Machine Learning Estimation of undrained shear strength of Bangkok clay [Tentative]

Viroon Kamchoom [a], Ankit Garg [b], Sai krishna Akash Ramineni [c], Thanu Harnpattanapanich [d,e], Phichet Ratanaprasatkul [f]

[a]Excellent centre for green and sustainable infrastructure, Department of Civil Engineering, School of Engineering, King Mongkut's Institute of Technology Ladkrabang (KMITL), Bangkok, Thailand
[b]University in Shantou, Shantou, China
[c]Undergraduate, Department of Civil Engineering at IIT Indore, Indore city, Madhya Pradesh, Postal code- 453552, India
[d] Geotech Pillar co., ltd, Bangkok, Thailand
[e] Thailand Underground and Tunnelling Group, The Engineering Institute of Thailand, Thailand
[f] Bureau of Engineering and Architectural Design, Royal Irrigation Department (RID), Bangkok, Thailand

Abstract

content...

Keywords:

1 Introduction

Undrained shear strength (Su), representing the maximum shear stress soil can sustain under undrained conditions, is critical for assessing the short-term stability of geotechnical structures such as foundations and slopes [1]. Its determination traditionally relies on laboratory tests like unconsolidated undrained triaxial and falling cone methods, which face limitations in replicating in-situ stress conditions and microstructure, particularly in complex deposits like Bangkok clay [1, 2]. While correlations with standard penetration test (SPT) and cone penetration test (CPT) data offer practical alternatives, these methods require site-specific calibration and exhibit substantial uncertainties, with SPT showing poor reliability in soft clays [2, 3, 4]. Hydrogeological factors, including pore pressure variations, can alter su ratios by over 10%, emphasizing the need for precise, context-sensitive analyses [5].

Bangkok clay presents one of the most challenging geotechnical materials encountered in modern engineering practice, characterized by complex mineralogical composition and highly variable engineering properties that have long troubled infrastructure development in Thailand and similar marine alluvial environments worldwide [6, 7]. Recent advances in machine learning technologies have emerged as promising solutions for accurately predicting the undrained shear strength of this problematic soil, offering significant improvements over traditional empirical methods and laboratory testing approaches [8, 9, 10]. Contemporary research demonstrates that sophisticated machine learning models, including hybrid stacking algorithms, random forest, support vector regression, and ensemble methods, can achieve prediction accuracies ranging from 51% to over 87% when properly calibrated with comprehensive geotechnical datasets [9, 10, 11]. The integration of advanced data preprocessing techniques, Bayesian optimization algorithms, and multi-algorithm approaches has shown particular promise in addressing the inherent variability and complex behavior of Bangkok clay, while providing more efficient and cost-effective alternatives to extensive field investigation programs [9, 10, 12].

The clay formation extends to depths of up to 80 meters beneath Bangkok city, with distinct stratifications: a soft, meta-stable layer (4–11 m depth), a stiff clay layer (12–24 m), and deeper hard clay layers, each exhibiting unique microstructural behaviors influenced by geological processes such as sea-level fluctuations and loading-unloading cycles [13]. The predominant clay mineral is high-swelling smectite, differentiating it from low-swelling smectite clays like Ariake, and contributing to its high plasticity (liquid limit: 100–140%, activity: 1.25–1.90) and significant volume change potential [14].These properties, combined with complex responses to chemical treatments and cation exchange, necessitate specialized soil improvement techniques such as Liquefied Stabilized Soil (LSS), Controlled Low Strength Material (CLSM), and liquefied rubber applications to address infrastructure challenges, particularly in pipeline construction [15].

The study title “Modeling Undrained Shear Strength of Sensitive Alluvial Soft Clay Using Machine Learning Approach” analyzed 111 geotechnical samples of sensitive alluvial clay from Egypt’s Nile Delta, using eight easily measurable soil properties as input features: water content (W_n), liquid limit (LL), dry unit weight (γ_d), plasticity index (PI), consistency index (CI), void ratio (e), specific gravity (G_s), and pocket penetration shear (q_p) [8]. These parameters were selected based on their direct physical relationships with undrained shear strength (USS)—for example, higher LL, PI and γ_d correlated with increased USS due to greater clay content and particle density, while W_n and e showed inverse relationships [8]. The dataset was normalized to address scale disparities, and feature importance analysis confirmed (q_p) as the strongest predictor, aligning with its empirical link to field vane shear test results [8].

