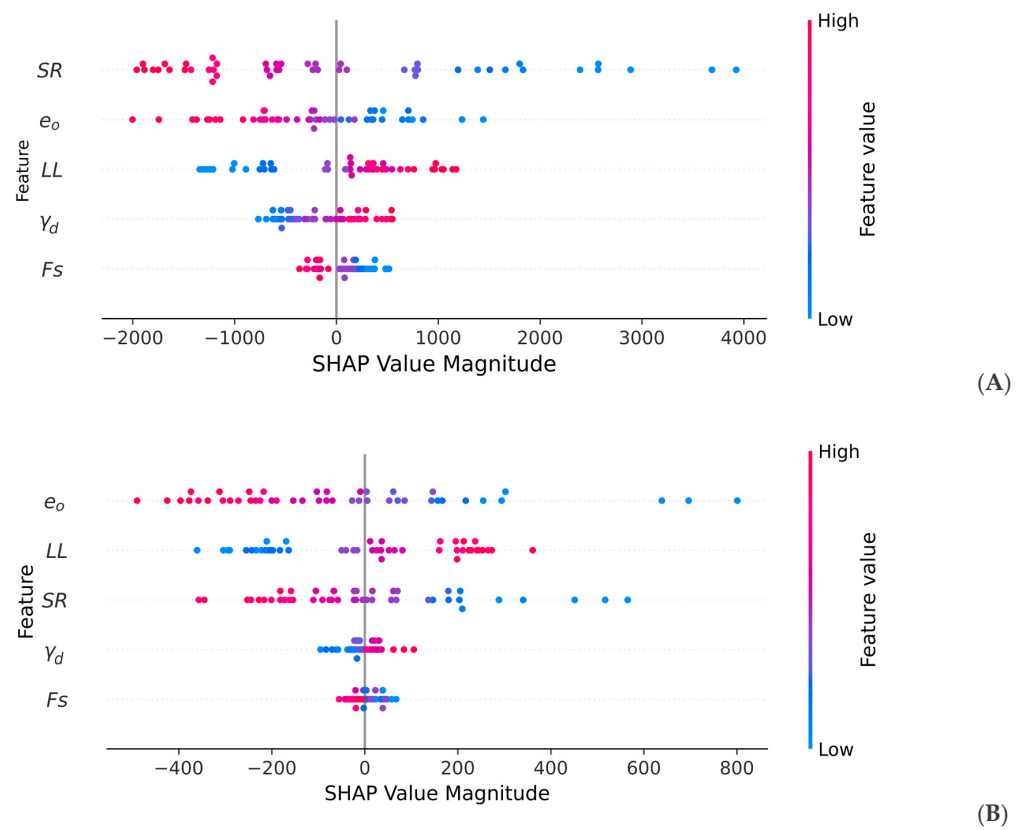


Figure 23. Relative importance of features with SHAP values in XGBoost model for (A) E_t and (B) q_u .

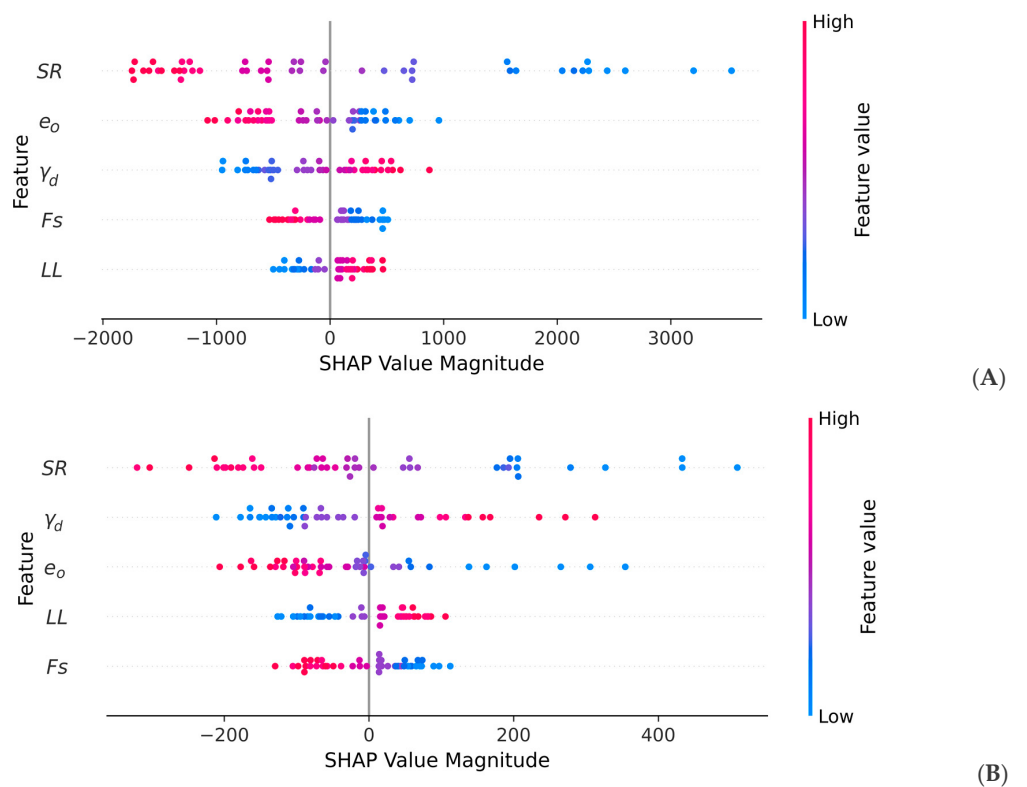


Figure 24. Relative importance of features with SHAP values in RFR model for (A) E_t and (B) q_u .

