Acta Geotechnica
https://doi.org/10.1007/s11440-026-03047-w

RESEARCH PAPER

Metamodeling-based geotechnical characterization of the shear strain dependency of shear modulus of clayey soils from an Italian database

Mojtaba Masoumi Shahrbabak[1,2] · Marco Uzielli[1] · Johann Facciorusso[1] · Claudia Madiai[1]

Received: 1 July 2025 / Accepted: 6 April 2026

Abstract
The quantitative modeling of the dependency of shear modulus on shear strain is a central aspect of the characterization of cyclic and dynamic properties of soils in the nonlinear elastic domain for earthquake geotechnical engineering applications such as site response analysis. This study investigates the dependency of normalized shear modulus on shear strain for clayey soils using the UNIFI database, which contains results from laboratory tests performed on soils from Italian geotechnical sites. A hybrid intelligent metamodeling framework is proposed within the Group Method of Data Handling (GMDH) paradigm. The approach combines the COMBI algorithm for systematic identification of candidate model structures with the shuffled complex evolution (SCE) algorithm for global optimization of model parameters, resulting in a hybrid COMBI-SCE modeling framework. The modeling procedure includes dependency analysis for the identification of influential predictors, generation of candidate model structures through COMBI, and parameter optimization using the SCE algorithm Query ID="Q3" Text="Please check the edit made in the article title. The predictive capability of the resulting models was evaluated using independent training and testing datasets and multiple statistical goodness-of-fit indicators. The results indicate that the most influential predictors of normalized shear modulus degradation in the analyzed dataset include the overconsolidation ratio, plasticity index, initial void ratio, mean confining stress, and shear strain amplitude. The resulting COMBI-SCE metamodel provides a compact analytical representation capable of capturing the nonlinear relationship between shear strain and normalized shear modulus while maintaining physical interpretability. The study demonstrates the potential of hybrid intelligent modeling approaches for developing predictive relationships for dynamic soil behavior and highlights the importance of incorporating site-specific data when developing data-driven geotechnical models.

Keywords Dynamic soil properties · Clays database · Shear modulus · Machine learning · Group method of data handling · Shuffled complex evolution

✉ Marco Uzielli
marco.uzielli@unifi.it

Mojtaba Masoumi Shahrbabak
mojtaba.masoumi.shahrbabak@edu.unige.it

Johann Facciorusso
johann.facciorusso@unifi.it

Claudia Madiai
claudia.madiai@unifi.it

[1] Department of Civil and Environmental Engineering, University of Florence, Florence, Italy

[2] Present Address: Department of Civil, Chemical and Environmental Engineering, University of Genoa, Genoa, Italy

1 Introduction

The characterization of soil behavior under cyclic and dynamic loading (hereinafter, "dynamic soil characterization") is a fundamental issue for many geotechnical earthquake engineering analyses. Among the central topics in cyclic dynamic characterization lies the investigation on the nonlinear shear strain dependency of both the shear modulus and damping ratio at low-to-medium strain amplitude in the nonlinear elastic domain. This investigation is traditionally pursued in two ways: directly, through laboratory experimental activities conducted using dynamic cyclic soil testing methods, and indirectly, through "transformation models" that relate dynamic parameters to easily measured and "routine" parameters such as plasticity index, overconsolidation ratio, and confining stress among others. The term "transformation models" can be used to encompass a variety of approaches, including analytical correlations as well as machine learning-based and data-driven approaches, among others.

Regarding the direct approach, resonant column (RC) and cyclic torsional shear (CTS) tests are commonly used for dynamic soil characterization at low-to-medium strain amplitude [20]. In general, laboratory tests provide precise and reliable measurements of dynamic soil parameters, provided that they are conducted correctly and in accordance with established standards when available. However, this approach entails the availability of undisturbed samples and involves high testing costs. Therefore, in routine geotechnical projects, it is seldom possible to obtain a large data set through the direct approach alone, making it difficult to obtain a comprehensive understanding of dynamic soil behavior across a wide range of soil types and, thus, to capture the inherent variability of the dynamic soil properties at the site.

Alternatively, indirect characterization based on empirical correlations relying primarily on RC test results and, to a lesser extent, on CTS tests provides a more cost-effective and faster means of estimating dynamic soil properties at small-to-medium strain amplitudes from easily and routinely measurable parameters. Over the past decades, extensive research has focused on identifying the key parameters influencing the dynamic behavior of soils, leading to the development of various correlation-based models to predict normalized shear modulus and damping ratio versus shear strain curves, incorporating different key parameters for coarse- and fine-grained soils ([6, 12, 15, 19, 29–32, 35, 39], among others). Among these parameters, the plasticity index is the most significant, strongly affecting shear modulus reduction and damping behavior [36]. Mean confining stress influences stiffness [14], though its effect decreases with increasing plasticity index, while overconsolidation ratio plays a secondary role [36]. Dynamic loading characteristics, such as frequency and the number of cycles, influence damping at small strain but have a limited impact on normalized modulus reduction curves [33]. Several studies have shown that the accuracy of existing models varies, with some providing better predictions for low strain levels and others for broader strain ranges. Moreover, residual analyses suggest that selecting a model requires considering both normalized and maximum shear modulus together, as they are correlated with each other.

While correlation-based models are useful in many practical applications, they typically include a limited number of parameters and may not be able to fully account for the complexity of the physical phenomena governing cyclic dynamic soil behavior. Also, they introduce significant uncertainties due to at least two reasons. First, transformation models are inherently affected by epistemic transformation uncertainty, stemming from the incomplete understanding of the exact relationships between the static and dynamic properties [18]. Recent studies have explored advanced machine learning algorithms for predicting geotechnical parameters such as undrained shear strength, demonstrating improved predictive capability compared with conventional empirical correlations [42]. Second, these models are frequently developed from local datasets comprising data from "source" sites (e.g., [5, 9, 17, 34], among others) which may not fully represent the conditions at the "target" site, and are thus unable to capture site-specific conditions related to soil fabric, stress and strain history, and other local factors. Similar limitations may also affect data-driven and machine learning-based predictive models. Although these approaches can capture complex nonlinear relationships among variables, their predictive capability often depends strongly on the representativeness of the training dataset. Consequently, models trained on datasets from specific geological or geotechnical contexts may exhibit limited generalization capability when applied to new sites, a problem commonly referred to as limited model transferability [43]. This issue highlights the importance of approaches that incorporate site-specific data when developing predictive models for geotechnical applications.

As an alternative to correlation-based estimation, indirect dynamic characterization can be pursued through the metamodeling of the source database. "Metamodeling" is the name commonly used to practice replicating the results of a model using a different model. Metamodeling approaches may provide an efficient approach to the identification of complex relationships between input and output variables in geotechnical datasets. In recent years, these approaches have increasingly relied on artificial intelligence (AI) and machine learning techniques capable of capturing nonlinear dependencies among multiple variables. Numerous applications can be found in geotechnical literature, including Group Handling of Data Method (GMDH) ([2, 10, 22, 23, 25, 26, 28, 38, 40]), artificial neural networks ([24, 27]), genetic programing [11], and fuzzy logic ([1]), among others. Recent reviews highlight the rapid expansion of machine learning applications in geotechnical engineering and their potential for modeling complex soil behavior ([3, 39, 43]). Genetic-programming approaches have been successfully applied to the prediction of soil–water characteristic curves and other soil properties [20], while machine learning-based predictive models have also been developed for problems such as the estimation of retaining wall displacements [21] and the prediction of undrained shear strength using ensemble learning algorithms [42]. Data-driven approaches have also been used to characterize similarities among geotechnical sites and improve the transferability of geotechnical knowledge under site-specific data constraints [13]. Stochastic and geostatistical approaches have been proposed to account explicitly for spatial variability and uncertainty in seismic ground response analyses ([22, 23]). These developments highlight the growing role of data-driven and hybrid modeling approaches in geotechnical engineering, particularly in problems where conventional analytical correlations may not adequately capture the complexity of soil behavior.

This study focuses on the metamodeling of the shear strain dependency of shear modulus. More specifically, metamodeling is aimed at identifying and optimizing predictive methods from a vast database of Italian clay sites comprising a set of "input" and "output" variables compiled through a variety of geotechnical testing methods. Input variables include "routine" geotechnical parameters (compositional properties, overconsolidation ratio, index properties, confining stress, void ratio, etc.), while output variables are dynamic geotechnical parameters measured during cyclic dynamic testing, i.e., shear modulus, damping ratio, and excess pore pressure. The study implements the combinational GMDH (hereinafter, COMBI) for the identification of the best-performing model and exploits the synergy between this predictive method and the shuffled complex evolution (SCE) optimization algorithm. COMBI was used to explore all possible models, excluding interdependent ones, to ensure compatibility with model selection. To identify interdependent models, a thorough dependency analysis was conducted. Once the best-performing model was determined, it was further refined by defining a hybrid algorithm (hereinafter referred to as COMBI-SCE) in two distinct approaches. In the first approach, only the coefficients of the COMBI polynomials were optimized. In the second approach, the COMBI polynomials were connected to the hyperbolic model by Yokota et al. [37], which is frequently used in the indirect modeling of shear strain dependency of shear modulus through the fitting to cyclic dynamic testing data. Outputs of this hybrid predictive-optimization algorithm are presented and assessed using goodness-of-fit statistics. In contrast to conventional GMDH implementations, where model structure identification and parameter estimation are typically performed using regression-based calibration procedures, the proposed COMBI-SCE framework couples combinatorial model structure discovery with global optimization of model parameters, thereby improving the robustness of the resulting predictive relationships. Results are discussed in light of current geotechnical knowledge. All metamodeling activities rely on a large dataset containing the results of conventional and dynamic cyclic laboratory testing on Italian clays.