Among five machine learning models tested—linear regression, regression trees, Gaussian process regression, ensemble trees, and support vector regression (SVR)—the fine Gaussian SVR achieved superior accuracy, with a testing-set R^2 of 0.96 and RMSE of 1.65 [8]. Hyperparameter optimization via Bayesian methods yielded an optimal kernel scale of 8.69 and box constraint (C) of 11.83, enabling the model to capture nonlinear interactions between features without overfitting [8]. The SVR’s performance significantly outperformed traditional empirical correlations (e.g., those based solely on PI or W_n), demonstrating its utility in bypassing costly field vane tests and disturbed laboratory sampling for USS estimation [8].

The research titled “Prediction of shear strength of soft soil using machine learning methods ” specifically investigates and compares the predictive performance of four machine learning methods—Particle Swarm Optimization Adaptive Network-based Fuzzy Inference System (PANFIS), Genetic Algorithm Adaptive Network-based Fuzzy Inference System (GANFIS), Support Vector Regression (SVR), and Artificial Neural Networks (ANN)—for estimating the shear strength of soft soils. The study utilizes 188 plastic clay soil samples collected from the Nhat Tan and Cua Dai bridge projects in Vietnam, with input variables including moisture content, clay content, liquid limit, plastic limit, plastic index, and consistency index. The models were trained and validated using these datasets, and their performance was assessed using Root Mean Square Error (RMSE) and correlation coefficient (R) [16].

The findings reveal that PANFIS achieved the highest prediction accuracy (RMSE = 0.038, R = 0.601), outperforming GANFIS (RMSE = 0.04, R = 0.569), SVR (RMSE = 0.044, R = 0.549), and ANN (RMSE = 0.059, R = 0.49). This comparative analysis demonstrates that PANFIS is a promising technique for predicting the strength of soft soils, providing a more effective and accurate alternative to other machine learning approaches tested in this context [16].

The study “Estimation of the undrained shear strength of sensitive clays using optimized inference intelligence system ” introduces two hybrid machine learning models, ANFIS-CA (adaptive neuro-fuzzy inference system with cultural algorithm) and ANFIS-PSO (adaptive neuro-fuzzy inference system with particle swarm optimization), to predict undrained shear strength using five input parameters: depth, effective vertical stress, natural water content, liquid limit, and plastic limit. Unlike traditional empirical models (e.g., Hansbo 1957, Chandler 1988) that rely on error-prone pre-consolidation pressure measurements [17, 18], this approach eliminates the need for pre-consolidation data, addressing sampling disturbance challenges common in sensitive clays.

The models were trained on 216 Finnish sensitive clay samples, with ANFIS-PSO achieving superior accuracy (R = 0.715) compared to ANFIS-CA (R = 0.6) [17]. This performance surpasses empirical correlations (R = 0.67–0.71) [18], demonstrating the viability of machine learning for structured clay prediction. Key innovations include metaheuristic optimization (PSO/CA) to refine ANFIS parameters and a focus on vertical effective stress as the most influential input variable (RF weight: 283.33) [17].

2 Material and Methods

2.1 Soil sample collection and testing

Soft Bangkok clay samples were collected from two different areas in the western part of Bangkok, Thailand as shown in Figure-1. The soil investigation was carried out in order to obtain the design parameters for the water drainage tunnel under Bang Nam Chued and Likij canals, which are part of a national scheme to improve the drainage system and mitigate the recurring inundation problem in the lower Chao Phraya River Basin. A total of 54 boreholes were used in this study. Boreholes with a diameter of 10 cm were drilled using a power auger to a depth of 2–3 meters, and then percussion wash boring was carried out throughout the borehole depth. The soft clay layer was found to vary within 12–15 m depth. The undisturbed samples were collected according to ASTM D1587 by using a Thin-Walled Shelby Tube with a 7.5-cm diameter and a 75-cm length. The samples were collected in soft to medium clay layers at every 1 m. The tube was pressed into the soil about 0.5 m depth and then twisted to take each soil sample out of the borehole. After that, each tube was waxed on both sides to prevent any moisture loss from the soil and transported to the laboratory.

