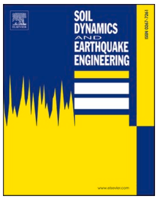
Contents lists available at ScienceDirect

Soil Dynamics and Earthquake Engineering

journal homepage: www.elsevier.com/locate/soildyn



Advancing V_s estimation from CPTu for engineering practice: A data-driven approach

Yuting Zhang [a,*], Héctor Marín-Moreno [b], Susan Gourvenec [a]

[a] *School of Engineering, University of Southampton, United Kingdom*
[b] *School of Ocean and Earth Science, University of Southampton, United Kingdom*

ARTICLE INFO

Keywords:
Shear wave velocity
Seismic piezocone tests (SCPTu)
Standard piezocone tests (CPTu)
Machine learning techniques
Integrated geoscience

ABSTRACT

Shear wave velocity, V_s, is a critical parameter for offshore site characterisation to estimate the small strain shear modulus, which is essential for subsequent geotechnical designs. Direct measurements of V_s are often sparse due to time and resource constraints, while indirect estimations of V_s based on empirical correlations can exhibit significant errors. This study presents the performance of 125 models with various combinations of standard piezocone tests (CPTu) input features (e.g., depth, z; sleeve friction resistance, f_s; corrected cone tip resistance, q_t; and pore pressure at the shoulder of the cone, u_2), CPTu and V_s data pairing methods, and prediction techniques (support vector regression (SVR), random forest regression (RFR), extreme gradient boosting regression (XGBR), deep neural network (DNN) and multiple linear regression (MLR)). To do this, we compile a seismic piezocone test (SCPTu) database from onshore and offshore sites across the globe (Netherlands, Austria, Germany, Nepal, and Taipei) and consider five different methods for pairing CPTu data (resolution of 0.02 m) and V_s data (resolution of 0.5 m and 1 m depending on the dataset). Two cases consider the more conventional downsampling of CPTu data to V_s data. The remaining three methods consider augmented V_s data to the resolution of CPTu measurements, to fully utilise all the CPTu data. Results indicate that data augmentation enhances predictive performance. Incorporating pore pressure as an input feature also improves model performance, particularly in cemented materials such as chalk. In contrast, the derived features have a negligible influence. The recommended model combines a DNN with four directly measured CPTu parameters (z, f_s, q_t, and u_2), and uses an augmentation method that assumes constant V_s values within each V_s interval. This model achieves a mean absolute error (MAE) of 37.3 m/s and a coefficient of determination (R^2) of 0.59.

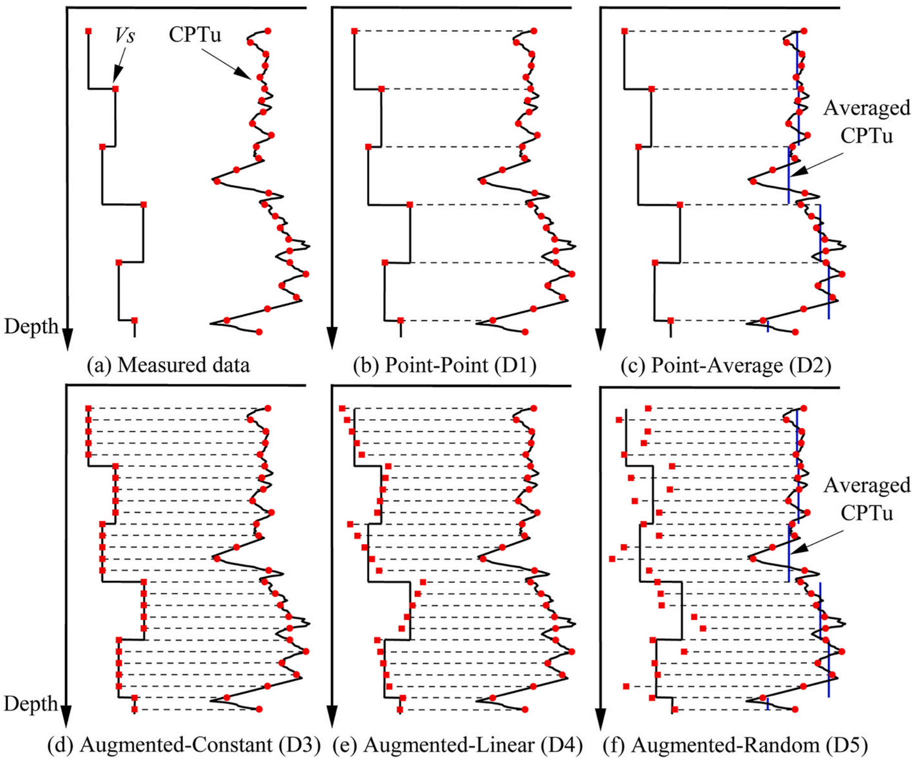
1. Introduction

Shear wave velocity, V_s, is a fundamental property of geomaterials that is adopted in design codes for site characterisation [1–5]. The small strain shear modulus is directly related to V_s based on elasticity theory, and it is a critical parameter utilised in various geotechnical design applications such as site response analysis [6], prediction of foundation settlement on soft clays [7], seismic pile foundation design [8], and design of monopile foundations for offshore wind turbines [9,10].

Direct measurements of V_s are typically obtained through laboratory or in-situ tests during offshore site investigations [11,12]. Laboratory tests such as bender element tests [13] require high-quality, undisturbed samples, which are particularly challenging to obtain for soft clays or granular deposits at offshore sites. Moreover, laboratory tests provide V_s values only at discrete depth locations. In-situ measurements of V_s are conducted using either non-intrusive or intrusive techniques. Non-intrusive methods, such as multichannel analysis of surface waves (MASW), rely on inversion analysis, which often leads to non-unique solutions [14]. MASW methods are most commonly employed onshore, although some applications in offshore are reported [15]. Alternatively, intrusive methods, such as seismic piezocone tests (SCPTu), which integrate V_s measurements with cone penetration test (CPT) and pore pressure (CPTu), are widely employed [16]. While the CPT/CPTu is commonly used across a broad range of projects, SCPTu is typically conducted at a limited number of CPT/CPTu locations (generally at around 10–15 % of the locations) due to the specialised equipment requirements, high costs, and time-consuming nature of drilling [17]. As a result, direct measurements of V_s obtained through laboratory or in-situ testing are generally sparse. Consequently, indirect estimations of V_s based on conventional CPT/CPTu data have become

[*] Corresponding author.
E-mail address: Yuting.Zhang@soton.ac.uk (Y. Zhang).

https://doi.org/10.1016/j.soildyn.2025.109972
Received 9 June 2025; Received in revised form 5 November 2025; Accepted 18 November 2025
Available online 22 November 2025



(a) Measured data (b) Point-Point (D1) (c) Point-Average (D2)

(d) Augmented-Constant (D3) (e) Augmented-Linear (D4) (f) Augmented-Random (D5)

Fig. 1. Concepts of CPTu and V_s data pairing methods. Red points indicate individual measurements, and paired data points are connected using dashed lines for visualisation.

essential for geotechnical designs, allowing for the generation of V_s profiles at unsampled locations without incurring additional testing costs.

Numerous empirical relationships between V_s and CPT/CPTu have been proposed in various functional forms, incorporating different CPT/CPTu parameters [e.g. Refs. [17–21]]. However, these empirical relationships often exhibit substantial prediction errors (i.e., the difference between predicted and measured values) when applied to newly collected datasets [22–24]. These discrepancies can be attributed to two primary factors: i) existing empirical relationships are typically developed based on regional or site-specific data [16,25], making them less generalisable across different soil conditions, and ii) the relationship between V_s and CPTu measurements is inherently complicated due to the differences in soil mechanics, strain regimes, and loading frequencies associated with these two types of measurements.

Machine learning (ML) techniques have demonstrated effectiveness in capturing complicated relationships and have been applied to develop CPT/CPTu-V_s correlations [e.g. Refs. [26–29]]. Notably, some of these studies have omitted pore pressure in V_s prediction, primarily due to the limited availability of pore pressure measurements [28]. Moreover, when constructing databases for training, validation, and testing, CPTu parameters are typically paired only at matching V_s depths or averaged over the V_s sampling intervals. However, the resolution of V_s measurements obtained via SCPTu is relatively low (e.g., 1 m intervals) compared to the resolution of CPT/CPTu tests (e.g., 0.02 m intervals). Therefore, a significant portion of high-resolution CPTu measurements is discarded, leading to a loss of valuable data and information that could potentially enhance ML model performance. Comparison analyses of CPT-V_s correlations from different ML techniques have been conducted, but the V_s used for training was mainly derived from an empirical equation rather than real measurements [e.g. Ref. [26]].

Our objective is to investigate the use of data-driven approaches for deriving CPTu-V_s correlations and provide guidance on best practices for their use in geotechnical engineering practice. To achieve this, we present an analysis of the performance of 125 models with various combinations of CPTu input features (depth, sleeve friction resistance, cone tip resistance, pore pressure, and their derived parameters), CPTu and V_s data pairing methods, and prediction techniques (support vector regression (SVR), random forest regression (RFR), extreme gradient boosting regression (XGBR), deep neural network (DNN), and multiple linear regression (MLR)). These ML techniques are selected for their simplicity and efficiency in addressing engineering problems. Because of the current lag in the widespread adoption of ML techniques in geotechnical engineering, simpler models are more appropriate for promoting understanding and acceptance among practitioners. Some of these techniques, such as RFR and SVR, have been adopted to correlate CPTu-V_s, and have demonstrated superior performance compared to traditional empirical correlations [e.g. Refs. [26,28]]. However, a comprehensive comparison between various ML techniques using real measurement data has not yet been conducted. This study aims to fill that gap by systematically evaluating and comparing the performance of various ML models and offering practical guidance for their application in geotechnical engineering practice. This study also introduces data augmentation strategies aimed at fully utilising high-resolution CPTu data to improve model performance.