2 Database and dataset preparation

2.1 Source database

The source database, hereinafter named "UNIFI database," contains the results of conventional "routine" geotechnical laboratory testing (unit weight γ_n, water content w, liquid limit w_L, plastic limit w_P, plasticity index $I_P = w_L - w_P$), void ratio e_0, preconsolidation stress $\sigma_p\prime$, and overconsolidation ratio OCR as well as dynamic and cyclic laboratory tests carried out with RC and CTS equipment at the Geotechnical Laboratory of the University of Florence on 173 undisturbed samples of fine-grained soils of both and high plasticity (Fig. 1) taken from more than 90 sites in central and northern Italy, at which latitude ω, longitude λ, and depth below ground level z were recorded. Tests were performed using equipment and following standardized procedures as specified by national and international standards and guidelines. The measurements thus obtained, although referring to different sites, are comparable in terms of testing apparatus and procedure. All cyclic dynamic tests were performed using the apparatus devised by Stokoe in its standard fixed-free configuration, suitably modified for CTS [13]. In many cases, CTS and RC tests were performed on the same specimen, obtaining CTS and, subsequently, RC measurements for each level of induced strain. Both CTS and RC allowed the measurement of the shear modulus (G), damping ratio (D), and excess pore pressure (Δu) for increasing amplitudes of induced shear deformation (γ). CTS tests also defined the stress–strain cycles used to determine the value of G and D for each step (i.e., level of applied strain). In addition, RC tests provided

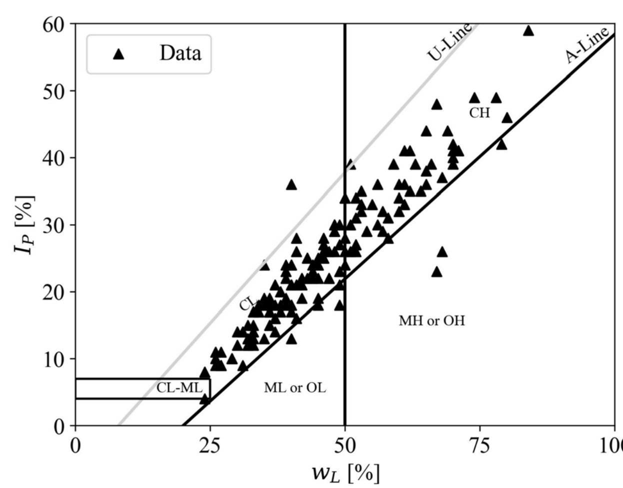

Fig. 1 Casagrande plasticity chart for samples of the UNIFI database

the initial shear modulus ($G_{0,c}$) and damping ratio ($D_{0,c}$), at small deformations (γ_c) at each step of the consolidation process. Further details of the equipment used, the testing apparatus, and the interpretation process are given in [8].

The structure of the UNIFI database is given in Table 1, which indicates the symbol, measurement units, and description of each of the 14 parameters along with their classification as "static" or "dynamic" (S/D) parameters and according to their role as input or output (I/O) variables in the metamodeling process.

Since data-driven approaches are used for predicting outcomes from relationships among different variables and for outlining the relevance of certain parameters among others, it must be noted that all the parameters of the UNIFI database listed in Table 1 were considered in the analyses also including those parameters which are not clearly related to the shear strain dependency of shear modulus (e.g., latitude and longitude).

Consolidation-phase measurements from RC testing were not included as they are not relevant to the modeling of shear strain dependency. The values of G_0 and D_0 pertaining to the last step of consolidation were taken as small-strain parameters for each of the samples.

2.2 Dataset construction and preprocessing

In the preprocessing phase, data from all tests in the UNIFI database were merged into a single dataset (termed "Step-1 dataset") for subsequent metamodeling purposes. Each of the 173 tests constituting the database samples displayed a single value of static parameters and multiple values of the dynamic parameters, in number corresponding to the number of levels of shear strain attained during testing. A new extended dataset (termed "Step-2 dataset") was obtained by expanding the Step-1 dataset. In this expanded dataset, each row refers to single values of shear strain and shear modulus. For example, Fig. 2a and Fig. 2b provides, respectively, the first row of the dataset prior and following its expansion, which results in 21 rows in the Step-2 dataset from the single row in the Step-1 dataset.

The expansion process resulted in an aggregate dataset of 3070 rows in the Step-2 dataset. The indexing numbers of the original samples were saved in the new datasets. Step-2 dataset entries were subsequently scrutinized, and incomplete rows (i.e., with missing values) were excluded. This exclusion process ensured that shear modulus could be estimated from the complete set of static parameters through metamodeling. Following the exclusion step, a further aggregate dataset comprising 1436 rows, hereinafter termed "Step-3 dataset," was obtained. Subsequently, Z-score and theoretical outlier testing were used for detecting samples that do not meet theoretical assumptions. More specifically, a significant inverse correlation between γ and G/G_0 is expected because shear modulus tends to decrease (initially linearly, subsequently nonlinearly) with increasing shear strain. This condition was met for all samples except for row number whose eight rows were excluded from the Step-3 dataset. In the Z-score outlier testing approach, each input variable X was normalized to the Z-score statistic Z using sample mean μ and standard deviation σ:

Table 1 Input and output variables in the UNIFI database

Parameter	Units	Description	S/D	I/O
z	[m]	In situ depth of sample below ground level	S	I
λ	[°]	Longitude	S	I
φ	[°]	Latitude	S	I
γ_n	[kN/m^3]	Unit weight	S	I
w_L	[%]	Liquid limit	S	I
I_P	[%]	Plasticity index	S	I
w	[%]	Water content	S	I
e_0	[–]	Initial void ratio	S	I
w_P	[%]	Plastic limit	S	I
OCR	[–]	Overconsolidation ratio	S	I
$\sigma_0\prime$	[kPa]	Mean confining pressure	S	I
γ	[%]	Shear strain amplitude induced during RC or CTS testing	D	I
G	[kPa]	Shear modulus measured during RC or CTS testing	D	O
D	[%]	Damping ratio measured during RC or CTS testing	D	O
Δu	[kPa]	Excess pore water pressure measured during RC or CTS testing	D	O

(a)

	z	φ	λ	ρ	w	I_P	w_P	w_L	OCR	e_o	$\sigma_o^{'}$	γ
1	49.8	44.74	10.35	20.06	23.00	26.00	25	51	nan	0.424	525	7.26e-05, 1.09e-04, 1.76e-04, ..., 6.91e-02, 1.58e-01, 4.01e-01

(b)

	row	z	φ	λ	ρ	w	I_P	w_P	w_L	OCR	e_o	$\sigma_o^{'}$	γ
1	1	49.8	44.74	10.35	20.06	23	26	25	51	nan	0.424	525	0.000073
2	1	49.8	44.74	10.35	20.06	23	26	25	51	nan	0.424	525	0.000109
3	1	49.8	44.74	10.35	20.06	23	26	25	51	nan	0.424	525	0.000176
...	...	...	...	...	...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...	...	...	...	...
19	1	49.8	44.74	10.35	20.06	23	26	25	51	nan	0.424	525	0.069126
20	1	49.8	44.74	10.35	20.06	23	26	25	51	nan	0.424	525	0.157515
21	1	49.8	44.74	10.35	20.06	23	26	25	51	nan	0.424	525	0.400621

Fig. 2 Example instance of the database expansion procedure: **a** first row of the Step-1 dataset; **b** rows resulting from its expansion

$$Z = \frac{X - \mu}{\sigma} \tag{1}$$

The outlier Z-score test detects outliers based on an integer threshold which can be calibrated in this test for each data row. The statistic Z was initially calculated for all variables. Subsequently, the maximum absolute value of Z among all variables was set as Z-score. In this study, among the 1428 samples in the extended dataset, 1243 samples exhibited $|Z|<3$ and 160 samples exhibited $3 \leq |Z|<6$. Based on a critical subjective assessment of the resulting datasets, the threshold of the Z-score statistic was set subjectively at 6. As a result, for instance, the 16 rows in the extended dataset associated with soil sample 78, for which Z =6.89, were excluded from the dataset. Overall, 24 data rows were removed to obtain the final "reference dataset," which included 1412 rows. The Z-score transformation also provides a standardized representation of the variables, which facilitates numerical stability during the metamodeling procedure.

2.3 Dependency analysis

Dependency analysis constitutes a fundamental preliminary step for effective data exploration, feature selection, model building, and interpretation in multivariate datasets. Such analysis involves the examination and quantification of the causal relationships and dependencies between variables within a dataset to enhance the comprehension of the structure and dynamics of the data. This analysis is beneficial to multiple purposes including dimensionality reduction, feature selection, optimal model configuration and interpretation, and, ultimately, the improvement of predictive performance through increased robustness, interpretability, and reduced overfitting by algorithms.

Dependency analysis requires the quantification of pairwise relationships between input variables. In this study, both the Pearson correlation coefficient (PCC) and Kendall's tau statistic were used to assess linear and monotonic dependencies, respectively. The Pearson correlation coefficient between two variables x and y represents the covariance of the two variables divided by the product of their standard deviations:

$$PCC_j = \frac{\sum_{i=1}^{Nj}\left(\vartheta_{ji} - \underline{\vartheta_{ji}}\right)\left(\widehat{\vartheta_{ji}} - \underline{\widehat{\vartheta_{ji}}}\right)}{\sqrt{\sum_{i=1}^{Nj}\left(\vartheta_{ji} - \underline{\vartheta_{ji}}\right)^2}\sqrt{\sum_{i=1}^{Nj}\left(\widehat{\vartheta_{ji}} - \underline{\widehat{\vartheta_{ji}}}\right)^2}}, \tag{2}$$

Kendall's tau test is a nonparametric test based on Kendall's tau statistic which measures the similarity in the rank ordering of data by two variables:

$$\tau_k = \frac{2}{N_j\left(N_j - 1\right)} \sum_{i=1}^{N_j} \sum_{k>i}^{Nj} \operatorname{sgn}\left(\left(\vartheta_{ji} - \vartheta_{jk}\right)\left(\widehat{\vartheta_{ji}} - \widehat{\vartheta_{jk}}\right)\right), \tag{3}$$

Both statistics range between − 1 and + 1, with values close to ± 1 indicating strong dependency. At this stage, these statistics are used exclusively to identify interdependent input variables. Dependency analysis was conducted using these two statistics to identify interdependent input variables and to exclude inadmissible combinations prior to metamodel construction. Outputs of PCC dependency analysis are provided in Fig. 3. The establishment of a threshold value of PCC for dependency assessment is subjective. According to previous studies (e.g., Blalock, 1963), a PCC value above 0.7 indicates a strong linear dependency between variables, whereas values between 0.6 and 0.7 suggest a weaker relationship and may imply statistical independence. Therefore, input variables with a PCC greater than 0.7 are considered dependent and are not included together in the models. All models containing more than one dependent variable are eliminated. Following such a selection process, the number of shortlisted models was reduced to 216. In Fig. 3, the row and column of PCC values pertaining to γ are not considered in the figure for improved visual representation because such parameter is included in all candidate models due to its high correlation with shear modulus (PCC = 0.66) and its fundamental role in the study, which investigates shear strain dependency. As shown in Fig. 3, the dendrogram divides input variables into five clusters. Variables in each cluster are mutually dependent and cannot be included in possible models simultaneously.