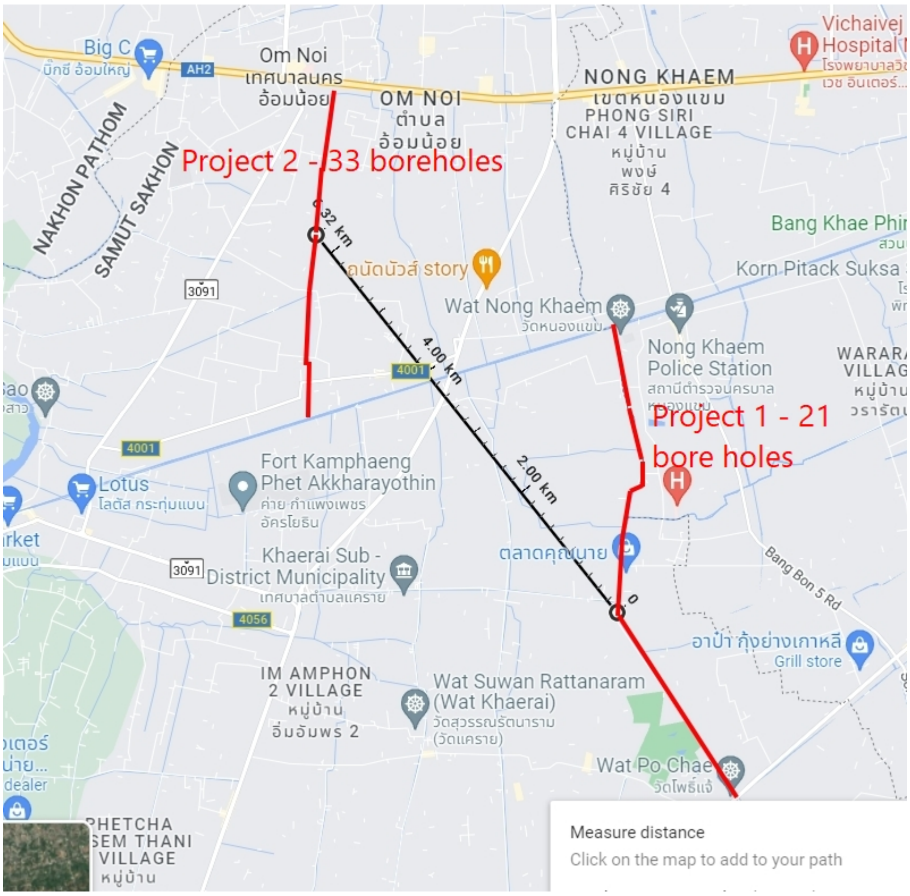

Figure 1: *Sample Site Location*

The laboratory tests were carried out in accordance with ASTM standards. Bulk unit weight can be calculated from the ratio between the weight of the soil sample and the volume of the soil being measured. The natural water content was calculated based on ASTM D2216, which is the ratio between the weight of soil moisture and the weight of dry soil. The Atterberg limits test according to ASTM D4318 is used to determine the soil index properties that depend on the soil composition, such as mineral compounds in the soil, soil consistency, etc. Grain size distribution was obtained based on ASTM D422. The soil sample was oven-dried at approximately 105 °C for 24 h, after which the dried soil was sieved through a sieve number of 200 (for particles smaller than 75 microns). The hydrometer test according to ASTM D2487 was conducted for soil classification. The limits, such as LL (Liquid limit), PL (Plastic limit), and PI (Plastic Index), together with the grain size distribution, were used to determine the classification of fine-grained soils (i.e., ASTM D2487). The Unconfined Compression Test, referred to as ASTM D21466, was conducted to determine the undrained shear strength of the soil.

2.2 Machine Learning Framework

2.2.1 Dataset Preparation and Preprocessing

Effective preprocessing is foundational in machine learning, as it significantly influences model precision and generalization capabilities [19, 20]. Key operations include data cleaning, feature rescaling, encoding, dimensionality handling, and addressing class distribution issues.

