CME LIVE OPEN ACCESS

Journal of
Advances in Civil and Mechanical Engineering

Research Article ISSN: 3067-2457

Modelling and Prediction of Soil Shear Strength Using Machine Learning

Tony Ndung'u Munene

Department of Civil Engineering, University of Nairobi, P.O. Box 30197, Nairobi, Kenya.

Corresponding Author: Tony Ndung'u Munene, Department of Civil Engineering, University of Nairobi, P.O. Box 30197, Nairobi, Kenya.

Received: 2025 Sep 22 **Accepted:** 2025 Oct 13 **Published:** 2025 Dec 31

Abstract

This study explores the application of machine learning algorithms to predict soil shear strength parameters. Using a dataset of 188 soil samples containing properties such as Atterberg limits, grain size distribution, maximum dry density, and optimum moisture content, eight different machine learning models were trained and evaluated. The models included Linear Regression, Random Forest (baseline and tuned), XGBoost (baseline and tuned), Neural Networks, LightGBM, and Support Vector Regression. Performance was assessed using Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R^2 metrics. Support Vector Regression achieved the best performance with RMSE of 0.053, MAE of 0.0262, and R^2 of 0.41. However, all models showed limited ability to explain variance in shear strength, with R^2 values below 0.5, indicating challenges in predicting this complex geotechnical parameter. The results demonstrate that while machine learning shows promise as an alternative to traditional testing methods, data quality and quantity remain critical limitations for practical implementation.

Keywords: Machine Learning, Soil Shear Strength, Geotechnical Engineering, Regression Analysis, Neural Networks, Random Forest

1. Introduction

The shear strength of soil represents one of the most fundamental parameters in geotechnical engineering, governing foundation design, slope stability analysis, and earth pressure calculations. Traditionally defined by the Mohr-Coulomb failure criterion:

$$\tau f = c' + \sigma' \tan(\phi') \quad (1)$$

where τf is shear strength, c' is effective cohesion, σ' is effective stress, and ϕ' is the angle of internal friction, this parameter is typically determined through laboratory testing methods including triaxial and direct shear tests [1].

Contemporary geotechnical practice faces increasing pressure to deliver rapid, cost-effective soil characterization for infrastructure projects. Traditional laboratory testing, while reliable, presents significant constraints including extended testing periods (16-24 hours for moisture content determination, up to one week for CBR testing), substantial costs for skilled technicians and equipment, and limited testing capacity that may inadequately represent site variability.

Machine learning presents a compelling alternative approach for soil parameter prediction. By leveraging historical testing data and readily measurable soil properties, ML algorithms can potentially provide rapid predictions while maintaining acceptable accuracy levels [2]. This capability becomes particularly valuable in developing countries experiencing rapid urbanization where traditional testing infrastructure may be limited.

Recent applications of machine learning in geotechnical engineering have shown promising results for various soil properties [3]. demonstrated successful prediction of soft soil shear strength using ensemble methods, while [4]. achieved high accuracy in predicting soil-geomembrane interface shear strength. However, these studies often focus on specific soil types or controlled laboratory conditions, limiting their broader applicability.

The primary objective of this research is to evaluate the effectiveness of various machine learning algorithms in predicting soil shear strength using commonly available soil properties. Specific aims include developing predictive models using multiple ML approaches, comparing algorithm performance across different metrics, and assessing the practical viability of ML-based predictions for geotechnical applications.

2. Materials and Methods

2.1. Data Collection and Sources

The dataset comprised historical soil test results from the University of Nairobi Civil Engineering Laboratory, Materials Laboratory, and the Materials Testing and Research Division (MTRD) laboratories in Nairobi. These records span several years of testing and encompass diverse soil types and conditions representative of Kenyan geological formations.

Shear strength determination followed standard procedures for triaxial compression tests and direct shear tests. Triaxial tests were conducted on cylindrical specimens (36 mm diameter, 76 mm length) using confining pressures ranging from 50-400 kPa. Direct shear tests utilized square specimens (60 mm × 60 mm × 25 mm) under normal stresses of 100-300 kPa.

