(Research article)
Journal of Engineering Technology Access

Determination of Clay Content by Applying Machine Learning with Hydrometer Testing and Specific Gravity Analyses

C. Sukkanon[1], J. Supakosol[1] and P. Chaipanna[1*]

[1]*Faculty of Industry and Technology, Rajamangala University of Technology Isan, Sakon Nakhon Campus*

pattanasak.ca@rmuti.ac.th

Received: 12/03/2025
Accepted: 07/05/2025
Published : 30/06/2025

Abstract

This study aims to analyze and compare hydrometer test results with fundamental soil properties while applying Machine Learning (ML), a branch of Artificial Intelligence (AI), to enhance the speed and accuracy of clay content prediction. The study utilized soil samples from Nakhon Phanom and Sakon Nakhon provinces, Thailand. The experimental process included specific gravity and hydrometer analysis. For ML model development, linear regression (LR) and random forest regressor (RFR) were compared to analyzing factors influencing clay content. The data evaluation was based on feature importance analysis and statistical correlation (Correlation Matrix). The application of 10-fold cross-validation ensured that the models did not suffer from overfitting and confirmed the stability of predictions when using hydrometer data from longer test durations. The results indicate that hydrometer readings at longer durations exhibit a strong correlation with clay content and significantly improve the prediction accuracy of LR and RFR. The highest R^2 values obtained were 0.93 for LR and 0.87 for RFR, demonstrating that longer hydrometer test durations lead to more accurate clay content predictions. ML method combined with the hydrometer readings at 180 minutes, the R^2 exceeds 0.75. Specifically, LR outperformed RFR at minute 240, suggesting that the linear model better explains data variance at this duration. This research concludes that incorporating ML with hydrometer test data significantly improves the accuracy of clay content predictions. The findings highlight the potential of ML applications in soil property analysis and geotechnical engineering design, leading to more efficient and reliable engineering solutions.

Keywords: Clay content; Hydrometer; Basic properties; Artificial Intelligence; Machine Learning

1. INTRODUCTION

The presence of high clay content in soils significantly affects the stability and integrity of engineering structures. Due to its high-water absorption and expansion properties, clayey soils undergo volumetric changes upon moisture variation, which can cause subsidence or swelling, leading to structural failures (Ural, 2018). Such soil behavior increases maintenance costs and necessitates corrective measures for infrastructure projects (Terzaghi et al., 1996).

Various methods exist for determining clay content, each with its own advantages and limitations. Sieve analysis is widely used for coarse-grained soils like sand and gravel, where particle size is determined using a series of sieves with different mesh sizes. However, sieve analysis is ineffective for particles smaller than 0.063 mm (Gee & Or, 2002), making it unsuitable for clay and silt.

Laser Diffraction Analysis is another advanced technique that measures particle size using light scattering principles. It provides rapid results with high accuracy and can analyze a broad range

(Research article)
Journal of Engineering Technology Access

of particle sizes, from microns to millimeters (Eshel et al., 2004). However, this method requires specialized equipment and may not be cost-effective for routine laboratory testing.

The hydrometer method, established in 1927, is a sedimentation-based technique widely used to determine the particle size distribution of fine-grained soils, particularly silts and clays. This method operates on the principle of sedimentation, where soil particles suspended in a liquid settle at velocities proportional to their size, density, and the fluid's viscosity, as described by Stokes' Law (Das & Sobhan, 2018). The hydrometer measures the relative density of the suspension over time, allowing for the calculation of particle size distribution. The hydrometer method is particularly effective for analyzing fine-grained soils where traditional sieve analysis is impractical. It provides a continuous particle size distribution curve, offering detailed insights into soil composition. Additionally, it is cost-effective and standardized, making it accessible for routine soil analysis. Despite its widespread use, the hydrometer method has inherent limitations. It assumes that soil particles are spherical and of uniform density, which is often not the case in natural soils. Clay particles, for instance, are typically plate-shaped, leading to deviations from theoretical settling velocities predicted by Stokes' Law. Moreover, the method requires precise temperature control, as fluid viscosity changes can significantly affect the settling rates. The presence of dispersing agents, such as sodium hexamethaphosphate, is necessary to prevent flocculation, however, achieving complete dispersion can be challenging.

