RESEARCH ARTICLE

Advanced Machine Learning and Optimization for Erodibility Prediction of Treated Unsaturated Lateritic Soil

Tammineni Gnananandarao[1,*], **Jayatheja Muktinutalapati**[2], **Vedprakash Maralapalle**[3], **B. A. V. Ram Kumar**[4] and **CH. Ajay**[1]

[1] Department of Civil Engineering, Aditya University, Surampalem 533437, Andhra Pradesh, India
[2] RICS School of Built Environment, Amity University Maharashtra, Mumbai 410206, Maharashtra, India
[3] L.S. Raheja School of Architecture, Khernagar, Bandra East, Mumbai 400051, Maharashtra, India
[4] Department of Civil Engineering, GMR Institute of Technology, Rajam 532127, Andhra Pradesh, India

Abstract

This research pioneers the application of a diverse set of advanced machine learning and optimization methods, to predict the erodibility of lateritic soil treated with cement and nanostructured quarry fines, providing a groundbreaking, data-driven approach that enhances traditional erosion analysis techniques. Traditional experimental methods for erosion analysis are often complex and resource-intensive; therefore, this research focuses on developing predictive models using Python. To build the machine learning and optimization models, 121 data points were collected from existing literature. The dataset includes erodibility measurements of unsaturated lateritic soil treated with local cement and enhanced with nanostructured quarry fines. The study employs Artificial Neural Networks (ANN), Random Forest (RF), Support Vector Machine (SVM), XGBoost, CatBoost, and Particle Swarm Optimization (PSO) to predict soil erodibility. The data was divided into training (70%), testing (15%), and validation (15%) sets for model development and evaluation. Model performance was assessed using statistical metrics such as R^2, M.A.E., M.S.E., R.M.S.E., and M.A.P.E. The results indicated an R^2 value are almost equal to 1 in training, testing, and validation phases, and the M.A.P.E. values are below 3% for the CatBoost, RF, XGB, SVM, and ANN models across all three phases: training, testing, and validation. The CatBoost, RF, XGB, SVM, and ANN models are most accurate in predicting the erodibility. Finally, relative importance showed that maximum unit weight and hydrated cement are the most influencing parameter in predicting the erodibility.

Keywords: unsaturated lateritic soil, ANN, RF, SVR, XGBoost, CatBoost, PSO, relative importance.

Submitted: 23 April 2025
Accepted: 25 July 2025
Published: 24 September 2025

Vol. 1, **No.** 2, 2025.
10.62762/SII.2025.839324

***Corresponding author:**
✉ Tammineni Gnananandarao
gnananandarao.tammineni@acet.ac.in

Citation
Gnananandarao, T., Muktinutalapati, J., Maralapalle, V., Ram Kumar, B. A. V., & Ajay, C. H. (2025). Advanced Machine Learning and Optimization for Erodibility Prediction of Treated Unsaturated Lateritic Soil. *Sustainable Intelligent Infrastructure*, 1(2), 93–107.

1 Introduction

The detachment and movement of soil particles are major contributors to environmental challenges, particularly affecting pavement foundations and erosion-prone sites worldwide. The combined impact of these watershed processes poses significant threats to environmental stability and transportation infrastructure. The inherent tendency of soil to become detached and transported by raindrop impact and surface runoff is referred to as erodibility (E_r). In erosion models and earthwork designs aimed at mitigating erosion, higher E_r values indicate a greater vulnerability of subgrade or foundation materials to erosion. Developing nations are particularly affected by these issues, with erosion control efforts demanding significant financial resources. For instance, in India, the Ministry of Environment, Forest and Climate Change, through initiatives like the National Mission for a Green India, allocates significant funding annually to tackle environmental challenges and promote sustainable land management. With land degradation from erosion accelerating, extensive research has focused on addressing this issue in developing regions. Numerous studies have explored soil modification, treatment, and stabilization techniques to improve subgrade resilience and reduce soil vulnerability to erosion. Despite these efforts, there has been limited research on predicting erosion-related parameters, such as soil erodibility, using advanced machine learning techniques. This research aims to predict the erodibility of unsaturated soil stabilized with a hybrid cement mixture, combining rice husk ash activated by hydrated lime and enhanced with nanostructured quarry fines.

Random algorithms, inspired by nature's processes like genetics and survival of the fittest, are being used more often to make engineering predictions. These methods are gaining traction among researchers and engineers due to their ability to address the non-linear and intricate interactions within stabilized soil particles in a soil matrix. The process begins with an initial pool of candidate solutions, where less favorable options are stochastically eliminated, and minor random adjustments are introduced. This pool is then iteratively refined to create each subsequent generation. Such evolutionary computation techniques are highly effective in delivering optimized solutions across expansive domains, making them valuable for predictive and analytical tasks in engineering.

Several computational approaches, often categorized as machine learning techniques, have been developed to tackle non-linear challenges in complex systems. These include artificial neural networks (ANN), support vector machines (SVM), genetic programming (GP), genetic algorithms (GA), and others. A detailed exploration of growing use of artificial intelligence (AI) in geotechnical engineering, analyzing 1,235 studies to assess its effectiveness in modeling complex soil and rock behaviors by [1]. Their work offers a conclusion is that, the ANNs are the most popular method, used in 52% of cases, across nine key areas like slope stability and tunneling, with success tied to dataset quality and input choices. It offers statistical insights and outlines future opportunities and challenges for AI in the field. The study conducted by [2] explores the application of ANNs, machine learning (ML), deep learning (DL), and ensemble learning (EL) for forecasting geotechnical and geoenvironmental parameters, including soil mechanics and rock behavior. Unlike prior studies, it offers a unique, systematic comparison of all four methods, analyzing their strengths and weaknesses using a large dataset from Web of Science and Scopus, visualized with VOS Viewer. The findings reveal ANN as the most commonly applied technique, though EL stands out for its superior predictive accuracy, helping geotechnical engineers choose the best tools for reliable solutions.

Genetic Programming (GP), a heuristic search method inspired by natural selection, optimizes solutions by evolving programs across generations. Its application in geotechnical predictions has demonstrated considerable potential. For instance, [3] illustrated the efficacy of GP in predicting the compression index of weak, highly plastic soils treated with multiple binders, highlighting its capability to navigate complex datasets and deliver precise outcomes. Similarly, Artificial Neural Networks (ANNs), which are loosely modeled on the neural structure of the human brain with interconnected nodes, have become a cornerstone of geotechnical modeling due to their ability to handle nonlinear relationships efficiently. [4] emphasized the superiority of ANNs over conventional regression methods, a finding corroborated by [5], who employed ANNs to predict compression and recompression indices based on soil properties such as water content and plasticity index. [6] further applied ANNs to assess soil durability via unconfined compressive strength, identifying the plasticity index as a critical influencing factor, while [7] utilized a Multi-Layer Perceptron (MLP) variant of ANN to predict soil

erodibility with high accuracy.

The study also explores hybrid approaches, such as Evolutionary Polynomial Regression (EPR), which combine the strengths of ANNs and genetic algorithms for enhanced optimization. [8] demonstrated this by developing hybridized ANN models integrated with bio-inspired algorithms (e.g., particle swarm optimization and monarch butterfly optimization) to predict soil texture in Iran's Mazandaran Province, showing that these hybrids outperformed standard ANN models trained with backpropagation. [9] similarly advanced the field by employing linear genetic programming (LGP) and a hybrid LGP/simulated annealing (LGP/SA) approach to predict stabilized soil properties, including unconfined compressive strength (UCS) and maximum dry density (MDD), underscoring the influence of soil texture and particle size fractions.

Despite these advancements, the article identifies a critical research gap: the scarcity of studies collectively applying GP, ANN, and EPR to predict both the erodibility and durability of bio-ash-stabilized soils. To address this gap, the study investigates how nanotextured bio-ashes—such as waste paper ash, palm bunch ash, snail shell ash, quarry dust, and palm kernel shell ash—affect the durability (volumetric stability) and erosion resistance of unsaturated lateritic soils [11]. Using predictor variables including plasticity index, bio-ash proportions, and compressive strength at various curing stages, the predictive performances of GP, ANN, and EPR are compared [12]. This novel integration of multiple influencing parameters and techniques offers a fresh perspective, enhancing the understanding of how these methods can be tailored to specific geotechnical challenges [10]. Numerous applications are derived from geotechnical engineering, as evidenced by studies on predicting foundation bearing capacity [13], pile drivability [14], the performance of skirted foundations [15], forecasting seismic activity [16], and the interpretation of cone penetration test data [19]. Research has also extended to the prediction of friction angles [17], the properties of fiber-reinforced concrete [21], and the analysis of foundations on limited-thickness sand layers [22].