Additionally, the dry unit weight emerges as another influential factor affecting q_u and E_t , demonstrating a positive correlation, as indicated in previous studies such as those by [106,107]. The sand content parameter shows a negative correlation, aligning with the findings of [108]. Conversely, the liquid limit variable demonstrates the least impact on both q_u and E_t , with a positive correlation, consistent with studies by [106,107].

These observations suggest that the models have effectively captured the relationships between each independent variable and q_u and E_t , in line with existing technical literature. A concise summary of the feature importance analysis conducted using the two approaches is presented in Table 10.

5. Conclusions

This study investigates the effect of sand in expansive soil on unconfined test parameters and presents a novel application of machine learning models to clarify the intricate behaviors of expansive soils. Focusing on the impact of sand content, saturation level, and dry density, the following conclusions were made:

- The modulus of elasticity (E_t) and unconfined compressive strength (q_u) are positively correlated with the soil's dry unit weight.
- Increasing sand content, while keeping dry unit weight constant, reduces modulus of elasticity and unconfined compressive strength due to expanded voids decreasing suction forces.
- A rise in the degree of saturation results in a notable but nonlinear decrease in both modulus of elasticity and unconfined compressive strength, becoming less pronounced at higher saturation levels.
- For samples at maximum dry unit weight, adding sand initially increases strength and elasticity up to a threshold of 30% sand content, beyond which the effect reverses.
- Optimal sand content for maximum strength and elasticity at full saturation is 30%.

Further into the study, equations derived from symbolic regression are introduced to predict the modulus of elasticity and unconfined compressive strength. These equations consider a range of factors including void ratio, degree of saturation, liquid limit, and sand percentage, thereby providing a robust predictive framework for soil properties.

Feature importance analysis reveals the following:

- Degree of saturation is the most influential variable, showing a negative correlation with q_u and E_t .
- Void ratio shows a significant negative correlation.
- Dry unit weight demonstrates a positive correlation.
- Sand content exhibits a negative correlation.
- Liquid limit has the least impact, with a positive correlation.

The ANN-GMDH model, integrating GMDH features, outperforms other models in predicting soil behavior. Both ANN and SVR models are effective, with ANNs excelling in capturing complex patterns.

This study's limitations arise from depending on a specific dataset tailored to a particular range for each parameter, as indicated in Table 4. Additionally, Atterberg limits should be positioned slightly above and near line A on the Casagrande chart. This commitment to achieving accurate predictions may, however, pose constraints on the generalizability of the findings.

Although the evaluated models produced reasonable results, there is still room for improvement. Future work could explore new methods such as deep learning or reinforcement learning. Additionally, enhancing the models' prediction capabilities might require incorporating other characteristics, such as the plasticity index, clay percentage, maximum dry density, and optimum saturation degree. Expanding the dataset and refining risk mitigation strategies related to expansive soils should be a focus of future research, contributing to the continuous advancement of geotechnical engineering practices. These enhancements aim to further solidify the applicability and effectiveness of the proposed

approach in real-world scenarios. In conclusion, while this study provides valuable insights, addressing these limitations via broader datasets and refined strategies will be crucial for improving the reliability and robustness of predictive models in geotechnical engineering.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/app14114819/s1>.

Author Contributions: Conceptualization, A.A., M.O.A. and R.P.R.; investigation, A.A. and M.O.A.; experimental work, A.A.; writing—original draft preparation, A.A., H.H.H. and C.L.; modeling, A.A., R.P.R., H.H.H. and C.L.; writing—review and editing, R.P.R., M.O.A. and A.A.; supervision, R.P.R. and M.O.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding. Funding for open access granted by Szechenyi István University (SZE).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