Despite advancements in soil analysis techniques, the hydrometer method remains widely used in geotechnical engineering due to its low cost, simplicity, and standardized procedures. The integration of hydrometer method and ML enhances data analysis efficiency, providing accurate predictions while reducing testing time (Vargas-Zapata et al., 2025; Zhu et al., 2018).

This research explores ML applications in soil analysis to predict clay content by investigating its relationship with specific gravity and hydrometer readings, developing and comparing Linear Regression and Random Forest Regressor models, and evaluating their performance with R2 to enhance accuracy and efficiency over traditional methods.

2. OBJECTIVES

1. To investigate the relationship between clay content and specific gravity combined with hydrometer readings at various time intervals.

2. To develop and compare ML models for clay content prediction using Linear Regression and Random Forest Regressor, evaluating their performance using the R^2 coefficient.

3. RESEARCH METHODOLOGY

3.1 Basic Soil Property Testing

The hydrometer analysis method, based on Stokes' Law, determines particle size distribution by measuring sedimentation rates in a fluid medium (ASTM D7928-17, 2017*)*. This test involves dispersing soil particles in a liquid and measuring fluid density at different depths over time using a hydrometer. Larger particles settle faster than smaller ones, allowing for particle size determination based on sedimentation rates. Stokes' Law (Das & Sobhan, 2018) defines the velocity of particle settling as:

$$d = \sqrt{\frac{18\mu Hy}{(G_s - G_w)gt}} \tag{1}$$

(Research article)
Journal of Engineering Technology Access

where d is particle diameter (mm), μ is viscosity of water (Pa·s), Hy is the effective depth (m) of hydrometer , G_s is specific gravity of soil particles, G_w is specific gravity of water, g is gravitational acceleration and t is the time elapsed (minute)

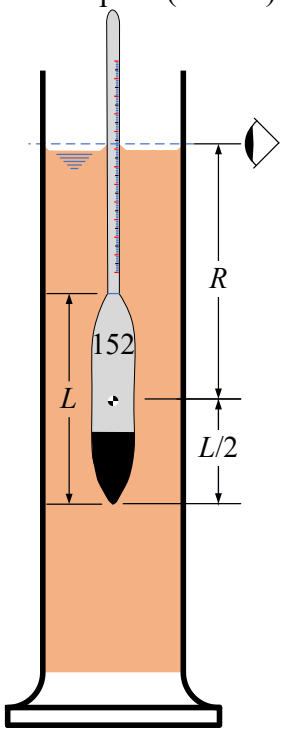

Figure 1 Hydrometer Reading

Soil samples (31 in total) were collected from Nakhon Phanom and Sakon Nakhon provinces, Thailand. Samples were sieved using a No. 40 sieve for specific gravity testing and a No. 200 sieve for hydrometer analysis, consist of both silt and clay. The hydrometer used, type 152, weighs 78 grams. Hydrometer readings (R) were recorded from 15 seconds ($Hy15s$) to 1440 minutes ($Hy1440m$) as shown in Figure 1. The effective depth (Hy) can be calculated as follows:

$$Hy = R + C_m \pm C_t - C_d \tag{2}$$

where R is the hydrometer reading, C_m is meniscus correction, C_t is temperature correction, C_d is dispersing agent correction.

3.2 Machine Learning (ML)

ML techniques were used to optimize soil property analysis. The study compared: Linear Regression (LR)**:** A simple predictive model assuming linear relationships. Random Forest Regressor (RFR)**:** An ensemble learning model that enhances prediction accuracy by averaging multiple decision trees.

3.3 Model Training and Validation

The dataset was divided using 10-Fold Cross-Validation (Scikit-learn, 2025) to ensure stability and prevent overfitting. The data distribution for Cross-Validation is shown in Figure 2.

