International Journal of
Science, Architecture, Technology and Environment

Hybrid Physics–Machine Learning Model for CPT-Based Settlement Prediction of Shallow Foundations

Clifford S. Njah[1*], T. K. S. Abam[2], S. Cliff-Teme[3], G. W. T. Jaja[4]

[1]*Ph.D. Candidate, Dept. of Civil Engineering, Rivers State Univ., Nkpolu-Oroworukwo, Port Harcourt, Nigeria (corresponding author). Email: cliffnjah1988@gmail.com*

[2]*Professor, Institute of Geosciences and Environmental Management, Rivers State Univ., Nkpolu-Oroworukwo, Port Harcourt, Nigeria.*

[3]*Professor, Dept. of Geology, Rivers State Univ., Nkpolu-Oroworukwo, Port Harcourt, Nigeria.*

[4]*Dept. of Civil Engineering, Rivers State Univ., Nkpolu-Oroworukwo, Port Harcourt, Nigeria.*

[*]*Corresponding author*

DOI: https://doi.org/10.63680/ijsate102677.04

Abstract

Classical settlement methods for shallow foundations were calibrated largely on temperate sedimentary soils and may perform poorly in lateritic and other tropical residual soils, whose behavior is governed by cementation and uneven weathering. Machine learning (ML) can capture such nonlinear behavior, but most ML models are opaque, which limits their adoption in design practice. This study developed the Hybrid Physics-Informed Settlement Model (HPISM), which predicts shallow-foundation settlement from eight cone penetration test (CPT)-derived parameters: embedment depth, corrected cone resistance, soil behavior type index, unit weight, overconsolidation ratio, foundation width, foundation length, and applied pressure. A database of 8,934 CPT records was compiled from four public repositories. Because these repositories contain no measured footing settlements, a reference settlement was computed for each record with the Schmertmann strain-influence method, and the data were split 70:30 into training and testing sets. Twelve supervised algorithms were trained and compared, and the four best were tuned. The tuned Extra Trees ensemble reproduced the reference settlements most accurately (test coefficient of determination 0.995, root-mean-square error 6.13 mm). SHapley Additive exPlanations (SHAP) ranked foundation width and applied pressure as the most influential predictors. Symbolic regression distilled the tuned ensemble into an explicit equation (test coefficient of determination 0.913), and a gradient boosting residual correction raised this to 0.990 (root-mean-square error 8.65 mm); a fully closed polynomial form suitable for hand or spreadsheet calculation retained a value of 0.973. Three worked examples compare the proposed models with the Terzaghi, Janbu, Meyerhof, and Schmertmann methods. The HPISM offers an inspectable, spreadsheet-ready alternative to a black-box model, although its use for lateritic ground requires verification against measured footing settlements.

Keywords: Shallow foundation settlement; Lateritic soils; Cone penetration test; Machine learning; Extra Trees; Explainable artificial intelligence; SHAP; Symbolic regression; Hybrid physics–machine learning model

Practical Applications

Engineers designing shallow foundations commonly estimate settlement from cone penetration test (CPT) results using hand methods developed decades ago for temperate soils. This study offers two complementary tools that use the same eight CPT-derived inputs routinely available from a site investigation: a tuned Extra Trees model, which gives the most accurate estimate when software is available, and the Hybrid Physics-Informed Settlement Model (HPISM), whose closed-form version can be evaluated in a spreadsheet and checked line by line by a reviewer. In preliminary design, the tools allow designers to test quickly how changing footing width, embedment depth, or bearing pressure affects predicted settlement, and to identify when a pad footing is unlikely to satisfy a 25 mm serviceability limit, so that a raft, ground improvement, or a deep foundation should be considered. The models were trained on settlements computed with the Schmertmann method and should be used only within the parameter ranges of the database. Before routine use in lateritic or other tropical residual soils, their predictions should be checked against settlements measured on local structures or load tests.

Introduction

A foundation must transfer a structure's loads safely into the ground and ensure stability and serviceability throughout the structure's design life. Among serviceability concerns, settlement is critical: excessive settlement can crack walls, distort finishes, damage services, and, in severe cases, lead to failure. Shallow foundations—isolated footings, strip footings, and rafts—are common because of their simplicity and cost-effectiveness. Their in-service performance depends on settlement under load, which is influenced by soil stiffness, stress history, foundation geometry, drainage, cementation, and soil uniformity (Bowles 1996; Terzaghi et al. 1996; Das and Sivakugan 2019).

The Cone Penetration Test (CPT) is a widely used site investigation method. It is fast, repeatable, and relatively inexpensive, and it provides a continuous ground profile from which corrected cone resistance, soil behavior type, friction ratio, unit weight, and stiffness can be derived (Lunne et al. 1997; Robertson and Cabal 2022). The CPT-based settlement methods developed by Schmertmann (1970, 1978), Meyerhof (1974), and Janbu (1967) significantly advanced practice. Their calibration, however, largely reflects temperate transported sedimentary deposits. The methods also depend on simplified stress-distribution or elastic-response assumptions, or on empirical correction factors (Table 1). These conditions may not represent the behavior of lateritic soils adequately. They are prevalent across tropical Africa, South America, and Southeast Asia, and their behavior is controlled primarily by cementation from iron and aluminum oxides, uneven weathering profiles, and mixed mineralogy. It is well established that applying methods developed for temperate soils directly to lateritic ground results in prediction errors (Das and Sobhan 2018; Okewale and Grobler 2022).