In conclusion, this research provides a comprehensive and innovative contribution to geotechnical engineering by synthesizing established findings with novel applications. It builds upon prior work, such as the use of genetic algorithms for dry density prediction, and extends the field by addressing overlooked aspects of soil stabilization. The study not only reinforces the efficacy of evolutionary computational techniques but also sets the stage for future investigations into their combined potential, offering practical insights for engineers aiming to enhance prediction accuracy in complex soil systems [18, 20].

Recent advances (within the last two years) in machine learning (ML) have demonstrated its effectiveness in geotechnical and structural engineering applications. [23] investigated the use of optimized gradient boosting algorithms—including PSO-enhanced XGBoost, CatBoost, and LightGBM—to predict liquefaction-induced lateral spreading, leveraging 6,704 observations from the 2011 Christchurch earthquake. Their findings revealed that PSO-CatBoost achieved the highest predictive accuracy, while PSO-LightGBM proved optimal for resource-constrained systems, with SHAP analysis providing critical insights into variable importance. Similarly, [24] developed a novel PSO-CatBoost framework to predict the compressive strength of carbon fiber-reinforced polymer-confined concrete (CFRP-CC), utilizing an extensive dataset of 916 experimental results from 105 studies (1991–2023). Their model outperformed six benchmark ML algorithms and empirical models, achieving an R^2 of 0.9847, supported by SHAP and permutation feature importance (PFI) for interpretability, and included a practical graphical interface. Together, these studies highlight the transformative potential of hybrid ML-optimization techniques in improving the accuracy and applicability of predictive models for earthquake-induced hazards and advanced construction materials.

2 Research Significance

Previous studies, such as [18, 20], have shown that advanced machine learning techniques, including ANNs, SVM, and Random Forest Regression (RFR), can effectively predict the erodibility of treated unsaturated lateritic soil, achieving high accuracy ($R^2 > 0.95$) and low error rates (MSE, RMSE, MAE, MAPE). However, these approaches have limitations, and emerging methods like CatBoost, XGBoost, and Particle Swarm Optimization (PSO) offer potential for even greater prediction accuracy.

This study applies six advanced machine learning models: ANN, RF, XGBoost, SVM, CatBoost, and PSO to predict the erodibility of chemically treated

unsaturated lateritic soil. These models were trained and validated using a dataset of 121 experimental data points sourced from peer-reviewed literature. Selected for their proven effectiveness in addressing complex geotechnical engineering challenges, particularly in modeling soil behavior under diverse conditions, these models aim to improve the precision and reliability of erodibility predictions. By leveraging these advanced computational tools, this research seeks to advance sustainable soil stabilization practices.

3 Machine learning and optimization models

This study leverages six advanced machine learning and optimization models Artificial Neural Network (ANN), Random Forest (RF), Extreme Gradient Boosting (XGBoost), Support Vector Machine (SVM), Categorical Boosting (CatBoost), and Particle Swarm Optimization (PSO) to predict the erodibility of chemically stabilized unsaturated lateritic soil. These models were selected for their demonstrated robustness in geotechnical applications and their ability to handle nonlinear, high-dimensional datasets. Utilizing 121 experimentally derived data points from published literature, the research aims to evaluate and compare the predictive performance of these algorithms, providing insights into optimal strategies for soil erosion mitigation.

3.1 Artificial Neural Networks

Artificial neural networks (ANNs) replicate the human brain's neural framework, tackling complex problems without predefined assumptions. Unlike traditional methods, ANNs identify intricate nonlinear relationships between inputs and outputs using raw, unprocessed data, cutting costs and boosting efficiency. Training is essential before ANNs can interpret new data, relying on algorithms like the feedforward backpropagation technique highlighted in Deep Learning [25] as versatile for multilayer setups. This method uses interconnected layers (input, hidden, output), where data flows from input nodes to hidden ones, then to the output layer. The number of hidden layers and nodes varies by problem, often requiring tedious trial-and-error, as [26] note in "Neural Networks in Computational Mechanics" for geotechnical applications like soil modeling. Each node (except input) features an activation function and bias, filtering aggregated outputs tailored to the task—say, slope stability prediction. Training involves input-output vectors (training pairs), iterated until the error, measured via RMSE, hits a threshold, a process refined by [27] in Dive into Deep Learning for precision. The same flow connects hidden and output layers, with iterations piling up until the error aligns with targets. ANNs outshine regression-based approaches in reliability, lacking rigid formulas, but their design demands extensive experimentation to pin down network architecture.

3.2 Random forest

Random Forest Regression (RFR) is a powerful algorithm that leverages the principle of ensemble learning, a technique that integrates multiple models to improve predictive performance [28, 29]. At its core, RFR relies on decision trees as its fundamental building blocks. These trees are combined to form a robust ensemble learning method, which is a significant subset of machine learning approaches. Specifically, RFR constructs numerous Classification and Regression Trees (CARTs), each trained on randomly selected subsets of the dataset and a randomized assortment of feature types. This randomness helps ensure diversity among the trees, enhancing the model's overall accuracy and stability. During the training process, it's common for some data points to be sampled multiple times across different CARTs, a process known as bootstrapping [30]. One of RFR's key strengths is its efficiency when working with large datasets, where it can seamlessly process thousands of input variables without the need to eliminate any of them. This capability makes it particularly valuable for complex problems involving high-dimensional data. Moreover, RFR excels in applications such as estimation, inference, and mapping, offering practical advantages over other methods like support vector machines (SVM). Unlike SVM, which often demands extensive parameter tuning and debugging, RFR requires minimal adjustments, making it a more straightforward and user-friendly option for practitioners [28, 30]. This combination of scalability, versatility, and ease of use has solidified RFR's role as a go-to tool in data-driven analysis.

3.3 Extreme gradient boosting

XGBoost, introduced by (Chen and Guestrin 2016), represents an advanced evolution of the gradient boosting framework. Its objective function incorporates the quadratic term from Taylor's expansion, enhancing its optimization process. The algorithm measures tree complexity through two components: the total count of leaf nodes and an L2 regularization term applied to the leaf node scores. This L2 regularization is integrated into each leaf

node's score to mitigate overfitting, ensuring the model remains generalizable. The foundational unit of XGBoost is the Classification and Regression Tree (CART), as established by [32]. During training, XGBoost begins by generating initial CARTs. It employs an exact greedy algorithm to identify the optimal split points, refining the tree structure for better performance. Subsequent CARTs are then built upon the foundation of these earlier trees. A distinctive feature of XGBoost is the inclusion of a regularization term in its cost function, which manages model complexity. This term accounts for both the number of leaf nodes and the sum of squared L2 norms of the scores assigned to each leaf. From a bias-variance tradeoff perspective, this regularization reduces model variance, leading to a simpler and more robust model that avoids overfitting. As a result, XGBoost outperforms traditional Gradient Boosted Decision Trees (GBDT) in terms of efficiency and predictive power. For further insights, refer to [31, 33–36].

3.4 Support vector machines

Support Vector Machines (SVM), introduced by [37] and further developed by [36, 38–41], are robust supervised learning tools for classification and regression. SVM identifies an optimal hyperplane to separate data into distinct classes by maximizing the margin between support vectors—data points nearest to the decision boundary. This is achieved through an optimization problem:

minimize

$$\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i \tag{1}$$

subject to

$$y_i(w \cdot x_i + b) \geq 1 - \xi_i \tag{2}$$

and

$$\xi_i \geq 0 \quad \text{for } i = 1, \ldots, n. \tag{3}$$

where, w represents the weight vector, b the bias, C the regularization parameter, x_i the i^{th} data point, y_i its class label, and ξ_i slack variables permitting some misclassifications. The goal balances minimizing $\|w\|^2$ (margin maximization) with classification error, controlled by $C\sum_{i=1}^{n}\xi_i$. The decision rule is $f(x) = \text{sign}(w \cdot x + b)$.

