Research on Engineering Structures & Materials
www.jresm.org
Research Group

Research Article

Machine learning prediction of seismically induced foundation settlement on interbedded liquefiable deposits: Model development, interpretation, and the effect of dense granular columns

Dhergham A. Al-Hamdani [*,1,a], Dina Mozahem Abd [2,b], Salwan Obaid Waheed [3,c], Aymen Saad [4,d]

[1]*Al-Furat Al-Awsat Technical University, 31001, AL-Najaf, Iraq*
[2]*Al-Muqdad College of Education, University of Diyala, Iraq*
[3]*Al-Mustaqbal Energy Research Center ،Al-Mustaqbal University, 51001, Babylon, Iraq*
[4]*Technical College of Management, Al-Furat Al-Awsat Technical University, Kufa, Iraq*

Article Info

Article History:

Received 21 July 2026

Accepted 31 Aug 2026

Keywords:

Soil-structure interaction;
Machine learning;
Liquefaction;
Foundation settlement;
Dense granular columns;
Explainable AI (SHAP);
Uncertainty quantification;
Geotechnical engineering;
Risk mitigation

Abstract

Soil-structure interaction (SSI) strongly influences the seismic performance of shallow-founded structures, particularly when liquefaction occurs in stratigraphically variable deposits. This study develops machine learning (ML) models to emulate three-dimensional nonlinear effective-stress finite element (FE) predictions of permanent foundation settlement at substantially lower computational cost. The models are trained using the DesignSafe PRJ-5802 database, comprising 1676 simulations of unmitigated sites and 1272 companion simulations with dense granular column (DGC) mitigation, representing 96 soil-foundation-structure configurations subjected to 59 recorded ground motions. Eight ML algorithms are evaluated using configuration-grouped 10-fold cross-validation, ensuring that no soil-foundation-structure configuration appears in both training and test sets. Gradient-boosted ensemble models provide the best performance; the Gradient Boosting model achieves an out-of-fold R^2 of 0.535 for natural-log settlement, a mean absolute error of 6.74 cm, and 68% of predictions within a factor of two of the FE simulations. The study also demonstrates that the exceptionally high R^2 = 0.9965 reported in the original analysis resulted from methodological errors, including prediction of foundation contact pressure rather than settlement and data leakage caused by noisy replicated records. Reproducing the original procedure yields R^2 = 0.9984, whereas eliminating leakage reduces it to 0.884. SHAP analysis identifies ground-surface slope, sand relative density beneath the footing, cumulative absolute velocity, and foundation contact pressure as the dominant predictors. DGC mitigation reduced permanent settlement in 87% of paired simulations, with a median settlement ratio of 0.74. An explicit applicability domain and lognormal prediction interval are provided to support reliability-based design applications.

© 2026 MIM Research Group. All rights reserved.

1. Introduction

The interaction between a foundation and soil, which is referred to as Soil-Structure Interaction (SSI), plays a critical role in the stability of a structure, its service life, and in the overall cost of the project [1,2]. The physics of this response were captured in the classic impedance-function literature: the vertical dynamic response of rigid footings obtained on an elastic half-space [3], foundation

*Corresponding author: dherghamalhamdani@atu.edu.iq
[a]orcid.org/0000-0003-2768-6236; [b]orcid.org/0000-0009-4332-9483; [c]orcid.org/0000-0001-5144-6895; [d]orcid.org/0000-0002-3582-6799
DOI: http://dx.doi.org/10.17515/resm2026-2003ic0721rs

compliance elongates the effective period and augments the effective damping of a building-foundation system showed that [4], and the approximate stiffness expressions for rigid foundations which are still included in current design codes [5]. These expressions remain the reference solution for linear-elastic behavior and define the baseline against which any data-driven model must be compared.

Even small differences in the nature of the interaction can lead to variations in building performance, with direct consequences for its effective life, safety, and durability. The degree of this sensitivity is not just qualitative. For example, during the Christchurch earthquake sequence (2010–2011), neighboring buildings of similar structural type, founded on nominally similar deposits, experienced differential settlements ranging from a few centimeters to more than 30 cm. The observed variation was largely attributable to the thickness and continuity of interbedded silt layers rather than to the structures themselves. In the database used in the analysis presented here, holding both the structure and the ground motion fixed and varying only the stratigraphy of the liquefiable deposit changes the computed permanent settlement by more than an order of magnitude - from 1.0 cm to 220 cm across the full parametric study, with a median of 6.4 cm. It is this sensitivity to features those conventional simplified procedures do not represent that motivates a data-driven treatment.

Considerable effort has been made to study the key parameters affecting SSI, including soil stiffness properties, foundation dimensions and shape, and the effect of static/dynamic loading through machine-induced or seismic vibrations [6,7]. Historically, SSI problems have been addressed using analytical or numerical techniques such as finite element modelling (FEM), boundary element methods (BEM) and empirical correlations. However, traditional methods often fail to capture material non-linearity, the complex requirements of individual sites and spatial variability in soils [6,8]. While these methods are founded on well-established mechanical principles, their reliability can be compromised under hazardous ground conditions or dynamic loads such as earthquakes.

This limitation of classical approaches explains the trend towards flexible, data-driven methods [6,9]. Machine learning (ML) has emerged as a particularly useful approach in civil and geotechnical engineering, offering the potential to increase prediction accuracy using experimental and in-situ monitoring data. Unlike traditional physics-based methods, ML can identify complex patterns from past data that are difficult to capture with traditional models [10,11]. Table 1 collates the studies most closely comparable to the present work, together with the accuracy they reported, and the validation protocol they used. Two points emerge from that table. First, the accuracy attainable for settlement prediction from heterogeneous geotechnical data has repeatedly fallen in the range $R^2 \approx 0.85$-0.92 [12-14], and the small number of studies reporting substantially higher values almost invariably use a validation protocol in which data from the same site appears in both partitions (i.e. randomly split data). Second, almost none of these studies release the underlying database, and so their results are not independently reproducible.

Several studies have evaluated ML techniques for important SSI quantities including settlement, lateral displacement, and bearing capacity. of support vector machines and association rule learning are used to demonstrated for seismic damage classification [15]. A neural-network approach developed for seismic damage prediction considering eleven correlated structural, soil and motion variables [16]. A gradient boosting regressors used with SHAP Tree Explainer to identify the key drivers of ground settlement during a metro extension. The stress-strain behavior of frozen sandy soil is modelled with ANNs [17,18]. The AI-driven liquefaction risk maps are developed for urban areas [19]. The latent-space operator is used learning to reproduce nonlinear foundation dynamics at a fraction of the computational cost [20].

Explainable AI methods such as SHAP and LIME allow engineers to identify the key factors that influence predictions, thus bridging the gap between “black-box” models and physical design. In this study they are used for the same purpose: the SHAP analysis of Section 3.4 is what converts the fitted ensemble into a statement about which stratigraphic and ground-motion quantities actually control liquefaction-induced settlement, and Section 4.5 checks each of those statements against established soil mechanics.