Table 1: Region of derivation and key limitations of selected classical settlement models

Settlement model	Region / basis	Key limitation
Terzaghi one-dimensional consolidation (1943)	Temperate — validated on Boston and Chicago clays	Assumes purely 1-D vertical strain; ignores 3-D pore-pressure effects and secondary (creep) compression
Skempton–Bjerrum correction (1957)	Temperate — London and Oslo	Correction factor depends on the pore-pressure parameter A, which is difficult to measure reliably
Schmertmann strain-influence method (1970, 1978)	Temperate — Florida sands	Strain-influence factors calibrated on sands only; sensitive to the assumed E_s–q_c relationship
Elastic theory (Boussinesq 1885; Poulos and Davis 1974)	Temperate — theoretical, generalized	Assumes a homogeneous, isotropic, linear-elastic half-space; ignores time-dependent consolidation
Janbu tangent modulus method (1967)	Temperate — Trondheim, Norway	Modulus number and stress exponent are empirical and require local calibration
Meyerhof CPT/SPT correction (1965, 1974)	Temperate — Nova Scotia, Canada	Applies a flat correction factor rather than a soil- or site-specific adjustment
Burland and Burbidge (1985)	Temperate — UK case histories	Statistically fitted mainly to UK/European data; less reliable for cemented or fines-rich soils
De Beer and Martens (1957)	Temperate — Belgium	Early, simplified constrained-modulus approach with no soil-behavior-type differentiation

Machine-learning (ML) methods can learn nonlinear relationships from observed data without requiring a constitutive equation to be specified in advance. Algorithms including Random Forest, XGBoost, CatBoost and Gradient Boosting have been applied successfully in geotechnical studies (Breiman 2001; Chen and Guestrin 2016; Iyeke et al. 2016; Dorogush et al. 2018). Their predictions can nevertheless be difficult to interpret in terms of familiar soil-mechanics principles, which may limit confidence in their use for engineering decisions (Rudin 2019). Explainable artificial intelligence provides tools for examining model behavior (Adadi and Berrada 2018; Arrieta et al. 2020; Molnar 2022). In particular, SHapley Additive exPlanations (SHAP) estimates the contribution of individual predictors to a model output (Lundberg and Lee 2017). Symbolic regression complements this analysis by searching data for explicit mathematical relationships without imposing a fixed equation beforehand (Koza 1992; Schmidt and Lipson 2009; La Cava et al. 2021; Cranmer 2023). Together, these techniques offer a route from a high-performing but difficult-to-interpret model to a settlement expression that engineers can inspect and calculate directly.

Four issues remain important in settlement prediction for lateritic soils. First, conventional methods may depend on assumptions that do not adequately represent tropical residual soils. Second, many ML applications emphasize predictive accuracy without providing sufficient engineering interpretation. Third, the combined use of ML, explainable AI and symbolic regression under physical constraints has received limited attention (Karniadakis et al. 2021; Willard et al. 2022). Finally, direct comparisons with established

settlement equations using the same dataset are uncommon. The present study is designed to address these issues.

Aim and Objectives

This study aims to develop a hybrid physics–machine learning model for predicting the settlement of shallow foundations on lateritic soils using geotechnical parameters derived from CPT data. The objectives are to: (i) construct an integrated CPT-derived database for settlement prediction in lateritic soils; (ii) analyze the statistical relationships between CPT-derived parameters and settlement; (iii) develop and compare multiple supervised machine learning models; (iv) optimize the hyperparameters of the most effective models; (v) interpret the optimized model using SHAP; (vi) derive an explicit settlement equation through symbolic regression; (vii) formulate the Hybrid Physics-Informed Settlement Model (HPISM) by integrating the symbolic equation with residual learning; (viii) validate the model on a withheld test set; and (ix) compare its performance with established conventional settlement prediction methods.

This investigation considers settlement of shallow foundations in lateritic soils and comparable tropical residual soils. Deep and piled foundations, bearing-capacity assessment and liquefaction are not analyzed, although they may be mentioned where needed to explain the underlying theory.

Materials and Methods

Data Sources and Integrated Database

Data for the predictors were assembled from four publicly available CPT sources: Kaggle, Mendeley Data, the Adapazarı CPT database, and the Premstaller Geotechnik database. They were chosen for the availability of CPT variables relevant to the study. The records were screened for quality, duplicates, and internal consistency, producing 8,934 complete observations. The soil behavior type index followed Robertson (2010), and unit weight and OCR were estimated from the CPT using the correlations of Mayne (2014). Because the repositories do not contain measured footing settlements, the target settlement S for each record was computed with the Schmertmann strain-influence method (Schmertmann 1970, 1978), using a creep factor C2 = 1.4 and a CPT-based constrained modulus (Robertson 2009). The ML models therefore learn to reproduce, and generalize from, this CPT-based reference settlement. Before model fitting, the dataset was divided at random into training and testing subsets in a 70:30 ratio: 6,253 observations for training and 2,681 for testing (Table 2). The test subset was withheld from training and hyperparameter tuning to avoid data leakage.