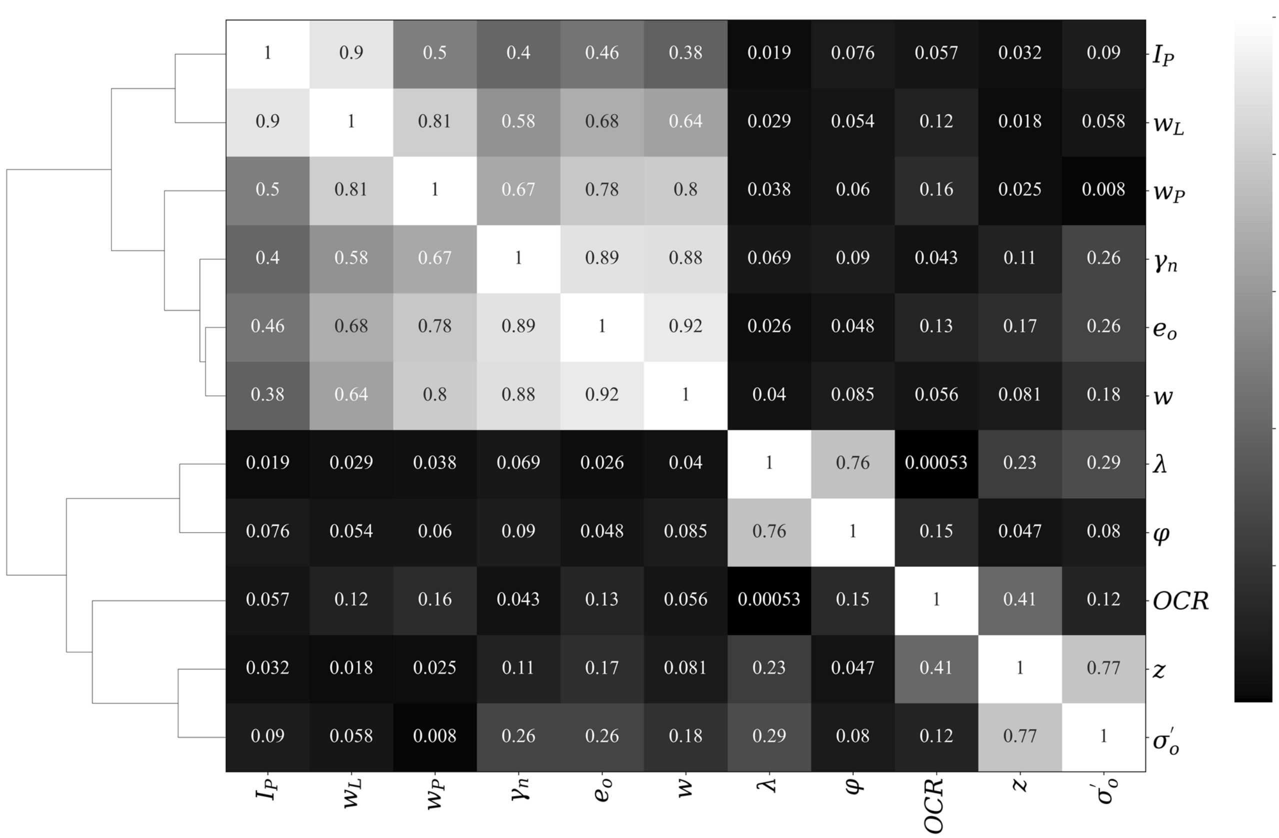

Fig. 3 PCC dependency test with dendrogram

3 Description of metamodeling approaches

Many geotechnical phenomena are characterized by significant complexity. From a modeling perspective, such complexity entails the identification of elaborate models involving many parameters. Conventional methods relying, say, on analytical formulations may not be able to satisfactorily replicate these phenomena and provide reliable engineering approaches. In such cases, metamodeling can provide a useful, effective, and efficient alternative. Metamodeling approaches typically rely on the synergy between predictive and optimization algorithms. Prediction algorithms implement suitable models to predict outputs. Optimization algorithms are aimed at improving the metamodeling performance of predictive algorithms through the enhanced numerical calibration of predictive model coefficients. Genetic algorithms (GA), particle swarm optimization (PSO), harmony search (HS), and SCE are examples of optimization algorithms. This study relies on the novel hybrid COMBI-SCE algorithm which exploits the synergy between the combinatorial group method of data handling (COMBI) prediction algorithm and the SCE optimization algorithm for the metamodeling of the shear strain dependency of the shear modulus. The integration of the COMBI algorithm with the SCE optimization procedure allows the simultaneous exploration of alternative model structures and parameter spaces, reducing the likelihood of convergence toward suboptimal local solutions that may occur when conventional regression-based calibration methods are used. The GMDH and the SCE are described individually in the following sections.

3.1 Combinatorial GMDH (COMBI)

GMDH is a well-known self-organizing prediction algorithm which is typically used in complex system identification and modeling that was initially proposed by [16]. GMDH has shown significant potential in addressing complex problems in geotechnical engineering. Reference [2] employed GMDH to perform predictions of soil compaction. References [22] and [23] developed an ensemble GMDH model to predict soil liquefaction potential using CPT-based data, showing improved performance over traditional methods. Reference [38] utilized GMDH neural networks to predict the pressure meter modulus (EM) from standard penetration test (SPT) data, demonstrating a strong correlation between predicted and measured values. Reference [25] created a GMDH predictive model for estimating rock material strength using non-destructive tests, outperforming other predictive models. Reference [28] modeled the mechanical behavior of cement-treated sand, highlighting the capability of GMDH to accurately predict stress–strain and pore pressure–strain behaviors.

COMBI is a specialized form of GMDH that can systematically explore different combinations of input features to identify the most relevant ones. COMBI provides a useful means in the selection of a subset of features that contribute most significantly to the model's accuracy, thereby improving its performance and interpretability. Moreover, COMBI effectively reduces the complexity of GMDH polynomials by evaluating combinations of polynomial terms and eliminating redundant or less significant terms.

COMBI uses the activation function built from the n-dimensional Kolmogorov–Gabor polynomial

$$f(x) = a_0 + \sum_{i_1=1}^{n} a_{i_1} x_{i_1} + \ldots + \sum_{i_1=1}^{n} \sum_{i_2=i_1}^{n} \ldots \sum_{i_p=i_{p-1}}^{n} a_{i_1 i_2 i_3 \ldots i_p} x_{i_1} x_{i_2} x_{i_3} \ldots x_{ip} \tag{4}$$

where a_i is the i-th polynomial coefficient, n is the dimension of the data set, and p is the degree of the i-th polynomial term. The primary advantage of COMBI is its capability to decompose a complex multivariate problem into several manageable subproblems, developed from subsets of Eq. 2 and referred to as partial models. Figure 4 illustrates the variable selection process in partial models for an n-dimensional problem, i.e., with n partial models. A bit value of "1" indicates that the variable is included in the subproblem, while a bit value of "0" indicates its exclusion.

In principle, given the binary character of the selection process, the number of possible partial models is 2^n. However, the "zero model," for which the binary attribute defining the inclusion of the variable is identically null for $x_{1\ldots}x_n$, is excluded and only the remaining $2^n - 1$ partial models are considered. In this study, three scenarios were defined to predict shear modulus, corresponding to polynomial activation functions of degree 2, 3, and 4, respectively. Since the UNIFI dataset comprises n = 12 input variables, a total of 6144 partial models (for each scenario, only γ included 2048 partial models) were generated. The possible partial models were calculated after carrying out dependency analysis. Models with significantly dependent variables were discarded on the basis of the outputs of such analysis.

x_1	x_2		x_{n-1}	x_n
0 or 1	0 or 1	. . . .	0 or 1	0 or 1

Fig. 4 Binary representation of variable selection in COMBI

3.2 Shuffled complex evolution

The SCE algorithm is an optimization algorithm which relies on a novel concept of complex shuffling [7]. This algorithm is based on a combination of random and deterministic approach clustering, systematic complex evolution of points, and competitive evolution. SCE has been applied for the optimization of geotechnical problems such as prediction of probability of liquefaction (Majeed et al., 2020) and optimal clay core dimension of earth fill dams (Mohammadi et al., 2013). SCE displays many advantages with respect to other optimization algorithms; namely: (1) SCE can solve complex problem with several major region of attraction and numerous local minima; (2) SCE has been used to solve problems with a non-continuous objective function and its derivatives; (3) SCE objective functions and parameters may exhibit varying degrees of sensitivity and great deal of interaction and compensation; and (4) in SCE, the response surface near the true solution is often not convex. The SCE algorithm was therefore adopted to provide a global optimization strategy for the calibration of COMBI polynomial coefficients. The steps of SCE may be described as follows:

Step SCE-1: Initialize SCE parameters such as number of complexes (p), number of points in each complex (m), number of variables (n), and sample size ($s = p \cdot m$) lower and upper bounds (LB andUB) of input variables were determined.

Step SCE-2: Generate an initial sample with s random variables.

Step SCE-3: Sort the generated sample in increasing order of function values and save it in a function values array.

Step SCE-4: Partition the array D into complexes A^1 to A^p according to Eq. (5).

Complexes array was constructed based on a rank-based interleaving scheme. For example, for division a 10-member population into 2 complexes, first-, third-, fifth-, seventh-, and ninth-rank population built first complex. Equation 5 provides complex divisions for s-member population.

$$A^k = \{x_j^k, f_j^k | x_j^k = x_{k+p(j-1)}, f_j^k = f_{k+p(j-1)} . j = 1, 2, ..., m\} \tag{5}$$

where $k + p(j-1)$ is the rank of function values in ascending order and j is the position of x^k and f^k in complex A^k

Step SCE-5: When the SCE algorithm reaches this step, the competitive complex evolution (CCE) is activated, and each complex is evaluated according to the following steps:

- Step CCE-1: Initialize CCE parameters such as number of parents (q) and evolution and shuffling termination criteria.
- Step CCE-2: Assign weights according to a triangular probability distribution according to Eq. (6):

$$\mathrm{pr}_l = \frac{2(m + l - 1)}{m(m + 1)} \tag{6}$$

where pr_l determines the probability of selecting the l-th member from a complex.

- Step CCE-3: Select parents randomly based on assigned weights in previous step and stored in a sub-complexwhere u_r and u_c generate new children during reflection and contraction process, respectively. In Eqs. (7) and (8), g is the centroid of all parents except the worst one (the parent with highest function value) and u_q is the worst parent.
- Step CCE-4: Update sub-complex using reflection, contraction, and mutation according to Eq. (7) and Eq. (8)

$$u_r = 2g - u_q \tag{7}$$

$$u_c = \frac{g + uq}{2} \tag{8}$$

where u_r and u_c generate new children during reflection and contraction process, respectively. In Eqs. (7) and (8), g is the centroid of all parents except the worst one (the parent with highest function value) and u_q is the worst parent.