- Senoon, A.-A.; Hussein, M. Correlation between Unconfined Compression Strength (UCS) and Index Properties of Soil in Assiut Governorate, Egypt. In Proceedings of the Fifteenth International Conference on Structural and Geotechnical Engineering, Cairo, Egypt, December 2018; Advances In Construction Techniques. Ain Shams University, Faculty of Engineering Structural Engineering Department: Cairo, Egypt, 2018. Available online: [https://staffsites.sohag-univ.edu.eg/uploads/470/1546201444%20-%20Correlation%20between%20Unconfined%20Compression%20Strength%20\(UCS\)%20and%20Index%20Properties%20of%20Soil%20in%20Assiut%20Governorate,%20Egypt.pdf](https://staffsites.sohag-univ.edu.eg/uploads/470/1546201444%20-%20Correlation%20between%20Unconfined%20Compression%20Strength%20(UCS)%20and%20Index%20Properties%20of%20Soil%20in%20Assiut%20Governorate,%20Egypt.pdf) (accessed on 29 April 2024).
- Khalid, U.; Ur-Rehman, Z.; Farooq, K.; Mujtaba, H. Prediction of Unconfined Compressive Strength from Index Properties of Soils. *Sci. Int.* **2015**, *27*, 4071–4075.
- Steinberg, M.L. *Controlling Expansive Soil Destructiveness by Deep Vertical Geomembranes on Four Highways*; Transportation Research Board: Washington, DC, USA, 1985; ISBN 0309039231. Available online: <https://trid.trb.org/View/272080> (accessed on 29 April 2024).
- Goodarzi, A.R.; Akbari, H.R.; Salimi, M. Enhanced Stabilization of Highly Expansive Clays by Mixing Cement and Silica Fume. *Appl. Clay Sci.* **2016**, *132*, 675–684. [\[CrossRef\]](#)
- Kolay, P.K.; Ramesh, K.C. Reduction of Expansive Index, Swelling and Compression Behavior of Kaolinite and Bentonite Clay with Sand and Class C Fly Ash. *Geotech. Geol. Eng.* **2016**, *34*, 87–101. [\[CrossRef\]](#)
- Salimi, M.; Ilkhani, M.; Vakili, A.H. Stabilization Treatment of Na-Montmorillonite with Binary Mixtures of Lime and Steelmaking Slag. *Int. J. Geotech. Eng.* **2018**, *14*, 295–301. [\[CrossRef\]](#)
- Nelson, J.; Miller, D.J. *Expansive Soils: Problems and Practice in Foundation and Pavement Engineering*; John Wiley & Sons: New York, NY, USA, 1997; ISBN 0471181145.
- Kaoua, F.; Derriche, Z.; Laradi, N. Contribution à l'étude de La Stabilisation Des Sols Gonflants Par Ajouts de Sable. *Algérie Equip. Rev. De L'école Natl. Des Trav. Publics* **1994**, *15*, 12–15.
- Bahia, L.; Ramdane, B. SAND: An Additive for Stabilization of Swelling Clay Soils. *Int. J. Geosci.* **2012**, *3*, 719–725. [\[CrossRef\]](#)
- Nagaraj, H.B. Influence of Gradation and Proportion of Sand on Stress–Strain Behavior of Clay–Sand Mixtures. *Int. J. Geo-Eng.* **2016**, *7*, 19. [\[CrossRef\]](#)
- Deng, Y.; Wu, Z.; Cui, Y.; Liu, S.; Wang, Q. Sand Fraction Effect on Hydro-Mechanical Behavior of Sand-Clay Mixture. *Appl. Clay Sci.* **2017**, *135*, 355–361. [\[CrossRef\]](#)
- Phanikumar, B.R.; Dembla, S.; Yatindra, A. Swelling Behaviour of an Expansive Clay Blended With Fine Sand and Fly Ash. *Geotech. Geol. Eng.* **2021**, *39*, 583–591. [\[CrossRef\]](#)
- Alnmr, A.; Ray, R. Investigating the Impact of Varying Sand Content on the Physical Characteristics of Expansive Clay Soils from Syria. *Geotech. Geol. Eng.* **2023**, *42*, 2675–2691. [\[CrossRef\]](#)
- Alnmr, A.; Ray, R.P.; Alsirawan, R. A State-of-the-Art Review and Numerical Study of Reinforced Expansive Soil with Granular Anchor Piles and Helical Piles. *Sustainability* **2023**, *15*, 2802. [\[CrossRef\]](#)
- Alnmr, A.; Ray, R.P.; Alsirawan, R. Comparative Analysis of Helical Piles and Granular Anchor Piles for Foundation Stabilization in Expansive Soil: A 3D Numerical Study. *Sustainability* **2023**, *15*, 11975. [\[CrossRef\]](#)
- Khemissa, M.; Mahamedi, A.; Mekki, L. Problematic Soil Mechanics in the Algerian Arid and Semi-Arid Regions: Case of M'sila Expansive Clays. *J. Appl. Eng. Sci. Technol.* **2015**, *1*, 37–41.