In this study, features were standardized using the *Standard Scaler* method, which normalizes input data by subtracting the mean (μ) and dividing by the standard deviation (σ). The transformation is represented as:

$$z = \frac{x - \mu}{\sigma} \qquad \text{(Eq. 1)}$$

To enhance generalization and limit overfitting, normalization was combined with a robust evaluation strategy. An initial 80/20 training-to-testing split was applied. Additionally, five-fold cross-validation was implemented within the training partition to validate model stability and reduce variance [21].

2.2.2 Regression Models Employed

A variety of regression techniques, ranging from linear to complex ensemble methods, were examined to estimate continuous outcomes [22, 23]. The approach included stages such as preprocessing, model training, hyperparameter optimization, and validation. Model selection was based on complexity, interpretability, and computational efficiency.

Table 1 provides a concise summary of the adopted models, their governing equations or loss functions, optimization procedures, and primary considerations.

Table 1: *Overview of Regression Techniques and Optimization Approaches*

Model	Equation / Loss Function	Optimization Strategy	Model Constraints
Linear Regression	$Y = \beta_0 + \sum_{i=1}^{n} \beta_i X_i + \epsilon$ (Eq. 2) OLS objective: $\min_{\beta} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$ (Eq. 3)	Closed-form: $\beta = (X^T X)^{-1} X^T y$ (Eq. 4)	• $X^T X$ must be invertible • Linearity and homoscedasticity
Decision Tree	Predicts by partitioning into leaf nodes	Recursive splitting minimizing node MSE	• Control depth and node size • Use pruning to reduce overfitting
Random Forest	$\hat{y} = \frac{1}{T} \sum_{t=1}^{T} h_t(x)$ (Eq. 5)	Bagging of trees on bootstrapped samples	• Number of estimators • Feature randomness
SVR	$\min_{w,b} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} (\xi_i + \xi_i^*)$ (Eq. 6)	Quadratic programming	• Kernel function (RBF, poly, linear) • Tuning parameter C
XGBoost	$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i)$ (Eq. 7)	Gradient tree boosting	• Learning rate • Regularization terms
AdaBoost	$\hat{y} = \sum_{m=1}^{M} \alpha_m h_m(x)$ (Eq. 8)	Sequential weighting based on errors	• Estimator count • Learning coefficient
Gradient Boosting	$r_i^{(m)} = y_i - F_{m-1}(x_i)$ (Eq. 9)	Function space gradient descent	• Max tree depth • Estimator count
KNN	$\hat{y} = \frac{1}{k} \sum_{i \in N_k(x)} y_i$ (Eq. 10)	Instance-based; no global model	• Neighbor count k • Distance metric (Euclidean)

Model	Equation / Loss Function	Optimization Strategy	Model Constraints
Gaussian Process	$\mu_* = K_*^T K^{-1} y$ (Eq. 11)	Maximize marginal likelihood	• Kernel (e.g., Matérn) • Hyperparameter tuning
Neural Networks	$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$ (Eq. 12) , $MAE = \frac{1}{n}\sum_{i=1}^{n}\lvert y_i - \hat{y}_i\rvert$ (Eq. 13)	Adam optimizer with backpropagation	• Layers: 3; Neurons: 32/layer • ReLU activation • Dropout regularization • Early stopping applied

Explanatory symbols: β_0 and β_i are regression intercept and coefficients, respectively. ϵ denotes residuals, $X^T X$ is the Gram matrix, y_i and $\hat{y}_i$ denote actual and predicted outcomes. In ensemble models, $h_t(x)$ or $f_t(x_i)$ represent weak learners. SVR includes slack variables ξ_i and a margin-control hyperparameter C. In KNN, $N_k(x)$ is the set of k nearest neighbors, and in Gaussian Processes, K and K_* denote training and cross-kernel matrices.

The study utilized Linear Regression as a foundational benchmark, with Decision Trees and Random Forests selected for their interpretability and non-linear capability. SVR with varying kernels gauged the impact of transformation functions. KNN offered a non-parametric contrast, while ANNs explored high-complexity dependencies. Ensemble approaches like XGBoost, AdaBoost, and GBoost were evaluated for their cumulative correction strategies. Gaussian Process Regression provided probabilistic predictions and uncertainty estimates, supporting robust model comparisons for the soil-biochar interaction task.