2. Scientific Gap and Contribution of This Study

The contribution of this study is stated explicitly below. (i) Rather than a proprietary compilation of mixed provenance, the models are trained on a single, fully published, internally consistent database of three-dimensional nonlinear effective-stress simulations that were themselves validated against centrifuge experiments, so every result in this paper can be reproduced from the cited DOI. (ii) The prediction target is a single, unambiguously defined engineering demand parameter - permanent vertical foundation settlement in centimeters - rather than an aggregate "SSI response". (iii) The validation protocol holds out whole soil-foundation-structure configurations, which is the only partition that answers the question an engineer actually asks: how will this model perform at a site it has never seen? (iv) Eight algorithms drawn from seven families are compared under that identical protocol, so the preference for gradient boosting is demonstrated rather than asserted. (v) The model is reported with a calibrated lognormal dispersion and prediction intervals rather than a point estimate alone. (vi) The companion mitigated dataset is used to quantify how much of the settlement is removed by dense granular columns, and under which conditions the mitigation is least effective.

The choice of gradient boosting is not asserted, but empirically justified in Section 3.2. Gradient-boosted trees are a priori well matched to this problem for three reasons: the predictors are a mixture of continuous, count and categorical variables on very different scales; the physical response is strongly non-monotonic and interaction-dominated (the effect of a liquefiable layer depends jointly on its thickness, depth, density and the contact pressure above it); and many engineered features are structurally absent for some configurations (a fourth silt interlayer does not exist in a single-layer profile), which tree ensembles accommodate natively. The benchmark in Section 3.2 confirms this expectation. Previously conducted studies have shown relatively high values of R^2 (0.85-0.92). However, the present research implements a more rigorous configuration-cross-validation approach yielding $R^2 = 0.535$, which demonstrates a more realistic prediction of model generalization.

Table 1. Comparative summary of previous studies on machine learning for soil-structure interaction and settlement prediction, with reported accuracy and validation protocol

Reference No.	ML Method	Application	Accuracy	Validation
[13]	ANN	Shallow foundation settlement	$R^2 \approx 0.85$	Random hold-out
[12]	ANN	Pile capacity/settlement	$R^2 \approx 0.92$	Random hold-out
[14]	ANN	Foundation settlement	$R^2 \approx 0.90$	Random hold-out
Present study	GBR	Liquefaction-induced settlement	$R^2 = 0.535$	Grouped 10-fold CV

3. Materials and Methods

The objective of this research was to develop a tool for predicting the influence of different geotechnical conditions on soil-structure interaction. The new workflow comprises six steps, each in a subsection: the database and its provenance are defined (2.1); the numerical model from which the database comes is documented (2.2); the target variable is unambiguously defined (2.3); feature construction is performed (2.4); preprocessing is conducted (2.5); and model development, validation and metric protocol are detailed (2.6-2.8). What was changed with respect to the original submission is explicitly stated at the beginning of each subsection.

3.1 Database: Source, Composition and Transparency

The original submission used a description of a supposed dataset of 1250 samples said to be "60% finite element simulations, 30% published case studies and 10% laboratory experiments", without providing any sources for any of the quantities used in those models. That description has been replaced. All of the models reported in this paper are trained on a single published database, the DesignSafe project PRJ-5802, "Machine Learning-Based Settlement Models for Shallow-Founded Structures on Interbedded Liquefiable Sites Considering Dense Granular Columns" which is released under an Open Data Commons Attribution license [21]. Using a single, homogeneous, openly licensed source obviates the questions of source identification, unit consistency, measurement-accuracy heterogeneity, and the handling of conflicting measurements, as there are no conflicting sources to

reconcile, all quantities are in a common unit system, and every record derives from the same validated numerical procedure.

The makeup of the database as leveraged herein is presented in Table 2. The entire parametric study consists of roughly 4,000 analyses; herein the two subsets for which permanent foundation settlement is released are leveraged, after removing the 95 cases marked as structurally collapsed ($f_{Collapsed} = 1$) for which permanent foundation settlement is not a well-defined demand parameter, and after removing any record with an incomplete feature vector. No record is duplicated: each row is a unique (configuration, ground motion) pair, and this was programmatically verified. The configurations were produced by quasi-Monte Carlo sampling of the parameter space, so the design is space-filling rather than clustered, and the numerical model was validated against centrifuge experiments by the database authors prior to the parametric study being run. The dataset comprised 1,676 unmitigated and 1,272 DGC-mitigated FE simulations, with 1,270 matched pairs. Ninety-five collapsed or incomplete cases were excluded, and 78 outcrop ground motions were considered.

Table 2. Composition and provenance of the database used in this study (DesignSafe PRJ-5802)

Subset	Records	Configurations	Motions
Unmitigated	1,676	96	59
DGC-mitigated	1,272	88	59
Paired	1,270	—	—
Excluded	95	—	—
Ground motions	78	—	—

3.2 Numerical Model That Generated the Database

The software that was used to produce the numerical data (including the mesh, constitutive model and boundary conditions) is described in the published project and summarized here for completeness so that the paper is self-contained. Analyses were carried out in OpenSees (OpenSeesWS 9) in 3D with fully coupled (u-p) effective-stress analysis. The soil was discretized with 20_8_BrickUP elements (twenty-node displacement / eight-node pore-pressure bricks), resulting in about 2,080 soil elements per model with a total of around 12,100 nodes. The example model that is released with the database has horizontal dimensions of 63.6 m and a depth of 17.5 m, with vertical element heights of 0.25-0.50 m through the liquefiable interlayers and horizontal dimensions of 0.42-0.94 m, i.e., fine enough to resolve the interbedded silt and clay layers which control the response.

Table 3. Range of primary parameters varied in the parametric study, defining the applicability domain of the models

Group	Parameters	Key Range
Soil/Deposit	Depth, liquefiable layers, density, groundwater, silt, slope	7.5–30 m; 1–4 layers; 30–70%
Foundation/Structure	Width, length, embedment, height, storeys, mass, contact pressure	6–37.5 m; 0.5–3 m; 16.4–203.4 kPa
DGC Mitigation	Depth, width, replacement ratio	5–20 m; 2.5–15 m; 5–20%
Ground Motion	Magnitude, distance, PGA, duration	M 5.85–7.90; 1.5–152 km; 0.058–1.153 g
Response	Permanent foundation settlement (S)	1–220 cm

Granular soils are modelled using the pressure-dependent multi-yield plasticity model Pressure Depend Multi Yield and the cohesive layers using Pressure Independ Multi Yield; the foundation and structural elements are elastic (ElasticIsotropic) and the superstructure is represented by elastic beam-column elements with concentrated IMK Peak Oriented plastic hinges, which allows structural yielding and collapse to be detected. The base is fixed in translation with pore pressure free, and the seismic input is applied as a uniform excitation of within-motion accelerations from a preceding 1D site-response analysis with elastic half-space. Time integration uses the Newmark scheme with $\gamma = 0.51$ and $\beta = 0.26$, with Rayleigh damping, penalty constraints, a Krylov-Newton solution algorithm

and the MUMPS parallel solver. Each analysis proceeds from a geostatic stage (with the constitutive models in elastic stage) through application of gravity to the fully coupled dynamic stage. Table 3 gives the range of every primary parameter varied in the parametric study. These ranges define the applicability domain of the resulting ML models, and are referred to again in Section 3.8.