- Step CCE-5: Check if the termination criteria of the evolution process have been met. If the criteria are satisfied, continue to CCE-6; otherwise, return to CCE-4
- Step CCE-6: Check if the termination criteria of the shuffling process have been met. If the criteria are satisfied, continue to SCE-6; otherwise, return to CCE-3

Step SCE-6: Replace the updated complex into the array D and sort it in ascending order.

Repeat Steps SCE-3 to SCE-6 until the convergence criterion is satisfied.

Summarizing: Shuffled complex evolution (SCE) has 2 sub-algorithms of SCE and competitive complex evolution (CCE). SCE provides the division of population in predefined complexes, and CCE improves each complex. In CCE, first, each complex shuffled randomly, and $n + 1$(number of unknown variables) populations were randomly selected (parents) and sorted based on their objective function in ascending order, and the worst population (u_q) was determined and excluded from parents.

Then, the centroid of other parents was calculated (g). Finally, a new child built using crossover and between u_q and g based on Eqs. 7 and 8 If the new child had a worse performance, the mutation would be carried out.

Using these values, during the reflection, contraction, and mutation processes, a new child is generated based on Eq. (7) and Eq. (8) under the following conditions:

- If the fitness function of the new child, generated from the reflection process (Eq. 7), is smaller than that of the worst parent, the new child replaces the worst parent.
- If the new child's fitness function is not better than the worst parent's, a new child is generated using the contraction process (Eq. 8).
- Similarly, if the fitness function of the child from contraction is better than that of the worst parent, the new child replaces the worst parent. Otherwise, a new child is generated through the mutation process (a new child is generated randomly).

The SCE-CCE algorithm is illustrated in Fig. 5. The section in the right column pertains to the competitive complex evolution (CCE) algorithm. In each iteration of the SCE algorithm, after partitioning the initial sample into complexes, the CCE algorithm is applied to each complex. Within this process, certain members of each complex, called parents, are updated through reflection, contraction, and mutation operations. Once all complexes have been updated, they are merged back together, and the stopping criteria are assessed to decide whether the optimization process should continue.

4 Metamodeling of strain dependency of shear modulus

Following the construction, preprocessing, and preliminary screening of the dataset described in Sect. 2, this section presents the metamodeling framework and discusses the corresponding results. The objective is to identify, calibrate, and evaluate predictive models for the shear strain dependency of the shear modulus of clayey soils, based on the admissible input variables identified a priori.

4.1 Metamodeling strategy and analysis scenarios

In this study, metamodeling of the shear strain dependency of shear modulus was conducted using COMBI and its hybridization with the SCE optimization algorithm. The metamodeling strategy relies on polynomial activation functions of increasing degree, combined with systematic feature selection and numerical calibration. Second-, third-, and fourth-degree polynomial activation functions were implemented. The performance of these activation functions was assessed comparatively. The notation i=2,3,4 was used to define the polynomial degree, while the notation i=1,2,3,…,n_{pm} was adopted to refer to the models, where n_{pm} is the number of possible models obtained through dependency analysis. In each analysis scenario, 75% of the database instances were used for training and 25% were used for testing.

4.2 Database partitioning

The partitioning of the dataset into training and testing subsets was performed using a characteristic-matrix-based approach. Previous studies suggest the use of this approach in the presence of interdependent input variables ([3, 21]). For each subdivision state, characteristic matrices of the training and testing datasets were computed and compared with that of the full dataset to ensure consistency in terms of selected statistical properties (minimum, maximum, mean, standard deviation, skewness, and 20%, 30%, 60%, and 90% percentiles). This procedure reduces the possibility of obtaining biased model evaluation. Details of the process are illustrated in Table 2.

4.3 Definition of performance assessment criteria

This section introduces the statistical indicators used to assess the predictive performance of the metamodels on training and testing datasets. Each model was applied to the datasets for each of the output variables, yielding ϑ_{ji}-sized samples of model predictions (i=1, 2…, N_j) and, correspondingly, of model factors Ψ_{ji} by calculating the ratio of the model-predicted value $\widehat{\vartheta_{ji}}$ to the corresponding measured value ϑ_{ji}

$$\Psi ji = \frac{\widehat{\vartheta_{ji}}}{\vartheta_{ji}} \tag{9}$$

The predictive performance of each model of the analysis scenario S_j was assessed through the calculation of seven quantitative goodness-of-fit statistics, each capturing a different aspect of predictive performance. These include: (a) coefficient of determination (Eq. 10); (b) mean of model factor (Eq. 11); (e) standard deviation of model factor (Eq. 12); (f) Pearson correlation coefficient (Eq. 2); and (g) Kendall's tau statistic (Eq. 3). Note that the latter two statistics were introduced in Sect. 2 as they have also been used in dependency analysis.

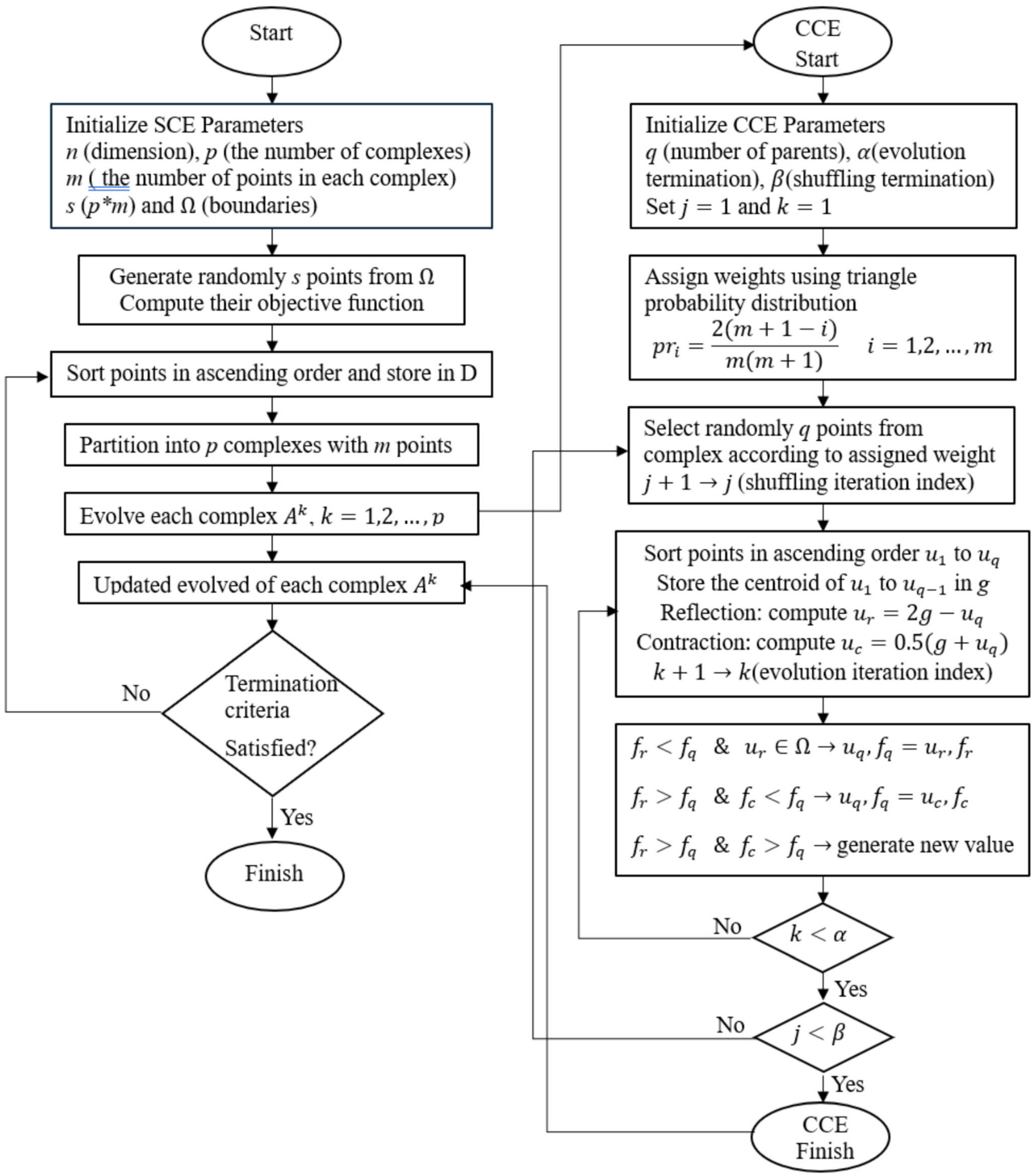

Fig. 5 Flowchart of the shuffled complex evolution (SCE) algorithm [7]

$$R_j^2 = 1 - \frac{\sum_{i=1}^{Nj}\left(\vartheta_{ji} - \widehat{\vartheta_{ji}}\right)^2}{\sum_{i=1}^{Nj}\left(\vartheta_{ji} - \underline{\underline{\vartheta_{ji}}}\right)^2} \quad (10)$$

$$\mu_{\Psi,j} = \frac{1}{N_j}\sum_{i=1}^{N_j}\Psi_{ji} \quad (11)$$

Table 2 Characteristics matrix of input variables

Parameter	Unit	min	max	mean	std	skew	10%	30%	60%	90%
z	[m]	1.80	40.3	11.93	8.69	1.41	3.25	5.75	12.3	23.75
φ	[°]	41.7	45.6	43.80	0.80	0.55	43.16	43.45	43.7	44.83
λ	[°]	10.03	13.69	12.22	1.01	–0.782	10.34	12.11	12.7	13.24
γ_n	[kN/m^3]	13.85	21.87	19.55	1.29	–1.436	18.09	19.19	19.9	20.94
I_P	[%]	9.00	59.00	24.70	10.2	1.656	12.3	18.00	26.0	39.00
w	[%]	14.08	93.06	27.96	10.6	3.477	18.99	23.26	28.0	37.56
w_L	[%]	26.00	164.0	48.53	18.1	3.034	32.00	38.00	49.0	68.00
OCR	[−]	1.00	9.4	2.606	1.91	1.783	1.00	1.000	2.50	5.50
e_0	[%]	0.41	2.248	0.674	0.23	3.890	0.45	0.568	0.68	0.85
$\sigma_0\prime$	[−]	40.0	450	190.3	106	0.873	80.00	120.0	200	350.0
γ	[kPa]	0.00	0.62	0.044	0.09	3.123	0.00	0.001	0.11	0.139

$$\sigma_{\Psi,j} = \sqrt{\frac{1}{N_j}\sum_{i=1}^{N_j}\left(\Psi_{ji} - \mu_{\Psi,j}\right)} \tag{12}$$

In Eq. (10), the sample mean ϑji is calculated as

$$\underline{\vartheta ji} = \frac{1}{N}\sum_{i=1}^{Nj}\vartheta ji \tag{13}$$

4.4 COMBI-based metamodeling outputs

The application of COMBI to the admissible subsets of input variables resulted in a total of 648 candidate models (216 subsets for each of the three polynomial degrees). For each model, performance metrics were averaged over 100 dataset partitioning states to ensure statistical robustness. The optimal state, which depends on both the polynomial degree and the variables included in each model as discussed in Sect. 4.1, was determined according to the similarity of characteristics matrix between training and testing datasets as resulting from the assessment of the proximity of eigenvalues.