17. Roy, T.K. Influence of Sand on Strength Characteristics of Cohesive Soil for Using as Subgrade of Road. *Procedia Soc. Behav. Sci.* **2013**, *104*, 218–224. [CrossRef]
18. Yao, Y.; Li, J.; Ni, J.; Liang, C.; Zhang, A. Effects of Gravel Content and Shape on Shear Behaviour of Soil-Rock Mixture: Experiment and DEM Modelling. *Comput. Geotech.* **2022**, *141*, 104476. [CrossRef]
19. Alnmr, A.; Alzawi, M.; Abdullah, S.; Ibraheem, J. Prediction by Improving the Characteristics of Compaction and Strength in the Expansive Soils According to Its Content of Sand. *Tishreen Univ. J. -Eng. Sci. Ser.* **2020**, *42*, 203–219.
20. Prasad, C.R.V.; Sharma, R.K. Influence of Sand and Fly Ash on Clayey Soil Stabilization. *IOSR J. Mech. Civ. Eng.* **2014**, *334*, 36–40.
21. Al Rawi, O.S.; Assaf, M.N.; Hussein, N.M. Effect of Sand Additives on the Engineering Properties of Fine Grained Soils. *ARPN J. Eng. Appl. Sci.* **2018**, *13*, 3197–3206.
22. Alnmr, A. Material Models to Study the Effect of Fines in Sandy Soils Based on Experimental and Numerical Results. *Acta Tech. Jaurinensis* **2021**, *14*, 651–680. [CrossRef]
23. Alnmr, A.; Ray, R.P. Review of the Effect of Sand on the Behavior of Expansive Clayey Soils. *Acta Tech. Jaurinensis* **2021**, *14*, 521–552. [CrossRef]
24. Alnmr, A.N.; Omran Alzawi, M.; Abdallah, S. Study the Effect of the Percentage of Fines on the Compressibility Behavior of Sandy Soils. *Tishreen Univ. J. Res. Sci. Stud.* **2017**, *4*, 2079–3001. Available online: <http://www.journal.tishreen.edu.sy/index.php/engscnc/article/view/3677/0> (accessed on 29 April 2024).
25. Alnmr, A.; Ray, R. Numerical Simulation of Replacement Method to Improve Unsaturated Expansive Soil. *Pollack Period.* **2023**, *18*, 41–47. [CrossRef]
26. Alnmr, A.; Alzawi, M.O.; Ray, R.; Abdullah, S.; Ibraheem, J. Experimental Investigation of the Soil-Water Characteristic Curves (SWCC) of Expansive Soil: Effects of Sand Content, Initial Saturation, and Initial Dry Unit Weight. *Water* **2024**, *16*, 627. [CrossRef]
27. Khan, F.S.; Azam, S.; Raghunandan, M.E.; Clark, R. Compressive Strength of Compacted Clay-Sand Mixes. *Adv. Mater. Sci. Eng.* **2014**, *2014*, 1–6. [CrossRef]
28. Mallikarjuna Rao, K.; Subba Rao, G.V.R. Influence of Coarse Fraction on Characteristics of Expansive Soil-Sand Mixtures. *Int. J. Geosynth. Ground Eng.* **2018**, *4*, 19. [CrossRef]
29. Fredlund, D.G.; Rahardjo, H. *Soil Mechanics for Unsaturated Soils*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 1993; ISBN 9780470172759.
30. Yao, Y.; Ni, J.; Li, J. Stress-Dependent Water Retention of Granite Residual Soil and Its Implications for Ground Settlement. *Comput. Geotech.* **2021**, *129*, 103835. [CrossRef]
31. Wong, J.T.F.; Chow, K.L.; Chen, X.W.; Ng, C.W.W.; Wong, M.H. Effects of Biochar on Soil Water Retention Curves of Compacted Clay during Wetting and Drying. *Biochar* **2022**, *4*, 4. [CrossRef]
32. Vanapalli, S.K.; Fredlund, D.G.; Pufahl, D.E.; Clifton, A.W. Model for the Prediction of Shear Strength with Respect to Soil Suction. *Can. Geotech. J.* **2011**, *33*, 379–392. [CrossRef]
33. Vanapalli, S.K.; Fredlund, D.G.; Pufahl, D.E. Comparison of Saturated-Unsaturated Shear Strength and Hydraulic Conductivity Behavior of a Compacted Sandy-Clay Till. In *Proceedings of the Proceedings of the 50th Canadian Geotechnical Conference*; Ottawa, ON, USA, 20 October 1997; pp. 20–22.
34. Ling, Q.; Zhang, Q.; Wei, Y.; Kong, L.; Zhu, L. Slope Reliability Evaluation Based on Multi-Objective Grey Wolf Optimization-Multi-Kernel-Based Extreme Learning Machine Agent Model. *Bull. Eng. Geol. Environ.* **2021**, *80*, 2011–2024. [CrossRef]
35. Liu, L.; Zhang, S.; Cheng, Y.M.; Liang, L. Advanced Reliability Analysis of Slopes in Spatially Variable Soils Using Multivariate Adaptive Regression Splines. *Geosci. Front.* **2019**, *10*, 671–682. [CrossRef]
36. Wang, H.; Zhang, L.; Yin, K.; Luo, H.; Li, J. Landslide Identification Using Machine Learning. *Geosci. Front.* **2021**, *12*, 351–364. [CrossRef]
37. Ray, R.; Kumar, D.; Samui, P.; Roy, L.B.; Goh, A.T.C.; Zhang, W. Application of Soft Computing Techniques for Shallow Foundation Reliability in Geotechnical Engineering. *Geosci. Front.* **2021**, *12*, 375–383. [CrossRef]
38. Wang, L.; Wu, C.; Tang, L.; Zhang, W.; Lacasse, S.; Liu, H.; Gao, L. Efficient Reliability Analysis of Earth Dam Slope Stability Using Extreme Gradient Boosting Method. *Acta Geotech.* **2020**, *15*, 3135–3150. [CrossRef] [PubMed]
39. Li, S.; Zhao, H.B.; Ru, Z. Slope Reliability Analysis by Updated Support Vector Machine and Monte Carlo Simulation. *Nat. Hazards* **2013**, *65*, 707–722. [CrossRef]
40. Kumar, M.; Samui, P. Reliability Analysis of Pile Foundation Using ELM and MARS. *Geotech. Geol. Eng.* **2019**, *37*, 3447–3457. [CrossRef]
41. Shen, H.; Li, J.; Wang, S.; Xie, Z. Prediction of Load-Displacement Performance of Grouted Anchors in Weathered Granites Using FastICA-MARS as a Novel Model. *Geosci. Front.* **2021**, *12*, 415–423. [CrossRef]
42. Li, T.Z.; Pan, Q.; Dias, D. Active Learning Relevant Vector Machine for Reliability Analysis. *Appl. Math. Model.* **2021**, *89*, 381–399. [CrossRef]
43. Kardani, N.; Aminpour, M.; Nouman Amjad Raja, M.; Kumar, G.; Bardhan, A.; Nazem, M. Prediction of the Resilient Modulus of Compacted Subgrade Soils Using Ensemble Machine Learning Methods. *Transp. Geotech.* **2022**, *36*, 100827. [CrossRef]
44. Yi, P.; Wei, K.; Kong, X.; Zhu, Z. Cumulative PSO-Kriging Model for Slope Reliability Analysis. *Probabilistic Eng. Mech.* **2015**, *39*, 39–45. [CrossRef]
45. Zhang, W.; Wu, C.; Tang, L.; Gu, X.; Wang, L. Efficient Time-Variant Reliability Analysis of Bazimen Landslide in the Three Gorges Reservoir Area Using XGBoost and LightGBM Algorithms. *Gondwana Res.* **2022**, *123*, 41–53. [CrossRef]