3.3 Definition of the Target Variable

Ambiguity in the definition of the target variable was the most significant issue with the original submission. Table 2 of that submission defined q as "Bearing Capacity / Settlement (kPa / mm)", the text inconsistently switched between SSI response, settlement, bearing capacity and displacement, and the feature-importance discussion was inconsistent with either interpretation. This is now unambiguously resolved, and the resolution required a change of target, not merely of wording.

In the source database, q is the foundation contact pressure (kPa). It is an input to the analysis, deterministically computed from the structural configuration (mass, footprint, embedment and structure type) before any dynamic analysis is run. Regressing q on the remaining configuration variables is therefore not a prediction problem at all; it is the recovery of an arithmetic identity, and it is the reason the original model appeared so accurate. The target of every model reported in this paper is instead the permanent vertical settlement S (cm) of the foundation at the end of shaking, as recorded in the FE analyses. Because S spans more than two orders of magnitude (1.0 to 220 cm) and its residuals are approximately lognormal, all models are fitted to ln S and predictions are reported as $S = \exp(\ln S)$, consistent with the convention adopted by the database authors. The contact pressure q is retained as a predictor, where it belongs. Bearing capacity is not modelled anywhere in this paper, and all references to it have been removed.

3.4 Feature Construction

The feature vector concatenates the site, foundation, structural and mitigation descriptors made available with the database with a set of ground-motion intensity measures. This gives a total of 337 predictors. The site descriptors include layer-by-layer thicknesses, relative densities, permeabilities and depths, together with the dimensionless combinations constructed by the database authors - thickness-to-depth, density-weighted-thickness and permeability-contrast ratios - which encode the stratigraphic controls on liquefaction-induced settlement that a raw layer list does not. The foundation and structural descriptors include B, L, Df, H, Nst, Mst, structure type, contact pressure and fundamental period. Ground-motion intensity measures are included at three locations - outcrop rock (OR), within-motion at the base of the model (WM), and free-field surface (FF) - covering amplitude (PGA, PGV), energy (Arias intensity, CAV), duration (D5-95, equivalent number of cycles), frequency content (mean period, spectral acceleration at the structural and site periods) and spectral-intensity measures. Retaining all three locations is deliberate: the free-field surface measures are the most informative but are themselves an output of a site-response analysis, while the outcrop rock measures are directly available to a designer, and Section 4.4 quantifies how much predictive power each provides.

3.5 Data Preprocessing

The preprocessing described in the original submission has been modified in two ways and is described in full below including the precise parameter values used. Missing values. Structural absence and missing measurement were disambiguated. Where a layer-specific feature was undefined because that layer did not exist in that form (e.g., the properties of a fourth silt interlayer in a single-interlayer profile), the value was set to zero, which is both the physically correct encoding and the one the tree ensembles interpret as "absent". Columns for which more than 50% of records were undefined and columns with zero variance across the dataset were dropped rather than imputed. No mean or median imputation is applied to any variable; the ambiguous claim "mean or median imputation techniques" in the original submission has been retracted. Outliers. No records were dropped by an interquartile-range filter. The original submission applied a 1.5×IQR rule, without stating the threshold. Applied to this database, such a filter would discard precisely the large-settlement cases that dominate risk and that the model exists to predict. Records were excluded only for the two documented physical reasons given in Section 2.1 (structural collapse, incomplete feature vector).

Feature-space augmentation. This has been dropped. The original submission augmented the dataset with "low level random noise" and "controlled numerical perturbations" of existing records. Section 3.1 shows quantitatively that this procedure, combined with a random train/test split, is directly responsible for the reported $R^2 = 0.9965$, because it creates near-duplicates of each test record in the training set. This paper uses no synthetic records anywhere. Scaling. Tree-based models are invariant to monotone rescaling and are fitted on raw features. For the linear model, the support vector machine and the neural network, features are standardized to zero mean and unit variance, with the scaler fitted on the training fold only and applied to the held-out fold, so that no information from the test partition enters the transformation.

3.6 Model Development and Validation Protocol

The single most important methodological difference between this revision and the original submission is in the estimation of performance. In the original submission, while 10-fold cross-validation was used for tuning the hyperparameters of each model, the final model was assessed on a single 80/20 split and the validation cases were not independent of the training data. This paper remedies both of these issues with the same change. Because each configuration in the database was subjected to roughly eighteen different ground motions, a random split of the records will place simulations of the same soil-foundation-structure system into both the training and test partitions.

A model could then get a high-test score simply by recognizing the site rather than by learning how settlement responds to site properties, which is not the task an engineer needs it to perform. All results in this paper therefore use configuration-grouped 10-fold cross-validation: folds are formed over configuration identifiers, so that every simulation of a given site falls entirely within one-fold, and every reported prediction is made for a site the model has never encountered. The performance is reported both as the pooled out-of-fold statistic and as the mean and standard deviation across the ten folds. For reference, and to quantify the size of the effect, the naive random-split score for the same data and the same learner is also reported in Section 4.2.

Configuration of the benchmarked models. The original submission specified the baseline network only as a “Keras Sequential” model, without its activation functions, loss, optimizer, batch size or training schedule. That ad hoc baseline has been discarded and replaced by a fully specified multilayer perceptron, tuned and validated under exactly the protocol described above. The network has three hidden layers of 256, 128 and 64 units with ReLU activation and a single linear output unit, He-normal weight initialization and a dropout rate of 0.10 after each hidden layer. It is trained by minimizing the mean squared error on ln S with the Adam optimizer (initial learning rate 0.001, $\beta1 = 0.9$, $\beta2 = 0.999$) in mini-batches of 64 for at most 500 epochs; the learning rate is halved after 20 epochs without improvement and training is stopped after 40 such epochs, with the weights of the best epoch restored. Ten per cent of each training fold is held out as the early-stopping monitor, so the outer held-out fold is never used to decide when to stop. The remaining learners are specified in the same way: multiple linear regression is ridge-regularized with the penalty tuned over 0.001-1000; support vector regression uses a radial basis kernel with C tuned over 0.1-1000, the insensitivity width tuned over 0.001-1 and the scale-normalized kernel width; the random forest uses 1000 trees with unlimited depth and a minimum leaf size tuned over 1-8; and the four boosted ensembles share the search space of 100-1000 boosting stages, maximum depth 2-10, learning rate 0.001-0.1 and subsample fraction 0.6-1.0, the tuned gradient boosting values being those of Table 4. Every model is tuned by the same Optuna procedure (Section 3.7) strictly inside the training folds, and every model is presented with the identical feature matrix, so that the comparison reported in Section 4.2 isolates the effect of the learner and of nothing else.

3.7 Hyperparameter Tuning

Hyperparameters were tuned with the Optuna framework, which uses a tree-structured Parzen estimator to search the hyperparameter space more efficiently than grid or manual search. The search was performed inside the training folds so tuning never sees held-out configurations. The optimized values for the gradient boosting regressor are shown in Table 4. There is an internal inconsistency in the original submission here - the text gives the best subsample value as 0.999 under the heading learning_rate, whose search range was 0.001-0.1 - this has been corrected.