(Research article)
Journal of Engineering Technology Access

3.4 Data Analysis and Interpretation

Table 1 summarizes the statistical properties of the clay content, specific gravity, and hydrometer readings at different time intervals, providing an overview of the variation in the soil samples used for analysis.

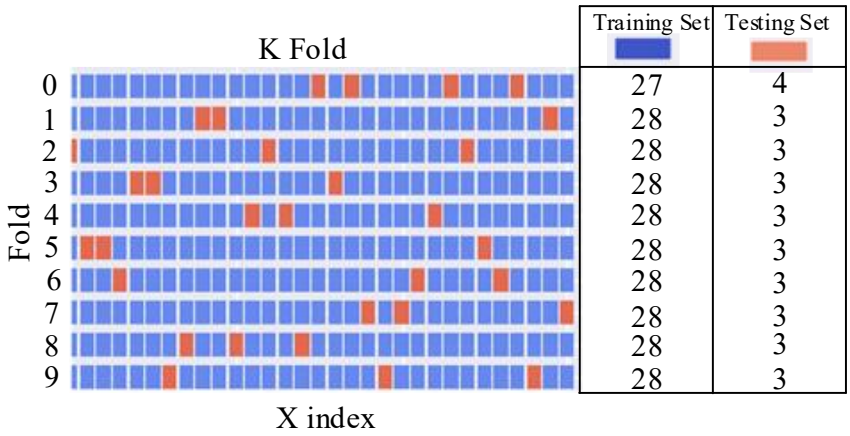

Figure 2 Cross-Validation

The mean clay content is 23.55% with a standard deviation of 14.04, indicating a moderately high variation in soil properties across samples. The minimum (9.09%) and maximum (61.95%) values suggest a significant disparity in clay content among samples, which could be attributed to variations in sampling locations. The high standard deviation implies that the dataset includes a mixture of different soil classifications, ranging from sandy silt to highly clayey soils.

The specific gravity values show a narrow range (Min = 2.55, Max = 2.75) with a low standard deviation (0.06). These values are within the expected range for typical clay and silt soils (2.6 - 2.8) (Holtz et al., 2011), confirming the dataset's reliability. Since G_s remains relatively stable, it may not be a dominant predictor variable in *ML* models but serves as a secondary feature to improve predictions.

As expected, the hydrometer readings decrease over time, demonstrating sedimentation of fine particles. Initial readings (*Hy15s* = 52.05% mean) are the highest due to suspended fine particles, whereas *Hy1440m* = 10.67% mean indicates the final settling phase. High standard deviations at earlier times (e.g., *Hy5m* = 6.20, *Hy10m* = 6.97) suggest substantial variation in soil suspension behavior among samples. At *Hy420m* and *Hy180m*, variability decreases, indicating that these time points may be more stable for modeling clay content.

Table 1 Statistical data for data analysis

Variables	Count	Mean	Std	Min	25%	50%	75%	Max
Clay	31	23.55	14.04	9.09	12.96	18.75	28.25	61.95
G_s	31	2.65	0.06	2.55	2.60	2.65	2.69	2.75
Hy15s	31	52.05	1.88	47.81	51.13	51.88	52.81	57.81
Hy30s	31	49.12	1.94	44.44	48.00	48.81	50.03	53.19
Hy1m	31	46.25	2.82	38.44	44.16	46.81	48.00	51.19
Hy2m	31	41.31	5.33	21.44	39.00	41.88	45.19	48.19
Hy5m	31	35.54	6.20	18.44	32.00	35.81	39.34	46.19
Hy10m	31	29.92	6.97	15.44	23.97	30.19	34.81	44.19
Hy20m	31	24.86	8.41	10.19	17.47	25.88	31.81	41.19
Hy40m	31	20.41	8.62	6.81	12.62	21.88	25.81	40.19