2. Database generation

2.1. Data collection

The database used in this study is compiled from five publicly available geotechnical datasets [30–34], collected from offshore (Netherlands and Germany) and onshore sites (Germany, Austria, Taipei, and Nepal). The database includes the curated measurements of CPTu parameters (depth, z; sleeve friction resistance, f_s; corrected cone tip resistance, q_t; and pore pressure at the shoulder of the cone, u_2) and the derived parameters (normalised friction ratio, F_r; normalised cone resistance, Q_t; normalised pore pressure, B_q; and soil behaviour type index, I_c) and V_s. While some original datasets already include the derived parameters, they may have been calculated using different

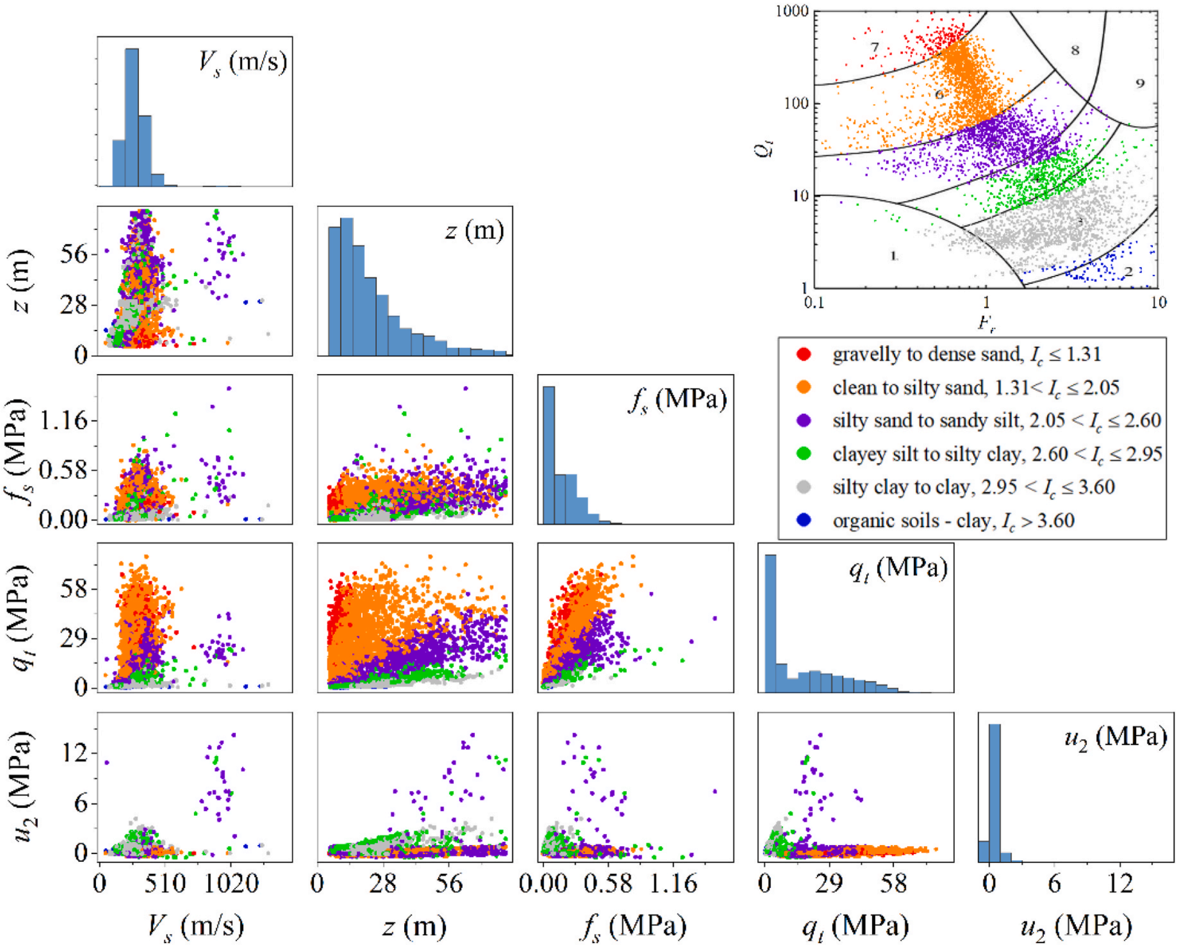


Fig. 2. Pairwise comparison and distribution of parameters and corresponding Robertson chart using the Point-Point pairing method.

methodologies. To ensure consistency across all data sources, these parameters are recalculated using the following equations [18].

$$B_q = \frac{u_2 - u_0}{q_t - \sigma_v} \tag{1}$$

$$Q_t = \frac{q_t - \sigma_v}{\sigma'_v} \tag{2}$$

$$F_r = \frac{f_s}{q_t - \sigma_v} \times 100\% \tag{3}$$

$$I_c = \sqrt{[3.47 - \log(Q_t)]^2 + [\log(F_r) + 1.22]^2} \tag{4}$$

where u_0 is the hydrostatic pore pressure, and σ_v and σ'_v are the total and effective vertical stress, respectively.

A preliminary filtering process is applied to the raw data, consisting of the following steps: i) only tests containing complete CPTu measurements (z, f_s, q_t, and u_2) are retained; and ii) V_s data recorded at shallow depths (e.g., less than 5 m) are excluded, as they may be unreliable due to refraction effects [35]. After filtering, the five CPTu and V_s data pairing methods described in Section 2.2 are applied to construct the databases (D1 to D5) described in Section 2.3.

2.2. CPTu and V_s data pairing methods

Typically, CPTu measurements are recorded with a depth resolution of 0.02 m, while V_s measurements are generally recorded at 1 m intervals. The recorded V_s values represent the average value within each 1 m interval, and the corresponding depth can be assigned either to the upper boundary or the midpoint of the interval. When the depth is assigned to the upper boundary, the visualisation of measurements appears as shown in Fig. 1 (a).

CPTu parameters (z, f_s, q_t, and u_2) are either directly paired at matching V_s depths (Fig. 1 (b), Point-Point method, D1) or averaged over the V_s sampling interval (Fig. 1 (c), Point-Average method, D2). The Point-Point method uses a single CPTu measurement within each V_s sampling interval and, consequently, valuable data and information are omitted. In contrast, the Point-Average method differs from the Point-Point method by averaging all CPTu values within the entire V_s interval rather than selecting one single measurement. This method also results in only one data pair per V_s interval, and the averaged CPTu values may not accurately represent the entire interval. The latter issue may arise when distinct parameter values are measured within the interval, which can impact model performance.

Data augmentation methods that fully utilise all CPTu measurements can be applied to increase the amount of V_s data. Fig. 1(d)–(f) illustrate the concepts of three V_s augmentation methods. The Augmented-Constant method (Fig. 1 (d), D3) assumes that V_s remains constant within each interval and pairs this value with each of the corresponding CPTu measurements in the interval (generally fifty).

The Augmented-Linear method (Fig. 1 (e), D4) assumes a linear trend of V_s within the interval [36]. The slope of the linear trend is determined based on V_s values at two consecutive intervals, while the mean value across the entire interval remains consistent with the measured V_s. Therefore, the augmented V_s generated by the Augmented-Linear method can be expressed as a function of depth, as shown in Eq. (5), given the measured depths and V_s values of two consecutive intervals,

Table 1
Magnitude and proportion of each soil category using five data pairing methods.

Soil type	CPTu and V_s data pairing methods and databases (D)				
	Point-Point D1	Point-Average D2	Augmented-Constant D3	Augmented-Linear D4	Augmented-Random D5
Gravelly to dense sand $I_c \leq 1.31$	173 (3.2 %)	166 (2.8 %)	6758 (2.7 %)	6758 (2.7 %)	6758 (2.7 %)
Clean to silty sand 1.31< $I_c \leq 2.05$	1908 (35.1 %)	2295 (38.3 %)	92,935 (36.8 %)	92,935 (36.8 %)	92,935 (36.8 %)
Silty sand to sandy silt 2.05 < $I_c \leq 2.60$	1205 (22.2 %)	1448 (24.2 %)	59,079 (23.4 %)	59,079 (23.4 %)	59,079 (23.4 %)
Clayey silt to silty clay 2.60 < $I_c \leq 2.95$	573 (10.5 %)	657 (11.0 %)	27,978 (11.1 %)	27,978 (11.1 %)	27,978 (11.1 %)
Silty clay to clay 2.95 < $I_c \leq 3.60$	1449 (26.6 %)	1319 (22.0 %)	58,504 (23.2 %)	58,504 (23.2 %)	58,504 (23.2 %)
Organic soils - clay $I_c > 3.60$	131 (2.4 %)	105 (1.8 %)	7201 (2.9 %)	7201 (2.9 %)	7201 (2.9 %)
All soils	5439 (100 %)	5990 (100 %)	252,455 (100 %)	252,455 (100 %)	252,455 (100 %)

$(z_i, V_{s,i})$ and $(z_{i+1}, V_{s,i+1})$. In Eq. (5), the term inside the brackets represents the linear interpolation between the two measured data points, while the term outside the brackets is an adjustment factor ensuring that the mean of the augmented values within the interval remains consistent with the measured value, $V_{s,i}$.

$$V_s(z) = \frac{2V_{s,i}}{V_{s,i} + V_{s,i+1}} \left[V_{s,i} + \frac{V_{s,i+1} - V_{s,i}}{z_{i+1} - z_i}(z - z_i) \right] \quad z_i \leq z < z_{i+1} \tag{5}$$

Rather than assuming a constant or linear trend for V_s, the Augmented-Random method (Fig. 1 (f), D5) treats V_s as a random variable that is correlated with CPTu measurements. For each depth, the CPTu measurement is compared to the average CPTu value within the entire V_s interval. Given that CPTu measurements consist of four parameters with significantly different scales, these parameters are first normalised to the range [0, 1], and equal weights are assigned to each parameter to compute the mean CPTu value. Following this normalisation, when the CPTu measurement is lower than the average CPTu measurement, a uniformly distributed random number between 0.8 and 1.0 is generated and multiplied by the measured V_s value to obtain the new V_s. Conversely, when it is higher, a random number between 1.0 and 1.2 is used. This process results in a series of randomly adjusted V_s values that are then paired with the corresponding CPTu data. For an interval containing N CPTu measurements, each measurement is represented as $\mathbf{C}_i = \left[z_i,\ f_{s,i},\ q_{t,i},\ u_{2,i}\right]$, containing four measured CPTu parameters. Each parameter is normalised within the interval using $C_i' = \frac{C_i - \min(\mathbf{C})}{\max(\mathbf{C}) - \min(\mathbf{C})}$. An equally weighted CPTu index at depth z_i is then computed as $I_i = \frac{1}{4}\sum_{j=1}^{4} C_{ij}'$. The mean CPTu index across the interval is expressed as $\bar{I} = \frac{1}{N}\sum_{i=1}^{N} I_i$. A random multiplier r_i is generated according to a uniform distribution between 0.8 and 1.0 if $I_i < \bar{I}$, and between 1.0 and 1.2 if $I_i \geq \bar{I}$, to ensure that the generated V_s values remain within physically plausible limits. The corresponding augmented V_s is calculated as $r_i \times V_{s,i}$. Other approaches are available for augmenting V_s between intervals. For example, V_s can be assumed to be depth-dependent, where an exponential stress-dependent gradient is used to consider increasing confining stress with depth [37]. While this method provides stronger engineering justification, it remains uncertain whether the assumed functional form fully captures realistic subsurface conditions. Alternatively, the Gaussian randomisation framework [38,39], with a specified coefficient of variation and vertical correlation length, can be used to model the spatial variability of V_s. However, estimating these statistical parameters from sparse V_s measurements may introduce additional uncertainties. In this study, the linear and random augmentation strategies are selected to explore the influence of basic variability in V_s, rather than to simulate physical soil variability rigorously. The databases generated using the different augmentation methods are publicly available, enabling readers to adopt one of the demonstrated methods or implement more physically based augmentation methods, as appropriate to their objectives.