Table 2: Summary of the integrated CPT-derived dataset

Characteristic	Value
Total observations	8,934
Predictor variables	8
Target variable	Settlement, *S*
Training observations (70%)	6,253
Testing observations (30%)	2,681
Missing values after preprocessing	0
Duplicate observations	0

Predictor Variables

Eight CPT-derived predictors were selected for their physical relevance: embedment depth (z, m), corrected cone resistance (qt, MPa), soil behavior type index (Ic), unit weight (γ, kN/m^3), overconsolidation ratio (OCR), foundation width (B, m), foundation length (L, m), and applied foundation pressure (q, MPa). The reference settlement (S, mm), computed as described above, is the response variable. Table 3 presents the descriptive statistics of the dataset.

Table 3: Descriptive statistics of the predictor variables and the target variable S (n = 8,934)

Variable	Mean	Median	Std. dev.	Min	Max
z (m)	2.266	2.340	0.511	0.240	3.000
q_t (MPa)	0.553	0.500	0.243	0.110	1.470
I_c (–)	3.039	3.000	0.265	2.228	3.882
γ (kN/m^3)	16.461	16.498	0.832	14.007	19.746
OCR (–)	8.215	7.908	3.666	1.084	15.000
B (m)	2.305	2.080	0.828	1.010	3.980
L (m)	1.818	1.800	0.427	1.200	2.500
q (MPa)	0.136	0.116	0.069	0.050	0.299
S (mm)	117.04	101.74	85.74	0.17	436.11

Machine Learning Model Development

Twelve supervised algorithms from the linear, regularized linear, tree-based, gradient-boosting, kernel, and neural-network families (Table 4) were implemented in scikit-learn (Pedregosa et al. 2011), following standard practice (James et al. 2021; Géron 2022), and trained using 10-fold cross-validation and then compared on the held-out test set. The four best-performing ensemble and boosting algorithms (Extra Trees, Random Forest, Gradient Boosting, and XGBoost) were subsequently tuned with RandomizedSearchCV and 5-fold cross-validation, evaluating 30 hyperparameter combinations for each model.

Table 4: Classification of the twelve supervised machine learning algorithms investigated

Category	Algorithms
Linear regression models	Linear Regression
Regularized linear models	Ridge, Lasso, Elastic Net Regression
Tree-based and bagging ensembles	Decision Tree, Random Forest, Extra Trees
Gradient boosting algorithms	Gradient Boosting, HistGradientBoosting, XGBoost, CatBoost
Kernel and neural-network models	Support Vector Regression (SVR), Multi-Layer Perceptron (MLP)

Explainable Artificial Intelligence: SHAP

The tuned model was interpreted using SHapley Additive exPlanations (SHAP), a feature-attribution method based on cooperative game theory (Shapley 1953). SHAP was used to estimate how much each predictor contributed to settlement predictions across the test set and for individual cases (Lundberg and Lee 2017). These results informed the subsequent analysis by identifying less influential variables, highlighting the strongest predictors and narrowing the search space for symbolic regression.

Symbolic Regression

Symbolic regression was carried out with PySR/gplearn using an evolutionary search based on genetic programming. The search combined arithmetic, logarithmic and inverse operations to identify a closed-form equation that approximates the tuned ML model (La Cava et al. 2021; Cranmer 2023). A complexity penalty was included to discourage unnecessarily complicated expressions and reduce overfitting. Rather than searching across every input variable, the procedure focused on predictors identified as important by SHAP; the selected expression was then chosen from the accuracy–complexity Pareto front.

Development of the Hybrid Physics-Informed Settlement Model

Settlement was initially represented as a general nonlinear function of the eight predictors:

$$S = f(z, \gamma, B, L, q, q_t, I_c, \mathrm{OCR}) \quad (1)$$

where the form of f(·) is not predefined but is determined by machine learning and symbolic regression. The search was guided by three monotonicity expectations from classical soil mechanics: settlement increases with applied pressure and decreases with corrected cone resistance and with overconsolidation ratio:

$$S \propto q, \quad S \propto \frac{1}{q_t}, \quad S \propto \frac{1}{\mathrm{OCR}} \quad (2)$$

The tuned machine learning model served as the reference prediction surface, $\hat{S}$ = FML(X). PySR then approximated this surface, rather than the reference settlements directly, using a SHAP-reduced set of predictors and minimizing a loss-plus-complexity objective. The resulting symbolic base equation, FSR(X), was combined with a Gradient Boosting (Friedman 2001) residual correction, R(X), trained on the systematic error of the base equation, ε(X) = Sref(X) − FSR(X). This yields the general hybrid form:

$$S = F_{\mathrm{SR}}(X) + R(X) \quad (3)$$

The numerical coefficients of FSR(X) were calibrated on the training data by minimizing the residual sum of squares:

$$\min J(\theta) = \sum_{i=1}^{n} \left(S_i - \hat{S}_i\right)^2, \quad \theta = [a_0, a_1, \ldots, a_n] \quad (4)$$

Where the equation must be solvable by hand or in a spreadsheet, the trained Gradient Boosting correction R(X) was converted into an explicit second-degree polynomial surrogate, g(X), in terms of the standardized predictors si = (xi − μi)/σi, where μi and σ i are the training-set mean and standard deviation of predictor xi:

$$g(X) = \beta_0 + \sum_i \beta_i s_i + \sum_i \beta_{ii} {s_i}^2 + \sum_{i<j} \beta_{ij} s_i s_j \quad (5)$$

This gives the final closed-form HPISM, with a constraint that ensures non-negative output:

$$S = \max[S_{\mathrm{base}}(X) + g(X), 0] \quad (6)$$

The complete 44-term coefficient tables for g(X), covering both the power-law and the symbolic base versions, are available from the corresponding author upon reasonable request (see Data Availability Statement).