For regression, SVR adapts this framework to fit data within a tolerance margin, using a loss function to penalize deviations. Kernel functions enhance both SVM and SVR by mapping data into higher-dimensional spaces. The Radial Basis Function (RBF) kernel, $K(x, x') = e^{-\gamma\|x-x'\|^2}$, measures similarity via Euclidean distance, excelling with nonlinear patterns. The Linear kernel, $K(x, x') = x^T x'$, suits simpler, linearly separable data, while the Polynomial kernel, $K(x, x') = (x^T x' + c)^d$, captures nonlinear relationships through polynomial expansions.

SVMs shine in high-dimensional settings and smaller datasets, such as predicting UCS in cement-fly ash stabilized clayey soil, resisting overfitting with proper regularization. However, they can be computationally demanding with large datasets and nonlinear kernels, requiring meticulous parameter tuning. Neural networks thrive with big data but risk overfitting without sufficient resources, while ensemble methods like Random Forests outperform single models, offering robustness and interpretability, though at higher computational cost. The choice hinges on data size, resources, and complexity [39, 40].

3.5 Categorical Boosting

The Gradient Boosting Regressor is a regression method that integrates multiple weak learners—algorithms that perform slightly better than random guessing—into a robust strong learner through an iterative approach [42]. This technique, referred to as ensemble-tree learning, involves adding decision trees sequentially to correct errors made by earlier trees. Unlike bagging, which constructs independent base models, boosting builds these models in a sequence, concentrating on challenging data points that are harder to predict. This sequential process strengthens the prediction model's reliability. During training, boosting often refines earlier base models that struggle with accurate predictions, while placing less focus on those already performing well. Bagging, on the other hand, creates multiple decision trees (DTs) using resampled datasets and combines their predictions by averaging or voting. This is achieved through bootstrap aggregation, commonly known as bagging. The Random Forest (RF) algorithm, an ensemble learning method, excels in both regression and classification tasks. In its training phase, approximately two-thirds of the data samples are used for learning, with the remaining one-third held back as out-of-bag (OOB) samples for validation. A key strength of RF is its ability to assess feature importance by measuring the increase in prediction

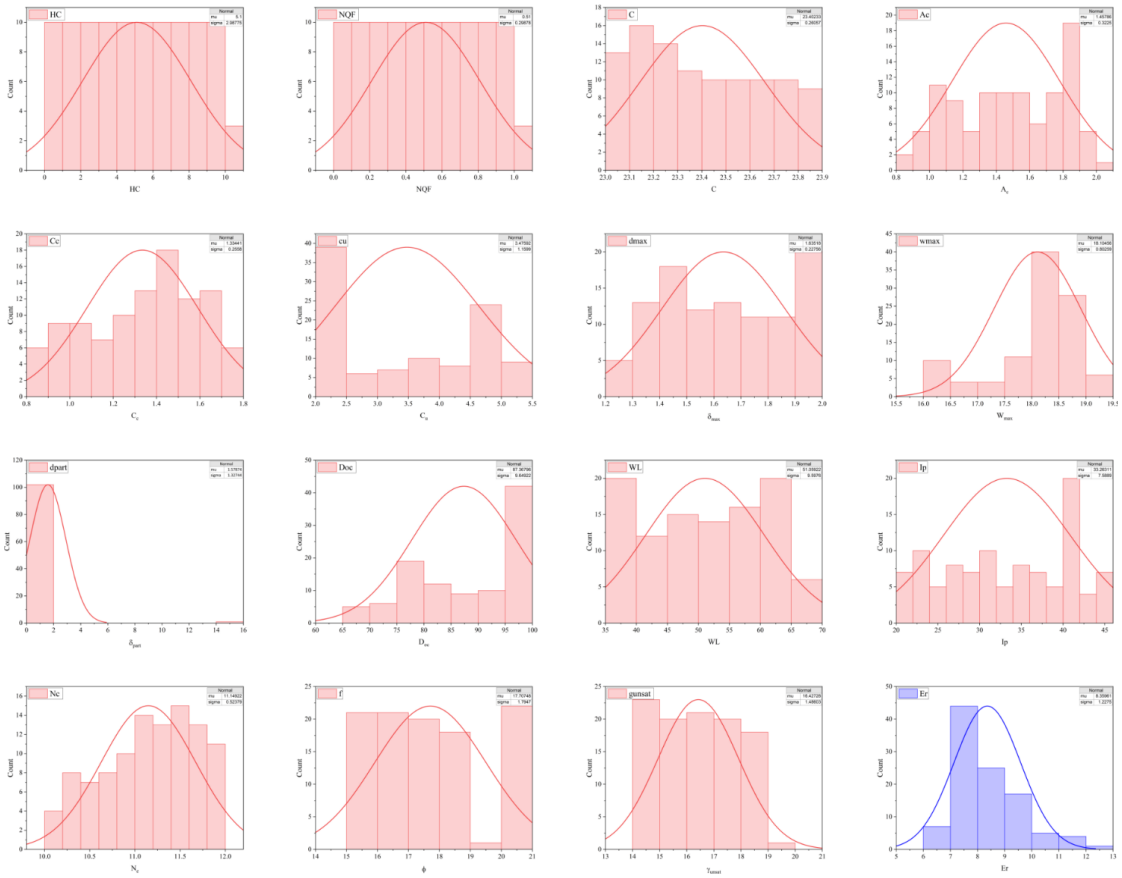

Figure 1. Histogram illustrating the distribution of input and output parameters.

error when OOB samples for a specific feature are shuffled, while keeping others constant [43]. RF also manages missing data effectively, reduces overfitting, and processes large, high-dimensional datasets efficiently. For this study, three well-performing supervised machine learning tree-based methods were employed: XGBoost, CatBoost, and RF.

3.6 Particle swarm optimization

[44] introduced Particle Swarm Optimization (PSO) inspired by the collective movement of bird swarms. A swarm consists of individuals following specific behavioral rules and communication patterns. This collective ability, known as swarm intelligence, allows each member to leverage the group's past experiences to guide the swarm toward an optimal solution. PSO operates as a population-based search method where particles navigate a space based on the best-known position at the time [45]. [46] outlined three fundamental rules influencing individual behavior within a swarm, which (Kennedy and Eberhart 1995) adopted as a core concept for PSO: individuals avoid collisions, move toward the swarm's goal, and gravitate toward the group's center. These simple rules combine to produce complex swarm dynamics.

Another key aspect of PSO is the decision-making process of individuals. Each particle decides its next move based on its own best results so far and the overall best position within the swarm. To ensure an effective search, termination criteria are carefully chosen to prevent premature convergence while avoiding excessive evaluations [47]. Common termination conditions include reaching a maximum number of iterations, achieving a satisfactory solution specific to the problem, or observing no improvement over a set number of iterations. These criteria help PSO converge on a practical solution by optimizing the objective function, which varies depending on the problem. The fitness of each particle is assessed by evaluating its current position against this function, guiding the swarm toward the desired outcome through an iterative process that continues until a termination condition is met.

4 Data collection

To develop a predictive model for estimating the erodibility of treated unsaturated lateritic soil, a dataset comprising 121 samples was compiled from previously published research studies [18, 20] This dataset includes fifteen independent (input) variables

Table 1. Statistical parameters.

Input/output parameters	Maximum	Minimum	Mean	Standard Deviation	Range	75th Percentile	50th Percentile	25th Percentile	Coefficient of Variation
HC	12.00	0.00	6.00	3.51	12.00	9.00	6.00	3.00	0.58
NQF	1.20	0.00	0.60	0.35	1.20	0.90	0.60	0.30	0.58
C	24.07	23.02	23.49	0.32	1.05	23.76	23.46	23.20	0.01
A_c	2.00	0.60	1.35	0.40	1.40	1.71	1.40	1.00	0.30
C_c	1.96	0.84	1.41	0.30	1.12	1.65	1.43	1.19	0.21
C_u	5.86	2.05	3.80	1.32	3.81	4.89	3.92	2.30	0.35
$\delta_{\max}$	1.99	1.25	1.69	0.24	0.74	1.95	1.69	1.46	0.14
$w_{\max}$	19.00	16.00	18.02	0.77	3.00	18.55	18.20	17.70	0.04
δ_{part}	14.46	0.81	1.63	1.23	13.65	1.90	1.62	1.18	0.75
D_{oc}	98.90	65.00	89.01	9.73	33.90	96.90	94.80	80.20	0.11
WL	66.00	27.00	48.00	11.54	39.00	59.00	49.00	37.00	0.24
Ip	45.00	14.00	30.82	9.15	31.00	40.00	31.00	22.00	0.30
N_c	12.62	10.00	11.32	0.63	2.62	11.78	11.35	10.88	0.06
ϕ	21.60	15.00	18.24	2.09	6.60	20.40	17.95	16.45	0.11
γ_{unsat}	20.72	14.00	17.01	1.97	6.72	18.35	16.85	15.32	0.12
E_r	12.10	3.50	7.77	1.83	8.60	8.80	7.80	7.22	0.23