2.3. Databases using different pairing methods

Fig. 2 illustrates pairwise scatter plots and the distributions of five parameters (z, f_s, q_t, u_2, and V_s) based on the database generated using the Point-Point method. Additionally, the data are mapped onto the Robertson's chart [18], shown in the upper right corner of the figure. The corresponding figures for the other four methods are provided in the supporting material (Figure S1 to Figure S4). From the scatter plots in Fig. 2, it is observed that depth (z) ranges from 5 to 80 m below ground level, covering a representative depth range for geotechnical applications. V_s ranges from 46 to 1310 m/s, with approximately 1 % of the data exceeding 600 m/s. According to the borehole report [40], data with V_s > 600 m/s correspond to chalk (generally considered a cemented material). The reliability of CPTu measurements, particularly in stiff or cemented soils, can vary depending on local geological conditions. In this case, the borehole report [40] associated with the CPTu measurements in chalk indicate no signs of technical issues such as partial penetration or tip underestimation.

A-priori outlier (here defined as datapoints that sit outside the range of most of the data in the database) filtering is not applied to the database to avoid introducing subjectivity into the analysis; therefore, all data are retained for model development. In Fig. 2, a clear correlation is observed between q_t and z, as well as between q_t and f_s, while no clear relationships are observed between V_s and z, f_s, q_t, or u_2. The Robertson chart indicates that the ranges of Q_t and F_r values for silty clay to clay ($2.95 < I_c \leq 3.60$), clayey silt to silty clay ($2.60 < I_c \leq 2.95$), and silty sand to sandy silt ($2.05 < I_c \leq 2.60$) are well represented. However, data for clean to silty sand ($1.31 < I_c \leq 2.05$) are concentrated in a relatively small region, and data for organic soils - clay ($I_c > 3.60$) and gravelly to dense sand ($I_c \leq 1.31$) is limited.

The magnitude and proportion corresponding to different soil types using the five data pairing methods are summarised in Table 1. The database generated using the Point-Average method contains slightly more data points compared to the Point-Point method. This is due to instances where CPTu measurements are not available at the exact depth where V_s is recorded, but measurements exist within the specified V_s interval. The distribution of soil types within the database reveals that gravelly to dense sand ($I_c \leq 1.31$) and organic soils - clay ($I_c > 3.60$) account for around 3 %, while clean to silty sand ($1.31 < I_c \leq 2.05$) represents the largest proportion, comprising approximately 37 % of the total data.

Table 2
Hyperparameters for various ML prediction techniques.

ML technique	Hyperparameters	Range	Optimised value
SVR	Width of the insensitive zone	[10^{-4}, 10]	8.70
	Kernel coefficient for RBF	[10^{-4}, 1]	0.98
	Regularisation parameter	[0, 500]	389
RFR	Tree depth	[2,10]	9
	Number of trees	[5, 100]	55
	Ratio of features considered per split	[0.1, 1]	0.97
	Minimum samples to split an internal node	[2,10]	8
	Minimum samples per leaf node	[1,5]	3
XGBR	Tree Depth	[2,8]	7
	Number of trees	[5, 100]	92
	Subsampling ratio for training	[0.5, 1]	0.86
	Subsampling ratio for features	[0.5, 1]	0.76
	Minimum loss reduction for split	[0, 1]	0.03
	Regularisation parameters	[0, 1]	0.80
DNN	Units in hidden layers	32, 64 and 128	[32, 128, 128] for three layers
	Learning rate	[10^{-6}, 10^{-2}]	0.01

Table 3
Tested cases in this study.

Group	Input feature	Database generation	Prediction technique
A	z, f_s, q_t	D1 to D5 with all datasets	MLR, SVR, RFR, XGBR and DNN
B	z, f_s, q_t, u_2	Same as above	Same as above
C	z, f_s, q_t, Q_t, F_r, I_c	Same as above	Same as above
D	z, f_s, q_t, u_2, Q_t, F_r, B_q, I_c	Same as above	Same as above
E	z, f_s, q_t, u_2, Q_t, F_r, B_q, I_c	D1 to D5 excluding Taipei dataset	Same as above

Table 4
Computational time required for training various prediction models.

		Computational time for training (s)				
		MLR	SVR	RFR	XGBR	DNN
Group D (Eight features)	Point-Point	0.004	10	6	1	156
	Augmented-Constant	0.02	96,180	487	4	8535
	Computational time multiplying factor	5	9618	81	4	55

3. Prediction techniques and performance evaluations

3.1. Prediction techniques

Five prediction techniques, namely MLR, SVR, RFR, XGBR, and DNN, are employed to establish the relationship between CPTu and V_s. A brief description of each technique is provided below. In these techniques, it is assumed that a database contains N pairs of CPTu and V_s measurement, $\{(x_1,y_1),(x_2,y_2),\ldots,(x_N,y_N)\}$. Here, $x_i \in R^n$ are the CPTu parameters used to predict V_s, R^n is the n-dimensional vector space, n is the number of CPTu parameters, and $y_i \in R$ is the measured V_s.

3.1.1. Multiple linear regression (MLR)

MLR is a statistical technique that predicts the value of a dependent variable based on the values of multiple independent variables [41]. However, due to its inherent assumptions, MLR may exhibit lower accuracy when applied to nonlinear multivariate engineering problems [42]. Despite these limitations, MLR remains a widely used approach in geotechnical engineering due to its mathematical simplicity and the ease of interpreting input variables. For example, to predict lateral spread displacement based on various factors (e.g., earthquake magnitude, the thickness of saturated granular layers, and particle size distribution), an MLR was developed using an extensive case history database and has since been widely adopted in engineering practice [43]. The MLR is described as [41]:

$$y_i = \beta_0 + \sum_{j=1}^{n} \beta_j x_{i,j} + \varepsilon_i \tag{6}$$

where β_0 is the intercept, β_1 to β_n are the slope coefficients associated with the n parameters used to predict V_s. ε_i is the error term corresponding to the i-th prediction. The $n+1$ coefficients are determined by minimising $\sum_{i=1}^{N} \varepsilon_i^2$.

3.1.2. Support vector regression (SVR)

SVR is a ML technique that is an extension of support vector machines (SVM) developed for regression problems, aiming to simultaneously minimise training errors while maximising the generalisation ability of the model [44]. Consequently, SVR may exhibit superior generalisation performance compared to artificial neural network models [45]. However, a major drawback of SVR is its computational complexity, which scales cubically with the number of training samples, making it computationally expensive and less practical for large-scale datasets [46]. Recently, SVR has been applied to various geotechnical engineering problems, including the prediction of the overconsolidation ratio of clay using piezocone data [47] and the capacity prediction of stone columns floating in soft clay [48].

SVR employs kernel functions to map input data into a higher-dimensional feature space, where linear regression can be effectively performed [44]:

$$f(x) = w^{\mathrm{T}}\phi(x) + b \tag{7}$$

where $f(x)$ is the linear function, w is the weight vector, T denotes the transpose, $\phi(x)$ is the kernel function, the most widely used radial basis function (RBF) is adopted here [49], and b is the bias.

In SVR, $f(x)$ is fitted to the data while allowing a certain level of tolerance (insensitive zone). δ denotes the width of the insensitive zone, where predictions are considered 'correct'. The main goal of SVR is to find $f(x)$ such that its deviation from the actual output y_i does not exceed δ, while simultaneously ensuring the function remains as flat as possible to minimise model complexity. This objective is formulated as the following optimisation problem [44]:

$$\begin{aligned} &\text{Minimise} \quad \frac{1}{2}w^{\mathrm{T}}w + C\sum_{i=1}^{N}(\xi_i + \xi_i^*) \\ &\text{Subject to} \quad \begin{cases} y_i - w^{\mathrm{T}}\phi(x_i) - b \le \delta + \xi_i \\ w^{\mathrm{T}}\phi(x_i) + b - y_i \le \delta + \xi_i^* \\ \xi_i, \xi_i^* \ge 0 \end{cases} \end{aligned} \tag{8}$$

where C is the penalty/regularisation parameter that controls model complexity, ξ_i and ξ_i^* are slack variables, which penalise prediction errors for training instances that fall outside the tolerance zone [48].

3.1.3. Random forest regression (RFR)

RFR is a tree-based ensemble ML technique that constructs multiple decision trees using independently sampled subsets of the original training dataset (bootstrap aggregation). Additionally, at each node split, only a randomly selected subset of features is considered, introducing further variability and reducing overfitting. The final prediction is obtained by averaging the predictions from all trees in the ensemble