Supporting soil characterization included grain size analysis following ASTM D422, Atterberg limits determination per ASTM D4318, compaction testing according to ASTM D698, and moisture content measurements using ASTM D2216.

2.2. Data Overview and Preprocessing

The initial dataset consisted of 218 rows and 21 columns. After removing 30 duplicate entries, the final dataset contained 188 samples.

Feature selection eliminated variables with insufficient variability (28mm and 20mm sieve fractions with 98% passing) and high correlation with the target variable (friction angle ϕ due to direct relationship with shear strength). A binary plasticity indicator was created based on Atterberg limits data.

Final predictor variables included:

- Liquid limit (LL)
- Plastic limit (PL)
- Plasticity index (PI)
- Linear shrinkage (LS)
- Plasticity modulus (PM)
- Maximum dry density (MDD)
- Optimum moisture content (OMC)
- Binary plasticity classification

Statistic	LL	PL	PI	LS	PM	MDD	OMC	Strength (kg/cm^2)	phi	Free Swell
Count	182	172	172	172	167	218	218	216	216	7
Mean	56.44	30.88	26.78	13.06	1992.04	1573.54	23.77	0.136	23.59	6.73
Std Dev	15.36	8.80	8.89	4.35	1022.27	330.60	8.86	0.096	2.88	6.41
Min	20.00	15.00	7.00	4.00	206.00	1.44	6.50	0.02	15.00	0.27
25th Percentile	46.00	24.00	21.00	10.00	1176.50	1425.50	17.30	0.110	22.00	0.37
Median (50th)	55.50	29.50	26.00	13.00	1855.00	1560.00	23.00	0.135	24.00	7.10
75th Percentile	66.00	37.25	31.00	16.00	2673.00	1777.00	30.25	0.150	25.00	12.35
Max	109.00	50.00	68.00	31.00	5758.00	2205.00	46.50	0.790	36.00	14.30

Table 1: Summary Statistics of Soil Properties

2.3. Exploratory Data Analysis

Target variable (shear strength) exhibited significant right skewness (skewness = 3.59, kurtosis = 16.17), necessitating logarithmic transformation to achieve approximately normal distribution.

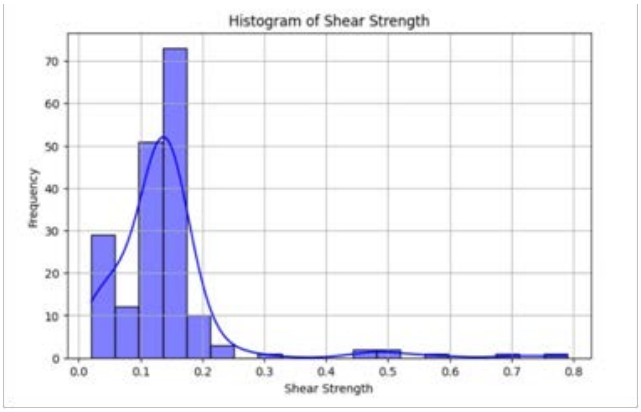

Figure 1: Data Distribution Prior to Log Transformation

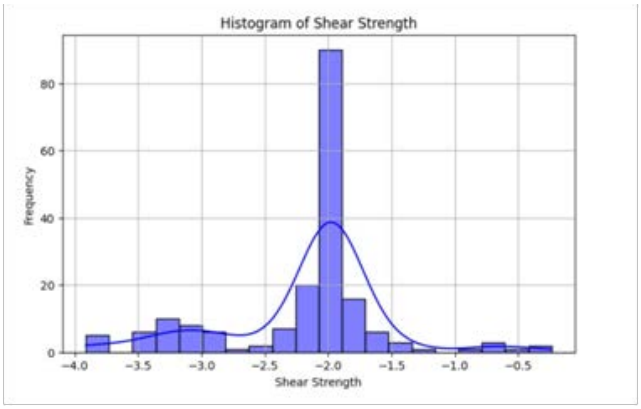

Figure 2: Data Distribution Post Log Transformation

Correlation analysis revealed weak linear relationships between predictor variables and shear strength, with the highest positive correlation being 0.17 with liquid limit.