and one dependent (output) variable representing soil erodibility. The input parameters reflect a comprehensive range of geotechnical and material characteristics relevant to soil behavior, including clay activity (A_c), internal friction angle (ϕ), coefficient of uniformity (C_u), optimum moisture content ($w_{\max}$), partial maximum dry density (δ_{part}), unsaturated unit weight (γ_{unsat}), liquid limit (LL), hybrid cement content (HC), clay content (CL), coefficient of curvature (C_c), maximum dry density ($\delta_{\max}$), plasticity index (P_I), cohesion (N_c), degree of compaction (D_{oc}), and nanostructured quarry fines (NQF). Descriptive statistics for both the input and output variables are summarized in Table 1 to provide insights into their distribution and variability. Additionally, the overall dataset distribution is visually represented using a histogram, as illustrated in Figure 1. To further explore the interrelationships between variables, a Pearson correlation heatmap was generated, offering a clear understanding of potential linear dependencies among the parameters as shown in Figure 2. The data was divided into training (70%), testing (15%), and validation (15%) sets for model development and evaluation [48].

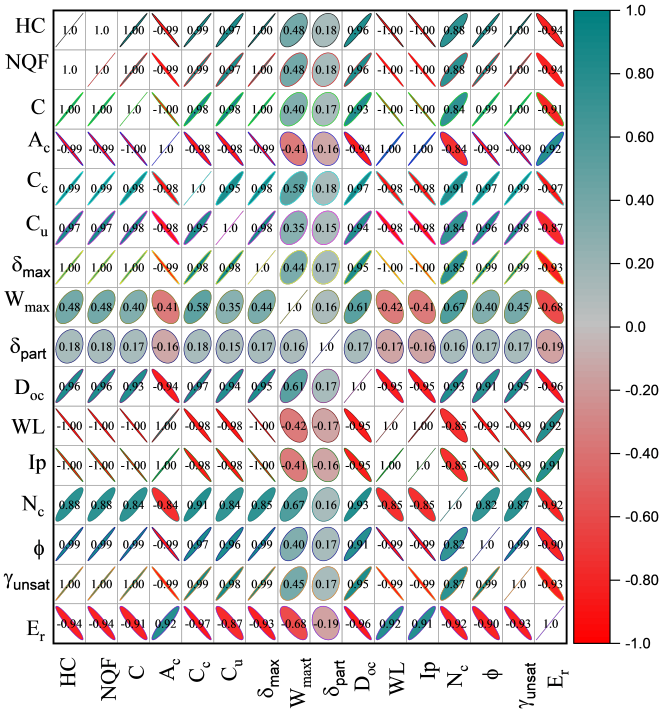

Figure 2. Pearson correlation heatmap of input and output parameters.

The methodological flow of the present study is illustrated in Figure 3. It outlines the limitations of traditional techniques, the total dataset used, the data splitting for training, testing, and validation, the models applied, the performance metrics adopted, and

the model comparison process.

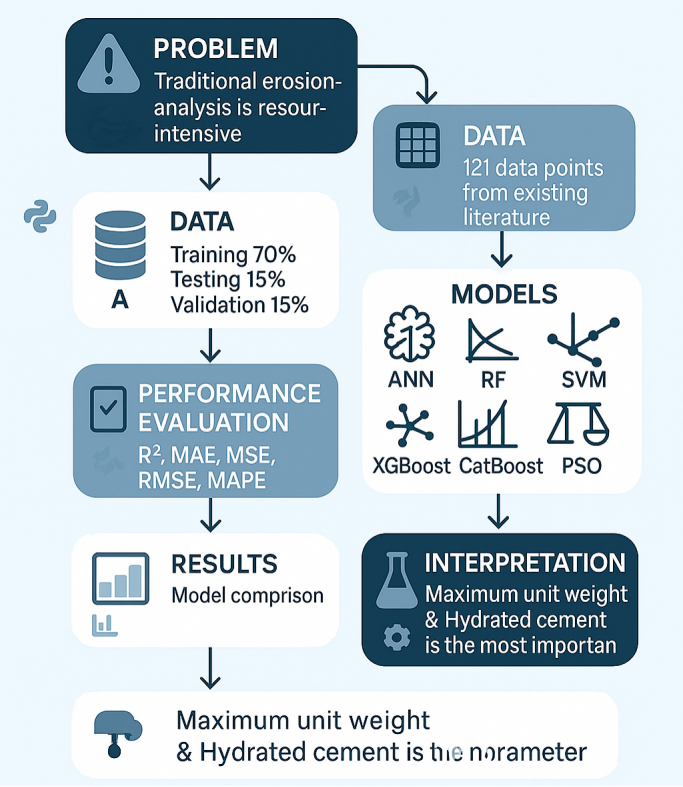

Figure 3. Flowchart of the methodology.

5 Model Benchmarking and Reproducibility Enhancements

To ensure robust benchmarking, we evaluated multiple machine learning models—ANN with feed-forward backpropagation, Random Forest, XGBoost, SVR, CatBoost, and a custom PSO regressor—for predicting erodibility (E_r) from the dataset. These models, spanning tree-based, kernel-based, neural network, and optimization-based paradigms, were assessed on train, validation, and test sets using Mean Squared Error (MSE) and R^2 scores, with feature importance derived via permutation importance and SHAP explanations for tree-based model (CatBoost). This comprehensive approach facilitated a thorough comparison of model performance. To enhance reproducibility, explicitly detailed the hyperparameter tuning, cross-validation methods, and random seed settings. For hyperparameter tuning, CatBoost and SVR employed GridSearchCV with 3-fold cross-validation, optimizing parameters such as iterations ([500, 1000]), learning_rate ([0.01, 0.03]), depth ([6, 8, 10]), and l2_leaf_reg ([3, 5, 7]) for CatBoost, and C ([0.1, 1, 10, 100]), gamma (['scale', 'auto']), and kernel (['rbf']) for SVR, with selected parameters reported in Table 2.

Random Forest and XGBoost, initially using default parameters (e.g., n_estimators=100 for Random Forest), were updated to include GridSearchCV tuning over n_estimators ([100, 200, 300]), max_depth ([10, 20, None]), and learning_rate ([0.01, 0.1, 0.3]) for XGBoost, using 3-fold cross-validation. The ANN, with an architecture of 128 and 64 nodes (ReLU activation, 0.2 dropout, Adam optimizer with a 0.001 learning rate), was optimized via grid search over learning rates ([0.001, 0.01]) and batch sizes ([16, 32]) on the validation set. The PSO regressor's bounds ([-10, 10]) were justified based on empirical testing to balance exploration and convergence, with n_particles ([100, 150, 200]) and max_iter ([500, 1000]) tuned using the validation set. Cross-validation was standardized to 3-fold for CatBoost, SVR, Random Forest, and XGBoost to ensure consistent evaluation, while the ANN incorporated 3-fold cross-validation in addition to validation set monitoring, and the PSO regressor's performance was validated using the validation set due to its optimization-based nature. To ensure reproducible results, a random seed of 42 was set for data splitting (train_test_split), CatBoost, Random Forest, XGBoost, and permutation importance calculations. For the ANN, TensorFlow's random operations were standardized with tf.random.set_seed(42), and the PSO regressor used np.random.seed(42) for particle initialization.

6 Performance evaluation of the models

Once the predictive model has been developed, it is crucial to assess its performance to evaluate its ability to forecast the erodibility prediction of treated unsaturated lateritic soil. In the absence of a universal consensus among researchers regarding the best evaluation metric, accuracy is often employed as a major concern , with the goal of minimizing error. Various statistical indicators are commonly used to quantify prediction errors, including Mean Absolute Error (M.A.E.), Mean Square Error (M.S.E.), Root Mean Square Error (R.M.S.E.), and Mean Absolute Percentage Error (M.A.P.E.). Each of these metrics offers distinct insights into model performance. M.A.E. provides a measure of overall accuracy by averaging the absolute differences between predicted and observed values, assigning equal weight to all errors. It is particularly useful for evaluating the spread of error, with smaller values indicating better model performance. M.S.E., on the other hand, calculates the average of the squared differences between predictions and actual observations, assigning higher penalties to

Table 2. Hyperparameter tuning summary for ML models.