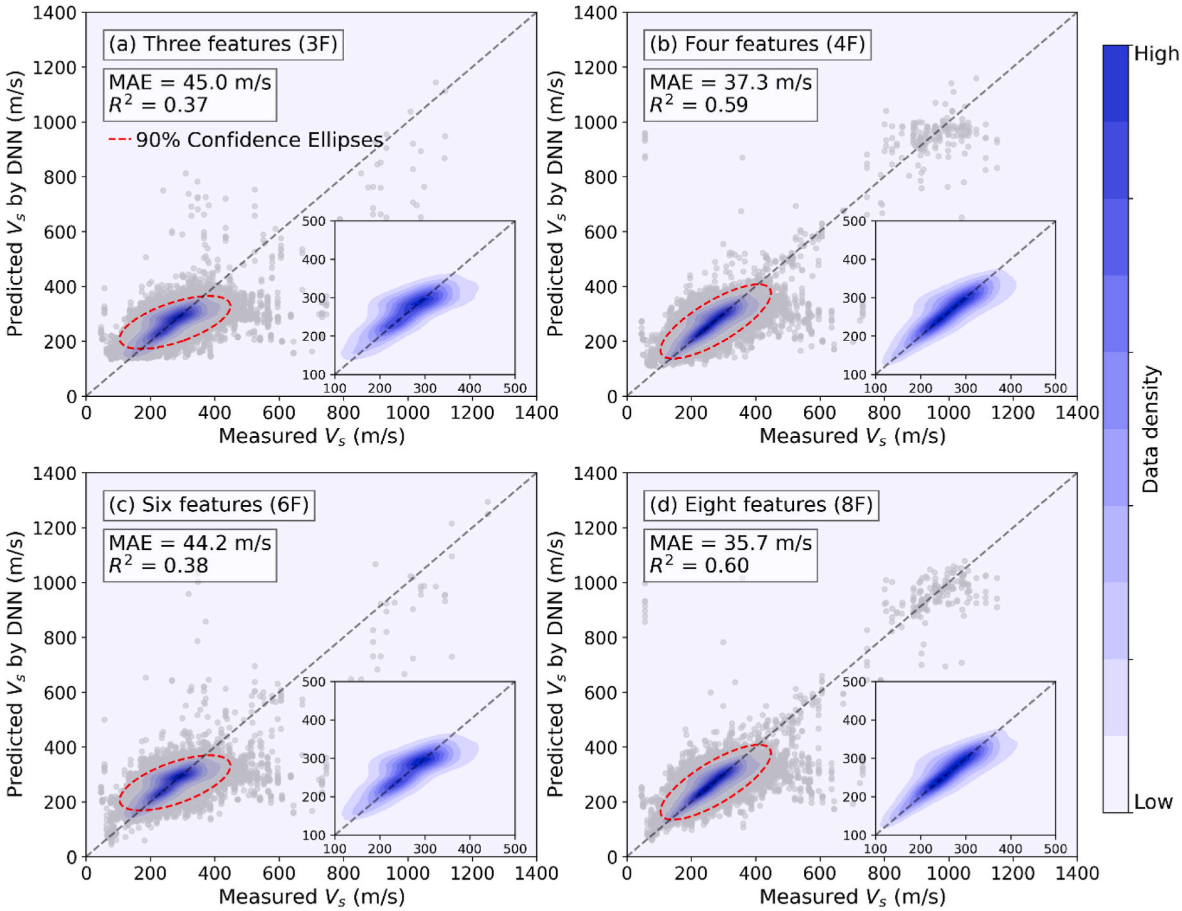
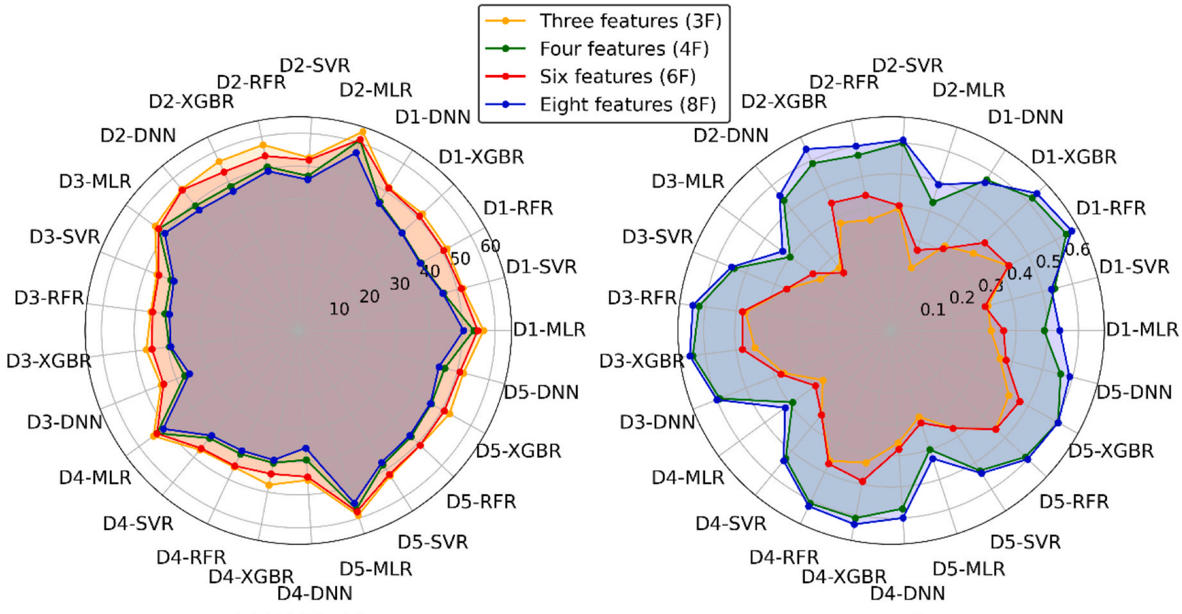


Fig. 3. Performance for different numbers of input features using the Augmented-Constant pairing method and DNN.



(a) MAE (m/s)

(b) R^2

Fig. 4. Comparison of (a) MAE and (b) R^2 for different input features. Note: D1-DNN represents the combination of Point-Point (D1) and DNN.

[29]. Compared to SVR, RFR generally exhibits superior performance in capturing complex, nonlinear relationships and is more computationally efficient for large datasets [50]. Recently, RFR has gained significant popularity in geotechnical engineering applications, such as the prediction of the bearing ratio of soils [50] and the assessment of pile drivability [51].

Mathematically, for an ensemble of K trees, the prediction $\hat{y}_i$ for input x_i is given by:

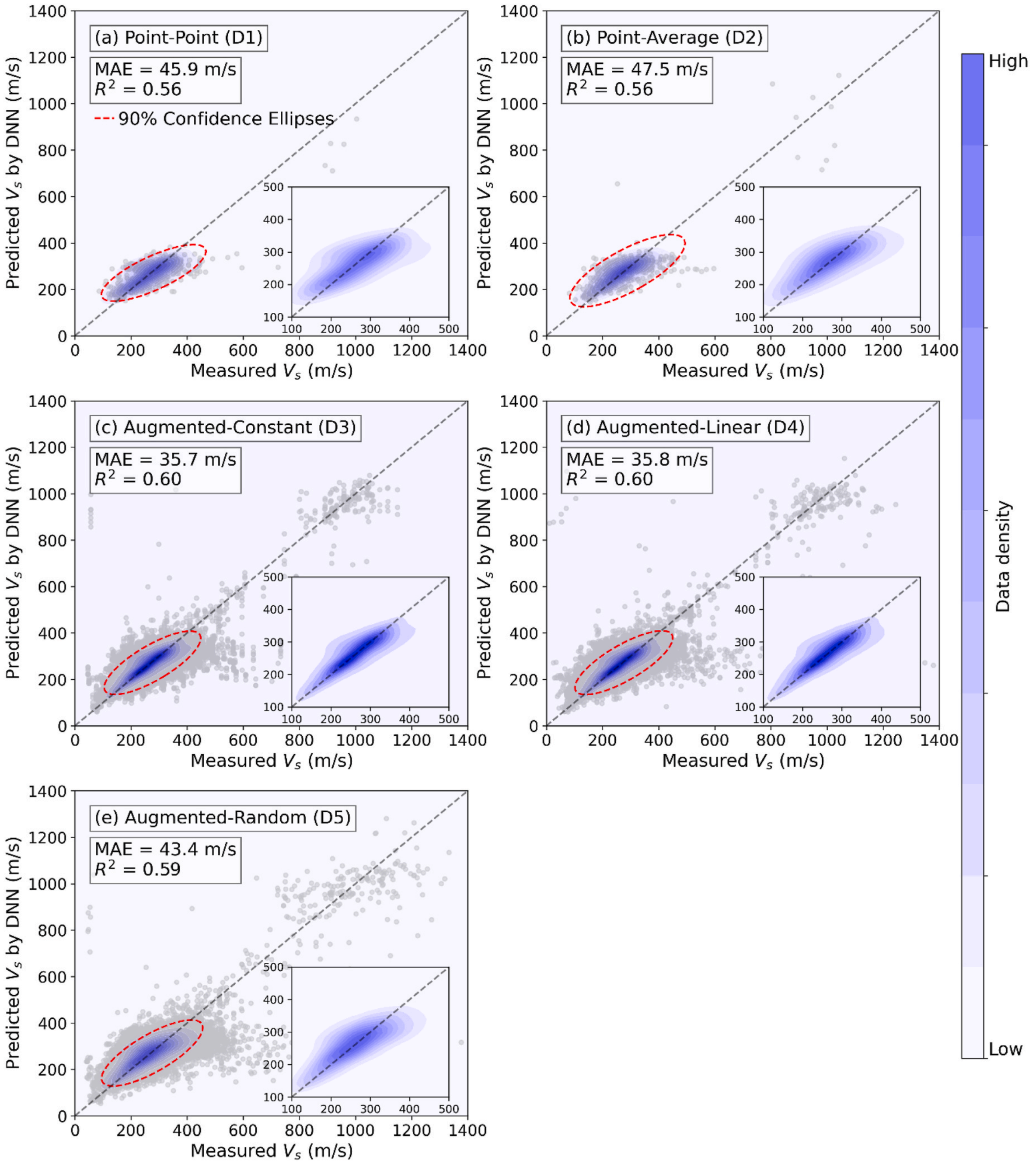


Fig. 5. Predictions with different CPTu and V_s data pairing methods using eight features and DNN.

$$\widehat{y}_i = \frac{1}{K}\sum_{k=1}^{K} f_k(x_i) \tag{9}$$

where $f_k(x_i)$ is the prediction from the k-th decision tree.

3.1.4. Extreme gradient boosting regression (XGBR)

XGBR is also a tree-based ensemble ML technique [52]. Unlike RFR, which employs bootstrap sampling to construct independent decision trees, XGBR utilises boosting, where trees are built sequentially, with each new tree correcting the errors of the previous one. This iterative learning process continuously enhances model performance by refining predictions based on previous outcomes [53]. Due to its efficiency and precision in regression tasks, as well as its ability to handle large datasets, XGBR has been applied to geotechnical applications, such as the stability assessments for braced excavations [54] and the prediction of Newmark sliding displacements [55].

The prediction by XGBR is calculated as:

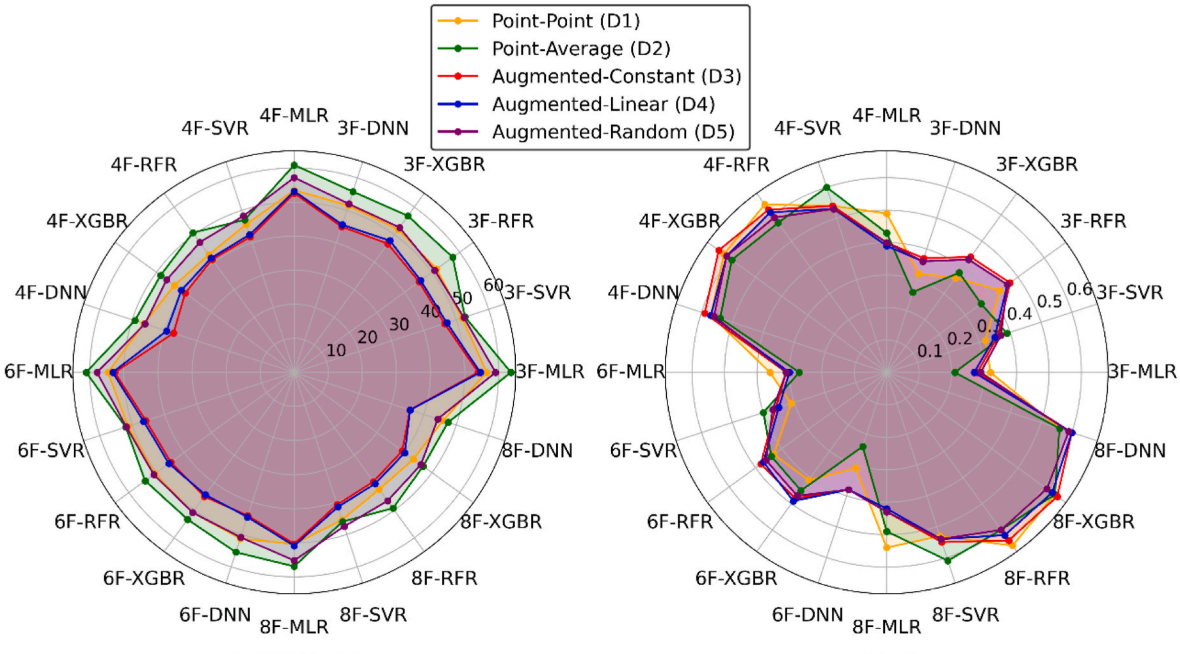
$$\widehat{y}_i = \sum_{m=1}^{M} f_m(x_i) = \widehat{y}_i^{M-1} + f_M(x_i) \tag{10}$$

where M is the number of trees, $\widehat{y}_i^{M-1}$ is the prediction based on the previous tree model, $f_M(\cdot)$ is the newly generated tree model.