(Research article)
Journal of Engineering Technology Access

Variables	Count	Mean	Std	Min	25%	50%	75%	Max
Hy80m	31	17.40	8.58	6.81	10.50	15.81	21.84	39.19
Hy180m	31	15.28	8.36	5.81	7.88	12.81	20.31	35.57
Hy240m	31	14.62	7.94	5.81	7.53	12.13	19.81	33.57
Hy420m	31	13.10	7.15	4.81	7.38	11.19	15.88	33.57
Hy1440m	31	10.67	6.00	4.50	7.03	8.81	12.65	32.57

3.5 Performance Measurement

In evaluating the predictive accuracy of *ML* models, particularly in regression problems, a key performance measures metric is commonly used. R-Squared (R^2) (Gao, 2024) metric quantifies the proportion of variance in the dependent variable (e.g. clay) that is explained by the independent variable (e.g. hydrometer readings as different times). It is defined as:

$$R^2 = 1 - \frac{\sum\left(y_i - y_i\right)^2}{\sum\left(y_i - \bar{y}_i\right)^2} \tag{3}$$

where y_i is actual observed values, y_i is predicted values from the model, $\bar{y}_i$ is mean of actual observed values

Interpretation of R^2 values:

$R^2 = 1$ Perfect prediction (model explains 100% of the variance)
$R^2 = 0$ Model does not explain any variance beyond the mean prediction
$R^2 < 0$ The model performs worse than a simple mean predictor.

4. RESEARCH RESULTS

4.1 Importance and Correlation of Feature

In Figure 3, the RFR model was utilized to evaluate the importance of various features in predicting clay content. The analysis showed that the most influential variables for predicting clay content were hydrometer readings taken at longer durations, specifically at *Hy420m*, *Hy1440m*, *Hy240m*, and *Hy180m*. These durations exhibited higher feature importance than the physical property of specific gravity (G_s), which was found to have a moderate effect. In contrast, the shorter hydrometer durations such as *Hy15s*, *Hy30s*, and *Hy1m* demonstrated low importance, indicating that they had a minimal contribution to accurate predictions of clay content.

The correlation matrix further supported these findings. It revealed a strong relationship between clay content and hydrometer readings at longer durations, especially at *Hy180m* with an R^2 value close to 0.9, as shown in Figure 4. This suggests that longer hydrometer test durations provide more reliable predictions for clay content. On the other hand, shorter durations like *Hy15s*, *Hy30s*, and *Hy1m* showed significantly weaker correlations, reinforcing the idea that selecting longer durations such as *Hy180m*, *Hy80m*, and *Hy240m* for feature inclusion can improve the model's predictive performance.

4.2 Accuracy of the Analysis

As shown in Figure 5, the evaluation of the ML models for predicting clay content was conducted through R^2, using 10-Fold Cross-Validation to reduce overfitting. The experimental

results indicate that the selection of hydrometer test durations directly affects the performance of the models. The comparison of two datasets is as follows:

- **Dataset 1**: Uses values from *Hy15s* to *Hy420m*
- **Dataset 2**: Uses values from *Hy10m* to *Hy420m*

It was found that the highest R^2 value occurred at *Hy180m*, where the RFR gave an R^2 value of 0.92 and LR gave an R^2 value of 0.85. This means that RFR explains the data variance best at *Hy180m*, whereas shorter durations, such as *Hy10m*, gave lower R^2 values, with some even being