4.4.1 Feature selection

In machine learning methods, feature selection is used for dimensionality reduction and determination of unimportant input variables. COMBI provides a powerful approach to feature selection by comparing different combinations of input features. In this approach, features are identified as significant if their inclusion leads to improved performance of the corresponding subsets. In this study, feature selection was implemented by first calculating the best state of training–testing subdivision as described in Sect. 4.1. In cases where the interaction between two significant variables may reduce model performance, only the most important variable is considered. The performance of each subset was evaluated using criteria such as R^2, PCC, Kendell's tau, as well as average, standard deviation, 10th and 90th percentiles, and Hist and Log percentage (related to the percentage of data within 20% accuracy) in the lognormal distribution and histogram for the testing dataset. Except for R^2, PCC, and Kendall's tau, the other criteria were calculated for the ratio of predicted to measured outputs. Figure 4 provides the average of R^2, PCC, and Kendall's tau for the predicted testing datasets of the possible subsets. For better illustration, the subsets were arranged by the number of included variables. (For instance, subsets numbers 197–216 have six variables.)

As shown in Fig. 6, subset 187, which included input variables λ, I_P, e_0, $\sigma_0\prime$, and γ, proves to be one of the worst-performing subsets. As previously mentioned, γ is the most

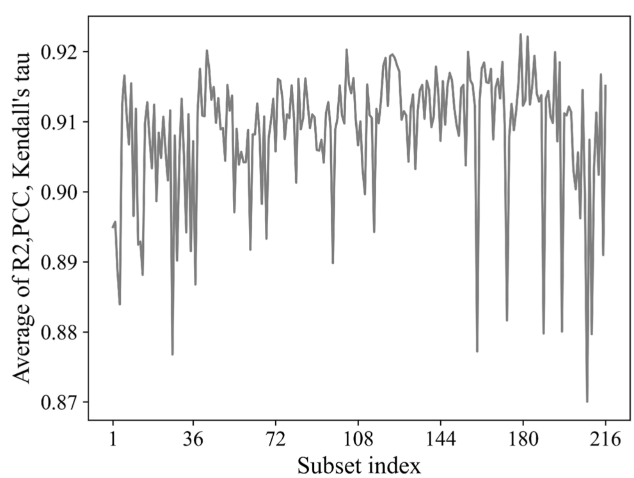

Fig. 6 Average of R^2, PCC, and Kendall's tau for possible subsets of scenario S$_4$

Table 3 Performance of models including λ,I_P, e_0, $\sigma_0\prime$, and γ

Eliminated variable	R^2	PCC	Kendall	mean	std	10%	90%	Hist	Log
λ	0.973	0.986	0.771	1.000	0.132	0.93	1.09	90.4%	87.0%
I_P	0.973	0.986	0.776	0.996	0.178	0.94	1.09	91.6%	74.1%
e_0	0.967	0.984	0.780	0.993	0.207	0.92	1.10	89.8%	66.3%
$\sigma_0\prime$	0.966	0.982	0.771	0.996	0.218	0.94	1.11	91.5%	64.2%
	0.969	0.985	0.7869	0.985	0.225	0.94	1.07	93.2%	62.5%

Table 4 Comparative predictive performance of the four subsets resulting from feature selection

Included variable(s)	R^2	PCC	Kendall	mean	std	10%	90%	Hist	Log
–	0.971	0.986	0.764	1.006	0.128	0.94	1.09	91.1%	88.3%
OCR	0.975	0.987	0.793	1.016	0.168	0.94	1.11	93.6%	77.5%
φ	0.967	0.983	0.777	0.996	0.206	0.94	1.08	90.7%	68.8%
φ, OCR	0.951	0.976	0.807	0.976	0.341	0.93	1.06	91.8%	44.9%

Table 5 Performance after inclusion of other input variable

Included variable(s)	R^2	PCC	Kendall	mean	std	10%	90%	Hist	Log
-	0.975	0.987	0.793	0.999	0.17	0.94	1.08	93.6%	76.5%
γ_n	0.960	0.979	0.762	1.005	0.246	0.94	1.09	92.5%	72.7%
w	0.948	0.979	0.771	0.999	0.228	0.95	1,08	90.6%	69.0%
w_L	0.967	0.983	0.763	1.004	0.163	0.95	1.09	89.7%	63.6%
w_P	0.965	0.975	0.736	1.016	0.178	0.91	1.10	88.7%	72.61%

Table 6 Comparative predictive performance of the COMBI approach by degree of polynomial activation function

Degree j	R^2	PCC	Kendall	mean	std	Pe10	Pe90	Hist	Log	out
2	0.909	0.953	0.707	0.995	0.269	0.92	1.23	81.4%	59.6%	5.9%
3	0.952	0.976	0.754	1.022	0.247	0.94	1.19	86.1%	63.8%	10%
4	0.975	0.989	0.793	1.016	0.168	0.94	1.11	93.2%	77.%	18%

important variable, so the poor performance of subset 187 can be attributed to the interaction between shear strain and one of the remaining input variables. To identify the variable with the worst interaction with shear strain, each variable in subset 187 except for shear strain was removed alternatively, and the performance of the resulting new subsets was evaluated. Table 3 reports the performance of these subsets.

As shown in Table 3, λ displays the worst interaction with γ, leading to the elimination of subsets containing this input variable. As a result, the number of possible subsets was reduced to 192. Additionally, the subset constructing from $I_P, e_0, \sigma_0\prime$, and γ performed well with respect to all the

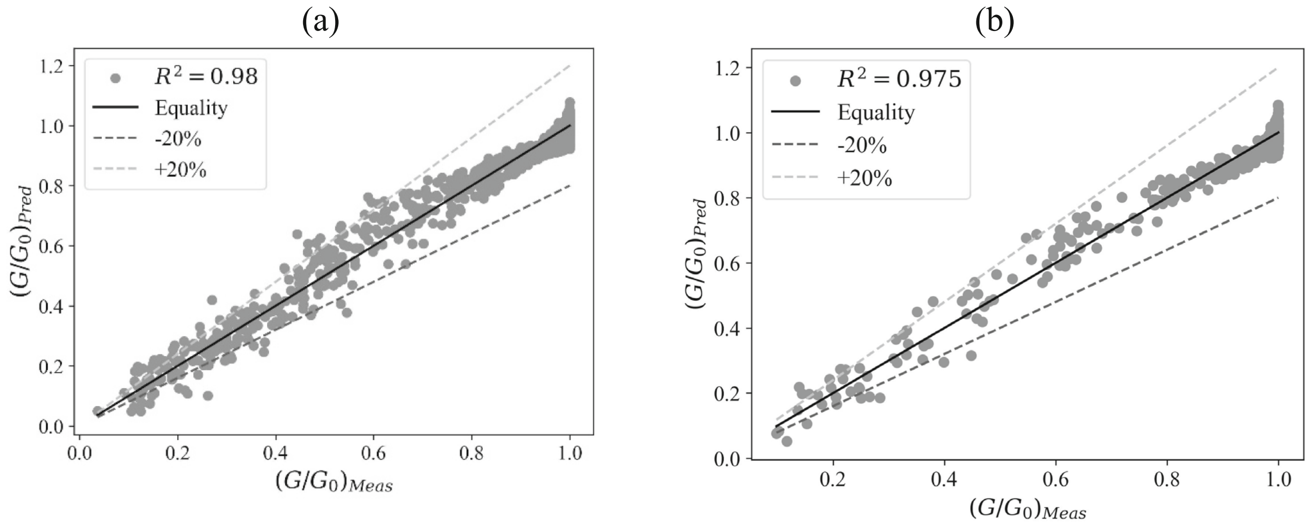

Fig. 7 Predicted vs. measured normalized shear modulus for the identification of the best subset: **a** training dataset; **b** testing dataset

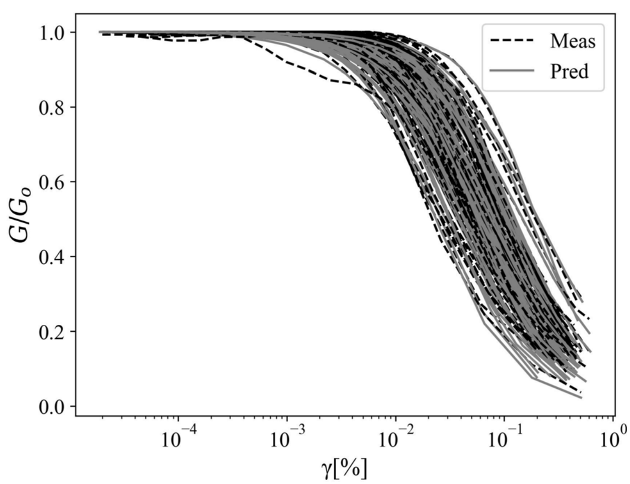

Fig. 8 COMBI-predicted shear strain dependency of modulus reduction for the best-performing model

criteria. Consequently, subsets lacking any of these variables were eliminated, reducing the number of possible subsets further to just four. In addition to the subsets containing the most important variables, the remaining subsets were constructed by incorporating OCR, ω, and both parameters jointly into these subsets. Table 4 provides a comparison between the resulting subsets.