Model Validation

Final model performance was assessed using the withheld 30% test subset (n = 2,681), which was not used during training or hyperparameter optimization. Evaluation measures included the coefficient of

determination (R^2), root-mean-square error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), and the a20 index—the proportion of predictions falling within ±20% of the reference settlement. Residual patterns and predicted-versus-measured values were also examined, and the outputs were checked for consistency with established settlement methods.

Results and Discussion

Comparative Performance of the Machine Learning Models

Performance differed markedly across the twelve algorithms. The linear models—Linear, Ridge and Lasso Regression—returned test R^2 values between 0.67 and 0.72, whereas the four leading ensemble methods achieved R^2 values of at least 0.99 after tuning (Table 5). This contrast indicates that the settlement response in the dataset is not adequately represented by a simple linear relationship with the CPT-derived inputs. Extra Trees produced the strongest results among the models tested. It is therefore referred to here as the tuned Extra Trees model (XT model) and is used as the reference model for the subsequent SHAP interpretation and symbolic-regression stages.

Table 5: Comparative test-set performance of the tuned ensemble models

Model	Test R^2	RMSE (mm)	MAE (mm)	a20 (%)
Tuned Extra Trees (XT model)	0.9950	6.13	2.13	97.84
XGBoost	0.9947	6.29	3.23	96.31
Random Forest	0.9932	7.17	3.48	95.79
Gradient Boosting	0.9926	7.48	4.41	94.03

SHAP-Based Interpretation

SHAP analysis of the tuned Extra Trees model gave the importance ranking B > q > qt > z > OCR > Ic > γ > L (Table 6; Fig. 1). Foundation width and applied foundation pressure are the primary determinants of the prediction. This agrees with classical foundation mechanics, in which wider footings and higher bearing pressures stress a larger soil zone beneath the foundation. Corrected cone resistance and embedment depth follow, reflecting soil stiffness and the confinement provided by overburden. The stress-history and classification variables (OCR, Ic, γ) and footing length contribute relatively little.

Table 6: Mean |SHAP| feature importance, tuned Extra Trees model (n = 2,681 test observations)

Rank	Feature	Mean |SHAP| (mm)
1	B — foundation width	69.98
2	q — applied foundation pressure	28.63
3	q_t — corrected cone resistance	16.09
4	z — embedment depth	10.69
5	OCR — overconsolidation ratio	7.43
6	I_c — soil behavior type index	4.99
7	γ — unit weight	1.64
8	L — footing length	1.48

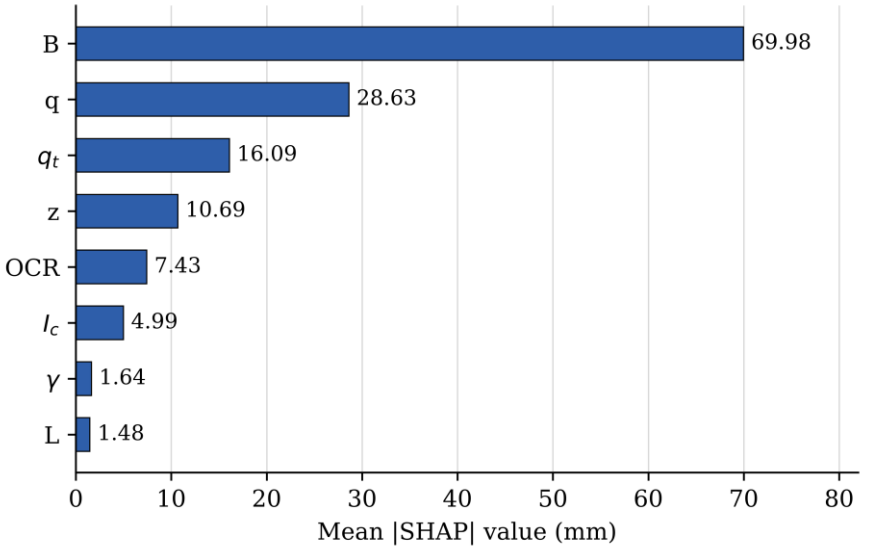


Fig. 1. Mean |SHAP| feature importance of the tuned Extra Trees model (n = 2,681 test observations).