46. Shi, C.; Wang, Y. Development of Subsurface Geological Cross-Section from Limited Site-Specific Boreholes and Prior Geological Knowledge Using Iterative Convolution XGBoost. *J. Geotech. Geoenviron. Eng.* **2021**, *147*, 04021082. [\[CrossRef\]](#)
47. Wang, M.-X.; Huang, D.; Wang, G.; Li, D.-Q. SS-XGBoost: A Machine Learning Framework for Predicting Newmark Sliding Displacements of Slopes. *J. Geotech. Geoenviron. Eng.* **2020**, *146*, 04020074. [\[CrossRef\]](#)
48. Alsirawan, R.; Sheble, A.; Alnmr, A. Two-Dimensional Numerical Analysis for TBM Tunneling-Induced Structure Settlement: A Proposed Modeling Method and Parametric Study. *Infrastructures* **2023**, *8*, 88. [\[CrossRef\]](#)
49. Alnmr, A. Impact of Soil Composition on Maximum Depth of Wetting in Expansive Soils. *Pollack Period.* **2023**, *19*, 85–92. [\[CrossRef\]](#)
50. Merouane, F.Z.; Mamoune, S.M.A. Prediction of Swelling Parameters of Two Clayey Soils from Algeria Using Artificial Neural Networks. *Math. Model. Civ. Eng.* **2018**, *14*, 11–26. [\[CrossRef\]](#)
51. Dutta, R.K.; Singh, A.; Gnananandarao, T. Prediction of Free Swell Index for the Expansive Soil Using Artificial Neural Networks. *J. Soft Comput. Civ. Eng.* **2019**, *3*, 47–62. [\[CrossRef\]](#)
52. Cho, S.E. Probabilistic Stability Analyses of Slopes Using the ANN-Based Response Surface. *Comput. Geotech.* **2009**, *36*, 787–797. [\[CrossRef\]](#)
53. Wang, Z.Z.; Xiao, C.; Goh, S.H.; Deng, M.-X. Metamodel-Based Reliability Analysis in Spatially Variable Soils Using Convolutional Neural Networks. *J. Geotech. Geoenviron. Eng.* **2021**, *147*, 04021003. [\[CrossRef\]](#)
54. Jalal, F.E.; Xu, Y.; Iqbal, M.; Javed, M.F.; Jamhiri, B. Predictive Modeling of Swell-Strength of Expansive Soils Using Artificial Intelligence Approaches: ANN, ANFIS and GEP. *J. Environ. Manag.* **2021**, *289*, 112420. [\[CrossRef\]](#)
55. Díaz, E.; Salamanca-Medina, E.L.; Tomás, R. Assessment of Compressive Strength of Jet Grouting by Machine Learning. *J. Rock Mech. Geotech. Eng.* **2023**, *16*, 102–111. [\[CrossRef\]](#)
56. Wu, K.; Meng, Q.; Li, R.; Luo, L.; Ke, Q.; Wang, C.; Ma, C. A Machine Learning-Based Strategy for Predicting the Mechanical Strength of Coral Reef Limestone Using X-Ray Computed Tomography. *J. Rock Mech. Geotech. Eng.* **2023**. [\[CrossRef\]](#)
57. Chang, Z.; Catani, F.; Huang, F.; Liu, G.; Meena, S.R.; Huang, J.; Zhou, C. Landslide Susceptibility Prediction Using Slope Unit-Based Machine Learning Models Considering the Heterogeneity of Conditioning Factors. *J. Rock Mech. Geotech. Eng.* **2023**, *15*, 1127–1143. [\[CrossRef\]](#)
58. Praks, P.; Brkić, D. Symbolic Regression-Based Genetic Approximations of the Colebrook Equation for Flow Friction. *Water* **2018**, *10*, 1175. [\[CrossRef\]](#)
59. Latif, Q.B.A.I.; Memon, Z.A.; Mahmood, Z.; Qureshi, M.U.; Milad, A. A Machine Learning Model for the Prediction of Concrete Penetration by the Ogive Nose Rigid Projectile. *Appl. Sci.* **2022**, *12*, 2040. [\[CrossRef\]](#)
60. Hamidia, M.; Ganjizadeh, A. Post-Earthquake Damage Evaluation of Non-Ductile RC Moment Frames Using Surface Crack Patterns. *Struct. Control Health Monit.* **2022**, *29*, e3024. [\[CrossRef\]](#)