2.2.3 Performance Assessment Criteria

A range of regression performance indicators were applied, including MAE, MSE, RMSE, MAPE, EVS, and both standard and adjusted R^2 scores, ensuring a multidimensional evaluation of model accuracy [24, 25]. Table 2 presents the metric formulations.

Table 2: *Metrics Used to Evaluate Model Performance*

Metric	Formula
Mean Absolute Error (MAE)	$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\lvert y_i - \hat{y}_i\rvert$ (Eq. 14)

Metric	**Formula**	
Root Mean Squared Error (RMSE)	$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$	(Eq. 15)
R-squared (R^2)	$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$	(Eq. 16)
Adjusted R^2	$\bar{R}^2 = 1 - \left(\frac{(1-R^2)(n-1)}{n-p-1}\right)$	(Eq. 17)

Here, $\bar{y}$ is the mean target value, n is the sample size, and p denotes the number of independent variables. These evaluation metrics collectively enabled a nuanced interpretation of model precision, variance capture, and generalizability.

3 Results and Discussion

4 Conclusion

Acknowledgments

The first author (V. Kamchoom) would like to thank the grant under Climate Change and Climate Variability Research in Monsoon Asia (CMONA) from National Research Council of Thailand (NRCT) and the National Natural Science Foundation of China (NSFC). The second author (A. Garg) thanks the grant XXX provided by XXX.

References

[1] C. P. Wroth. The interpretation of in situ soil tests. *Géotechnique*, 34(4):449–489, 1984.

[2] Soumyaranjan Mishra and Retnamony G. Robinson. Laboratory investigation on quasi-static penetration testing using spt sampler in soft clay bed. *Geotechnical Testing Journal*, 42(4):985–1005, 09 2018.

[3] Mintae Kim, Osman Okuyucu, Ertuğrul Ordu, Seyma Ordu, Özkan Arslan, and Junyoung Ko. Prediction of undrained shear strength by the gmdh-type neural network using spt-value and soil physical properties. *Materials*, 15(18), 2022.

[4] Alejandro Sepúlveda, Fausto A. Canales, Carlos Acosta, and Jose Duque and. Correlations for undrained shear strength of fine-grained soils with spt-values: literature review and uncertainty analysis. *International Journal of Geotechnical Engineering*, 19(5):315–327, 2025.

[5] Ingrid Belcavello and Lucas Deleon. Influence of hydrogeological conditions on estimation of undrained shear strength by cptu tests: case study in a tropical soil. In *Proceedings of the 7th International Conference on Geotechnical and Geophysical Site Characterization*, volume 18, page 21, 2024.

[6] Hengchhorn Phai and Amin Eisazadeh. Compaction properties of rice husk ash-lime-bangkok clay mixtures. In *Advanced Materials Research IX*, volume 803 of *Key Engineering Materials*, pages 331–337. Trans Tech Publications Ltd, 6 2019.

[7] Pithan Pairojn. The increasing of undrained shear strength and shear modulus of soft bangkok clay by silica powder using unconfined compression test with bender element. *GEOMATE Journal*, 18(69):118–123, May 2020.

[8] Mohamed B. D. Elsawy, Mohammed F. Alsharekh, and Mahmoud Shaban. Modeling undrained shear strength of sensitive alluvial soft clay using machine learning approach. *Applied Sciences*, 12(19), 2022.

[9] Selçuk Demir and Emrehan Kutlug Sahin. The effectiveness of data pre-processing methods on the performance of machine learning techniques using rf, svr, cubist and sgb: a study on undrained shear strength prediction. *Stochastic Environmental Research and Risk Assessment*, 38(8):3273–3290, jun 2024.

[10] Chenghang Zhang and Mingyue Chen. Prediction of undrained shear strength utilizing a hybrid stacking model enhanced by bayesian optimization algorithm. *Transportation Research Record*, 0(0):03611981241278354, 0.