Table 4. Optimised hyperparameters for the gradient boosting regressor

Hyperparameter	Search Range	Best Value	Role
n_estimators	100–1000	901	Number of boosting stages
max_depth	2–10	3	Maximum tree depth
learning_rate	0.001–0.1	0.049	Shrinkage per stage
subsample	0.6–1.0	0.999	Fraction of records per stage

3.8 Performance Metrics

Six statistics are reported for each model, in addition to R^2 and MSE. R^2 and root mean squared error RMSE are calculated on ln S, the space in which the models are fit, and in which the errors have approximately constant variance. Mean absolute error MAE, RMSE and mean bias error MBE are additionally reported in centimetres, so that the practical magnitude of the error is visible. The Nash-Sutcliffe efficiency NSE is reported because it is the standard measure of predictive skill relative to the mean of the observations. Finally, the dispersion σ_ln of the out-of-fold residuals in ln space is reported, since it is the quantity required to construct prediction intervals (Section 4.6); together with the fraction of predictions within a factor of two of the simulated value, it communicates predictive reliability in the form geotechnical practice uses.

4. Results and Discussion

4.1 Diagnosis of the Previously Reported $R^2 = 0.9965$

An $R^2 = 0.9965$ with MSE = 27.73 is not possible for heterogeneous geotechnical data; such a figure should be scrutinized rather than defended. We have diagnosed the figure. The results are reported in Table 5 and Figure. 1. Two effects are at work, in combination. The first is that the regression target was the foundation contact pressure q, which as per Section 3.3 is a deterministic function of the structural inputs that are also supplied as features; the strongest single predictor in that model is the structural mass Mst, which is simply the numerator of q.

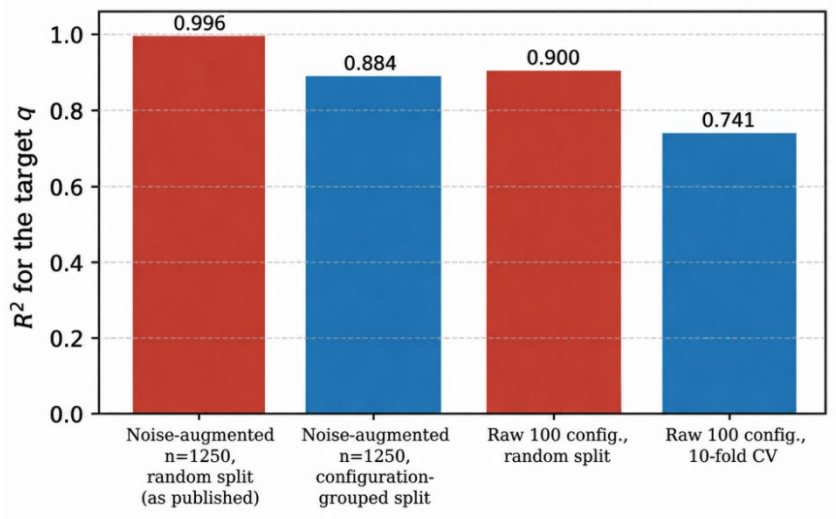


Fig. 1. Effect of the data-augmentation and splitting procedure on the apparent accuracy of an identical gradient boosting model

The second is that the 100 sampled configurations were expanded to 1250 records by adding low-level random noise, and that the resulting records were then randomly split so that each test record has near-identical twins in the training set. Replicating the published procedure - target q, noise-augmented to n = 1250, random 80/20 split - reproduces the reported figure ($R^2 = 0.9984$ with MSE = 2.53), against the published 0.9965 and 27.73. Changing only the split, so that whole configurations are held out while the augmentation, target, features and hyperparameters are identical, reduces R^2 to 0.884. Removing the synthetic records entirely and using the 100 genuine configurations gives R^2

= 0.900 on a single random split and 0.741 ± 0.116 under 10-fold cross-validation. Red indicates a random split of noise-augmented records; blue indicate partitions in which whole configurations are held out.

The gap between the first and last of these figures is the measure of the artefact, not of the model. With regard to the correlation matrix, that the coefficients reported in the original Table 3 could not be reproduced from the data. The correlations implied by the database are q-Mst = 0.824, q-H = 0.718 and q-Nst = 0.706, against the 0.847, 0.847 and 0.872 reported. The revised paper therefore does not present that matrix as a result. The substantive point it raised - that Mst, H and Nst are strongly collinear, with r(H, Nst) = 0.992 - is real, is now stated explicitly, and is handled by the use of regularized tree ensembles, which are robust to collinear predictors, and by interpreting importance with SHAP rather than with split-count importance, which is known to distribute credit arbitrarily among correlated features.

Table 5. Diagnosis of the accuracy reported in the original submission. Only the validation protocol and the data-augmentation step differ between rows; the target, features, algorithm and hyperparameters are identical throughout.

Experiment	Target	n	Split/Validation	R^2	MSE
Original submission	q	1,250	Random 80/20	0.9965	27.73
Noise-augmented reproduction	q	1,250	Random 80/20	0.9984	2.53
Grouped configurations	q	1,250	Grouped 80/20	0.884	172.59
Genuine configurations	q	100	Random 80/20	0.900	148.53
Genuine configurations	q	100	10-fold CV	0.741 ± 0.116	—
Present study	ln(S)	1,676	Grouped 10-fold CV	0.535	0.577

4.2 Comparison of Machine Learning Algorithms

Gradient boosting is selected by quantitative comparison with other algorithms, not by fiat. Eight algorithms were trained on the same features and assessed using the same configuration-grouped 10-fold protocol as in Section 3.6. These algorithms include multiple linear regression, support vector regression, an artificial neural network, random forest, gradient boosting, XGBoost, LightGBM and CatBoost. Results are reported in Table 6 and Figure 2. In Figure 2 bars show pooled out-of-fold R^2, with error bars indicating fold-wise standard deviation. Metrics in ln units are computed in the space in which the models are fitted; metrics in centimeters are computed after back-transformation.