Symbolic Regression Equation and the HPISM

The first candidate, Hybrid A, used a calibrated power-law equation fitted directly to the reference settlement [Eq. (7)]. Its test R^2 of 0.4427 showed that this fixed functional form did not capture the response well. Symbolic regression was then applied to the predictors prioritized by SHAP, producing the nested expression in Eq. (8); its component terms, T1–T8, are defined below the equation. The resulting expression followed the tuned model more closely, with test R^2 = 0.9133, RMSE = 25.55 mm, and MAE = 19.81 mm. Because Eqs. (7) and (8) were obtained by unconstrained regression, they are not dimensionally homogeneous; they are empirical expressions valid only with the units used in Table 3 (z, B, and L in m; qt and q in MPa; γ in kN/m^3).

$$S = 2.999 \times 10^{-5} \cdot z^{-0.8434} \cdot q_t^{-1.6743} \cdot I_c^{-6.0013} \cdot \gamma^{7.8054} \cdot \mathrm{OCR}^{-0.5543} \cdot B^{2.7995} \cdot L^{-0.0039} \cdot q^{0.9235} \quad (7)$$

$$S = B\gamma + q(B - q_t)\left[I_c T_1 + L\gamma + 2\gamma + q(B-L)T_2 + z + T_3 + \frac{q(B-L)T_4}{\mathrm{OCR}} + \frac{1}{\mathrm{OCR}}\right] \quad (8)$$

where the nested sub-terms are:

$$T_1 = \frac{B\gamma}{q_t} + B + L\gamma + (B - z)T_5$$

$$T_2 = I_c\left(\frac{I_c\gamma}{z} + \ln|B|\right) - L + \mathrm{OCR}$$

$$T_3 = (B - z)T_7$$

$$T_4 = \frac{BI_c\gamma}{{q_t}^2} - L + \mathrm{OCR}$$

$$T_5 = \gamma + z - (B - z)T_6$$

$$T_6 = \gamma + z - \frac{(\gamma + z)}{(B - z)\ln|B \cdot \mathrm{OCR}|}$$

$$T_7 = \gamma + z - (B - z)T_8$$

$$T_8 = \gamma + z + \frac{(\gamma + z)}{\left(B \cdot \mathrm{OCR} - \frac{q_t}{L}\right)\ln|L|}$$

Genetic-programming symbolic regression often yields deeply nested expressions rather than simple sums of terms; this complexity is the price of its higher accuracy relative to the compact power-law form. Applying

a Gradient Boosting residual correction to this symbolic base equation gives the HPISM, designated Hybrid C [Eq. (3)]. In its semi-closed form, where the residual term is supplied by the trained Gradient Boosting model, it achieves a test R^2 of 0.9901, an RMSE of 8.65 mm, and a MAPE of 13.35%. In its fully closed polynomial form [Eq. (6)], which permits manual calculation, it achieves an R^2 of 0.9725 and an RMSE of 14.38 mm. The corresponding hybrid built on the power-law base (Hybrid B) gave R^2 = 0.9853 in semi-closed form and 0.8930 in fully closed form (Tables 7 and 8). Residual correction therefore recovered most of the accuracy lost in each simplification: RMSE decreased by 66% for Hybrid C (from 25.55 to 8.65 mm) and by 84% for Hybrid B (from 64.77 to 10.51 mm).

Table 7: Comparative validation performance of the investigated settlement prediction models

Model	Test R^2	RMSE (mm)	MAE (mm)
Tuned XT model	0.9950	6.13	2.13
Hybrid C — HPISM (symbolic + GBR residual)	0.9901	8.65	5.29
Hybrid B (power-law + GBR residual)	0.9853	10.51	5.72
Symbolic regression equation alone	0.9133	25.55	19.81
Hybrid A (power-law base only)	0.4427	64.77	31.46

Table 8: Fidelity and accuracy cost of the closed-form polynomial surrogates

Correction	Surrogate fidelity to GBR (train R^2)	Fully closed test R^2	Fully closed RMSE (mm)
Hybrid B	0.8444	0.8930	28.38
Hybrid C — HPISM	0.7766	0.9725	14.38

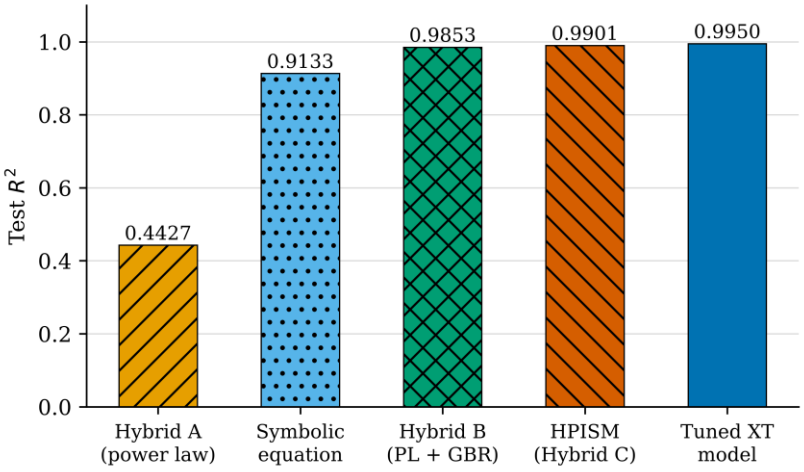


Fig. 2. Test-set R^2 of the investigated settlement prediction models.

The two proposed models serve different practical needs (Fig. 2). The tuned Extra Trees model gives the best predictive performance, exceeding the semi-closed HPISM by ΔR^2 = 0.0049 and reducing RMSE by 2.52 mm. Its main limitation is interpretability: it does not provide a single explicit equation that can be reviewed

directly. The HPISM gives up a measurable amount of accuracy in exchange for a more transparent and auditable formulation; in its simplified form, it can be evaluated without specialist ML software.