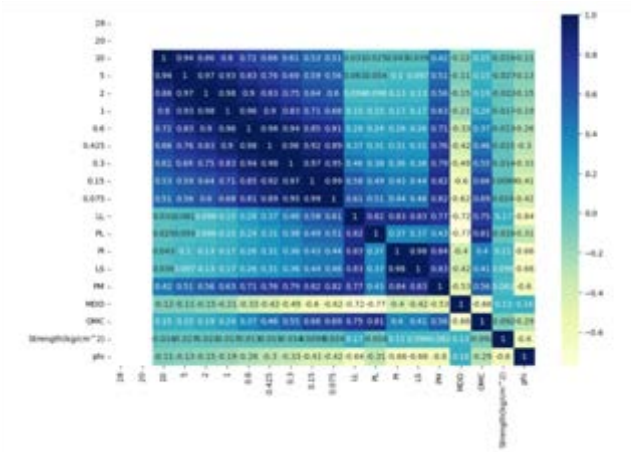

Figure 3: Correlation Heat Map

Predictor variables underwent min-max scaling to standardize ranges using:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \qquad (2)$$

where X' is the scaled value, X is the original value, X_{min} is the minimum value, and X_{max} is the maximum value.
The dataset was randomly divided into training (70%, n=132) and testing (30%, n=56) subsets using stratified sampling.

2.4. Machine Learning Implementation

Eight algorithms were implemented representing diverse ML approaches:

Linear Regression: Baseline linear model assuming linear relationships between predictors and target variable, fitted using ordinary least squares method.

Random Forest: Ensemble method combining multiple decision trees. Baseline model used 100 estimators with default parameters. Hyperparameter tuning via Optuna optimization yielded
Optimal Parameters: n estimators=184, max depth=12, min samples split=9.

XGBoost: Gradient boosting framework building sequential decision trees. Baseline used default parameters (100 estimators, max depth=6, learning rate=0.1). Optimization yielded: n estimators=950, max depth=7, learning rate=0.131, subsample=0.872.

Neural Networks: Shallow network with single hidden layer (30 neurons) using ReLU activation, compiled with Adam optimizer and mean squared error loss function. Training conducted over 100 epochs.

LightGBM: Efficient gradient boosting variant optimized for speed, configured with 200 estimators, 20 leaves, and learning rate of 0.004.

Support Vector Regression: Implemented with RBF kernel, optimized parameters: C=17.85, gamma='scale'.

Performance evaluation utilized three metrics:
Root Mean Squared Error: $RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$
Mean Absolute Error: $MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$
Coefficient of determination: $R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$

3. Results

3.1. Model Performance Comparison

Performance evaluation across all algorithms revealed consistent patterns in predictive capability and limitations. Table 2 presents comprehensive performance metrics for all implemented models.

Model	RMSE	MAE	R^2
Linear Regression	0.0611	0.0300	0.229
Random Forest (Baseline)	0.0631	0.0285	0.179
Random Forest (Tuned)	0.0617	0.0268	0.215
XGBoost (Baseline)	0.0677	0.0304	0.054
XGBoost (Tuned)	0.0614	0.0288	0.224
Neural Network	0.0600	0.0275	0.264
LightGBM	0.0638	0.0350	0.161
Support Vector Regression	0.0530	0.0262	0.410

Table 2: Machine Learning Model Performance

3.2. Algorithm-Specific Analysis

Support Vector Regression demonstrated superior performance across all metrics, achieving the lowest RMSE (0.0530) and highest R^2 value (0.410). The radial basis function kernel effectively captured non-linear relationships between soil properties and shear strength.

Linear Regression performed competitively despite its simplicity (RMSE=0.0611, R^2=0.229), suggesting underlying linear relationships between certain soil properties and shear strength. Tree-based Methods showed moderate improvement through hyperparameter tuning. Random Forest R^2 increased from 0.179 to 0.215, while XGBoost exhibited dramatic improvement from 0.054 to 0.224.