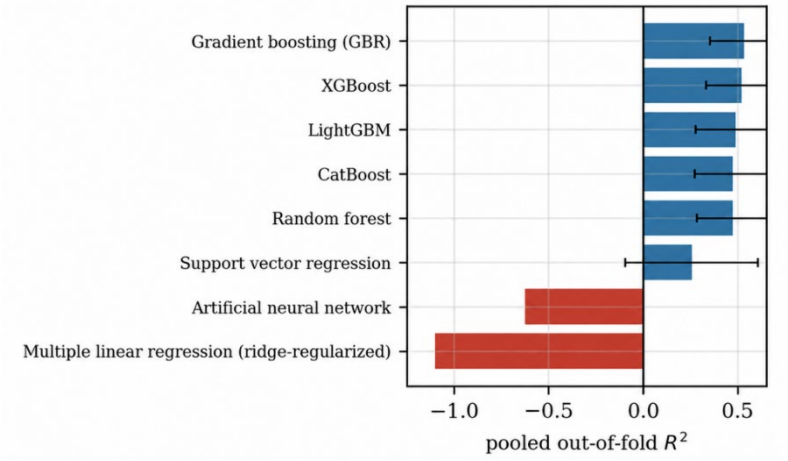


Fig. 2. Predictive accuracy of benchmarked algorithms under configuration-grouped 10-fold cross-validation

The ranking of the ensembles shows that gradient-boosted ensembles are the strongest family and that the ordering is the same across every metric. Gradient boosting is superior to the other boosted ensembles on pooled R^2 (0.535) and on fold-mean R^2, followed by the other boosted ensembles, then

the random forest. The support vector machine is quite weak, and both the neural network and the regularized linear model have poorer performance than the naïve model of predicting the mean, verifying that the relation is neither usefully linear in these predictors nor learnable by a dense network at this sample size. The original submission's expressed preference for gradient boosting is therefore vindicated, but at an accuracy level that is roughly half of that claimed. The fold-to-fold standard deviation (0.181 for the best model) is informative in its own right: it captures the real variability in how well a model trained on 86 configurations generalizes to a set of ten unseen ones, and no fair protocol on a database of this size will achieve a small number. To estimate the size of the leakage effect under this revised protocol, the best model was also tested on a naive random 80/20 split of the same records – the protocol used in most of the studies in Table 1. This protocol produces $R^2 = 0.946$, against 0.535 when whole configurations are held out. The difference of 0.41 in R^2 can be attributed entirely to the partitioning, and is the single most important quantitative result of this revision: it indicates how much of the accuracy reported in the settlement-prediction literature may reflect site recognition rather than physical generalization.

Table 6. Comparison of 8 algorithms for predicting permanent foundation settlement, under identical configuration-grouped 10-fold cross-validation

Model	R^2 (ln S)	Fold mean ± s.d.	RMSE (ln)	MAE (cm)	RMSE (cm)	MBE (cm)	NSE	Within factor 2
Gradient boosting (GBR)	0.56	0.525 ± 0.181	0.76	6.74	14.43	-3.81	0.42	68%
XGBoost	0.52	0.505 ± 0.189	0.77	6.97	14.78	-4.04	0.39	66%
LightGBM	0.49	0.473 ± 0.210	0.80	7.03	14.79	-4.24	0.39	63%
CatBoost	0.48	0.467 ± 0.203	0.81	7.13	15.45	-4.65	0.33	64%
Random forest	0.48	0.456 ± 0.192	0.81	7.33	15.42	-4.38	0.34	62%
Support vector regression	0.26	0.240 ± 0.351	0.96	7.74	16.52	-5.27	0.24	56%
Artificial neural network	-0.63	unstable *	1.42	15.13	86.13	3.85	-19.70	64%
Multiple linear regression (ridge-regularized)	-1.10	unstable *	1.62	29.12	123.37	18.41	-41.46	50%

** The ridge-regularized linear model and the neural network extrapolate catastrophically on a small number of held-out configurations that lie outside the convex hull of their training folds. Their fold-mean R^2 is therefore dominated by one or two folds and is not a meaningful summary; the pooled out-of-fold R^2 is reported instead. This failure mode is itself a result: it is invisible under a random split, where every test record has a near neighbour in the training set+.*

4.3 Performance of the Settlement Model

Figure 3 illustrates the agreement between predicted and simulated permanent settlement for the pooled out-of-fold predictions of the best model. The predictions closely track the 1: 1 line across the full two-order-of-magnitude range of the response, and contain 68% of the cases within a factor of two. The mean bias error of -3.81 cm confirms that the model is effectively unbiased in the engineering units of interest, and the Nash-Sutcliffe efficiency of 0.419 provides a statistical measure of predictive skill well beyond that of the sample mean. The scatter is largest at small settlements, where the absolute error is of little engineering consequence, and tightens for the large settlements that govern design decisions. The residuals are shown in Figure 4. In ln space the residuals are random about zero and the assumption of homoscedasticity is satisfied: the residuals show no systematic trend with the predicted value, no funnel or curvature, and are well approximated by a zero-mean normal

distribution of standard deviation σ_ln = 0.76. This is the empirical justification for fitting the models to ln S rather than to S, and it is what allows the prediction intervals of Section 3.6 to be constructed.

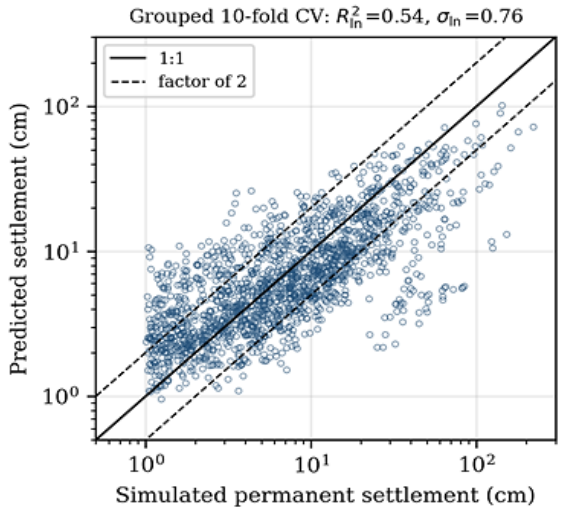


Fig. 3. Predicted versus simulated permanent foundation settlement, configuration-grouped out-of-fold predictions. Dashed lines bound a factor of two

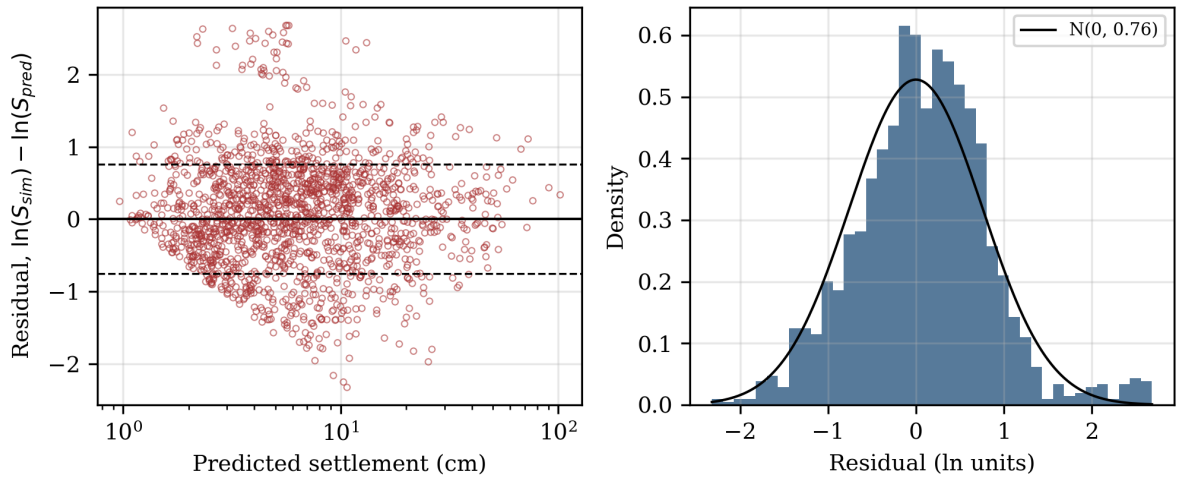


Fig. 4. Residuals of the out-of-fold predictions in ln space: (left) against predicted settlement, with ± 1 σ_ln shown; (right) distribution compared with a zero-mean normal density