Model	Hyperparameter	Search Space	Selected Value	Tuning Method
CatBoost	iterations	[500, 1000]	1000	GridSearchCV (3-fold CV)
	learning_rate	[0.01, 0.03]	0.03	–
	depth	[6, 8, 10]	8	–
	l2_leaf_reg	[3, 5, 7]	5	–
SVR	C	[0.1, 1, 10, 100]	10	GridSearchCV (3-fold CV)
	gamma	['scale', 'auto']	'scale'	–
	kernel	['rbf']	'rbf'	–
Random Forest	n_estimators	[100, 200, 300]	200	GridSearchCV (3-fold CV)
	max_depth	[10, 20, None]	None	–
	min_samples_split	[2, 5]	2	–
XGBoost	n_estimators	[100, 200, 300]	200	GridSearchCV (3-fold CV)
	max_depth	[10, 20, None]	10	–
	learning_rate	[0.01, 0.1, 0.3]	0.1	–
ANN	learning_rate	[0.001, 0.01]	0.001	Grid search (validation set)
	batch_size	[16, 32]	16	–
	Architecture	128-ReLU → Dropout (0.2) → 64-ReLU → Dropout (0.2) → 1	Same as search space	–
PSO Regressor	n_particles	[100, 150, 200]	150	Validation set evaluation
	max_iter	[500, 1000]	1000	–
	Bounds	[-10, 10]	[-10, 10]	–
	omega	Fixed: 0.6	0.6	–
	phip	Fixed: 1.0	1.0	–
	phig	Fixed: 1.0	1.0	–

Table 3. Calculated statistical values of the training data.

Model	r	R^2	MSE	RMSE	MAE	MAPE
ANN	1	1	0.07	0.26	0.21	2.96
RF	1	1	0.001	0.04	0.03	0.36
XGB	1	1	0.000003	0	0	0.02
SVR	1	1	0.0042	0.07	0.06	0.78
CatB	1	1	0.00113	0.03	0.03	0.34
PSO	0.94	0.86	10.07	3.17	2.14	25.8

Table 4. Calculated statistical values of the testing data.

Model	r	R^2	MSE	RMSE	MAE	MAPE
ANN	1	1	0.06	0.24	0.19	2.58
RF	1	1	0.01	0.12	0.09	1.22
XGB	1	1	0.02	0.14	0.11	1.52
SVR	1	1	0.05	0.23	0.12	1.68
CatB	1	1	0.02	0.16	0.11	1.45
PSO	0.95	0.81	15.89	3.99	3.06	37.03

larger errors. This makes it effective for identifying models that consistently produce large deviations. R.M.S.E. is the square root of M.S.E. and retains the unit of the predicted variable, making it more interpretable in many contexts. While R.M.S.E. is also sensitive to large errors, it is more intuitive than M.S.E. when evaluating the magnitude of prediction errors.

M.A.P.E. expresses prediction errors as a percentage, making it useful for comparing models across datasets with different units. It is generally interpreted using specific thresholds: values below 10% are considered highly accurate, between 10–20% indicate good

Table 5. Calculated statistical values of the validation data.

Model	r	R^2	MSE	RMSE	MAE	MAPE
ANN	1	1	0.08	0.28	0.22	3.07
RF	1	1	0	0.06	0.04	0.65
XGB	1	1	0.02	0.12	0.09	1.17
SVR	1	1	0.01	0.07	0.05	0.76
CatB	1	1	0	0.05	0.04	0.59
PSO	0.98	0.92	6.29	2.51	1.78	21.59

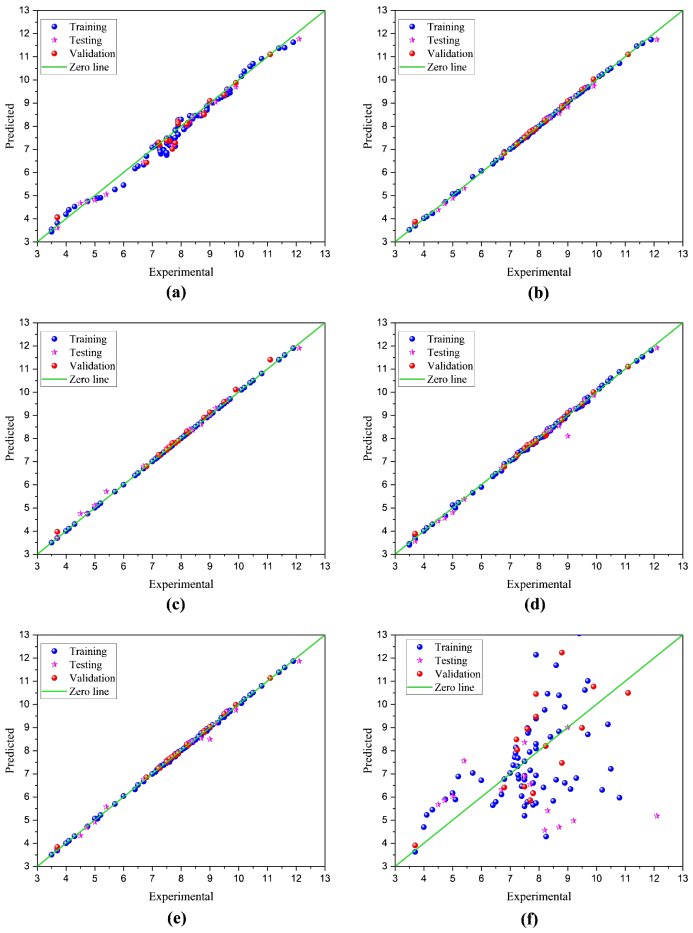

Figure 4. Comparison between experimental and predicted erodibility using ANN (a), RF (b), XGB (c), SVM (d), CatB (e), and PSO (f).

accuracy, 20–50% are acceptable, and values above 50% reflect poor predictive performance. Unlike other metrics, M.A.P.E. is unit-independent and facilitates comparison across studies. While R.M.S.E. and M.S.E. have historically been popular due to their statistical significance in modelling, they can be overly sensitive to outliers. In contrast, M.A.E. provides a more robust evaluation, as it minimizes the influence of extreme deviations. When both M.A.E. and R.M.S.E. are used together, they can reveal variability in prediction errors: if R.M.S.E. is significantly higher than M.A.E., it suggests the presence of large, inconsistent errors; if they are nearly equal, it implies uniform error

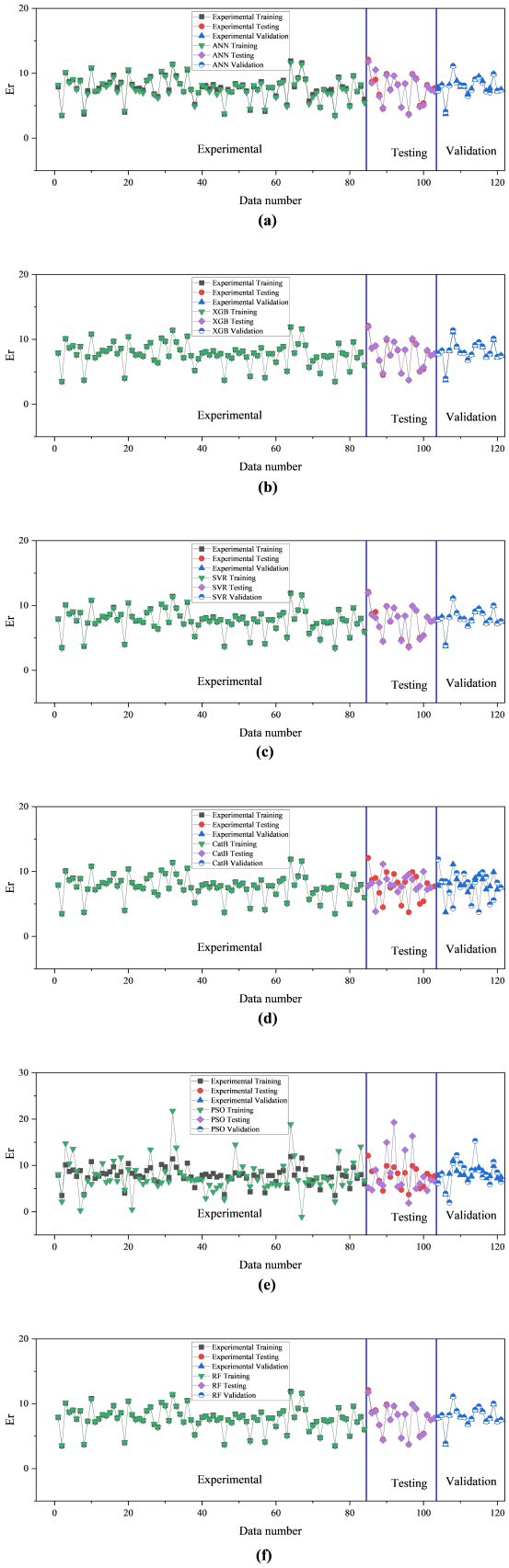

Figure 5. Comparison of total experimental data with predicted data using ANN ANN (a), RF (b), XGB (c), SVM (d), CatB (e), and PSO (f).