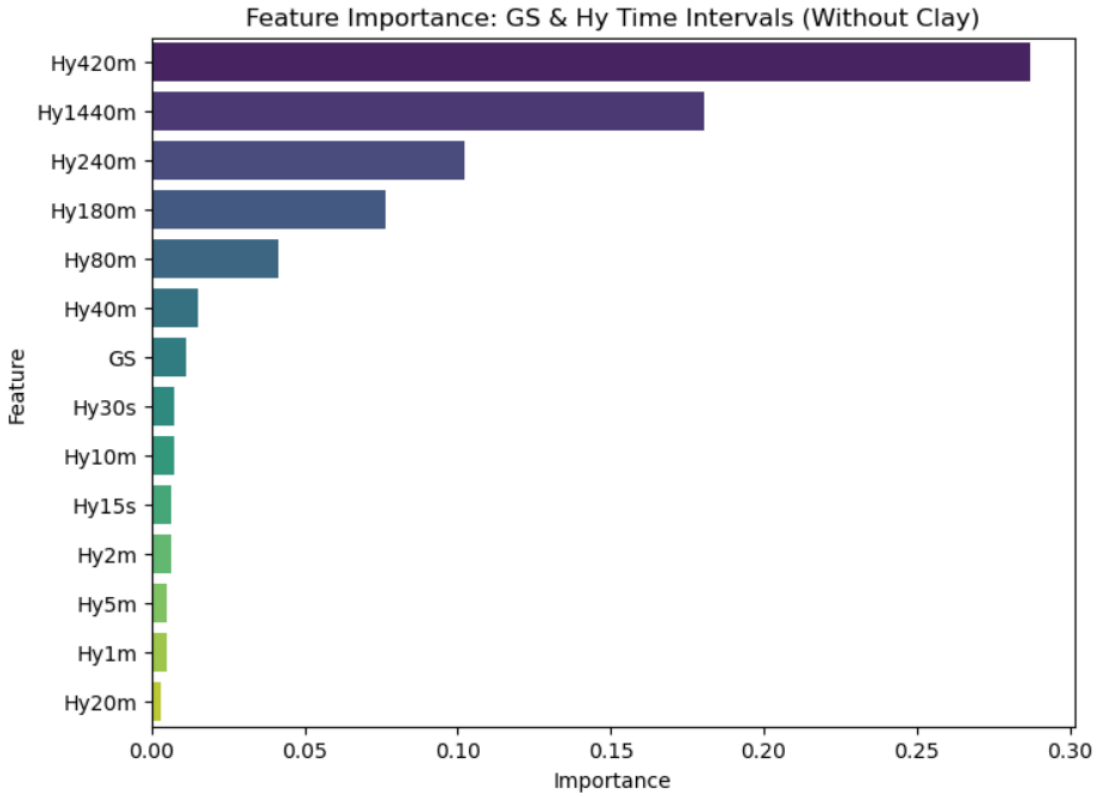

Figure 3 Feature Importance

negative, indicating poor prediction accuracy. Longer durations, such as Hy180m, allow for complete sedimentation of fine particles, leading to a more measurement of clay content. This is supported by stdud

Using data starting from *Hy10m* onwards gave higher prediction accuracy compared to datasets with shorter durations like *Hy15s* or *Hy30s*, which showed lower and unstable R^2 values. The selection of *Hy180m* as the most appropriate duration was based on a combination of three key factors:

1. Feature Importance (Figure 3)
   - Although *Hy420m* and *Hy1440m* had the highest feature importance, *Hy180m* also had a high importance value.
   - Selecting *Hy180m* reduced the testing time without sacrificing model accuracy.
2. Correlation Matrix (Figure 4)
   - *Hy180m* had a high correlation with clay content (0.83), demonstrating that this time still accurately reflects the soil properties without needing a longer duration.
3. R^2 Comparison Graph (Figure 5)
   - *Hy180m* showed the highest R^2 within the appropriate duration (0.92 for RFR and 0.85 for LR). R^2 values greater than 0.7 are acceptable in research (Musafar et al., 2023).

- It reduced testing time from *Hy420m* while maintaining high accuracy.
- At 240 minutes, the data followed a more linear trend, which made LR better suited to explain the variance in clay content, RFR, while poerful for non-linear data did not perform as well at this duration due to linear nature of the data.