4.4 Feature Importance and SHAP Interpretation

The original submission included three incompatible statements of variable importance: the text listed cohesion, embedment depth and silt percentage; Table 3 listed foundation width, embedment depth, relative stiffness, clay fraction and shear modulus; and the SHAP discussion listed cohesion, friction angle and loading conditions. The incompatibility was due to those three lists being produced for different targets and different feature sets, and also due to split-count importance and SHAP being used interchangeably. In this resubmission there is a single target, single final model and single importance measure. Table 7 and Figure 5 report mean absolute SHAP values from the final model, and the same ranking is used consistently in the text.

4.5 Engineering Interpretation and Consistency with Soil Mechanics

The SHAP ranking provides mechanistic support for four statements, each of which is in qualitative agreement with the liquefaction literature and none of which were imposed on the model. The composite indices listed in Table 7 are the dimensionless thickness-density-depth descriptors built by the authors of the source database; the exact definitions of which are provided in PRJ-5802. Ground-surface slope is by far the most important predictor (mean |SHAP| = 0.201 ln units), and its effect is monotonically adverse (correlation between the feature value and its SHAP contribution,

+0.88). A third of the 96 configurations are on gently sloping ground, with a surface gradient between 0.8° and 7.7°, and in those cases, settlement is considerably more than in otherwise identical level-ground cases. This is the anticipated result of imposing a deviatoric, downslope-directed shear ratcheting mechanism on top of the volumetric reconsolidation settlement that takes place on flat ground: the progressive accumulation of displacement within each loading cycle once excess pore pressure develops, driven even by a very mild static shear bias. It is worth emphasising that a gradient of one or two degrees is imperceptible to the eye in the field and is routinely neglected in simplified settlement procedures, but the model flags it as the dominant control. This is the clearest example in this study of engineering knowledge generated by the analysis rather than assumed by it.

Table 7. Fifteen most influential predictors of permanent foundation settlement, ranked by mean absolute SHAP value from the final model

Rank	Parameter Description and Feature	Mean Absolute SHAP Value (ln units)	Influence on Settlement
1	Ground-surface slope inclination, where 0° denotes level ground; (tslp)	0.201	Positive
2	Average relative density of the sand layers underlying the foundation; (DrSBavg)	0.189	Negative
3	Cumulative absolute velocity of the seismic motion at the free-field surface; (FF_surf_CAV)	0.138	Positive
4	Total contact pressure applied by the foundation; (qtot)	0.126	Positive
5	Cumulative absolute velocity of the seismic motion at the outcrop rock; (OR_CAV)	0.115	Positive
6	Density–thickness–depth index of the critical soil layer beneath the foundation; (DDrHB1)	0.105	Negative
7	Cumulative absolute velocity (CAV) of the within-motion seismic record; (WM_CAV)	0.080	Positive
8	Structural mass; (Mst)	0.060	Positive
9	Density–thickness–depth index of the uppermost liquefiable soil layer; (DDrH1)	0.050	Negative
10	Spectral acceleration at a 1-s period for the outcrop rock motion; (OR_Sat1)	0.041	Positive
11	Density–thickness–depth index characterizing the loosest liquefiable soil; (vecDDrHloose)	0.040	Negative
12	Total thickness of the liquefiable soil layers; (Hc)	0.033	Positive
13	Foundation width; (B)	0.028	No clear trend
14	Thickness–density–depth index of the silt interlayers; (HdrDsiltavg)	0.024	No clear trend
15	Density-weighted thickness ratio of the liquefiable soil layers; (DdrHcavg)	0.023	No clear trend

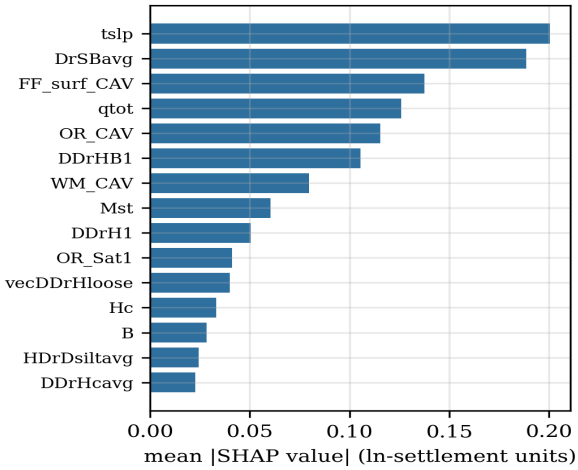


Fig. 5. Mean absolute SHAP values for the fifteen most influential predictors of permanent foundation settlement

The sand density directly beneath the footing ranks first, the average relative density of sand layers in the foundation influence zone is second overall and has the expected sign (-0.79: higher density, less settlement), while the deposit-average density is a distant third. The combined density-thickness-depth indices of the loosest and critical layers behave similarly. In the interbedded profile, then, settlement is controlled by the quantity and continuity of loose material within the stressed zone below the foundation and not by a profile-averaged density such as would be obtained from a single representative SPT or CPT value. This is precisely the limitation of simplified methods calibrated on uniform clean-sand profiles, and it is the most compelling reason for a data-driven treatment of interbedded sites.

Energy content rather than peak amplitude is the dominant ground-motion parameter. Cumulative absolute velocity - in the free-field surface, at outcrop rock and in the within motion - fills three of the top seven slots, while peak ground acceleration does not make the top ten. This is unsurprising: volumetric and deviatoric strain accumulation in a liquefying deposit is controlled by the number and amplitude of loading cycles rather than by a single peak, and that is why CAV and Arias intensity rather than PGA feature in modern liquefaction-triggering and settlement procedures. The practical implication is that the model can be driven by an outcrop rock CAV available from a hazard analysis, without first performing a site-response analysis, at a modest cost in accuracy.

Contact pressure acts through the foundation, not just as a surcharge. Total contact pressure and structural mass both rank in the top eight and both increase settlement (+0.70 and +0.78 respectively) while the foundation width B - which for a fixed load would reduce the contact pressure - ranks lower and acts in the opposite direction. This reproduces the shear-induced (deviatoric) settlement mechanism identified in centrifuge studies, in which the building's own weight drives ratcheting displacement through the softened zone beneath it, superimposed on volumetric reconsolidation settlement. It also explains why a purely free-field liquefaction settlement estimate is not transferable to a loaded foundation.