The objective function of XGBR is given by:

$$\Gamma = \sum_{i=1}^{M} L(y_i, \widehat{y}_i) + \Omega(f_M) \tag{11}$$

where $L(y_i, \widehat{y}_i)$ is the loss function, and $\Omega(f_k)$ is the regularisation function that penalises model complexity to prevent overfitting.



(a) MAE (m/s) (b) R^2

Fig. 6. Comparison of (a) MAE and (b) R^2 for different CPTu and V_s data pairing methods. Note: 3 F-DNN represents the combination of three features (3 F) and DNN.

3.1.5. *Deep neural network for regression (DNN)*

DNN is a subset of ML techniques that consists of multiple interconnected layers, including i) an Input Layer, which receives raw data, such as CPTu parameters; ii) Hidden Layers, composed of multiple neurons that transform the inputs using weights, biases, and activation functions; and iii) an Output Layer, which generates the final predictions, such as V_s. DNN has emerged as a powerful ML technique in geotechnical applications, demonstrating superior predictive capabilities compared to traditional ML techniques, such as SVR [56]. However, DNN generally requires large datasets for effective training.

During the training process, the network parameters (weights and biases) are optimised by minimising a predefined loss function. Mathematically, a fully connected DNN model with T layers is expressed as:

$$\begin{cases} \text{Input} \quad \text{layer} : h_0 = x \\ \text{Hidden} \quad \text{layer} : h_t = \sigma(w_t h_{t-1} + b_t) \quad \text{for} \quad t = 1, 2, \ldots, T-1 \\ \text{Output} \quad \text{layer} : \widehat{y} = w_T h_{T-1} + b_T \end{cases} \tag{12}$$

where h_t is the output of the t-th layer, $\sigma(\cdot)$ is the activation function, such as the rectified linear unit (ReLU) [57], which enhances the nonlinearity of the network. w_t and b_t are the weights and biases of the t-th layer, respectively.

3.2. *Hyperparameters, loss function, and performance metrics*

The performance of the four ML techniques, SVR, RFR, XGBR and DNN, is highly sensitive to their respective hyperparameters. Therefore, Bayesian optimisation [58] is adopted to efficiently tune these hyperparameters. The predefined initial ranges for the hyperparameters and the optimal hyperparameters for the model using four input features (z, f_s, q_t, and u_2) and the Augmented-Constant pairing method are listed in Table 2. Additionally, the maximum number of optimisation iterations is set to 50, with an early stopping criterion defined as no performance improvement over 5 consecutive iterations.

In this study, DNN models with three, four, and five hidden layers are tested and compared. However, increasing the number of hidden layers does not yield any notable improvements in prediction performance. Therefore, a DNN architecture with three hidden layers is adopted.

The generated databases (D1 to D5) are randomly divided into training, validation and testing datasets with a ratio of 80:10:10. The loss function adopted is the mean absolute error (MAE), defined in Eq. (13), as it is less sensitive to outliers than the mean squared error. Model performance is evaluated using MAE and the coefficient of determination (R^2), defined in Eq. (14). Specifically, MAE quantifies the average magnitude of prediction errors, where a lower MAE indicates superior model performance. In contrast, R^2 measures the proportion of variation in the data that the model can capture, where a higher value of R^2 means better prediction. Eq. (14) shows that the maximum R^2 value is 1, when the model fully explains all the variation in the data, and R^2 can be negative when the performance of the model is worse than the arithmetic mean.

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N} |\widehat{y} - y_i| \tag{13}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N} (\widehat{y} - y_i)^2}{\sum_{i=1}^{N} (y_i - \overline{y})^2} \tag{14}$$

where $\widehat{y}$ and y_i are the values predicted and measured values, respectively. $\overline{y}$ is the arithmetic mean value of measurements, $\overline{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$.

4. Investigated cases

Table 3 summarises the examined cases in this study. Groups A to D, comprising a total of 100 models, explore the influence of various factors on prediction performance, including input features, CPTu and V_s data pairing methods, and prediction techniques. Group E, consisting of an additional 25 models, evaluates the generalisation capability of the models. Group A and Group C exclude u_2 for the prediction, whereas Group B and Group D incorporate it to explore the impact of u_2 on prediction performance. If u_2 is found to have an insignificant effect, there is potential to leverage a large volume of onshore CPT data, where u_2 is often unavailable, to develop the CPTu/CPT-V_s correlation. The derived CPTu parameters, as described in Section 2.1, are considered in Group C and Group D, but not in Group A and Group B, to assess whether

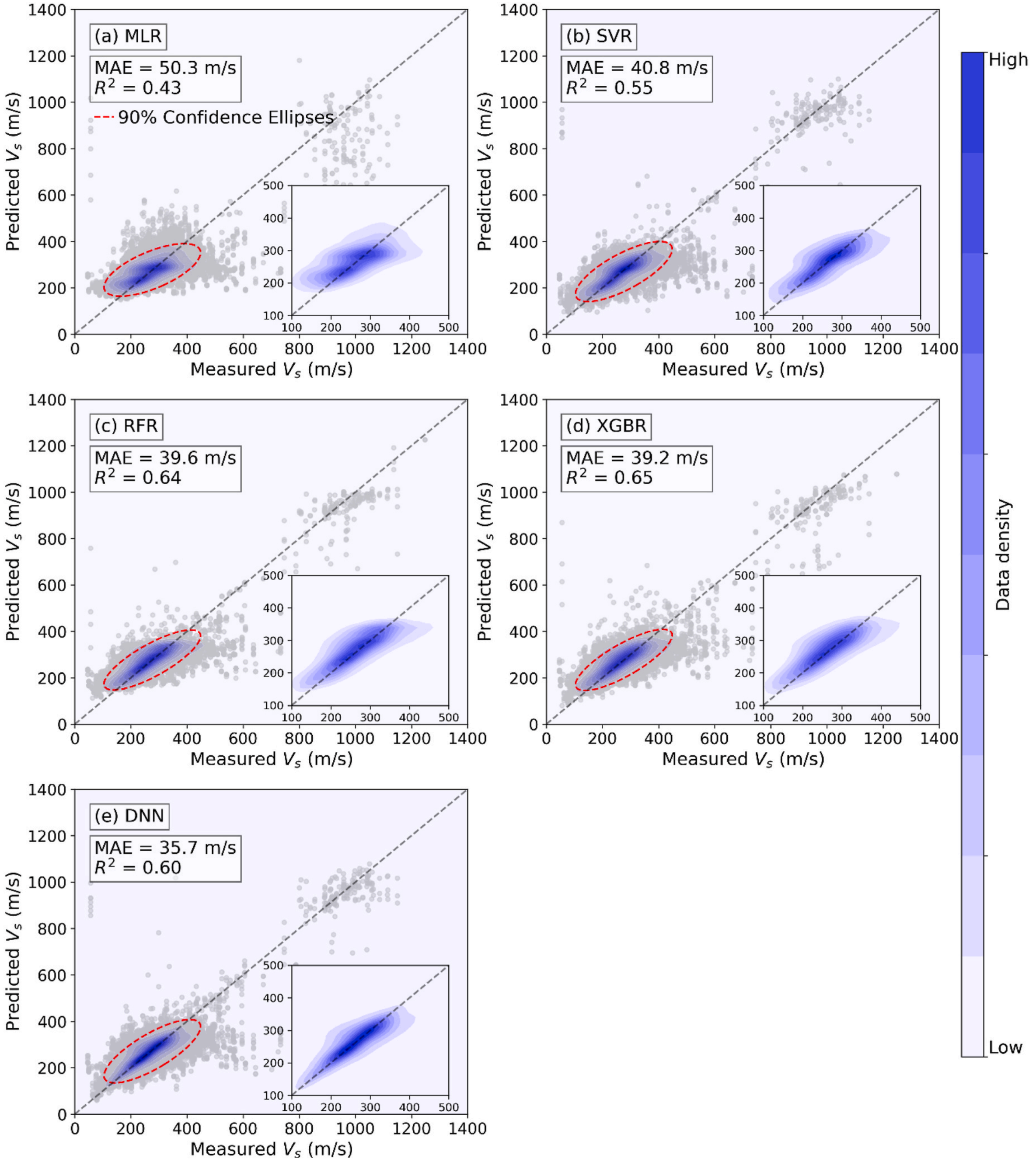
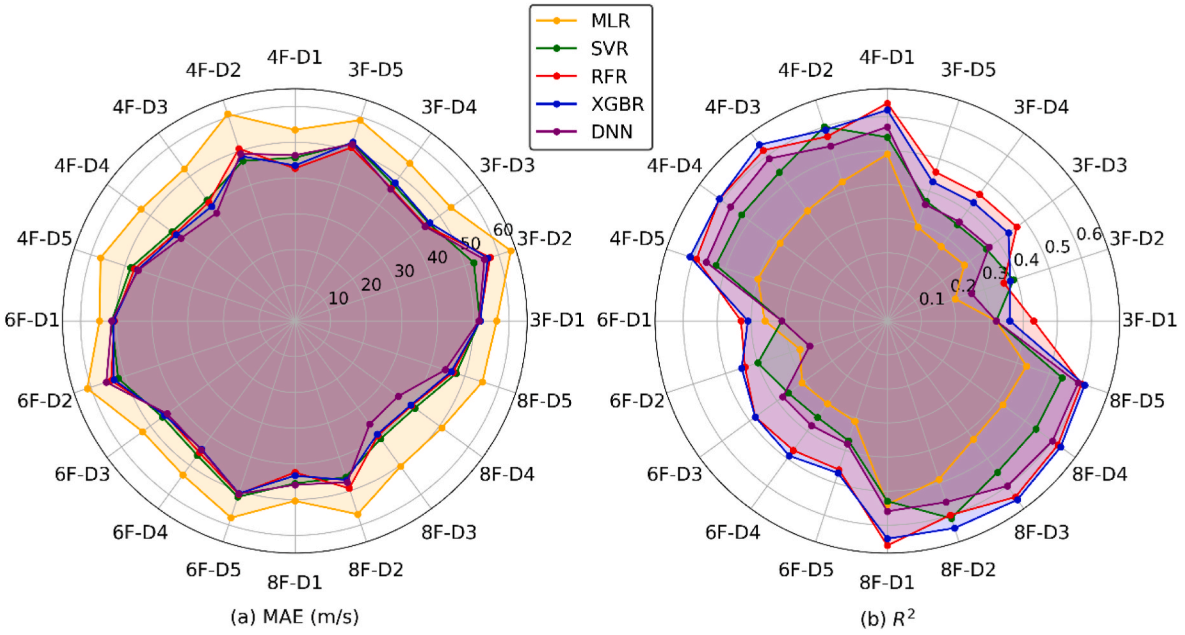


Fig. 7. Performance of five prediction techniques based on eight features and the Augmented-Constant pairing method.

incorporating them improves prediction performance, as they can be easily obtained from the measured CPTu parameters. It is noted that Group C does not include B_q, as its computation requires u_2, which is not used in this group. Within each group, all the CPTu and V_s data pairing methods and associated databases (D1 to D5) and prediction techniques (MLR, SVR, RFR, XGBR and DNN) are used to assess their impact on predictive performance. In Group E, data collected from Taipei are excluded when constructing the CPTu-V_s models. The models are then applied to predict V_s from CPTu measurements at the Taipei dataset to assess the generalisation capacity of the models and investigate an example of a site where direct V_s measurements are unavailable.