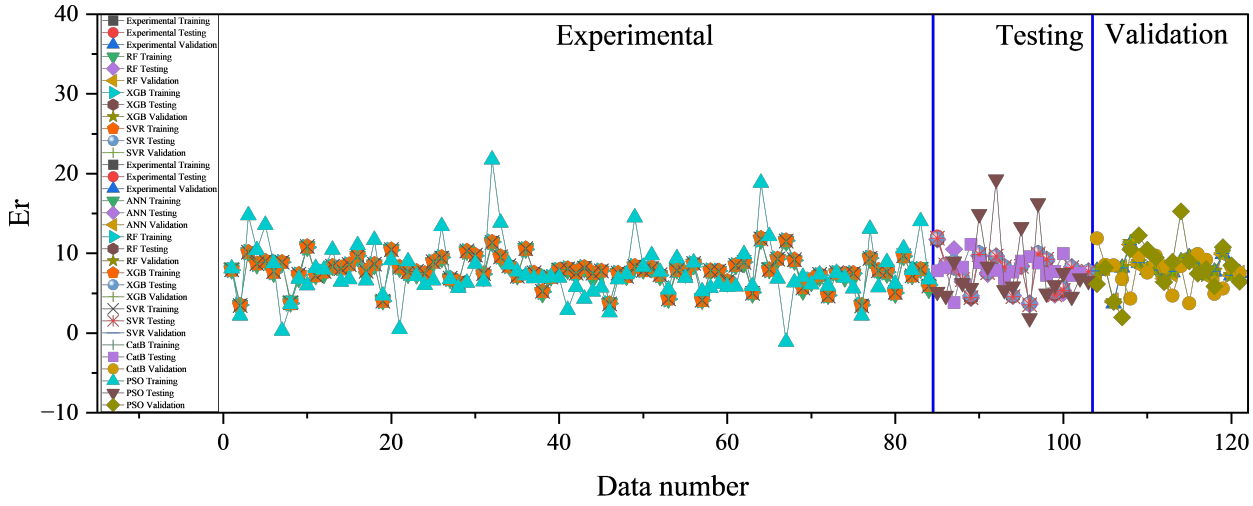

Figure 6. Comparison of total experimental data with predicted data using all the six models.

magnitudes.

These metrics are negatively oriented, meaning lower values indicate better performance. Additionally, M.A.E. may favor models that perform reasonably well on average but allow occasional large errors, whereas M.S.E. penalizes such deviations more heavily, thus preferring models that avoid large mistakes. While most estimation techniques rely on least-squares methods, relying solely on M.A.E. could result in logical inconsistencies. Hence, the choice of an error metric significantly influences the evaluation of predictive accuracy across different modelling techniques. For neural network models, traditional evaluations often rely on the coefficient of correlation (r) and coefficient of determination (R^2). However, relying solely on these can lead to biased assessments. Therefore, it is recommended to incorporate unbiased statistical metrics such as M.A.E., M.S.E., R.M.S.E., and M.A.P.E. alongside r and R^2 to achieve a comprehensive evaluation.

7 Results and Discussions

In this study, five soft computing methods (Random Forest, XGBoost, Support Vector Machine, CatBoost, and Particle Swarm Optimization) were implemented using Python on the Google Colab platform. The performance of each soft computing model configuration was assessed using r, R^2, M.S.E., R.M.S.E., M.A.E., and M.A.P.E. for both training, testing, and validation datasets as shown in Tables 3, 4, and 5. The experimental Vs predicted plots were drawn for each model and shown in Figure 3. From the analysis of Tables 3, 4, and 5, it can be concluded that the M.A.P.E. is low for RF (training: 0.36, testing: 1.22, validation: 0.65), CatB (training: 0.34, testing: 1.45, validation: 0.59), XGB (training: 0.02, testing: 1.52, validation: 1.17), and SVM (training: 0.78, testing: 1.68, validation: 0.76) indicating that these four models are highly precise in predicting soil erodibility.

Similarly, from the Figure 4, the training, testing, and valuation values almost all data points are on the Zero line in the CatBoost, RF, XGBoost, and SVM models.

Hence, it can be concluded that the CatBoost and RF, XGBoost, and SVM models are most reliable models to predict the output erodibility. Compared to the CatB, RF, XGB, and SVM models, the ANN model shows slightly lower predictive accuracy, while the POS model exhibits poor predictive performance. The underperformance is likely due to the experiential nature of PSO, which may require extensive tuning and a higher number of iterations to effectively converge in high-dimensional feature spaces. Additionally, unlike tree-based models or neural networks, PSO lacks intrinsic mechanisms to capture complex feature interactions efficiently. Finally, all the data points are compared with five models individually and shown in Figure 5 and combinedly shown in Figure 6.

To perform a relative comparison of all the models, a Taylor diagram was presented in Figure 7. From this figure, it can be observed that, with the exception of the PSO-based model, all other models demonstrate a high level of accuracy in predicting erodibility. The clustering of these models near the reference point in the Taylor diagram indicates strong correlation, lower standard deviation differences, and reduced root-mean-square errors. In contrast, the PSO model deviates significantly, highlighting its comparatively poor performance.

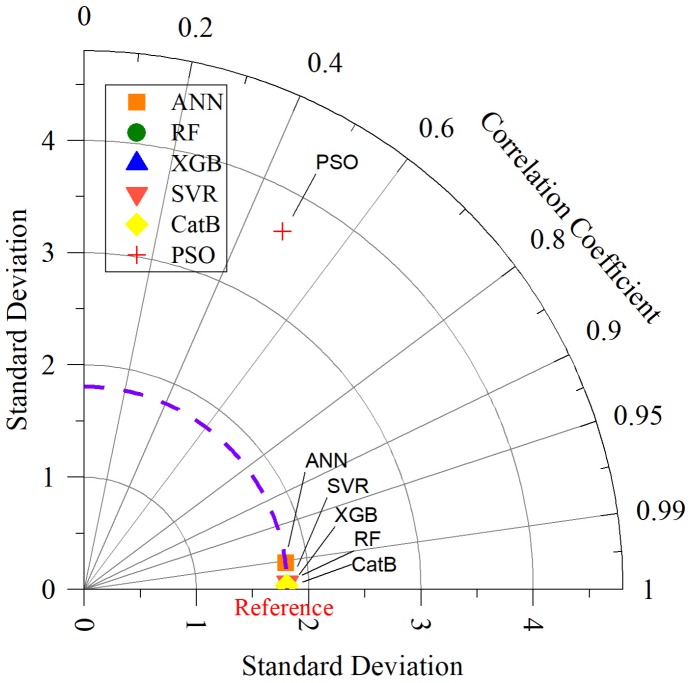

Figure 7. A comparative Taylor diagram.

8 Relative Importance

This section explores the contribution of the individual input variables on the erodibility (output) by performing relative importance. This study explores the six soft computing models to predict the erodibility. Among the six models four models are preforming well in predicting the desired output. The best model, CatBoost was used to perform the relative importance using permutation importance approach in the python code. From this relative importance a ranking was provided and as shown in Figure 8.