Comparison with Classical Settlement Prediction Methods

Table 9 contrasts the classical approaches with the symbolic equation, tuned XT model and HPISM in terms of accuracy, interpretability and practical use. In this comparison, the conventional methods are less accurate than the ML models, whereas the Extra Trees ensemble does not offer the transparency of an explicit equation. The HPISM provides a compromise between these two requirements, combining strong predictive performance with a formulation that can be inspected.

Table 9: Qualitative comparison of settlement prediction models (authors' assessment based on Tables 7 and 8)

Model	Predictive accuracy	Interpretability	Engineering applicability
Schmertmann / Terzaghi / Janbu / Meyerhof	Moderate	High	High
Symbolic regression equation	Good	Very high	Very high
Tuned XT model	Excellent	Low	Moderate
HPISM	Excellent	High	High

Worked Application Examples

Three worked examples illustrate the application of the five prediction routes: a typical site at the database mean, a favorable scenario on firm ground with a light load, and an unfavorable scenario on soft ground with a heavy load (Tables 10 and 11; Fig. 3).

Table 10: Input parameters for the three worked application examples

Variable (unit)	Ex. 1 Typical	Ex. 2 Firm / light	Ex. 3 Soft / heavy
Embedment depth, z (m)	2.26	2.30	1.80
Corrected cone resistance, q_t (MPa)	0.554	0.68	0.45
Soil behavior type index, I_c (–)	3.04	2.75	3.35
Unit weight, γ (kN/m^3)	16.5	17.2	15.7
Overconsolidation ratio, OCR (–)	8.26	10.0	6.0
Foundation width, B (m)	2.30	1.80	2.70
Foundation length, L (m)	1.82	1.60	2.10
Applied foundation pressure, q (MPa)	0.135	0.100	0.180

Table 11: Predicted settlement (mm) for the three worked examples, five modeling routes

Case	Hybrid A	Symbolic eqn	Hybrid B	HPISM	Tuned XT model
Ex. 1: typical site	80.7	127.7	123.9	135.3	135.1
Ex. 2: firm, light load	49.6	54.2	38.6	43.5	38.4
Ex. 3: soft, heavy load	129.7	265.2	82.6	308.7	263.4

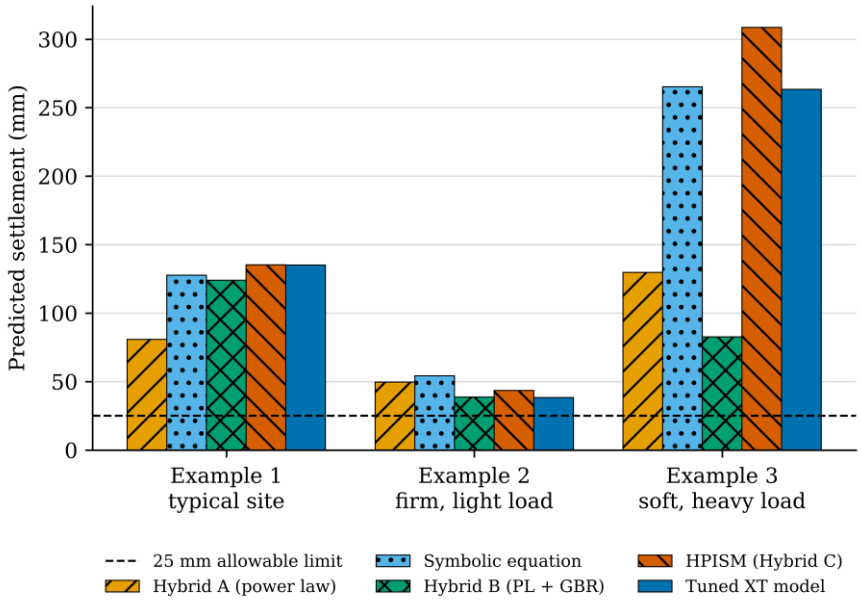


Fig. 3. Predicted settlement for the three worked application examples using the five modeling routes (dashed line: 25 mm allowable limit).

In the favorable case (Example 2), all five routes agree closely, within a narrow range of 38.4–54.2 mm, although even these values exceed the 25 mm serviceability limit typically applied to isolated footings. For the typical (Example 1) and unfavorable (Example 3) cases, the routes diverge more; in Example 3, for instance, they span from 82.6 mm (Hybrid B) to 308.7 mm (HPISM). Because of its higher validated accuracy, the tuned XT model provides the primary estimate, while the symbolic equation and the HPISM serve as transparent, manually workable checks. In all three examples the predicted settlement exceeds the 25 mm limit, and in Examples 1 and 3 it does so by a wide margin, indicating the need for ground improvement, a raft, or a deep foundation.

Across the three examples, the tuned XT model and the HPISM generally returned larger settlements than the Terzaghi and Janbu methods and also exceeded the Meyerhof estimates. One possible explanation is that approaches calibrated on stiffer temperate soils may underpredict deformation in soft, compressible deposits. The Schmertmann method gave an estimate close to the proposed models for the typical case (approximately 132–135 mm), which is expected because the reference settlements used for training were computed with that method; the predictions diverged in the unfavorable case.