The training process is conducted on a personal desktop equipped with RAM (64 GB) and 13th Gen Intel(R) Core(TM) i9-13900KF 3.00 GHz processor. The computational time required for training is summarised in Table 4, based on eight input features (Group D). Notably, the training time for each of the non-augmented databases (Point-Point and Point-Average) and for each of the augmented databases (Augmented-Constant, Augmented-Linear and Augmented-Random) is comparable due to the similar amount of data within their respective groups. Therefore, only the training times for the Point-Point and Augmented-Constant methods are presented in Table 4. Regardless of the database, MLR exhibits the shortest training time (less than 1s), as it only requires the estimation of a limited number of coefficients (e.g., nine coefficients when eight input features are considered). All prediction models using the Augmented-Constant method require longer training times because of the substantial increase in data volume. When the Point-Point method is used, training for SVR, RFR, and XGBR is completed within 10 s, whereas DNN requires the longest training time, at 156 s; however, this remains computationally efficient. When the Augmented-Constant method is utilised, SVR exhibits the highest



(a) MAE (m/s)

(b) R^2

Fig. 8. Comparison of (a) MAE and (b) R^2 for different prediction techniques. Note: 3 F-D1 represents the combination of three features (3 F) and Point-Point (D1).

computational demand, requiring more than 26 h for training due to the substantial increase in data volume. In contrast, XGBR demonstrates exceptional efficiency, with its training time increasing only marginally from 1 to 4 s despite the database expanding nearly 50-fold.

5. Results and discussions

Figs. 3, 5, 7 and 9 illustrate scatter plots of predicted versus measured V_s for the testing dataset. In these figures, the red ellipse denotes the 90 % confidence region, which represents the area where 90 % of the predicted and true V_s data pairs are expected to fall, generated through principal component analysis. The direction of the major axis of this ellipse indicates the principal direction of data. If this axis aligns closely with the 1:1 line, the model demonstrates strong predictive performance, whereas any misalignment signifies reduced performance. Additionally, data density contours are displayed, where darker regions indicate a higher concentration of data points.

Figs. 4, 6 and 8 illustrate radar charts of performance, including MAE and R^2, for the testing dataset of the 100 models considered in Groups A to D. While these figures display the same set of results, they are organised differently to separately highlight the effects of input features, CPTu and V_s data pairing methods and prediction techniques, respectively. In addition, a summary table that includes MAE and R^2 for all cases is provided in the supporting material (Table S1).

5.1. Prediction performance for different input features

This subsection evaluates the influence of input features on model performance. The scatter plots for predictions based on the Augmented-Constant method and DNN are shown in Fig. 3, while the radar charts of MAE and R^2 for all cases are shown in Fig. 4. A comparison between Fig. 3 (a) (excluding u_2) and Fig. 3 (b) (including u_2), as well as between Fig. 3 (c) (excluding u_2 and associated derived parameter, B_q) and Fig. 3 (d) (including u_2 and associated derived parameter, B_q), demonstrates that incorporating u_2 significantly enhances predictive performance. This conclusion is drawn based on three key observations: i) the major axis of the confidence ellipse in Fig. 3 (b) and (d) aligns more closely with the 1:1 line; ii) the data density contours in Fig. 3 (b) and (d) show that the predicted values are more concentrated around the 1:1 line with reduced scatter; iii) the inclusion of u_2 results in a lower MAE and an increase in R^2 values, demonstrating enhanced predictive capability. This trend is consistent across all cases, as illustrated in Fig. 4.

Additionally, Fig. 3 (b) shows that soils with $V_s > 600$ m/s present better agreement between predicted and measured values compared to those in Fig. 3 (a), which may be a key factor contributing to the significant improvement observed when u_2 is included. When data with $V_s > 600$ m/s are excluded from training, the improvement resulting from the inclusion of u_2 is less pronounced but remains evident. This highlights the importance of including u_2 as an input feature during training, particularly when cemented materials, such as chalk in this case, are present. The results obtained using only data with $V_s <$ 600 m/s for training are provided in the supplementary material (Fig. S5 and Fig. S6).

Based on a comparison between Fig. 3 (a) and (c), or Fig. 3 (b) and (d), as well as the MAE and R^2 shown in Fig. 4, it is observed that incorporating derived CPTu parameters has a minor effect on prediction performance as the MAE and R^2 values change only marginally. For instance, when all eight CPTu parameters are considered (Fig. 3 (d)), the MAE decreases only slightly, from 37.3 m/s with four features (Fig. 3 (b)) to 35.7 m/s, while R^2 increases from 0.59 to 0.60. While the model exhibits only a slight improvement in predictive performance when using eight features, the training time is approximately doubled compared to the four-feature case. Therefore, the recommended configuration is to use four input features. Nonetheless, if all trained models are made available, users may still choose the eight-feature model for prediction.

5.2. Prediction performance for different CPTu and V_s data pairing methods

The impact of CPTu and V_s data pairing methods on model performance is explored in this section. The scatter plots for predictions obtained using eight input features and the DNN are presented in Fig. 5, while the radar charts of MAE and R^2 for all cases are shown in Fig. 6. Fig. 5 illustrates that predictions generated by Point-Point and Point-Average methods, i.e. with the smaller databases, are more scattered compared to the other data pairing methods. Additionally, Fig. 5 (b) and Fig. 6 (a) indicate that the Point-Average method exhibits the highest MAE among the five data pairing methods. This may be attributed to the presence of very distinct CPTu measurements (very low or high values),

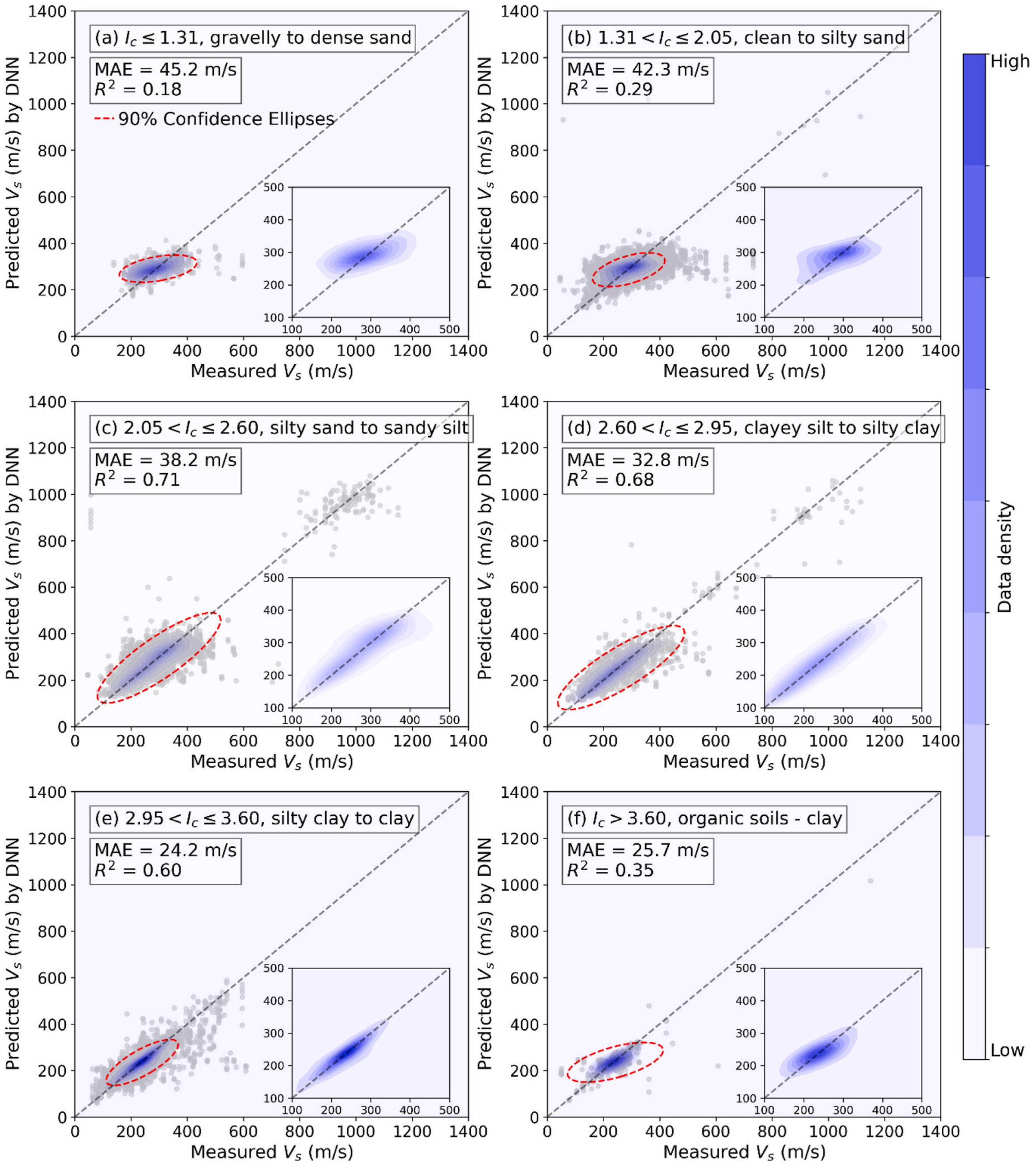


Fig. 9. Predictions with different soil types using eight features, Augmented-Constant pairing method, and DNN.

making the averaged CPTu unrepresentative of the entire interval and thereby reducing the accuracy. Conversely, the contour plots in Fig. 5 show that predictions generated using Augmented-Constant and Augmented-Linear methods are more concentrated along the 1:1 line, and Fig. 6 (a) demonstrates that these two methods yield similar MAE values, which are generally lower than those obtained using the non-augmented databases. This improvement can be attributed to the expanded databases, which mitigate the influence of individual CPTu-V_s pairings, thereby enhancing overall model performance.