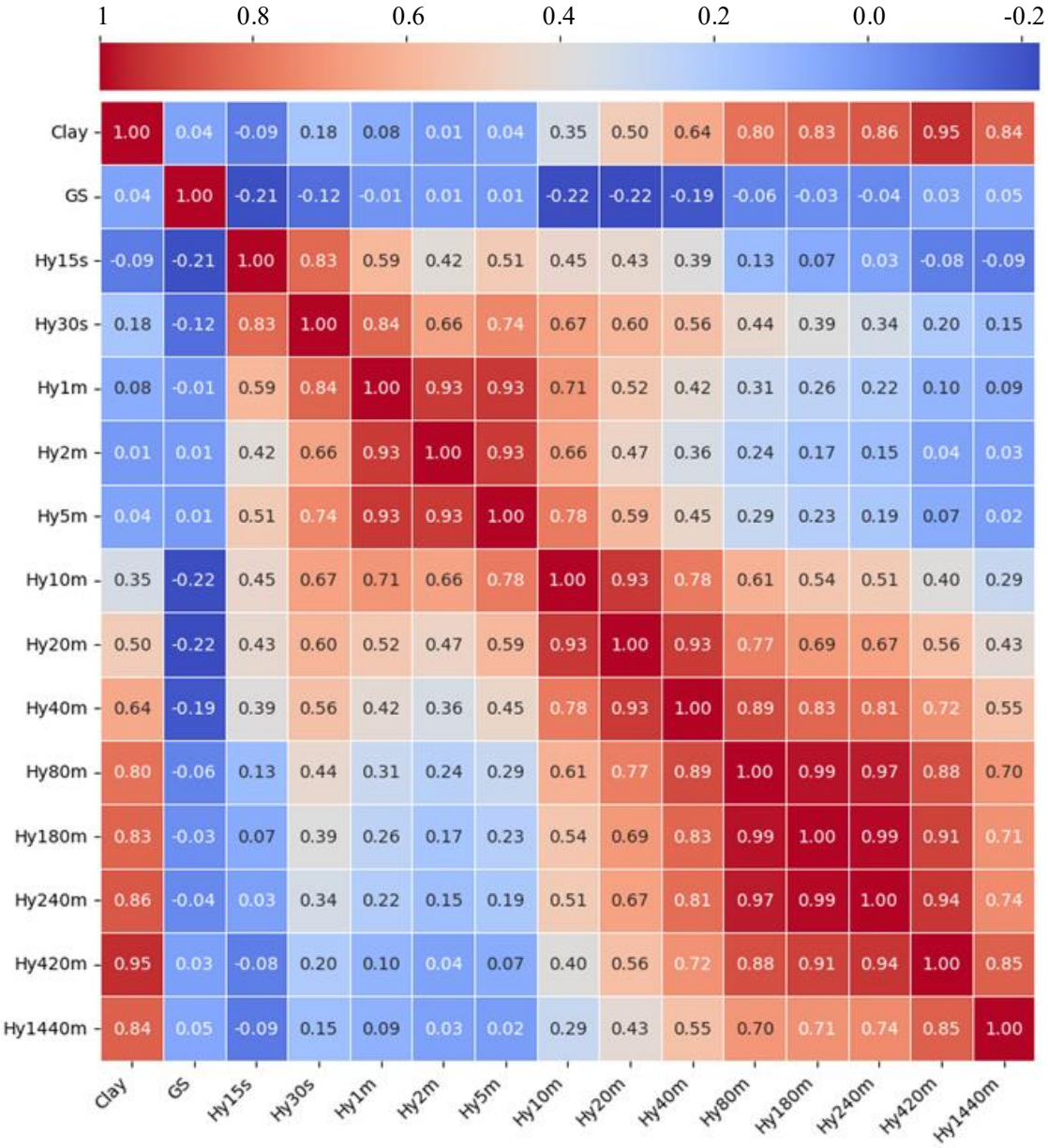

Figure 4 Correlation Matrix

Model Tuning and Reducing Testing Duration The results showed that using *Hy180m* reduced the testing duration compared to *Hy420m* while maintaining the highest accuracy. Choosing the period from *Hy10m* to *Hy180m* as input variables (Feature Selection) helped reduce model complexity and increased testing speed without compromising prediction accuracy. Additionally, using 10-Fold Cross-Validation ensured that the models were not overfitting and could predict new data accurately.