According to Table 4, incorporating these parameters improves predictive performance. Therefore, the final model was constructed using the variables I_P, $\sigma_0\prime$, and γ. In Tables 3 and 4, the effects of including $\lambda, \varphi, OCR, I_P, e_0$, and $\sigma_0\prime$ were investigated. For the analysis of other parameters (γ_n, w, w_L, and w_p), the performance of the best model after their inclusion was calculated. Based on dependency analysis, by inclusion of a new variable, its corresponding dependent variable was excluded. Table 5 provides the changes in the best model performance after inclusion of each variable. According to Table 5, the inclusion of new parameters did not improve the best model.

4.4.2 Comparison of scenarios

As mentioned previously, the reference subset was selected as final models based on the outputs of COMBI feature selection. Models pertaining to second-, third-, and fourth-degree polynomials have 21,56 and 126 coefficients, respectively, consistently with linear regression theory. Table 6 illustrates comparatively the performance of polynomial activation functions of varying degrees based on the values of the goodness-of-fit statistics (columns 2–10) and on the percentage of outlier predictions (column

Table 7 Comparative predictive performance of the hybrid COMBI-SCE approach by degree of polynomial activation function

Degree	R^2	PCC	Kendall	mean	std	10%	90%	Hist	Log	out
2	0.913	0.955	0.711	1.001	0.259	0.93	1.24	81.4%	59.7%	0.0%
3	0.960	0.977	0.769	1.014	0.189	0.94	1.18	87.9%	74.9%	0.0%
4	0.980	0.989	0.806	0.998	0.089	0.94	1.11	94.6%	97.9%	0.0%

Table 8 Example COMBI feature selection for a three-variable problem

model	Binary code	Included variables	Interdependency	Transfer function	R^2	*MSE*
0	000	-	No	$y = a_0$	0.000	–0.008
1	001	γ	No	$y = a_0 + a_3\gamma$	0.685	0.718
2	010	$\sigma_0\prime$	No	$y = a_0 + a_2\sigma_0\prime$	0.007	–0.015
3	011	$\sigma_0\prime, \gamma$	No	$y = a_0 + a_2\sigma_0\prime + a_3\gamma$	0.688	0.718
4	100	z	No	$y = a_0 + a_1 z$	0.009	–0.016
5	101	z, γ	No	$y = a_0 + a_1 z + a_3\gamma$	0.688	0.717
6	110	$z, \sigma_0\prime$	Yes		–	–
7	111	$z, \sigma_0\prime, \gamma$	Yes		–	–

11). This table provides the performance of COMBI-SCE for the three polynomial degrees with respect to a set of selected statistics; namely coefficient of determination (R^2), Pearson correlation coefficient (PCC), Kendall's tau (Kendall), sample mean, sample standard deviation, and 10- and 90-th sample percentiles, and the percentage of predictions falling within the ± 20% error in the sample's relative frequency histogram and best-fit lognormal distribution (columns "Hist" and "Log," respectively) and out-of-range data ("out" column).

The results indicate a systematic improvement in predictive performance with increasing polynomial degree. In particular, fourth-degree polynomial activation functions provided the highest values of R^2, PCC, and Kendall's tau, together with reduced dispersion of the model factor and increased percentages of predictions within the target accuracy range. Figure 7 shows the predicted and output values of normalized shear modulus G/G_0 for the j=4 activation function for the training and testing datasets. Almost all outputs for the training and testing datasets fall within the ± 20% prediction error.

Figure 8 plots COMBI-predicted shear modulus reduction curves obtained with the best-performing model, demonstrating an excellent overall predictive performance. However, COMBI does not inherently allow sequential prediction which would reflect the physical continuity in shear strain dependency of shear modulus. For these reasons, and despite the excellent predictive performance of COMBI, a hybrid approach that combines COMBI with SCE (COMBI-SCE) was implemented as described in the following.

Table 9 Initial population generated within predefined boundary in the defined range [-1,1] for steps SCE-2 and SCE-3

Sample	a_0	a_2	a_3	f	Rank
1	0.204	0.588	− 0.085	1.181	3
2	− 0.260	− 0.006	0.803	3.053	13
3	0.915	0.904	0.165	2.824	12
4	− 0.411	-0.78	0.401	2.741	11
5	0.32	0.702	0.458	2.445	10
6	− 0.396	0.017	− 0.315	0.729	1
7	− 0.417	0.142	− 0.042	1.108	2
8	− 0.937	0.305	0.225	2.350	9
9	− 0.136	− 0.833	0.118	2.042	7
10	− 0.631	− 0.818	− 0.663	1.454	5
11	0.758	0.784	− 0.889	1.464	6
12	0.49	− 0.122	0.926	3.684	14
13	− 0.257	− 0.95	− 0.818	1.320	4
14	0.678	− 0.93	− 0.105	2.723	8

Table 10 First shuffling iteration of the CCE sub-complex

Sample	a_0	a_2	a_3	f	Rank
6	− 0.396	0.017	− 0.315	0.729	1
10	− 0.631	− 0.818	− 0.663	1.454	2
2	− 0.260	− 0.006	0.803	3.053	4
4	− 0.411	− 0.78	0.401	2.741	3

Table 11 Calculation of the evolution part of the CCE sub-complex

Sample	a_0	a_2	a_3	f	Position
g	− 0.479	− 0.572	− 0.192	–	–
u_q	− 0.260	− 0.006	0.803	–	
u_r	− 0.698	− 1.048	− 1.048	–	–
Adjusted u_r	− 0.698	0.247	− 0.717	0.852	2

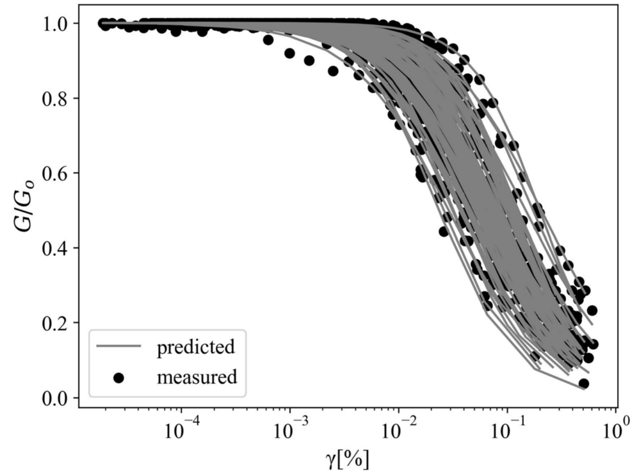

Fig. 9 Measured and COMBI-SCE predicted values of shear strain-dependent shear modulus reduction through the calibration of hyperbolic model coefficients

4.5 Hybrid COMBI-SCE metamodeling outputs

In COMBI, the vector of unknown weights W, which refers to the set coefficients a_i appearing in Eq. (4), was calculated using least-squares estimation (LSE). Reference [4] indicated that the LSE approach is prone to becoming trapped in local minima, performs poorly in identifying nonlinear systems, and is not suitable in the presence of significant data noise or significant percentage of out-of-range predictions, i.e., predictions ranging outside the ± 20% range with respect to measured values. Moreover, Therefore, the application of GMDH alone may not yield fully satisfactory results. Previous studies have been aimed at improving traditional GMDH by exploiting its synergy with other optimization algorithms. Reference [26] paired GMDH with harmony search to improve the suction caisson uplift capacity results. Reference [2] used a genetic algorithm to improve GMDH outcomes in soil compaction parameter prediction. In this study, the COMBI-SCE framework was applied using two complementary modeling strategies. In the first approach, the algorithm was used to calibrate the coefficients of COMBI polynomials and, thereby, construct a direct predictive metamodel for normalized shear modulus as a function of shear strain and soil parameters. In the second approach, the COMBI-SCE framework was used to estimate the parameters of the modified hyperbolic modulus reduction model proposed by Yokota et al. (1981). In this formulation, the metamodel predicts the parameters governing the modulus reduction relationship, and the normalized shear modulus curve is obtained by evaluating the analytical model over the range of shear strain values. While the first approach provides pointwise predictions of $G/G_{\max}$, the second approach yields a continuous modulus reduction curve through the analytical formulation of the Yokota-type model. The use of the SCE algorithm in the present study therefore provides a global optimization strategy that improves the calibration of the COMBI polynomial coefficients and enhances the robustness of the resulting metamodel.

4.5.1 SCE-based calibration of COMBI polynomial coefficients

In the first mode of application of the hybrid algorithm, the SCE algorithm is employed to calibrate the polynomial coefficients corresponding to the best-performing model identified through the feature selection process. The calibration of polynomial coefficients involved setting a combination of several criteria as the objective function of the SCE. This objective function optimized the COMBI polynomials based on randomly generated coefficients. Table 7 provides the performance of COMBI-SCE for the three polynomial degrees with respect to a set of selected statistics.

As illustrated in Table 7, the integration of COMBI with SCE significantly enhances overall performance compared to COMBI metamodeling, as evidenced by the elimination of outlier predictions as well as by increases in R^2, Pearson correlation coefficient (PCC), Log percentage, and Kendall's tau with respect to the COMBI predictions (Table 5). Additionally, the reduction in standard deviation reflects greater consistency in predictions. Minor changes, including a slight decrease histogram percentage and a small increase in both percentiles and average errors, outweighed by the substantial improvements in key performance metrics, demonstrating the efficacy of the hybrid COMBI-SCE approach. For better understanding this hybrid model, a

Table 12 Comparative predictive performance of the hybrid COMBI-SCE approach by degree of polynomial activation function for the metamodeling of the Yokota et al. hyperbolic model

Degree	R^2	PCC	Kendall	mean	std	10%	90%	Hist	Log	out
2	0.981	0.991	0.811	0.994	0.079	0.92	1.07	95.7%	99.1%	0.0%
3	0.987	0.993	0.825	0.993	0.091	0.95	1.03	94.1%	97.3%	0.0%
4	0.995	0.998	0.836	0.984	0.063	0.94	1.02	96.9%	100%	0.0%

three-variable example is presented in Table 8, which illustrates COMBI feature selection for a model with three input variables z, $\sigma_0\prime$, and γ.