Neural Networks achieved competitive performance (R^2=0.264) despite simple architecture, suggesting limited complexity in the dataset for deep learning benefits.

LightGBM underperformed expectations (R^2=0.161), potentially due to insufficient dataset size for this algorithm's optimization strategy.

Visual analysis of predicted versus actual values revealed consistent clustering patterns across all models, with predictions concentrating around mean values rather than capturing the full range of shear strength variations. This pattern indicates systematic underfitting.

4. Discussion

4.1. Machine Learning Viability Assessment

The results demonstrate both potential and limitations of machine learning approaches for soil shear strength prediction. While all models achieved reasonable error rates (RMSE less than 0.07), consistently low R^2 values (maximum 0.41) indicate substantial challenges in explaining variance in this complex geotechnical parameter.

Traditional empirical correlations for shear strength prediction typically achieve R^2 values of 0.6-0.8 when developed for specific soil types and geological conditions [5]. The lower performance observed likely reflects the heterogeneous nature of the dataset, encompassing diverse soil types without stratification by geological origin or soil classification.

4.2. Data Quality and Quantity Implications

Limited predictive performance highlights critical data requirements for successful ML implementation:

Dataset Size: With only 188 samples, the dataset may be insufficient for complex algorithms to identify subtle patterns. Typical ML applications require hundreds to thousands of samples for reliable pattern recognition [6].

Feature Completeness: Standard geotechnical parameters may not fully capture factors governing shear strength. Additional parameters such as stress history, soil structure characteristics, and mineralogy could improve predictive capability.

Data Homogeneity: Combining diverse soil types without stratification may obscure soilspecific relationships that could be captured by targeted models.

4.3. Practical Implementation Considerations

Despite limitations, results suggest potential applications:

Preliminary Screening: Models could provide rapid initial estimates for project feasibility assessment, with detailed testing reserved for critical applications.

Quality Assurance: ML predictions could flag unusual test results requiring verification, enhancing laboratory quality control procedures.

Resource Optimization: In regions with limited testing facilities, ML models could guide sampling strategies and prioritize testing efforts.

4.4. Comparison with Previous Studies

Achieved R^2 values of 0.51 using SVM for soil moisture prediction, while [8] reported good performance for slope stability computations using neural networks. The lower performance in this study may be attributed to the more complex nature of shear strength prediction and dataset limitations [7,8].

Recent comparative studies by indicate that SVM generally outperforms other algorithms for geotechnical applications, consistent with our findings. However, their review emphasizes the importance of dataset quality and size, which aligns with our observed limitations [2].

4.5. Methodological Limitations

Several factors may have constrained model performance:

Target Variable Distribution: Right-skewed distribution of shear strength values, even after transformation, may have biased model training toward central tendencies.

Limited Features: Absence of critical parameters such as in-situ stress conditions, loading history, and soil moisture content.

Cross-Validation Approach: Holdout validation strategy, while appropriate for dataset size, provides limited insight into model stability across different data partitions.

4.6. Future Research Directions

The development of machine learning applications in geotechnical engineering faces significant methodological and practical challenges that extend beyond simple model selection or parameter tuning. A critical limitation in current research is the systematic underutilization of extensive datasets that already exist within commercial geotechnical laboratories and educational institutions worldwide. Many established testing facilities have accumulated decades of soil testing data across diverse geological formations and project types, yet these valuable resources remain largely untapped due to data

standardization challenges, proprietary concerns, and lack of collaborative frameworks for data sharing.

The fragmentation of existing datasets represents a fundamental barrier to advancing predictive modeling capabilities. Commercial laboratories typically maintain project-specific databases optimized for internal quality control rather than research applications, while academic institutions often focus on controlled experimental conditions that may not reflect the variability encountered in practice. This disconnect between industry data repositories and research initiatives has resulted in machine learning models trained on limited, often homogeneous datasets that fail to capture the full spectrum of soil behavior encountered across different geological settings and construction contexts.