61. Naser, M.Z. Heuristic Machine Cognition to Predict Fire-Induced Spalling and Fire Resistance of Concrete Structures. *Autom. Constr.* **2019**, *106*, 102916. [\[CrossRef\]](#)
62. Rezaei, H.; Zarfam, P.; Golafshani, E.M.; Amiri, G.G. Seismic Fragility Analysis of RC Box-Girder Bridges Based on Symbolic Regression Method. *Structures* **2022**, *38*, 306–322. [\[CrossRef\]](#)
63. Gan, L.; Wu, H.; Zhong, Z. Integration of Symbolic Regression and Domain Knowledge for Interpretable Modeling of Remaining Fatigue Life under Multistep Loading. *Int. J. Fatigue* **2022**, *161*, 106889. [\[CrossRef\]](#)
64. Ben Chaabene, W.; Nehdi, M.L. Genetic Programming Based Symbolic Regression for Shear Capacity Prediction of SFRC Beams. *Constr. Build. Mater.* **2021**, *280*, 122523. [\[CrossRef\]](#)
65. Amanabadi, S.; Vazirinia, M.; Vereecken, H.; Vakilian, K.A.; Mohammadi, M.H. Comparative Study of Statistical, Numerical and Machine Learning-Based Pedotransfer Functions of Water Retention Curve with Particle Size Distribution Data. *Eurasian Soil Sci.* **2019**, *52*, 1555–1571. [\[CrossRef\]](#)
66. Ikizler, S.B.; Aytekin, M.; Vekli, M.; Kocabaş, F. Prediction of Swelling Pressures of Expansive Soils Using Artificial Neural Networks. *Adv. Eng. Softw.* **2010**, *41*, 647–655. [\[CrossRef\]](#)
67. Bachir, R.; Mohammed, A.M.S.; Habib, T. Using Artificial Neural Networks Approach to Estimate Compressive Strength for Rubberized Concrete. *Period. Polytech. Civ. Eng.* **2018**, *62*, 858–865. [\[CrossRef\]](#)
68. Najjar, Y.M.; Basheer, I.A.; McReynolds, R. Neural Modeling of Kansas Soil Swelling. *Transp. Res. Rec. J. Transp. Res. Board* **1996**, *1526*, 14–19. [\[CrossRef\]](#)
69. Najjar, Y.M.; Basheer, I.A. *Modeling of Soil Swelling via Regression and Neural Network Approaches*; Kansas Department of Transportation: Topeka, KS, USA, 1998. Available online: <https://trid.trb.org/View/473791> (accessed on 29 April 2024).
70. Doris, J.J.; Rizzo, D.M.; Dewoolkar, M.M. Forecasting Vertical Ground Surface Movement from Shrinking/Swelling Soils with Artificial Neural Networks. *Int. J. Numer. Anal. Methods Geomech.* **2008**, *32*, 1229–1245. [\[CrossRef\]](#)
71. Ashayeri, I.; Yasrebi, S. Free-Swell and Swelling Pressure of Unsaturated Compacted Clays; Experiments and Neural Networks Modeling. *Geotech. Geol. Eng.* **2009**, *27*, 137–153. [\[CrossRef\]](#)
72. Erzin, Y.; Güneş, N. The Prediction of Swell Percent and Swell Pressure by Using Neural Networks. *Math. Comput. Appl.* **2011**, *16*, 425–436. [\[CrossRef\]](#)
73. Ikeagwuani, C.C. Estimation of Modified Expansive Soil CBR with Multivariate Adaptive Regression Splines, Random Forest and Gradient Boosting Machine. *Innov. Infrastruct. Solut.* **2021**, *6*, 199. [\[CrossRef\]](#)
74. Eyo, E.U.; Abbey, S.J.; Lawrence, T.T.; Tetteh, F.K. Improved Prediction of Clay Soil Expansion Using Machine Learning Algorithms and Meta-Heuristic Dichotomous Ensemble Classifiers. *Geosci. Front.* **2022**, *13*, 101296. [\[CrossRef\]](#)