(Research article)
Journal of Engineering Technology Access

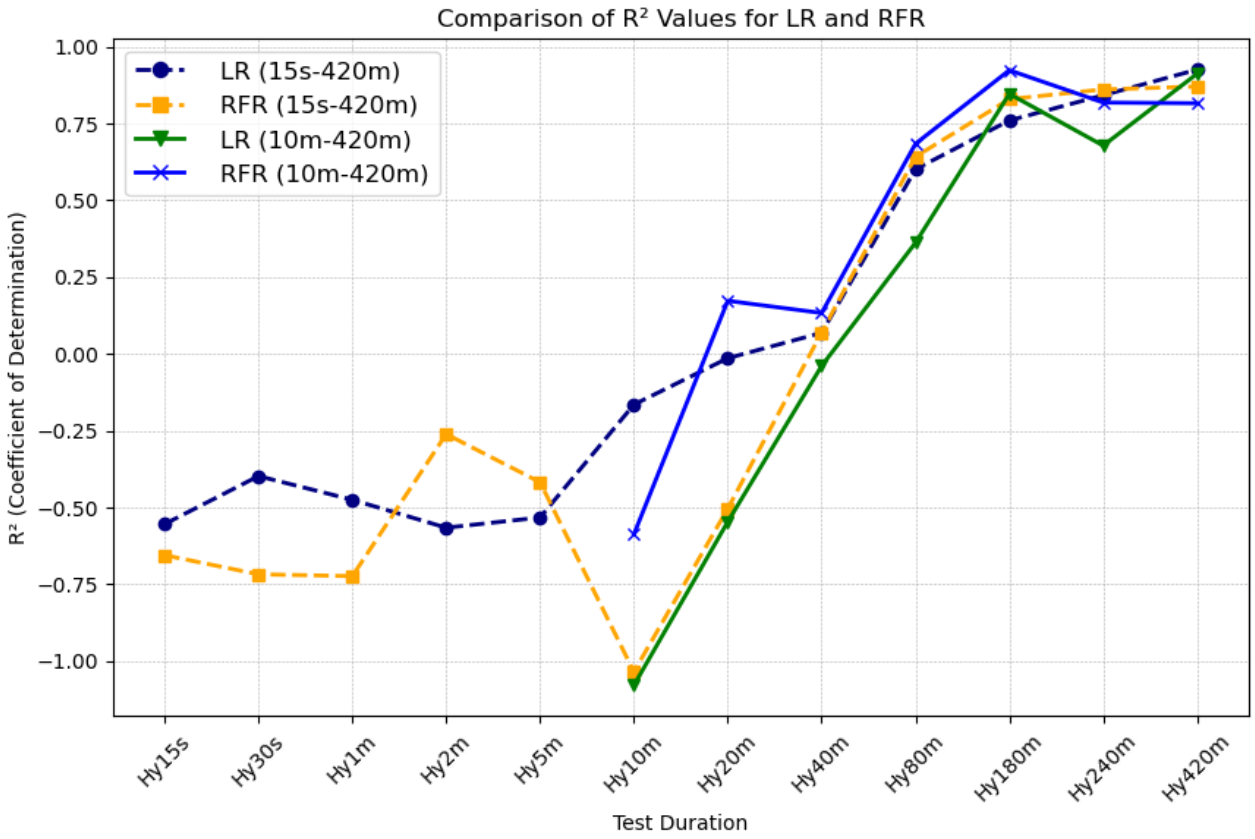

Figure 5 Performance comparison of LR and RFR models

5. CONCLUSIONS

This study investigates the integration of ML techniques with hydrometer testing to enhance the prediction of clay content in soils. The research compares two ML models LR and RFR to determine their effectiveness in predicting clay content using specific gravity (Gs) and hydrometer readings at various time intervals. Key Findings:

- Hydrometer Readings Influence Prediction Accuracy: Longer hydrometer test durations (e.g., *Hy180m*, *Hy240m*, *Hy420m*) showed a stronger correlation with clay content. The highest R^2 values were 0.87 for LR and 0.93 for RFR, demonstrating the importance of selecting the optimal test duration. The ML method combined with Hy180m readings resulted in R^2 exceeding 0.75, making it an effective balance between test duration and prediction accuracy.
- Model Performance Comparison: RFR outperformed LR in handling complex data and reducing prediction errors. LR performed best at *Hy240m*, suggesting it is better suited for cases where a linear relationship is dominant. Shorter hydrometer durations (e.g., *Hy15s*, *Hy30s*) showed low feature importance and weak correlation with clay content.
- Feature Selection and Model Optimization: Removing low-correlation variables improved model efficiency while reducing overfitting risks. 10-Fold Cross-Validation ensured stable predictions.