Engineering Implications

The HPISM uses CPT-derived inputs to estimate settlement within a single framework and reduces the sequence of layering and correction-factor calculations associated with conventional procedures. During preliminary design, engineers can use it to examine how changes in footing width, embedment depth and applied pressure affect predicted settlement. Its explicit form also allows implementation in a spreadsheet or design-office program. Application should remain within the predictor ranges represented in Table 3. Unit weight requires particular attention because the closed-form correction is sensitive to its value, whether obtained from CPT correlations or laboratory testing. The model addresses serviceability settlement of shallow foundations only; it is not intended for bearing-capacity analysis, liquefaction assessment or deep-foundation design.

Conclusions

The work developed and tested the Hybrid Physics-Informed Settlement Model (HPISM) to estimate shallow-foundation settlement in lateritic soils using eight CPT-derived parameters. Twelve supervised algorithms were assessed using 8,934 CPT records with Schmertmann-computed reference settlements. The ensemble approaches achieved test R^2 values of at least 0.99, compared with 0.67–0.72 for linear regression. Among the tested algorithms, the tuned Extra Trees model—termed the XT model—performed best, with test R^2 = 0.9950 and RMSE = 6.13 mm. SHAP placed foundation width and applied pressure at the top of the feature ranking, followed by corrected cone resistance, embedment depth, OCR, soil behavior type index, unit weight and foundation length ($B > q > q_t > z > OCR > I_c > \gamma > L$). The ranking was used to guide symbolic regression, which produced an explicit nested equation with test R^2 = 0.9133. Adding a Gradient Boosting residual correction yielded the HPISM, with test R^2 = 0.9901 and RMSE = 8.65 mm. A fully closed polynomial form was also obtained; it can be evaluated manually or in a spreadsheet, with test R^2 = 0.9725.

The main outputs are: (i) an integrated CPT-derived dataset with reference settlements; (ii) a comparative assessment of twelve supervised learning algorithms; (iii) a SHAP-based account of predictor importance; (iv) a symbolic-regression equation evaluated on a held-out test set; and (v) a hybrid model that combines ensemble predictions with an explicit design expression through residual correction. In the worked examples, the HPISM estimates differed from those obtained using Terzaghi, Janbu, Meyerhof and Schmertmann. This highlights the importance of checking conventional methods against data representative of tropical residual soils, particularly where the ground is soft and compressible.

Where prediction accuracy is the main priority, the tuned Extra Trees model is the preferred option; the HPISM is more suitable when an explicit and reviewable estimate is required. Because the training targets were computed with the Schmertmann method rather than measured, the models currently reproduce a CPT-based settlement estimate, and their accuracy for real foundations on lateritic soils must be confirmed against settlements measured on instrumented footings or plate load tests. Before either model is used routinely outside the database's represented conditions, its predictions should be checked against local field measurements. Further work should test the models on independent datasets from different geological settings, extend the approach to deep and piled foundations, and examine alternatives that incorporate physics-informed neural networks (Karniadakis et al. 2021).

Data Availability Statement

Some or all data, models, or code that support the findings of this study are available from the corresponding author upon reasonable request. These include the integrated CPT database with computed reference settlements, the trained models, and the 44-term coefficient tables for the closed-form correction g(X).

References

Adadi, A., and M. Berrada. 2018. "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)." IEEE Access 6: 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052.

Arrieta, A. B., N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, R. Chatila, and F. Herrera. 2020. "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI." Inf. Fusion 58: 82–115. https://doi.org/10.1016/j.inffus.2019.12.012.

Boussinesq, J. 1885. Application des potentiels à l'étude de l'équilibre et du mouvement des solides élastiques. Paris: Gauthier-Villars.

Bowles, J. E. 1996. Foundation analysis and design. 5th ed. New York: McGraw-Hill.

Breiman, L. 2001. "Random forests." Mach. Learn. 45 (1): 5–32. https://doi.org/10.1023/A:1010933404324.

Burland, J. B., and M. C. Burbidge. 1985. "Settlement of foundations on sand and gravel." Proc. Inst. Civ. Eng. 78 (6): 1325–1381. https://doi.org/10.1680/iicep.1985.1058.

Chen, T., and C. Guestrin. 2016. "XGBoost: A scalable tree boosting system." In Proc., 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, 785–794. New York: ACM. https://doi.org/10.1145/2939672.2939785.

Cranmer, M. 2023. "Interpretable machine learning for science with PySR and SymbolicRegression.jl." Preprint, arXiv. https://doi.org/10.48550/arXiv.2305.01582.

Das, B. M., and N. Sivakugan. 2019. Principles of foundation engineering. 9th ed. Boston: Cengage Learning.

Das, B. M., and K. Sobhan. 2018. Principles of geotechnical engineering. 9th ed. Boston: Cengage Learning.

De Beer, E. E., and A. Martens. 1957. "Method of computation of an upper limit for the influence of the heterogeneity of sand layers on the settlement of bridges." In Vol. 1 of Proc., 4th Int. Conf. on Soil Mechanics and Foundation Engineering, 275–282. London: Butterworths.

Dorogush, A. V., V. Ershov, and A. Gulin. 2018. "CatBoost: Gradient boosting with categorical features support." Preprint, arXiv. https://doi.org/10.48550/arXiv.1810.11363.