According to Table 8, the best model is number 3 with transfer function $y = 0.008 + 0.033\sigma_0\prime - 0.821\gamma$. To improve COMBI results, the coefficients appearing in the transfer function were calibrated using SCE optimization algorithm using the following objective function

$$f(a_0, a_2, a_3) = \frac{\sum_{i=1}^{N} \left(G/G_{0,i} - a_0 - a_2\sigma'_{0,i} - a_3\gamma_i\right)^2}{\sum_{i=1}^{N} \left(G/G_{0,i} - \underline{\underline{G/G_0}}\right)^2} \quad (14)$$

where $\underline{G/G_0}$ is the mean of measured G/G_0. As an illustrative example, SCE is applied for the calibration of a 3-coefficient model. According to Step SCE-1, the input parameters of SCE must be defined. For this example, SCE parameters can be defined as follows: p=2, m=7, $s = p \bullet m$ =14. Additionally, input variables of each population (a_0, a_2, and a_3) were generated within the interval [− 1, 1] of which $LB = -1$ and $UB = 1$ are the lower and upper bounds, respectively, of the range for the construction of the initial population. Following the definition of initial

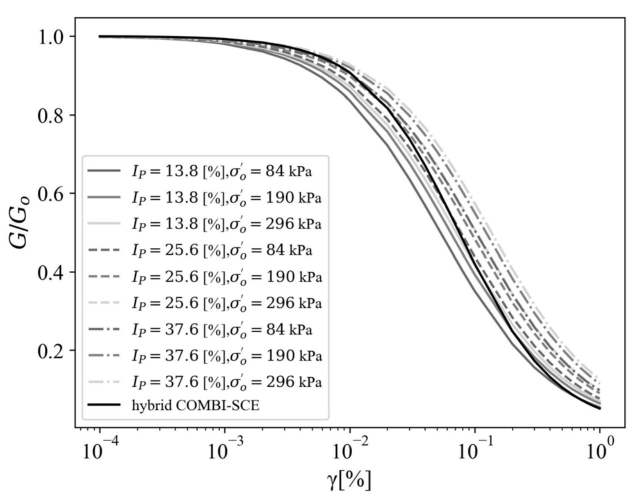

Fig. 10 COMBI-predicted shear strain dependency of shear modulus ratio reduction for the best-performing model compared with measured values and Darendeli (2001) curves for clayey soils

parameters, populations were randomly generated, and their function and ranks based on ascending order were calculated and are provided in Table 9 for steps SCE-2 and SCE-3.

According to Eq. 5 and Table 9 (SCE-4), odd-rank populations (i.e., samples 6, 1, 10, 9, 4, and 2) were selected to build the first complex. In SCE-5, each complex was transferred to the CCE procedure and evaluated as follows:

CCE-1: define parameters such as number of parents (q), evolution (α), and shuffling termination criteria (β). In this example, four samples were selected as parents and the required steps of both shuffling and evolution steps were defined 1.

CCE-2: Assign weights based on Eq. 6. For this case, weights were calculated as 7/28, 6/28, 5/28. 4/28, 3/28, 2/28, and 1/28. Weights represent the chance of selection probability by the relative position in each complex.

CCE-3: Generate a sub-complex, selected complex in SCE-4 shuffled and position of samples changed to 1, 4, 10, 6, 8, 2, and 9, respectively. Based on assigned weights, samples numbers 6, 10, 2, and 4 were selected for building sub-complex. Table 10 provides the sub-complex selected from 7-member complex in the first shuffling iteration of the CCE sub-complex.

CCE-4: According to Table 10, the worst sample (i.e., the sample with the maximum value of f) and the centroid of other samples were calculated as u_q and g, respectively. Table 11 provides the details of calculation in the evolution part of the CCE sub-complex. In this table, only the position was defined only for adjusted u_r.

According to Table 11, the adjusted value was substituted with sample position 2 in the original complex and initial population. Step CCE-4 was repeated α times. Once the evolution termination criterion was met, the updated samples were placed back within their position within the complex, and shuffling termination creation was checked. If this condition was satisfied, the CCE algorithm terminated; if not, the procedure returned to Step CCE-3.

4.5.2 Hybrid optimization of hyperbolic model coefficients

The polynomial activation function given in Eq. (4) has 126 coefficients for the fourth-degree polynomial. Consequently, the calibration process relying on COMBI-SCE as illustrated in the previous section is computationally expensive. Moreover, the use of a COMBI goal function does not consistently ensure the monotonic decrease in G/G_0 with increasing shear strain, even though instances of non-monotonic predicted behavior are very limited. To ensure the latter condition while also reducing the number of coefficients in the goal function, an alternative approach relying on the metamodeling of the well-established modified hyperbolic model by Yokota et al. [41] is adopted. Such model expresses the strain-dependent degradation of shear modulus as a function of two model coefficients α_G and β_G:

$$G/G_0 = \frac{1}{1 + \alpha_G \gamma^{\beta_G}} \tag{15}$$

Figure 9 shows the source dataset and the COMBI-SCE metamodeling outputs of the Yokota et al. model. Visual examination of the figure attests to the excellent performance of the metamodeling algorithm, whereby the few instances of less-than-optimal prediction visible in Fig. 8 are no longer present. Table 12 summarizes the comparative assessment of the predictive performance of polynomial activation functions using the hybrid COMBI-SCE approach for the metamodeling of the Yokota et al. hyperbolic model.

Table 12 shows that the connection of SCE with the Yokota et al. model increased predictive performance for all goodness-of-fit criteria. For example, PCC and R^2 values ranged above 0.99 for all polynomial degrees, indicating a strong correlation and linear relationship between predicted and measured output variable. Additionally, the average of related data is close to 1 and standard deviation is close to 0.01, indicating low scatter in predictive performance.

Comparison between the predictive performance statistics as given in Tables 7 and 12 highlights the benefit of applying the hybrid COMBI-SCE approach to the Yokota et al. model coefficients. Moreover, in the second model of application, the monotonic reduction in shear modulus with increasing shear strain is ensured. Because the Yokota-type formulation explicitly defines the dependency of normalized shear modulus on shear strain, the resulting predictions form a continuous modulus reduction curve that is consistent with the expected physical behavior of dynamic soil properties. For improved comparison between three models for fourth-degree polynomials, the mean squared prediction error was calculated for training and testing datasets:

$$\mathrm{MSE} = \frac{1}{N_j}\sum_{i=1}^{N_j}\left(\vartheta_{ji} - \widehat{\vartheta_{ji}}\right)^2 \tag{16}$$

For the training phase, the MSE values of COMBI, COMBI-SCE, and hybrid COMBI-SCE with Yokota model were calculated 0.114, 0.111, and 0.0335, respectively. For testing phase, the corresponding MSE values were calculated 0.119, 0.117, and 0.0553, respectively. The results show that hybrid COMBI-SCE with Yokota model achieves the least MSE in both training and testing phases. Figure 10 plots comparatively the best-performing COMBI-based output with the empirical curves proposed obtained from the model proposed by Darendeli (2001) for clayey soils which are also included in Fig. 10 for a comparison. In Darendeli's model, the normalized shear modulus G/G_0 is expressed as a function of shear strain through a modified hyperbolic relationship, with reference strain and curve shape mainly controlled by I_P, $\sigma_0\prime$, and *OCR*. For the comparison purposes, Darendeli-type curves were computed by varying I_P and over values corresponding to the mean ± one standard deviation of the UNIFI database (I_P = 13.8, 25.6, and 37.4%; $\sigma_0\prime$= 84, 190, and 296 kPa), while *OCR* was fixed at its mean value (*OCR* = 2.6). The comparison shows that Darendeli's curves are in very good agreement with the COMBI experimental data and generally fall within the envelope of the COMBI predictions, particularly in the small- to intermediate-strain range. Differences observed at larger strain levels reflect the different nature of the two modeling approaches.

5 Concluding remarks

This study investigated the dependency of normalized shear modulus on shear strain for clayey soils using the UNIFI database, which includes laboratory test results obtained from Italian sites. The analysis was conducted using a hybrid intelligent metamodeling framework based on the Group Method of Data Handling (GMDH) paradigm. Within this framework, the COMBI algorithm, a combinatorial variant of GMDH, was used to identify suitable predictive model structures by systematically exploring combinations of input variables and polynomial representations. In order to improve parameter calibration, the COMBI approach was integrated with the shuffled complex evolution (SCE) global optimization algorithm, resulting in the hybrid COMBI-SCE framework adopted in this study. This hybridization, which was applied to both the "direct" and "indirect" approaches for shear modulus reduction characterization, allows the simultaneous identification of model structure and optimization of model parameters and therefore extends conventional GMDH implementations that rely primarily on regression-based calibration procedures. The proposed modeling procedure involved several stages. First, dependency analysis was carried out to identify potentially influential predictors among the available soil parameters. Candidate model structures were then generated using the COMBI algorithm within the GMDH framework. The coefficients of these candidate models were subsequently optimized using the SCE algorithm. The predictive capability of the resulting metamodels was evaluated using independent training and testing datasets together with several goodness-of-fit indicators.

The results indicate that the most influential predictors of normalized shear modulus degradation in the analyzed dataset include the overconsolidation ratio, plasticity index, initial void ratio, mean confining stress, and shear strain amplitude. These findings are consistent with established geotechnical understanding of the factors controlling the nonlinear dynamic behavior of fine-grained soils. The resulting COMBI-SCE metamodel provides a compact analytical representation capable of reproducing the nonlinear relationship between shear strain and normalized shear modulus while preserving physical interpretability. Moreover, because the Yokota-type formulation used in the indirect-mode application of COMBI-SCE explicitly defines the dependency of normalized shear modulus on shear strain, the resulting predictions form a continuous modulus reduction curve consistent with the expected physical behavior of dynamic soil properties.

The results obtained from the analysis of the UNIFI database demonstrate that the proposed hybrid modeling framework can successfully capture the dependency of normalized shear modulus degradation on both stress-state variables and soil index properties. Compared with conventional regression-based transformation models, the COMBI-SCE approach allows a more flexible representation of nonlinear dependencies while preserving an explicit analytical formulation of the predictive relationship. From a broader perspective, the study also highlights the importance of addressing the limited transferability of predictive models in geotechnical engineering. Both empirical correlations and data-driven models may exhibit reduced predictive capability when applied to sites that differ from those represented in the calibration dataset.

The main contributions of this study can be summarized as follows. First, a hybrid intelligent metamodeling framework based on the integration of the COMBI algorithm within the GMDH paradigm and the SCE global optimization algorithm was developed for the prediction of normalized shear modulus degradation in clayey soils. Second, the proposed COMBI-SCE framework enables the systematic identification of predictive model structures together with the global optimization of model parameters, thereby improving the robustness of the resulting predictive relationships compared with conventional regression-based approaches. Third, the application of the proposed methodology to the UNIFI database provides insight into the relative importance of key geotechnical parameters controlling normalized shear modulus degradation. Future work may focus on extending the proposed methodology to larger and more heterogeneous geotechnical databases and on investigating the applicability of the COMBI-SCE framework to other dynamic soil parameters and geotechnical prediction problems.

Acknowledgements This study was partially supported by DPC-ReLUIS Research Project 2022-2024.