The study confirms that integrating ML with hydrometer analysis significantly improves clay content prediction accuracy, reducing testing time while maintaining reliability. The findings support ML applications in geotechnical engineering for more efficient and precise soil property analysis.

(Research article)
Journal of Engineering Technology Access

6. FUTURE RESEARCH DIRECTIONS

- Expanding ML models to include Deep Learning (e.g., ANN, CNN, RNN) for more complex soil behavior prediction.
- Develop a software tool or web application that integrates ML models for real-time predictions of clay content in the field.
- Integration of additional soil properties such as Atterberg limits, compaction characteristics, and mineral composition to improve predictive models.
- Developing explainable AI techniques to enhance model interpretability for engineering applications.

7. ACKNOWLEDGEMENTS

The authors would like to express their gratitude for the financial support from the Fundamental Fund for the fiscal year 2025 (FF68/SKC059). Special thanks are extended to colleagues and research assistants for their invaluable help in field data collection and analysis, as well as for their technical and academic support, which have been crucial in the successful completion of this research.

8. REFERENCES

[1] ASTM International. (2017). *ASTM D7928-17, Standard test method for particle-size distribution (gradation) of fine-grained soils using the sedimentation (hydrometer) analysis*. https://doi.org/10.1520/D7928-17

[2] Das, B. M., & Sobhan, K. (2018). *Principles of Geotechnical Engineering* (9th ed.). Boston, MA: Cengage Learning.

[3] Eshel, G., Levy, G. J., Mingelgrin, U., & Singer, M. J. (2004). Critical review of laser diffraction analysis in particle size measurement. *Soil Science Society of America Journal, 68*(3), 736–743. https://doi.org/10.2136/sssaj2004.7360

[4] Gao, J. (2024). R-Squared (R2)–How much variation is explained?. *Research Methods in Medicine & Health Sciences*, *5*(4), 104-109.

[5] Gee, G. W., & Or, D. (2002). Particle-size analysis using sieve and hydrometer methods. In J. H. Dane & G. C. Topp (Eds.), *Methods of soil analysis: Part 4 physical methods* (pp. 255–293). Madison, WI: Soil Science Society of America.

[6] Holtz, R.D., Kovacs, W.D., Sheahan, T.C., (2011). An Introduction to Geotechnical Engineering, Pearson Education, Upper Saddle River, N.J.

[7] Muzafar, S. A., Ali, K. N., Kassem, M. A., & Khoiry, M. A. (2023). Civil engineering standard measurement method adoption using a structural equation modelling approach. *Buildings*, *13*(4), 963.

[8] Scikit-learn. (2025). Cross-validation evaluation estimator performance. Scikit-learn. Retrieved March 10, 2025, from https://scikit-learn.org/stable/modules/cross_validation.html

[9] Terzaghi, K., Peck, R. B., & Mesri, G. (1996). *Soil mechanics in engineering practice* (3rd ed.). New York, NY: Wiley.

[10] Ural, N. (2018). The Importance of Clay in Geotechnical Engineering. InTech. doi: 10.5772/intechopen.75817

[11] Vargas-Zapata, M., Medina-Sierra, M., Fernando Galeano-Vasco, L., & Fernando Cerón

[12] Muñoz, M. (2025). Development of Machine Learning Models for Predicting Soil Texture Variables through Hyperspectral Imaging. IntechOpen. doi: 10.5772/intechopen.1009853

[13] Zhu, Q., Lin, H., & Tang, J. (2018). Application of machine learning in soil property prediction. *Computers and Electronics in Agriculture, 154*,448–457. https://doi.org/10.1016/j.compag.2018.09.020