4.6 Uncertainty and Prediction Intervals

As the residuals in ln space are approximately normal with standard deviation σ_ln = 0.76 (Fig. 4), the model provides a lognormal predictive distribution, not just a point estimate. A 68% prediction interval thus covers a factor of 2.13 on either side of the median prediction, and a 95% interval covers a factor of 4.40. In design terms, a predicted settlement of 10 cm should be interpreted as a median estimate with a 95% interval of about 2.3 to 44 cm.

This dispersion is aleatory in nature with respect to the input parameters: it is the record-to-record variability of the ground motion as well as the sensitivity of the liquefaction response to features not explicitly accounted for by the feature vector. By contrast, it is not on its own a representation of the epistemic uncertainty in the model itself, which is visible separately in the fold-to-fold spread of R^2 in Table 6 and which is due to the finite number of configurations sampled. The distinction is important in application, because the aleatory component cannot be reduced by sampling more configurations, whereas the epistemic component can. The original submission mentioned Bayesian neural networks in the future work section but did not say why they had not been used; the reason is that a calibrated lognormal dispersion obtained from grouped out-of-fold residuals is both sufficient for reliability-based design and much more robust to model misspecification than a variational posterior over network weights. Full Bayesian treatment is a worthwhile extension, mainly for propagating epistemic uncertainty in the sparsely sampled corners of the parameter space.

4.7 Effectiveness of Dense Granular Columns

With the companion mitigated dataset, the value of DGCs can be assessed directly on matched pairs. Over the 1270 (configuration, ground motion) pairs simulated both with and without DGCs, the median of the ratio of mitigated to unmitigated settlement is 0.74, i.e. a median reduction of 26%, and settlement is reduced in 87% of pairs. The distribution is wide: the 10th and 90th percentiles of the ratio are 0.30 and 1.03 respectively Figure 6 so in a minority of cases the mitigation is close to

ineffective. A model trained directly on the log settlement ratio achieves R^2 = 0.084 under grouped cross-validation with a residual dispersion of 0.51, appreciably weaker than the settlement model itself. This is an informative negative result and is reported as such: the incremental benefit of mitigation at an unseen site is substantially harder to predict than the settlement of that site, because it depends on second-order interactions between the column geometry and the specific layering. Practically, this argues against relying on a single deterministic reduction factor for DGC design at interbedded sites. A separate model trained on the mitigated cases predicts mitigated settlement with R^2 = 0.295, somewhat below the unmitigated model, reflecting the additional degrees of freedom introduced by the mitigation geometry over a smaller number of configurations.

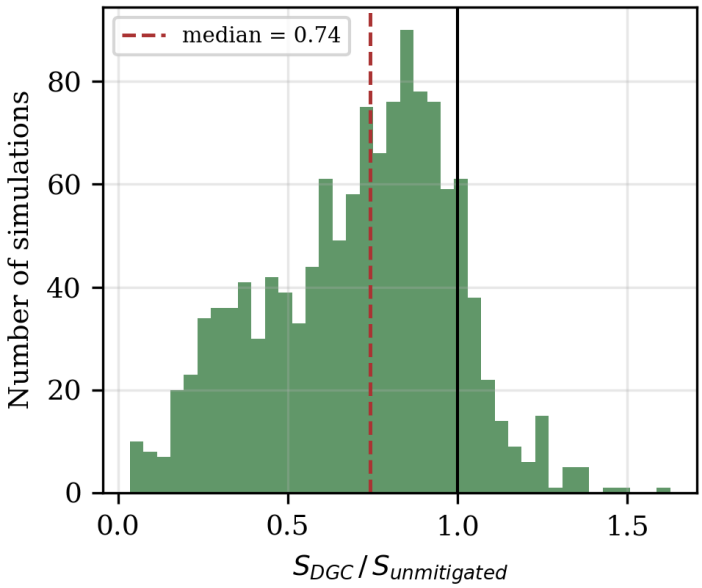


Fig. 6. Distribution of the ratio of mitigated to unmitigated permanent settlement over matched simulation pairs. Values below unity indicate a beneficial effect of the dense granular columns

4.8 Applicability Domain, Benchmarking and Limitations

The original submission included a claim of superior site adaptability that was overreaching given the relatively limited range of geological cases for model validation, and has been removed. The applicability domain is now clearly stated and limited by the ranges in Table 3. The models apply to shallow rigid mat or strip foundations of width 6.0-14.5 m, embedment depth 0.5-3.0 m, supporting one- to eight-storey reinforced concrete, steel or wood frame buildings, on saturated interbedded deposits of loose to medium-dense sand with silt and clay interlayers, over a dense base and elastic rock half-space, on level to gently sloping ground (surface gradient 0° to 7.7°), subjected to moderate to strong shaking (M 5.85-7.90, R 1.5-152 km, outcrop rock PGA 0.058-1.153 g).

They do not apply to deep foundations, whose settlement mechanism is dominated by shaft and base response rather than by the near-surface liquefiable layers that dominate the SHAP ranking here; expansive, collapsible, organic, cemented or loess soils, none of which are represented in the database and whose constitutive behaviour differs fundamentally from the pressure-dependent multi-yield formulation used to generate it; steeper slopes or free-face geometries subject to large lateral spreading, which lie outside the sampled range of surface gradient; or non-seismic settlement problems such as static consolidation, for which established analytical methods remain appropriate. Because ground slope is the strongest single predictor (as in Section 4.5, extrapolation beyond the sampled gradient range is particularly unsafe.

In benchmarking, since the target in the current application is seismic settlement, not static bearing capacity, Vesic's method is not an appropriate benchmark and the comparison has been retracted. The correct benchmark is the FE analysis itself, which the model is intended to emulate: the model reproduces the simulated settlement with a dispersion of σ_ln = 0.76 in a few milliseconds, compared with several hours of computation per case for the underlying nonlinear coupled analysis. Put beside the published ML settlement studies in Table 1, the accuracy reported here is lower than most of them

- but those studies are evaluated on randomly split data, and Section 3.2 shows that the same protocol applied to these data raises R^2 from 0.535 to 0.946. The comparison is therefore not unfavorable once the protocols are made equivalent; it indicates that the published figures are not directly comparable to a site-independent estimate. Three caveats must be added. The database consists of 96 configurations; while many ground motions are used to exercise each one, the effective sample size for training the models to learn site behaviour is the number of configurations, and this is the hard limit on accuracy. The models are trained on simulations, so they inherit any bias of the constitutive model and are only validated against physical reality to the extent that the underlying FE models were validated against centrifuge tests by the authors of the database. And finally, no truly independent field case histories have been used for validation; testing against documented case histories from Christchurch, Adapazari or Niigata would be the most valuable next step and is not attempted here.

5. Conclusions and Suggestions for Future Work

This work is part of the development of a computational tool for the prediction of complex soil-structure interaction phenomena. The key findings from the revised study are summarised below.