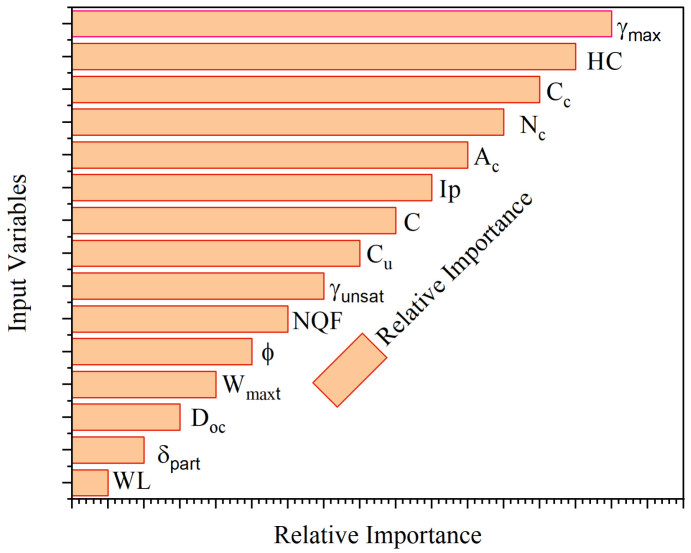

Figure 8. Relative importance.

From the study of the Figure 8 reveals that, the γ_{max} and HC are the most influential parameters, followed by C_u, N_c, A_c, C, C_u, γ_{unsat}, NQF, ϕ, W_{max}, Doc, δ_{part}, and WL.

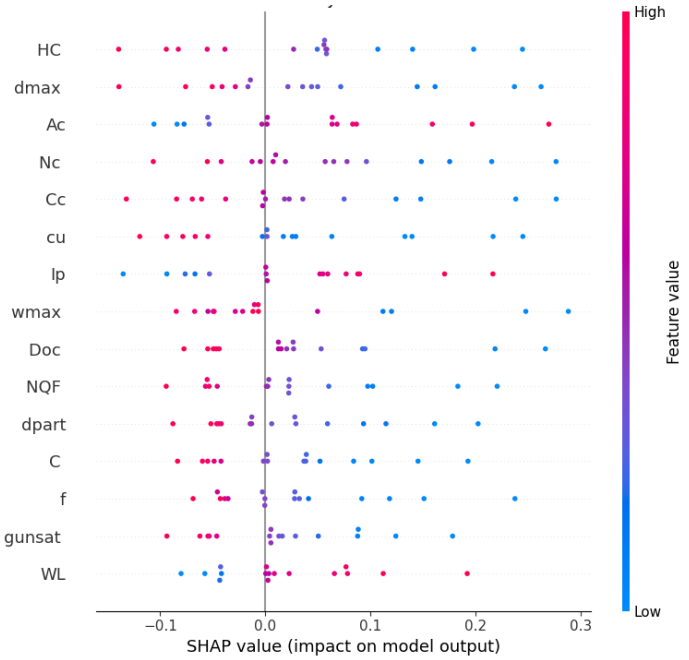

Figure 9. Feature significance derived from SHAP analysis for CatBoost.

Likewise, the SHAP analysis outcomes for the CatBoost model are shown in Figure 9. This figure reveals similarities with Figure 8, as the top five parameters are identical, though their rankings differ. The positions of the remaining nine input parameters vary significantly. Nevertheless, all 14 input parameters influence the output prediction to some extent, so none were excluded from the modeling process.

9 Conclusion

The erodibility of treated unsaturated lateritic soil was estimated using six advanced soft computing approaches, namely ANN, RF, XGB, SVM, CatBoost, and PSO. Based on the model's performance, the following conclusions were drawn:

- The datasets were utilized in various soft computing approaches including ANN, RF, XGB, SVM, CatBoost, and PSO to estimate soil erodibility based on several input variables.
- Model effectiveness was assessed through various error metrics, including, r, R2, M.A.E., M.S.E., R.M.S.E., and MAPE, all of which indicated strong predictive performance.
- The calculated M.A.P.E. values are below 3% for

the CatBoost, RF, XGB, SVM, and ANN models across all three phases: training, testing, and validation.

- Relative importance showed that γ_{max} and HC are the most influencing parameter in predicting the erodibility.

To enhance the prediction of this environmental challenge, it is advisable to explore alternative machine learning methods such as EPR, GP, GEP, and ANFIS. Additionally, expanding the dataset to include at least 300 or more samples is recommended for improved model reliability and performance.

Data Availability Statement

Data will be made available on request.

Funding

This work was supported without any funding.

Conflicts of Interest

The authors declare no conflicts of interest.

Ethical Approval and Consent to Participate

Not applicable.

References

[1] Baghbani, A., Choudhury, T., Costa, S., & Reiner, J. (2022). Application of artificial intelligence in geotechnical engineering: A state-of-the-art review. *Earth-Science Reviews, 228*, 103991. [Crossref]

[2] Yaghoubi, E., Yaghoubi, E., Khamees, A., & Vakili, A. H. (2024). A systematic review and meta-analysis of artificial neural network, machine learning, deep learning, and ensemble learning approaches in field of geotechnical engineering. *Neural Computing and Applications, 36*(21), 12655-12699. [Crossref]

[3] Onyelowe, K. C., Shakeri, J., Amini-Khoshalann, H., Salahudeen, A. B., Arinze, E. E., & Ugwu, H. U. (2021). Application of ANFIS hybrids to predict coefficients of curvature and uniformity of treated unsaturated lateritic soil for sustainable earthworks. *Cleaner Materials, 1*, 100005. [Crossref]

[4] Moges, G., McDonnell, K., Delele, M. A., Ali, A. N., & Fanta, S. W. (2023). Development and comparative analysis of ANN and SVR-based models with conventional regression models for predicting spray drift. *Environmental Science and Pollution Research, 30*(8), 21927-21944. [Crossref]

[5] Kurnaz, T. F., Dagdeviren, U., Yildiz, M., & Ozkan, O. (2016). Prediction of compressibility parameters of the soils using artificial neural network. *SpringerPlus, 5*(1), 1801. [Crossref]

[6] Le, H. A., Nguyen, T. A., Nguyen, D. D., & Prakash, I. (2020). Prediction of soil unconfined compressive strength using Artificial Neural Network Model. *Vietnam Journal of Earth Sciences, 42*, 15342. [Crossref]

[7] Kouchami-Sardoo, I., Shirani, H., Esfandiarpour-Boroujeni, I., Besalatpour, A. A., & Hajabbasi, M. A. (2020). Prediction of soil wind erodibility using a hybrid Genetic algorithm–Artificial neural network method. *Catena, 187*, 104315. [Crossref]

[8] Taghizadeh-Mehrjardi, R., Emadi, M., Cherati, A., Heung, B., Mosavi, A., & Scholten, T. (2021). Bio-inspired hybridization of artificial neural networks: An application for mapping the spatial distribution of soil texture fractions. *Remote Sensing, 13*(5), 1025. [Crossref]

[9] Alavi, A. H., Gandomi, A. H., & Mollahasani, A. (2012). A genetic programming-based approach for the performance characteristics assessment of stabilized soil. In *Variants of evolutionary algorithms for real-world applications* (pp. 343-376). Berlin, Heidelberg: Springer Berlin Heidelberg. [Crossref]

[10] Firoozi, A. A., Firoozi, A. A., Aati, K., Khan, A. H., & Vambol, V. (2025). Integrating the Fourth Industrial Revolution into Geotechnical Engineering: Transformations, Challenges, and Future Directions. *Ecological Questions, 36*(1), 1-41. [Crossref]

[11] Abdul Wahab, N., A Rashid, A. S., Roshan, M. J., Horpibulsuk, S., & Razali, R. (2025). Climate Change Impact on Unsaturated Cement-Stabilized Laterite Soil Using Suction-Controlled Testing. *Geotechnical and Geological Engineering, 43*(1), 11. [Crossref]

[12] Sharma, A., & Sharma, A. (2024). Strength prediction of construction demolition waste and pine needle fibre stabilized soil using artificial neural network. *Multiscale and Multidisciplinary Modeling, Experiments and Design, 7*(3), 1975-1991. [Crossref]

[13] Fattahi, H., Ghaedi, H., Malekmahmoodi, F., & Armaghani, D. J. (2024). Optimizing pile bearing capacity prediction: Insights from dynamic testing and smart algorithms in geotechnical engineering. *Measurement, 230*, 114563. [Crossref]

[14] Bertalot, D., Rattley, M. J., Burbury, D., Abyaneh, S., Mclean, R., & Francis, M. (2023, September). Performance assessment of driveability predictions for different pile geometries and soil conditions. In *SUT Offshore Site Investigation and Geotechnics* (pp. SUT-OSIG). SUT. [Crossref]

[15] Albusoda, B. S. (2024). Skirted foundation, performance, mechanism, and limitations: A review study. *Journal of Engineering, 30*(10), 102-121. [Crossref]