75. Benbouras, M.A.; Petrisor, A.I. Prediction of Swelling Index Using Advanced Machine Learning Techniques for Cohesive Soils. *Appl. Sci.* **2021**, *11*, 536. [CrossRef]
76. Chen, Y.; Xu, Y.; Jamhiri, B.; Wang, L.; Li, T. Predicting Uniaxial Tensile Strength of Expansive Soil with Ensemble Learning Methods. *Comput. Geotech.* **2022**, *150*, 104904. [CrossRef]
77. ASTM D2487-17e1; Standard Practice for Classification of Soils for Engineering Purposes (Unified Soil Classification System). ASTM International: West Conshohocken, PA, USA, 2017. [CrossRef]
78. AASHTO Standard Method of Test for The Classification of Soils and Soil Aggregate Mixtures for Highway Construction Purposes, Test Designation M145-91, Standard Specifications for Transportation Materials and Methods of Sampling and Testing. 2002. Available online: https://store.accuristech.com/standards/aashto-m-145-91-2021?product_id=1320718 (accessed on 29 April 2024).
79. Casagrande, A. Plasticity Chart for the Classification of Cohesive Soils. *Transactions Am. Soc. Civ. Eng.* **1948**, *113*, 901–930. [CrossRef]
80. Holtz, R.D.; Kovacs, W.D. Kansas Geotechnical Survey. *Curr. Res. Earth Science. Bull.* **1981**, *244*.
81. Alnmr, A.; Ray, R.; Omran Alzawi, M. A Novel Approach to Swell Mitigation: Machine-Learning-Powered Optimal Unit Weight and Stress Prediction in Expansive Soils. *Appl. Sci.* **2024**, *14*, 1411. [CrossRef]
82. Hasanipanah, M.; Faradonbeh, R.S.; Amnieh, H.B.; Armaghani, D.J.; Monjezi, M. Forecasting Blast-Induced Ground Vibration Developing a CART Model. *Eng. Comput.* **2017**, *33*, 307–316. [CrossRef]
83. Monjezi, M.; Baghestani, M.; Shirani Faradonbeh, R.; Pourghasemi Saghand, M.; Jahed Armaghani, D. Modification and Prediction of Blast-Induced Ground Vibrations Based on Both Empirical and Computational Techniques. *Eng. Comput.* **2016**, *32*, 717–728. [CrossRef]
84. Singh, R.; Umrao, R.K.; Ahmad, M.; Ansari, M.K.; Sharma, L.K.; Singh, T.N. Prediction of Geomechanical Parameters Using Soft Computing and Multiple Regression Approach. *Measurement* **2017**, *99*, 108–119. [CrossRef]
85. Tso, G.K.F.; Yau, K.K.W. Predicting Electricity Energy Consumption: A Comparison of Regression Analysis, Decision Tree and Neural Networks. *Energy* **2007**, *32*, 1761–1768. [CrossRef]
86. Rodriguez-Galiano, V.; Sanchez-Castillo, M.; Chica-Olmo, M.; Chica-Rivas, M. Machine Learning Predictive Models for Mineral Prospectivity: An Evaluation of Neural Networks, Random Forest, Regression Trees and Support Vector Machines. *Ore Geol. Rev.* **2015**, *71*, 804–818. [CrossRef]
87. Friedman, J.H. Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
88. Bardhan, A.; Kardani, N.; Alzo'ubi, A.K.; Roy, B.; Samui, P.; Gandomi, A.H. Novel Integration of Extreme Learning Machine and Improved Harris Hawks Optimization with Particle Swarm Optimization-Based Mutation for Predicting Soil Consolidation Parameter. *J. Rock Mech. Geotech. Eng.* **2022**, *14*, 1588–1608. [CrossRef]
89. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; ACM: New York, NY, USA, 2016.
90. Lai, V.; Ahmed, A.N.; Malek, M.A.; Afan, H.A.; Ibrahim, R.K.; El-Shafie, A.; El-Shafie, A. Modeling the Nonlinearity of Sea Level Oscillations in the Malaysian Coastal Areas Using Machine Learning Algorithms. *Sustainability* **2019**, *11*, 4643. [CrossRef]
91. Wu, J.; Liu, H.; Wei, G.; Song, T.; Zhang, C.; Zhou, H. Flash Flood Forecasting Using Support Vector Regression Model in a Small Mountainous Catchment. *Water* **2019**, *11*, 1327. [CrossRef]
92. Ibrahim Ahmed Osman, A.; Najah Ahmed, A.; Chow, M.F.; Feng Huang, Y.; El-Shafie, A. Extreme Gradient Boosting (Xgboost) Model to Predict the Groundwater Levels in Selangor Malaysia. *Ain Shams Eng. J.* **2021**, *12*, 1545–1556. [CrossRef]
93. Ali, I.; Alharbi, O.M.L.; Alothman, Z.A.; Badjah, A.Y.; Alwarthan, A.; Basheer, A.A. Artificial Neural Network Modelling of Amido Black Dye Sorption on Iron Composite Nano Material: Kinetics and Thermodynamics Studies. *J. Mol. Liq.* **2018**, *250*, 1–8. [CrossRef]
94. Zhao, L.; Guo, Y.; Mohammadian, E.; Hadavimoghaddam, F.; Jafari, M.; Kheirollahi, M.; Rozhenko, A.; Liu, B. Modeling Permeability Using Advanced White-Box Machine Learning Technique: Application to a Heterogeneous Carbonate Reservoir. *ACS Omega* **2023**, *8*, 22922–22933. [CrossRef]
95. Bogachev, D.A.; Volkovich, Y.M.; Sosenkin, V.E.; Podgornova, O.A.; Kosova, N.V. Prediction of Permeability Using Group Method of Data Handling (GMDH) Neural Network from Well Log Data. *Energies* **2020**, *13*, 551. [CrossRef]
96. Li, D.; Armaghani, D.J.; Zhou, J.; Lai, S.H.; Hasanipanah, M. A GMDH Predictive Model to Predict Rock Material Strength Using Three Non-Destructive Tests. *J. Nondestruct. Eval.* **2020**, *39*, 81. [CrossRef]
97. Ivakhnenko, A.G. Polynomial Theory of Complex Systems. *IEEE Trans. Syst. Man Cybern.* **1971**, *1*, 364–378. [CrossRef]
98. Dubčáková, R. Eureka: Software Review. *Genet. Program. Evolvable Mach.* **2011**, *12*, 173–178. [CrossRef]
99. Lizama, E.; Morales, B.; Somos-Valenzuela, M.; Chen, N.; Liu, M. Understanding Landslide Susceptibility in Northern Chilean Patagonia: A Basin-Scale Study Using Machine Learning and Field Data. *Remote Sensing* **2022**, *14*, 907. [CrossRef]
100. Zeini, H.A.; Al-Jeznawi, D.; Imran, H.; Bernardo, L.F.A.; Al-Khafaji, Z.; Ostrowski, K.A.; Kazmi, S.; Zeini, H.A.; Al-Jeznawi, D.; Imran, H.; et al. Random Forest Algorithm for the Strength Prediction of Geopolymer Stabilized Clayey Soil. *Sustainability* **2023**, *15*, 1408. [CrossRef]
101. Scikit-Learn Developers Scikit-Learn. Machine Learning in Python. Available online: https://scikit-learn.org/stable/modules/grid_search.html#randomized-parameter-search (accessed on 21 January 2024).