In Fig. 6 (b), it is observed that Augmented-Linear and Augmented-Random methods achieve similar R^2 across all cases, both of which are slightly lower than those obtained by the Augmented-Constant method. Moreover, augmented databases achieve higher R^2 than those obtained by non-augmented databases when RFR, XGBR and DNN are utilised. The apparent inconsistency in the performance of the Augmented-Linear method when evaluated using different metrics (i.e., MAE and R^2) arises because the MAE and R^2 assess different aspects of model performance. While MAE measures the average magnitude of prediction error, R^2 quantifies the proportion of variance in the observed data that is captured by the model. Although the Augmented-Linear method effectively reduces prediction error (as indicated by lower MAE), it may also introduce additional variability through interpolation. This added variability can limit the model's ability to explain variance in the target variable, resulting in a slightly lower R^2.

Based on these findings, it is concluded that the three augmentation methods have the potential to improve prediction performance by fully utilising high-resolution CPTu data. Furthermore, they outperform the Point-Average method, which also leverages high-resolution CPTu data but does so by averaging CPTu parameters. Within these data augmentation methods, assuming a constant V_s across each interval

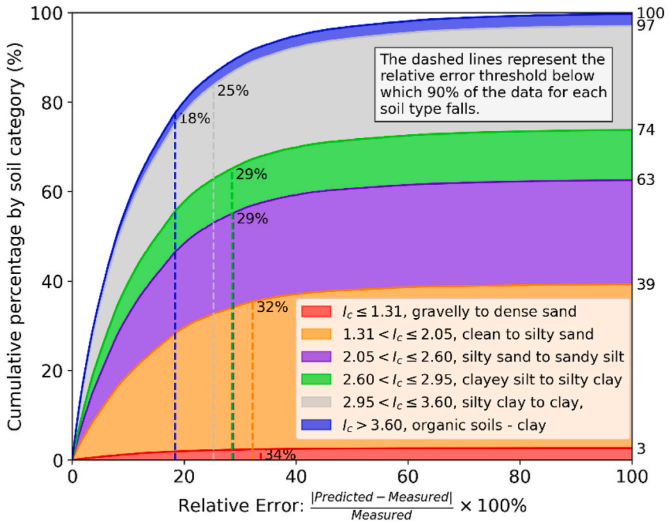


Fig. 10. Relationship between cumulative percentage of data and relative error for each soil type using eight features, Augmented-Constant pairing method, and DNN.

(Augmented-Constant database) introduces the least uncertainty into the augmented databases, leading to the best predictive performance, whereas assuming randomly varying values within each interval (Augmented-Random database) introduces the highest uncertainty, resulting in the poorest predictive performance. Future research could examine the impact of these uncertainties on geotechnical design and assess augmentation methods with stress-dependent logic to minimise uncertainties [e.g. Ref. [59]].

5.3. Prediction performance for five prediction models

Here we assess the impact of different techniques on predictive performance. The scatter plots for predictions obtained using eight input features and a database generated by the Augmented-Constant method are presented in Fig. 7, while the radar charts of MAE and R^2 for all cases are shown in Fig. 8. Additionally, MLR's performance using eight input features across various database generation methods is provided in the supplementary material (Fig. S7). Fig. 7 illustrates that predictions generated by MLR are more scattered, while Fig. 8 (a) indicates that MLR exhibits the highest MAE among the five prediction techniques. The two tree-based techniques, RFR and XGBR, display similar data distributions (Fig. 7), and also yield comparable MAE and R^2 values (Fig. 8). Additionally, RFR and XGBR consistently achieve higher R^2 values compared to the other techniques. Fig. 8 (a) demonstrates that SVR, RFR, XGBR and DNN generally exhibit similar MAE values. However, for 8 F-D3 and 8 F-D4, DNN achieves a lower MAE than the other techniques.

Based on these observations and the computational time discussed in Section 4, MLR is the simplest and fastest prediction model. However, its predictive performance is considerably lower than that of more advanced ML techniques. Specifically, MLR results in a 41 % higher MAE and 28 % lower R^2 compared to DNN. These results indicate that MLR is unsuitable for developing CPTu-V_s correlations. SVR demonstrates moderate predictive performance among the five techniques, although its computational demands significantly increase with the increase in database size. Thus, SVR is not recommended for use in scenarios where database augmentation is employed. RFR and XGBR are the third and second fastest models, respectively, and both consistently achieve the highest R^2 across all cases in this study. However, these tree-based models are prone to overfitting when trained on small databases, such as those generated using Point-Point and Point-Average methods, particularly if the initial parameter ranges are not appropriately selected [55]. To mitigate overfitting, it is recommended that a trial-and-error approach be used to determine appropriate initial ranges for hyperparameters. Based on the study presented in this paper, DNN is applicable across different database sizes, balancing both training time and predictive accuracy, making it a robust choice for constructing CPTu-V_s correlations.

5.4. Prediction performance in different soil types

Here we evaluate the predictive performance across different soil types categorised by soil behaviour index, I_c. We use the DNN for this analysis as our results evidence that it is the most accurate, robust and efficient technique from those analysed above. Fig. 9 illustrates the scatter plots for predictions obtained using DNN with eight input features and the Augmented-Constant method. In Fig. 9 (a) and (f), the principal axes of the ellipses do not align well with the 1:1 line indicating poor predictive performance for gravelly to dense sands ($I_c \leq 1.31$) and organic soils - clay ($I_c > 3.60$). This is primarily attributed to the insufficient amount of data for these two soil types, each accounting for less than 3 % of the entire database (see Table 1). However, it is noted that clean sands to silty sands (Fig. 9 (b)), despite representing the largest data subset, exhibit poorer predictive performance compared to the silty clays, clayey silts and clays shown in Fig. 9 (d) and (e). This can be attributed to i) the reduced information obtained from CPTu measurements in sandy soils compared to clays (as u_2 generally adds little information in sands); ii) the data imbalance, where the clean to silty sand data are primarily concentrated within a narrow region on the Robertson chart, whereas other soil types are more widely distributed. The poor predictive performance may also result from the complicated force chains and associated stress distribution in sand-dominated soils that develop due to for example, variations in grain shapes and orientation and grain-to-grain contacts. Additionally, it is observed that although the number of data points in Fig. 9 (c) and 9 (e) are comparable, Fig. 9 (e) represents clay-dominated soils, which exhibit significantly lower deviations between predicted and measured values, compared to the sand-clay mixed soil type in Fig. 9 (c).

Fig. 10 presents the relative error, defined as the absolute difference between the predicted and measured V_s normalised by the measured V_s, for each soil type. Based on Fig. 10, the number of predictions within a given relative error threshold can be determined. For example, for silty clays to clays ($2.95 < I_c \leq 3.60$), 90 % of predictions have relative errors below 25 %, providing a quantitative measure of prediction uncertainty for different soil types.

In addition, the predictive performance of the model across different soil depths is evaluated using DNN with eight input features and the Augmented-Constant method (Fig. S8). Results show that the highest MAE occurs at depths greater than 50 m. At this range, several data points exhibit extremely low measured values ($V_s < 50$ m/s), but unusually high predicted values ($V_s > 800$ m/s). This discrepancy may be attributed to the presence of thin soil layers that are captured by high-resolution CPTu measurements but missed by the low-resolution V_s measurements, which are averaged over larger depth intervals. Conversely, the predictions within the 10–20m and 20–30m depth ranges show relatively low MAE, indicating better model performance at these intermediate depths.

Furthermore, predictive performance across different soil regions (sub-datasets) is assessed using DNN with eight input features and the Augmented-Constant method (Fig. S9). Interestingly, model performance does not appear to be directly correlated with the size of the dataset. For instance, the Taipei dataset yields better prediction accuracy than the RVO dataset, despite the latter containing a significantly larger number of data points.

5.5. Generalisation ability of the models

The previous sections have assessed model performance using da-

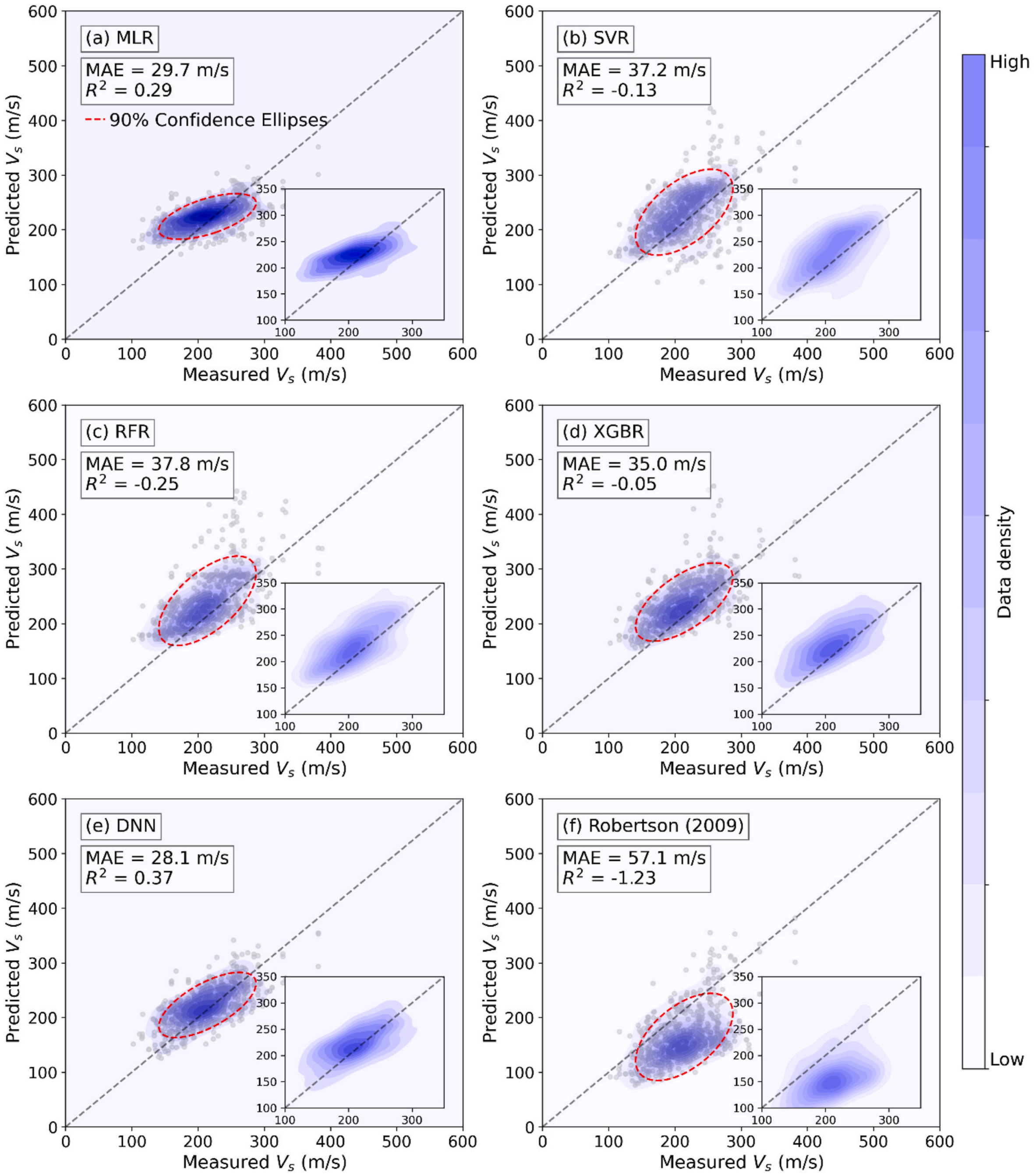


Fig. 11. Generalisation performance for different prediction techniques using eight features, Point-Point pairing method without Taipei dataset.