Future research must address the systematic integration of these dispersed data sources through the development of standardized data collection protocols and collaborative platforms that enable secure data sharing while protecting commercial interests. This requires establishing common data formats, quality control procedures, and metadata standards that allow meaningful aggregation of testing results from multiple sources without compromising data integrity or confidentiality.

Beyond data integration challenges, current modeling approaches suffer from oversimplified feature representations that inadequately capture the complex interdependencies governing soil behavior. The reliance on basic index properties and standard test parameters overlooks the wealth of derived parameters and interaction terms that could significantly enhance model performance. For instance, stress history effects, loading rate dependencies, and environmental factors such as temperature and moisture fluctuations are rarely incorporated into predictive models despite their known influence on geotechnical properties.

The development of physics-informed machine learning frameworks represents a particularly promising avenue for addressing the interpretability limitations that currently hinder the adoption of ML approaches in geotechnical practice. Traditional black-box models, while potentially accurate, fail to provide the mechanistic insights that engineers require for confident decisionmaking in critical infrastructure projects. Incorporating established geotechnical principles into model architectures through constraint-based learning or hybrid physics-ML approaches could bridge this gap between predictive accuracy and engineering understanding.

5. Conclusions

This study evaluated eight machine learning algorithms for predicting soil shear strength using standard geotechnical parameters. Support Vector Regression achieved the best performance (RMSE=0.053, R^2=0.41), followed by Neural Networks and tuned ensemble methods. However, all models showed limited ability to explain variance in shear strength, with R^2 values below 0.5.

Key findings include:

- Machine learning algorithms can achieve reasonable prediction accuracy for soil shear strength, with SVR showing superior performance
- Dataset size and quality significantly limit model performance, with 188 samples proving insufficient for complex pattern recognition
- Traditional geotechnical parameters alone may not capture all factors governing shear strength
- Hyperparameter optimization substantially improves ensemble method performance
- Current ML approaches serve better as screening tools rather than replacements for laboratory testing

The results suggest that while machine learning shows promise for geotechnical applications, successful implementation requires larger, more comprehensive datasets and potentially physicsinformed approaches that incorporate domain knowledge. For practical applications, ML models could serve as valuable screening tools and quality assurance measures, but should complement rather than replace traditional laboratory testing methods.

Future work should focus on expanding datasets, incorporating additional soil parameters, and developing soil-type-specific models to improve predictive capability and practical applicability in geotechnical engineering practice.

Acknowledgements

The authors acknowledge the University of Nairobi Civil Engineering Laboratory and Materials
Testing and Research Division for providing access to historical soil testing data.

References

1. Venkatramaiah, C. (2006). Geotechnical engineering: New Age International. *New Delhi.*
2. Shao, W., Yue, W., Zhang, Y., Zhou, T., Zhang, Y., Dang, Y., ... & Chao, Z. (2023). The application of machine learning techniques in geotechnical engineering: A review and comparison. *Mathematics, 11*(18), 3976.

3. Pham, B. T., Hoang, T. A., Nguyen, D. M., & Bui, D. T. (2018). Prediction of shear strength of soft soil using machine learning methods. *Catena, 166,* 181-191.
4. Chao, Z., Shi, D., Fowmes, G., Xu, X., Yue, W., Cui, P., ... & Yang, C. (2023). Artificial intelligence algorithms for predicting peak shear strength of clayey soil-geomembrane interfaces and experimental validation. *Geotextiles and Geomembranes, 51*(1), 179-198.
5. Smith, I. (2014). *Smith's elements of soil mechanics*. John Wiley & Sons.
6. Heung, B., Ho, H. C., Zhang, J., Knudby, A., Bulmer, C. E., & Schmidt, M. G. (2016). An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping. *Geoderma, 265*, 62-77.
7. Ahmad, S., Kalra, A., & Stephen, H. (2010). Estimating soil moisture using remote sensing data: A machine learning approach. *Advances in water resources, 33*(1), 69-80.
8. Kumar, S., & Basudhar, P. K. (2018). A neural network model for slope stability computations. *Géotechnique Letters, 8*(2), 149-154.