[16] Abdulrasool, A. S., Abbas, S. A., & Abdulrahman, S. M. (2021). Improving Bearing Capacity by Skirted Foundation: A Review Study. *Iraqi Journal of Civil Engineering, 15*(1), 66-71. [Crossref]

[17] Shahani, N. M., Ullah, B., Shah, K. S., Hassan, F. U., Ali, R., Elkotb, M. A., ... & Tag-Eldin, E. M. (2022). Predicting angle of internal friction and cohesion of rocks based on machine learning algorithms. *Mathematics, 10*(20), 3875. [Crossref]

[18] Rehman, M. A., Abd Rahman, N., Ibrahim, A. N. H., Kamal, N. A., & Ahmad, A. (2024). Estimation of soil erodibility in Peninsular Malaysia: A case study using multiple linear regression and artificial neural networks. *Heliyon, 10*(7). [Crossref]

[19] Tadayon, A., Eslami, A., & Fakharian, K. (2024). Geotechnical behavior of conical and skirted foundations through experimental and numerical assessment. *Transportation Infrastructure Geotechnology, 11*(4), 2501-2522. [Crossref]

[20] Nawaz, M. N., Nawaz, M. M., Awan, T. A., Jaffar, S. T. A., Jafri, T. H., Oh, T. M., ... & Azab, M. (2023). A sustainable approach for estimating soft ground soil stiffness modulus using artificial intelligence. *Environmental Earth Sciences, 82*(23), 579. [Crossref]

[21] Kazemi, F., Shafighfard, T., & Yoo, D. Y. (2024). Data-driven modeling of mechanical properties of fiber-reinforced concrete: a critical review. *Archives of Computational Methods in Engineering, 31*(4), 2049-2078. [Crossref]

[22] Awolusi, T. F., Finbarrs-Ezema, B. C., Chukwudulue, I. M., & Azab, M. (2024). Application of artificial intelligence (AI) in civil engineering. In *New Advances in Soft Computing in Civil Engineering: AI-Based Optimization and Prediction* (pp. 15-46). Cham: Springer Nature Switzerland. [Crossref]

[23] Demir, S., & Sahin, E. K. (2023). Predicting occurrence of liquefaction-induced lateral spreading using gradient boosting algorithms integrated with particle swarm optimization: PSO-XGBoost, PSO-LightGBM, and PSO-CatBoost. *Acta Geotechnica, 18*(6), 3403-3419. [Crossref]

[24] Khodadadi, N., Roghani, H., De Caso, F., El-kenawy, E. S. M., Yesha, Y., & Nanni, A. (2024). Data-driven PSO-CatBoost machine learning model to predict the compressive strength of CFRP-confined circular concrete specimens. *Thin-walled structures, 198*, 111763. [Crossref]

[25] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature, 521*(7553), 436-444. [Crossref]

[26] Ghaboussi, J. (2010). Advances in neural networks in computational mechanics and engineering. In *Advances of soft computing in engineering* (pp. 191-236). Vienna: Springer Vienna. [Crossref]

[27] Zhang, A., Lipton, Z. C., Li, M., & Smola, A. J. (2023). *Dive into deep learning*. Cambridge University Press.

[28] Breiman, L. (2001). Random forests. *Machine learning, 45*(1), 5-32. [Crossref]

[29] Cutler, A., Cutler, D. R., & Stevens, J. R. (2012). Random forests. In *Ensemble machine learning* (pp. 157-175). Springer, New York, NY. [Crossref]

[30] Zhang, W., Wu, C., Li, Y., Wang, L., & Samui, P. (2021). Assessment of pile drivability using random forest regression and multivariate adaptive regression splines. *Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards, 15*(1), 27-40. [Crossref]

[31] Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794). [Crossref]

[32] Breiman, L., Friedman, J., Olshen, R. A., & Stone, C. J. (2017). *Classification and regression trees*. Chapman and Hall/CRC. [Crossref]

[33] Wang, L., Wu, C., Tang, L., Zhang, W., Lacasse, S., Liu, H., & Gao, L. (2020). Efficient reliability analysis of earth dam slope stability using extreme gradient boosting method. *Acta Geotechnica, 15*(11), 3135-3150. [Crossref]

[34] Zhang, W. G., Li, H. R., Wu, C. Z., Li, Y. Q., Liu, Z. Q., & Liu, H. L. (2021). Soft computing approach for prediction of surface settlement induced by earth pressure balance shield tunneling. *Underground Space, 6*(4), 353-363. [Crossref]

[35] Zhang, W., Zhang, R., Wu, C., Goh, A. T. C., Lacasse, S., Liu, Z., & Liu, H. (2020). State-of-the-art review of soft computing applications in underground excavations. *Geoscience Frontiers, 11*(4), 1095-1106. [Crossref]

[36] Zhang, W., Li, Y., Wu, C., Li, H., Goh, A. T. C., & Liu, H. (2022). Prediction of lining response for twin tunnels constructed in anisotropic clay using machine learning techniques. *Underground space, 7*(1), 122-133. [Crossref]

[37] Boser, B. E. (1992). Proceedings of the 5th annual ACM workshop on computational learning theory. (No Title), 144.

[38] Vapnik, V. (2013). *The nature of statistical learning theory*. Springer science & business media.

[39] Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning, 20*(3), 273-297. [Crossref]

[40] Navia-Vázquez, A., Gutierrez-Gonzalez, D., Parrado-Hernández, E., & Navarro-Abellan, J. J. (2024). Distributed support vector machines. *IEEE Transactions on Neural Networks, 17*(4), 1091-1097. [Crossref]

[41] Nagaraju, T. V., Bahrami, A., Prasad, C. D., Mantena, S., Biswal, M., & Islam, M. R. (2023). Predicting California bearing ratio of lateritic soils using hybrid machine learning technique. *Buildings, 13*(1), 255. [Crossref]

[42] Freund, Y., Schapire, R., & Abe, N. (1999). A short introduction to boosting. *Journal-Japanese Society For*

Artificial Intelligence, 14(771-780), 1612.

[43] Catani, F., Lagomarsino, D., Segoni, S., & Tofani, V. (2013). Landslide susceptibility estimation by random forests technique: sensitivity and scaling issues. *Natural Hazards and Earth System Sciences, 13*(11), 2815-2831. [Crossref]

[44] Kennedy, J., & Eberhart, R. (1995, November). Particle swarm optimization. In *Proceedings of ICNN'95-international conference on neural networks* (Vol. 4, pp. 1942-1948). IEEE. [Crossref]

[45] Hossain, M. S., & El-Shafie, A. (2014). Evolutionary techniques versus swarm intelligences: application in reservoir release optimization. *Neural Computing and Applications, 24*(7), 1583-1594. [Crossref]

[46] Reynolds, C. W. (1987, August). Flocks, herds and schools: A distributed behavioral model. In *Proceedings of the 14th annual conference on Computer graphics and interactive techniques* (pp. 25-34). [Crossref]

[47] Engelbrecht, A. P. (2007). *Computational intelligence: an introduction* (Vol. 2). Hoboken, NJ, USA: John Wiley & Sons. [Crossref]

[48] Ebid, A. E., Deifalla, A. F., & Onyelowe, K. C. (2023, December). Data utilization and partitioning for machine learning applications in civil engineering. In *International conference on advanced technologies for humanity* (pp. 87-100). Cham: Springer Nature Switzerland. [Crossref]

Tammineni Gnananandarao received the PhD degree in Civil Engineering from National Institute of Technology, Hamirpur, India, in 2021. (Email: anandrcwing@gmail.com; gnananandaraot@adityauniversity.in)

Jayatheja Muktinutalapati received the PhD degree in Civil Engineering from Birla Institute of Technology and Science - Pilani, India, in 2012. (Email: jayathejam@gmail.com; jmuktinutalapati@ricssbe.edu.in)

Dr. Vedprakash Maralapalle received the PhD degree in Civil engineering from NMIMS University, Mumbai in 2023. (Email: civilved@gmail.com)

B A V Ram Kumar received the PhD degree in Civil Engineering from Pondicherry University, Pondicherry, India, in 2024. (Email: ramkumar.bav@gmrit.edu.in)

Dr. CH. Ajay received the Ph.D degree in Civil Engineering from Andhra University, Visakhapatnam, Andhra Pradesh, India, in 2023. (Email: ajay.chappa91@gmail.com ; ajayc@adityauniversity.in)