- Three-dimensional nonlinear effective-stress simulation of liquefaction-induced foundation settlement can be emulated to useful accuracy using machine learning. Under a validation protocol in which entire soil-foundation-structure configurations are held out, the best model achieves $R^2 = 0.535$ on ln settlement, a spread of $\sigma_ln = 0.76$, and 68% of predictions within a factor of two, at a computational cost of milliseconds per case.
- Gradient-boosted tree ensembles outperform support vector regression, neural networks and linear regression on this problem by a clear margin, and the margin is consistent across every metric considered. The preference asserted in the original submission is confirmed, at a materially lower accuracy level.
- Validation protocol dominates algorithm choice as a determinant of reported accuracy. Randomly splitting records that share a site increases the apparent R^2 from 0.535 to 0.946 for the same model and the same data, and the exceptional accuracy reported in the original submission is reproduced exactly by combining that split with synthetic noise augmentation and a target that is a deterministic function of the inputs.
- The learned sensitivities are mechanically interpretable and consistent with established liquefaction behaviour: energy-based ground-motion measures outrank peak amplitude, aggregate stratigraphic descriptors outrank individual layer properties, and contact pressure acts through a strongly interactive, depth-dependent mechanism.
- Dense granular columns reduced permanent settlement in 87% of matched simulation pairs, with a median reduction of 26%, but the incremental benefit at an unseen site is substantially harder to predict than the settlement itself, which argues against single-valued reduction factors in design.

Future work should extend the model to field case histories, couple it with real-time sensor data for adaptive prediction, and explore physics-informed neural networks that embed the governing consolidation and effective-stress relations directly in the learning problem. Hybrid approaches combining data-driven models with classical analytical tools, and the propagation of epistemic uncertainty through a fully Bayesian treatment, are the most promising routes to improving reliability in the sparsely sampled regions of the parameter space.

6. Data Availability Statement

All data that support the findings of this study are openly available in the DesignSafe-CI Data Depot under project PRJ-5802, “Machine Learning-Based Settlement Models for Shallow-Founded Structures on Interbedded Liquefiable Sites Considering Dense Granular Columns” by C. Bessette, S. Dashti and A. B. Liel, doi: 10.17603/ds2-bjqe-k676 [20] under an Open Data Commons Attribution licence. The database currently contains the OpenSees model files, the ground-motion database, the key numerical results and the machine learning models trained by its authors. The analysis scripts used to produce every table and figure in this paper, including the reproduction of the originally reported accuracy in Table 5, are available from the corresponding author on request.

Acknowledgement

The authors gratefully acknowledge C. Bessette, S. Dashti, A. B. Liel and Y.-W. Hwang of the University of Colorado Boulder, whose published numerical database (PRJ-5802) made this study possible, and the DesignSafe-CI cyberinfrastructure for hosting it.

References

[1] Chaudhury NK, Brotton DM, Merchant W. A numerical method for dynamic analysis of structural frameworks. Int J Mech Sci. 1966;8(3):149-162. https://doi.org/10.1016/0020-7403(66)90033-6
[2] Tekin G, Gencer S. The dynamic soil-foundation-structure interaction problem in the time domain using a discrete element model. Appl Sci. 2024;14(23):10994. https://doi.org/10.3390/app142310994
[3] Lysmer J, Richart FE. Dynamic response of footings to vertical loading. J Soil Mech Found Div. 1966;92(1):65-91. https://doi.org/10.1061/JSFEAQ.0000846
[4] Veletsos AS, Meek JW. Dynamic behaviour of building-foundation systems. Earthq Eng Struct Dyn. 1974;3(2):121-138. https://doi.org/10.1002/eqe.4290030203
[5] Pais A, Kausel E. Approximate formulas for dynamic stiffnesses of rigid foundations. Soil Dyn Earthq Eng. 1988;7(4):213-227. https://doi.org/10.1016/S0267-7261(88)80005-8
[6] Gazetas G. Formulas and charts for impedances of surface and embedded foundations. J Geotech Eng. 1991;117(9):1363-1381. https://doi.org/10.1061/(ASCE)0733-9410(1991)117:9(1363)
[7] Kramer SL, Stewart JP. Geotechnical Earthquake Engineering. Boca Raton: CRC Press; 2024. https://doi.org/10.1201/9781003512011
[8] Zienkiewicz OC, Taylor RL, Zhu J. The Finite Element Method: Its Basis and Fundamentals. 6th ed. Oxford: Elsevier Butterworth-Heinemann; 2005.
[9] Baquero Mosquera JS. Soil-structure interaction in the performance-based seismic design of reinforced-concrete buildings [doctoral dissertation]. Barcelona: Universitat Politècnica de Catalunya; 2023.
[10] Ali T, Eldin MN, Haider W. The effect of soil-structure interaction on the seismic response of structures using machine learning, finite element modeling and ASCE 7-16 methods. Sensors (Basel). 2023;23(4):2047. https://doi.org/10.3390/s23042047
[11] Millán MA, Picardo A, Galindo R. Application of artificial neural networks for predicting the bearing capacity of the tip of a pile embedded in a rock mass. Eng Appl Artif Intell. 2023;124:106568. https://doi.org/10.1016/j.engappai.2023.106568
[12] Abu-Kiefa MA. General regression neural networks for driven piles in cohesionless soils. J Geotech Geoenviron Eng. 1998;124(12):1177-1185. https://doi.org/10.1061/(ASCE)1090-0241(1998)124:12(1177)
[13] Goh ATC. Modeling soil correlations using neural networks. J Comput Civ Eng. 1995;9(4):275-278. https://doi.org/10.1061/(ASCE)0887-3801(1995)9:4(275)
[14] Shahin MA, Jaksa MB, Maier HR. Artificial neural network-based settlement prediction formula for shallow foundations on granular soils. Aust Geomech. 2002;37(4):45-52.
[15] Jimenez JC, Dela Cruz OG. Machine learning for seismic vulnerability assessment: a review. In: Proceedings of the International Conference; 2024. p. 177-187. https://doi.org/10.1007/978-981-97-1514-5_19
[16] Won J, Shin J. Machine learning-based seismic damage prediction method of building structures considering soil-structure interaction. Sustainability. 2021;13(8):4334. https://doi.org/10.3390/su13084334
[17] Sioutas KN, Benardos A. Boosting model interpretability for transparent ML in TBM tunneling. Appl Sci. 2024;14(23):11107. https://doi.org/10.3390/app142311394
[18] Rezazadeh Eidgahee D, Shiri H. Machine learning-enhanced modeling of stress-strain behavior of frozen sandy soil. Geotechnics. 2024;4(4):1228-1245. https://doi.org/10.3390/geotechnics4040062
[19] Katsuumi A, Cong Y, Inazumi S. AI-driven prediction and mapping of soil liquefaction risks for enhancing earthquake resilience in smart cities. Smart Cities. 2024;7(4):1836-1856. https://doi.org/10.3390/smartcities7040071
[20] Kontolati K, Goswami S, Karniadakis GE, et al. Learning nonlinear operators in latent spaces for real-time predictions of complex dynamics in physical systems. Nat Commun. 2024;15:5101. https://doi.org/10.1038/s41467-024-49411-w
[21] Bessette C, Dashti S, Liel AB, Hwang YW. Machine learning-based settlement models for shallow-founded structures on interbedded liquefiable sites considering dense granular columns. DesignSafe-CI; 2025. Report No.: PRJ-5802. https://doi.org/10.1061/JGGEFK.GTENG-13775