Friedman, J. H. 2001. "Greedy function approximation: A gradient boosting machine." Ann. Stat. 29 (5): 1189–1232. https://doi.org/10.1214/aos/1013203451.

Géron, A. 2022. Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow. 3rd ed. Sebastopol, CA: O'Reilly Media.

Iyeke, S. D., E. O. Eze, J. O. Ehiorobo, and S. O. Osuji. 2016. "Estimation of shear strength parameters of lateritic soils using artificial neural network." Niger. J. Technol. 35 (2): 260–269.

James, G., D. Witten, T. Hastie, and R. Tibshirani. 2021. An introduction to statistical learning: With applications in R. 2nd ed. New York: Springer.

Janbu, N. 1967. Settlement calculations based on the tangent modulus concept. Bulletin No. 2. Trondheim, Norway: Norwegian Institute of Technology.

Karniadakis, G. E., I. G. Kevrekidis, L. Lu, P. Perdikaris, S. Wang, and L. Yang. 2021. "Physics-informed machine learning." Nat. Rev. Phys. 3 (6): 422–440. https://doi.org/10.1038/s42254-021-00314-5.

Koza, J. R. 1992. Genetic programming: On the programming of computers by means of natural selection. Cambridge, MA: MIT Press.

La Cava, W., P. Orzechowski, B. Burlacu, F. O. de França, M. Virgolin, Y. Jin, M. Kommenda, and J. H. Moore. 2021. "Contemporary symbolic regression methods and their relative performance." In Proc., Neural Information Processing Systems Track on Datasets and Benchmarks.

Lundberg, S. M., and S.-I. Lee. 2017. "A unified approach to interpreting model predictions." In Advances in neural information processing systems 30, 4765–4774. Red Hook, NY: Curran Associates.

Lunne, T., P. K. Robertson, and J. J. M. Powell. 1997. Cone penetration testing in geotechnical practice. London: Blackie Academic & Professional.

Mayne, P. W. 2014. "Interpretation of geotechnical parameters from seismic piezocone tests." In Proc., 3rd Int. Symp. on Cone Penetration Testing (CPT'14), 47–73. Las Vegas, NV.

Meyerhof, G. G. 1965. "Shallow foundations." J. Soil Mech. Found. Div. 91 (SM2): 21–31.

Meyerhof, G. G. 1974. "Penetration testing outside Europe: General report." In Vol. 2.1 of Proc., European Symp. on Penetration Testing, 40–48. Stockholm, Sweden.

Molnar, C. 2022. Interpretable machine learning: A guide for making black box models explainable. 2nd ed. https://christophm.github.io/interpretable-ml-book/.

Okewale, I. A., and H. Grobler. 2022. "Investigations into suitability of tropical clay for engineering applications." Innovative Infrastruct. Solutions 7 (1): 85.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay. 2011. "Scikit-learn: Machine learning in Python." J. Mach. Learn. Res. 12: 2825–2830.

Poulos, H. G., and E. H. Davis. 1974. Elastic solutions for soil and rock mechanics. New York: Wiley.

Robertson, P. K. 2009. "Interpretation of cone penetration tests—A unified approach." Can. Geotech. J. 46 (11): 1337–1355. https://doi.org/10.1139/T09-065.

Robertson, P. K. 2010. "Soil behaviour type from the CPT: An update." In Vol. 2 of Proc., 2nd Int. Symp. on Cone Penetration Testing (CPT'10), 575–583. Huntington Beach, CA.

Robertson, P. K., and K. L. Cabal. 2022. Guide to cone penetration testing for geotechnical engineering. 7th ed. Signal Hill, CA: Gregg Drilling & Testing.

Rudin, C. 2019. "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead." Nat. Mach. Intell. 1 (5): 206–215. https://doi.org/10.1038/s42256-019-0048-x.

Schmertmann, J. H. 1970. "Static cone to compute static settlement over sand." J. Soil Mech. Found. Div. 96 (3): 1011–1043.

Schmertmann, J. H. 1978. Guidelines for cone penetration test: Performance and design. Rep. No. FHWA-TS-78-209. Washington, DC: Federal Highway Administration.

Schmidt, M., and H. Lipson. 2009. "Distilling free-form natural laws from experimental data." Science 324 (5923): 81–85. https://doi.org/10.1126/science.1165893.

Shapley, L. S. 1953. "A value for n-person games." In Contributions to the theory of games II, edited by H. W. Kuhn and A. W. Tucker, 307–317. Princeton, NJ: Princeton University Press.

Skempton, A. W., and L. Bjerrum. 1957. "A contribution to the settlement analysis of foundations on clay." Géotechnique 7 (4): 168–178. https://doi.org/10.1680/geot.1957.7.4.168.

Terzaghi, K. 1943. Theoretical soil mechanics. New York: Wiley.

Terzaghi, K., R. B. Peck, and G. Mesri. 1996. Soil mechanics in engineering practice. 3rd ed. New York: Wiley.

Willard, J., X. Jia, S. Xu, M. Steinbach, and V. Kumar. 2022. "Integrating scientific knowledge with machine learning for engineering and environmental systems." ACM Comput. Surv. 55 (4): 66. https://doi.org/10.1145/3514228.