102. Lundberg, S.M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 4768–4777.
103. Zhang, W.; Wu, C.; Zhong, H.; Li, Y.; Wang, L. Prediction of Undrained Shear Strength Using Extreme Gradient Boosting and Random Forest Based on Bayesian Optimization. *Geosci. Front.* **2021**, *12*, 469–477. [[CrossRef](#)]
104. Rosone, M.; Airò Farulla, C.; Ferrari, A. Shear Strength of a Compacted Scaly Clay in Variable Saturation Conditions. *Acta Geotech.* **2016**, *11*, 37–50. [[CrossRef](#)]
105. Han, Z.; Vanapalli, S.K. Stiffness and Shear Strength of Unsaturated Soils in Relation to Soil-Water Characteristic Curve. *Géotechnique* **2016**, *66*, 627–647. [[CrossRef](#)]
106. Yan, X.Q.; Zhou, C.Y.; Rashed Hama, A.; Tahir, T.A.; Jamal Ali, B.; Professor Nihad Bahaaldeen Salih, A. Geotechnical Characteristics Correlations for Fine-Grained Soils. *IOP Conf. Ser. Mater. Sci. Eng.* **2020**, *737*, 012099. [[CrossRef](#)]
107. Mohammed, A.S. Property Correlations and Statistical Variations in the Geotechnical Properties of (CH) Clay Soils. *Geotech. Geol. Eng.* **2018**, *36*, 267–281. [[CrossRef](#)]
108. Cabalar, A.F.; Mustafa, W.S. Behaviour of Sand–Clay Mixtures for Road Pavement Subgrade. *Int. J. Pavement Eng.* **2017**, *18*, 714–726. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.