Author contributions Marco Uzielli conceptualized the study, coordinated the research framework, participated in the planning and assessment of the metamodeling activities, participated in the drafting of the manuscript, and supervised all phases of the work. Mojtaba Masoumi developed the machine learning models, performed all metamodeling activities, and participated in the drafting of the manuscript. Johann Facciorusso provided the source dataset and contributed to the drafting of the introductory and concluding sections. Claudia Madiai contributed to the drafting of the introductory and concluding sections and provided the funding for the conduction of the research. All authors contributed to manuscript revision; moreover, they read and approved the final version.

Funding Open access funding provided by Università degli Studi di Firenze within the CRUI-CARE Agreement. Funding was provided by Dipartimento della Protezione Civile, Presidenza del Consiglio dei Ministri, DPC-ReLUIS 2022-2024, Claudia Madiai.

Data availability Data sets generated during the current study are available on reasonable request.

Declarations

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the study.

References

1. Adoko AC, Yagiz S (2019) Fuzzy inference system-based for TBM field penetration index estimation in rock mass. Geotech

Geol Eng 37:1533–1553. https://doi.org/10.1007/s10706-018-0706-5
2. Ardakani A, Kordnaeij A (2019) Soil compaction parameters prediction using GMDH-type neural network and genetic algorithm. Eur J Environ Civ Eng 23(4):449–462. https://doi.org/10.1080/19648189.2017.1304269
3. Armaghani DJ, Asteris PG, Hasanipanah M, Tarinejad R, Huynh VV (2022) Application of machine learning techniques in geotechnical engineering: a state-of-the-art review. Eng Geol 303:106657. https://doi.org/10.1016/j.enggeo.2022.106657
4. Bowden GJ, Maier HR, Dandy GC (2002) Optimal division of data for neural network models in water resources applications. Water Resour Res 38(2):2-1–2-11. https://doi.org/10.1029/2001WR000266
5. Chen G, Wang X, Li X (2015) Fundamentals of complex networks: models, structures and dynamics. John Wiley & Sons
6. Ciancimino A, Lanzo G, Alleanza GA, Amoroso S, Bardotti R, Biondi G, Cascone E, Castelli F, Di Giulio A, d'Onofrio A, Foti S, Lentini V, Madiai C, Vessia G (2020) Dynamic characterization of fine-grained soils in Central Italy by laboratory testing. Bull Earthq Eng 18:5503–5531. https://doi.org/10.1007/s10518-019-00611-6
7. Darendeli MB (2001). Development of a new family of normalized modulus reduction and material damping curves. PhD Dissertation, University of Texas at Austin
8. Duan Q, Sorooshian S, Gupta V (1992) Effective and efficient global optimization for conceptual rainfall-runoff models. Water Resour Res 28(4):1015–1031. https://doi.org/10.1029/91WR02985
9. Facciorusso J (2021) An archive of data from resonant column and cyclic torsional shear tests performed on Italian clays. Earthq Spectra 37(1):545–562. https://doi.org/10.1177/8755293020936692
10. Facciorusso J, Madiai C, Vannucchi G (2016) The 2012 Emilia earthquake (Italy): geotechnical characterization and ground response analyses of the paleo-Reno river levees. Soil Dyn Earthq Eng 86:71–88. https://doi.org/10.1016/j.soildyn.2016.04.017
11. Farahani S, Barari A (2023) A simplified procedure for the prediction of liquefaction-induced settlement of offshore wind turbines supported by suction caisson foundation based on effective stress analyses and an ML-based Group Method of Data Handling. Earthq Eng Struct Dyn 52(15):5072–5098. https://doi.org/10.1002/eqe.4000
12. Fatehnia M, Amirinia G (2018) A review of genetic programming and artificial neural network applications in pile foundations. Int J Geoeng 9(1):2. https://doi.org/10.1186/s40703-017-0067-6
13. Han L, Jiang M, Zhang W, Yang L (2024) Similarity characterization of geotechnical engineering sites under the site-specific data scenario. Bull Eng Geol Environ 83(12):504. https://doi.org/10.1007/s10064-024-03990-6
14. Hardin BO, Drnevich VP (1972) Shear modulus and damping in soils: design equations and curves. J Soil Mech Found Div 98(7):667–692. https://doi.org/10.1061/JSFEAQ.0001760
15. Isenhower WM (1979) Torsional simple shear/resonant column properties of San Francisco Bay mud. University of Texas, Austin, TX
16. Ishibashi I (1992) Discussion to "effect of soil plasticity on cyclic response" by M. Vucetic and R. Dobry. J Geotech Eng 118(5):830–832
17. Ishibashi I, Zhang X (1993) Unified dynamic shear moduli and damping ratios of sand and clay. Soils Found 33(1):182–191. https://doi.org/10.3208/sandf1972.33.182
18. Ivakhnenko AG (1968) The group method of data handling; a rival of the method of stochastic approximation. Sov Autom Control 1(3):43–55
19. Iwasaki T, Tatsuoka F, Takagi Y (1978) Shear moduli of sands under cyclic torsional shear loading. Soils Found 18(1):39–56. https://doi.org/10.3208/sandf1972.18.39
20. Johari A, Habibagahi G, Ghahramani A (2006) Prediction of soil-water characteristic curve using genetic programming. J Geotech Geoenviron Eng 132(5):661–665. https://doi.org/10.1061/(ASCE)1090-0241(2006)132:5(661)
21. Johari A, Khaleghi M, Rahmati A (2016) A genetic-based model to predict maximum lateral displacement of retaining wall in granular soil. Sci Iran 23(1):54–65. https://doi.org/10.24200/sci.2016.2097
22. Johari A, Amjadi AH, Heidari A (2021) Application of shear modulus: Stochastic nonlinear ground response analysis via a case study site in Shiraz city. Sci Iran 28(4):2070–2086. https://doi.org/10.24200/sci.2021.55997.4507
23. Johari A, Amjadi AH, Heidari A (2022) Stochastic nonlinear ground response analysis considering existing boreholes locations by the geostatistical method. Bull Earthq Eng 20:6539–6565. https://doi.org/10.1007/s10518-022-01409-8
24. Kishida T (2016) Comparison and correction of modulus reduction models for clays and silts. J Geotech Geoenviron Eng. https://doi.org/10.1061/(ASCE)GT.1943-5606.0001627
25. Kokusho T, Yoshida Y, Esashi Y (1982) Dynamic properties of soft clay for wide strain range. Soils Found 22(4):1–18. https://doi.org/10.3208/sandf1972.22.4_1
26. Kramer SL (2024). Geotechnical Earthquake Engineering (2nd ed.). CRC Press. ISBN 9781032842745.
27. Kuhn M, Johnson K (2013) Applied predictive modeling. Springer New York, New York, NY. https://doi.org/10.1007/978-1-4614-6849-3
28. Kurnaz TF, Kaya Y (2019) SPT-based liquefaction assessment with a novel ensemble model based on GMDH-type neural network. Arab J Geosci 12(15):1–14. https://doi.org/10.1007/S12517-019-4640-5
29. Kurnaz TF, Kaya Y (2019) A novel ensemble model based on GMDH-type neural network for the prediction of CPT-based soil liquefaction. Environ Earth Sci 78:339. https://doi.org/10.1007/s12665-019-8344-7
30. Kurnaz TF, Erden C, Kökçam AH, Dağdeviren U, Demir AS (2023) A hyperparameterized artificial neural network approach for prediction of the factor of safety against liquefaction. Eng Geol 319:107109. https://doi.org/10.1016/j.enggeo.2023.107109
31. Li D, Armaghani DJ, Zhou J, Lai SH, Hasanipanah M (2020) A GMDH predictive model to predict rock material strength using three non-destructive tests. J Nondestruct Eval 39(4):1–14. https://doi.org/10.1007/S10921-020-00725-X
32. Masoumi-Shahrbabak M, Khanjani MJ, Qaderi K (2016) Uplift capacity prediction of suction caisson in clay using a hybrid intelligence method (GMDH-HS). Appl Ocean Res 59:408–416. https://doi.org/10.1016/j.apor.2016.07.005
33. Moayedi H, Hayati S (2018) Applicability of a CPT-based neural network solution in predicting load-settlement responses of bored pile. Int J Geomech 18:06018009. https://doi.org/10.1061/(ASCE)GM.1943-5622.0001125
34. Mola Abasi H, Khajeh A, Jamshidi Chenari R (2021) Use of GMDH-type neural network to model the mechanical behavior of a cement-treated sand. Neural Comput Appl 33(22):15305–15318. https://doi.org/10.1007/S00521-021-06157-6
35. Oztoprak S, Bolton MD (2013) Stiffness of sands through a laboratory test database. Géotechnique 63(1):54–70. https://doi.org/10.1680/geot.10.P.078
36. Rollins KM, Singh M, Roy J (2020) Simplified equations for shear-modulus degradation and damping of gravels. J Geotech Geoenviron Eng 146(9):04020076. https://doi.org/10.1061/(ASCE)GT.1943-5606.0002300

37. Seed HB, Idriss IM (1970). Soil moduli and damping factors for dynamic response analyses. Report No. EERC 70–10, Earthquake Engineering Research Center, Univ. of California, Berkeley, California.
38. Seed HB, Wong RT, Idriss IM, Tokimatsu K (1986) Moduli and damping factors for dynamic analyses of cohesionless soils. J Geotech Eng 112(11):1016–1032. https://doi.org/10.1061/(ASCE)0733-9410(1986)112:11(1016)
39. Shahin MA, Jaksa MB, Maier HR (2023) Artificial intelligence applications in geotechnical engineering: a review of recent developments. Comput Geotech 156:105242. https://doi.org/10.1016/j.compgeo.2023.105242
40. Shibuya S, Mitachi T, Fukuda F, Degoshi T (1995) Strain rate effect on shear modulus and damning of normally consolidated clay. Geotech Test J 18(3):365–375
41. Yokota K, Imai T, Kono M (1981) – Dynamic deformation characteristics of soil determined by laboratory tests. OYO Technical Report, n. 3.
42. Zhang W, Wu C, Zhong H, Li Y, Wang L (2021) Prediction of undrained shear strength using extreme gradient boosting and random forest based on Bayesian optimization. Geosci Front 12(1):469–477. https://doi.org/10.1016/j.gsf.2020.03.007
43. Zhang W, Li DQ, Cao ZJ, Phoon KK (2022) Machine learning in geotechnical engineering: applications and future trends. Géotechnique 72(11):1011–1037. https://doi.org/10.1680/jgeot.21.00027