[11] Binh Thai Pham, Chongchong Qi, Lanh Si Ho, Trung Nguyen-Thoi, Nadhir Al-Ansari, Manh Duc Nguyen, Huu Duy Nguyen, Hai-Bang Ly, Hiep Van Le, and Indra Prakash. A novel hybrid soft computing model using random forest and particle swarm optimization for estimation of undrained shear strength of soil. *Sustainability*, 12(6), 2020.

[12] Huajian Yang, Zhikui Liu, Yuantao Li, Haixia Wei, and Nengsheng Huang. Catboost–bayesian hybrid model adaptively coupled with modified theoretical equations for estimating the undrained shear strength of clay. *Applied Sciences*, 13(9), 2023.

[13] Krit Saowiang and Pham Huy Giao. Sea-level related engineering geology and intrinsic compression behaviour of bangkok clays. *GEOMATE Journal*, 17(59):144–153, Jul. 2019.

[14] Masami Ohtsubo, Kazuhiko Egashira, Tatsuya Koumoto, and Dennes T. Bergado. Mineralogy and chemistry, and their correlation with the geotechnical index properties of bangkok clay: Comparison with ariake clay. *Soils and Foundations*, 40(1):11–21, 2000.

[15] Sideth Prum, Nalinee Jumnongpol, Chutima Eamchotchawalit, Pisit Kantiwattanakul, Vannee Sooksatra, Thanatip Thanatip Jarearnsiri, and Somsak Passananon. Guideline for backfill material improvement for water supply pipeline construction on bangkok clay, thailand. *Proceedings of the World Congress on Civil, Structural, and Environmental Engineering*, April 2019.

[16] Binh Thai Pham, Le Hoang Son, Tuan-Anh Hoang, Duc-Manh Nguyen, and Dieu Tien Bui. Prediction of shear strength of soft soil using machine learning methods. *CATENA*, 166:181–191, 2018.

[17] Quoc Anh Tran, Lanh Si Ho, Hiep Van Le, Indra Prakash, and Binh Thai Pham. Estimation of the undrained shear strength of sensitive clays using optimized inference intelligence system. *Neural Computing and Applications*, 34(10):7835–7849, January 2022.

[18] Wengang Zhang, Chongzhi Wu, Haiyi Zhong, Yongqin Li, and Lin Wang. Prediction of undrained shear strength using extreme gradient boosting and random forest based on bayesian optimization. *Geoscience Frontiers*, 12(1):469–477, 2021.

[19] Sotiris B Kotsiantis, Dimitris Kanellopoulos, and Panagiotis E Pintelas. Data preprocessing for supervised leaning. *International journal of computer science*, 1(2):111–117, 2006.

[20] Salvador García, Sergio Ramírez-Gallego, Julián Luengo, José Manuel Benítez, and Francisco Herrera. Big data preprocessing: methods and prospects. *Big Data Analytics*, 1(1), September 2016.

[21] G. Kesavaraj and S. Sukumaran. A study on classification techniques in data mining. In *2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT)*, pages 1–7, 2013.

[22] Van Trieu Vy Nguyen, Song Thanh Quynh Le, and Duc Duy Nguyen. Artificial intelligence models for overall equipment effectiveness prediction: A case study in an assembly manufacturing company. In *2024 7th International Conference on Green Technology and Sustainable Development (GTSD)*, pages 67–71, 2024.

[23] Sowmya Christina, S Sowjanya, Ch. Lakshmhyma, L Prathiba, and Md Shaik Amzad Basha. Data-driven insights into student performance: Benchmarking machine learning models for grade prediction using regression and classification approaches. In *2025 International Conference on Intelligent Systems and Computational Networks (ICISCN)*, pages 1–6, 2025.

[24] Nidhi Haribhau Gaikwad, Arghya Mitra, Jangilwar Payal, Soumyabrata Das, and Ritesh Kumar Keshri. Impact of data preprocessing on wind forecasting using machine learning techniques. In *2025 Fourth International Conference on Power, Control and Computing Technologies (ICPC2T)*, pages 1–6, 2025.

[25] Ankur Kumar, Asim Ali Khan, and Jaspreet Singh. Enhancing the diagnosis of cardiovascular disease: A comparative examination of support vector machine and artificial neural network models utilizing extensive data preprocessing techniques. *WSEAS TRANSACTIONS ON COMPUTERS*, 23:318–327, December 2024.