tabases compiled from all five datasets. In this subsection, we assess the generalisation ability of some of the models using databases that exclude all or part of the Taipei dataset, which contains 636 data points when generated using the Point-Point method. The trained models are then tested by predicting V_s values using the remaining data (i.e., data not used in training) from the Taipei dataset. Fig. 11 illustrates the prediction results that exclude the Taipei dataset during training, obtained from five prediction techniques and using eight input features and the Point-Point method, alongside predictions from a widely used empirical CPTu-V_s correlation (Eq. (15)) proposed by Robertson [18]. The empirical correlation is found to generally underestimate V_s, producing significant errors and negative R^2. It is also observed that MLR and DNN yield lower MAE and higher R^2 compared to SVR, RFR, and XGBR. However, the principal axis of MLR deviates significantly from the 1:1 line, suggesting potential bias in its predictions. In contrast, predictions from DNN are symmetrically distributed along the 1:1 line, demonstrating the best generalisation performance among all five techniques. Furthermore, predictions generated by SVR, RFR, and XGBR exhibit a general tendency toward overestimation and negative R^2, indicating that the generalisation capacity of these techniques for the Taipei dataset is worse than using an arithmetic mean. This suggests that these models may not be reliable to accurately predict V_s conditions that are unseen during training.

$$V_s = \left[10^{(0.55I_c+1.68)} \times \frac{q_t - \sigma_v}{p_a}\right]^{0.5} \tag{15}$$

where p_a is the atmospheric pressure.

To further assess the impact of site-specific data on model perfor-

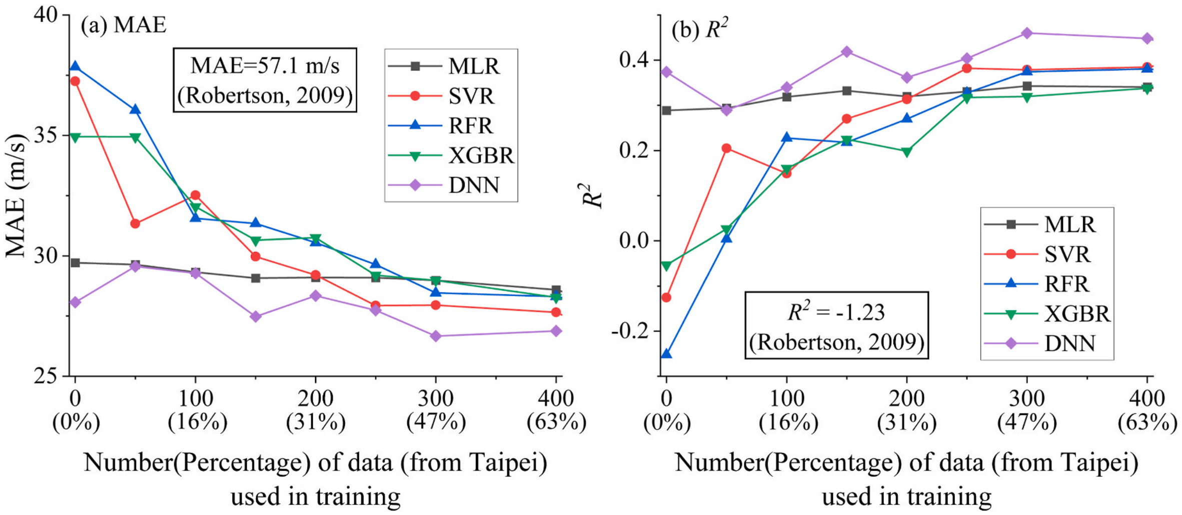


Fig. 12. Generalisation performance using different numbers of data from the Taipei dataset in training (eight features, Point-Point pairing method, and DNN).

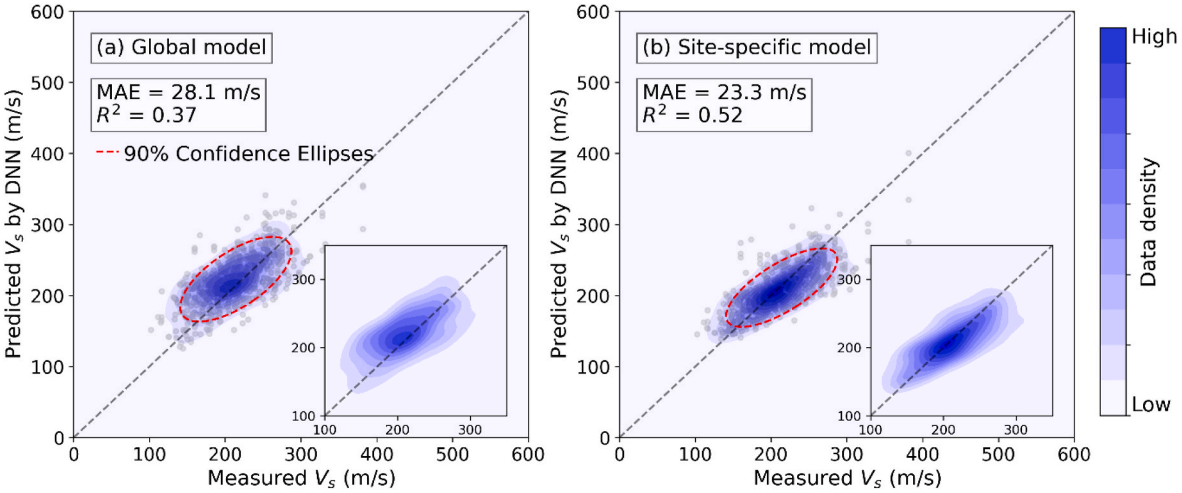


Fig. 13. Performance on Taipei dataset using global model (use all the data except Taipei dataset in training) and site-specific model (use only Taipei dataset in training).

mance, an additional test was conducted by incorporating varying portions of the Taipei dataset into the training process, as shown in Fig. 12. From the MAE and R^2 values in Fig. 12, it is observed that the prediction performance of MLR remains nearly unchanged, regardless of the amount of Taipei data included. This is mainly because MLR assigns equal weight to all data points, and a small number of additional data points (e.g., 400) exert a negligible influence on the model compared to the large-scale dataset initially used for training. Conversely, ML techniques (SVR, RFR, XGBR and DNN), which dynamically assign varying weights to individual data points, exhibit a stronger capacity to learn from site-specific data. Consequently, as more Taipei data is incorporated into training, these models demonstrate improved prediction performance, as evidenced by decreasing MAE and increasing R^2. For example, when RFR is used, MAE decreases from 37.8 m/s to 28.5 while R^2 increases from -0.25 to 0.37. It is also noted that MAE and R^2 reach a plateau when approximately 50 % of the Taipei dataset is utilised for training.

To further assess model generalisation, the trained model is tested using an external dataset from Christchurch, New Zealand [17]. The model's performance on this dataset is relatively poor, as reflected in the MAE and R^2 (Fig. S11). This can be attributed to the fact that the Christchurch data predominantly represent clean to silty sand ($1.31 < I_c \leq 2.05$), which shows relatively poor predictive performance even during model training (see Fig. 9). In contrast, the Taipei dataset primarily consists of silty clay to clay ($2.95 < I_c \leq 3.60$), which shows good predictive performance during model training (see Fig. 9). These results suggest that the model exhibits stronger generalisation for soil types that are well represented and better learned during training.

A comparison analysis, presented in Fig. 13, is conducted between the site-specific model (trained on only the Taipei dataset) and the global model (trained on all data excluding the Taipei dataset). Results indicate that the site-specific model outperforms the global model. This is expected when sufficient local data are available, as models trained on site-specific data often generalise better within the same geotechnical context. However, in situations where site-specific data are limited, training a dedicated model may not be feasible or may result in significant bias. In such cases, a more practical approach is to enhance a global model by integrating limited site-specific data. In this study, we explored a straightforward method by directly adding a subset of the Taipei data to the global training dataset to retrain the model, which led to improved predictive performance.

6. Conclusions

This paper assesses statistical and ML-based techniques to establish CPTu-V_s correlations, where the CPTu measurements are served as input features, while shear wave velocity, V_s, is the output. The effects of various factors on the prediction performance have been investigated, including the number and selection of input features, database size or data pairing methods, prediction techniques, and soil type. Additionally, the generalisation ability of the different predictive techniques has been evaluated. The main conclusions are summarised as follows:

1) The incorporation of pore pressure as an input feature enhances prediction accuracy, particularly in cemented materials such as chalk. In contrast, the inclusion of derived parameters has an insignificant effect on model performance.
2) For the conventional measurement resolutions of CPTu (every 0.02 m) and V_s data (every 1 m), the proposed V_s data augmentation methods provide up to a 50-fold increase in dataset size (here 46-fold), without requiring additional data collection. This leads to improved prediction accuracy compared to conventional data pairing methods in which CPTu data is downsampled to V_s resolution. Moreover, these augmentation methods are simple, efficient, and easily applicable in engineering practice.
3) The three augmentation methods inherently introduce uncertainties due to their underlying assumptions. Among them, assuming a constant V_s value within each V_s interval results in the lowest uncertainty, leading to the lowest MAE in predictions. The opposite occurs when assuming random V_s values within each V_s interval. 4) For the investigated augmented databases, DNN exhibits the lowest MAE, while RFR or XGBR achieve the highest R^2, demonstrating their capability for establishing CPTu-V_s correlations. Conversely, MLR exhibits significant prediction errors, making it unsuitable for constructing CPTu-V_s correlations. SVR requires extensive computational time (e.g., exceeding 26 h and approximately 11 times, 197 times, and 24,045 times longer than DNN, RFR, and XGBR, respectively), rendering it impractical for large databases.
5) Prediction accuracy varies across soil types. In general, sand-dominated soils exhibit greater prediction errors than clay-dominated soils, which can be attributed to the complicated stress distribution in sand-dominated soils caused by variations in grain shapes and orientation and grain-to-grain contacts.
6) Based on the results, the recommended model for developing CPTu-V_s correlations should utilise four measured features, a database generated using the augmentation method that assumes constant V_s values within each V_s interval and pairs them with each CPTu data point throughout the interval, and the DNN prediction technique.
7) Future research that naturally follows from this study may include adding uncertainty quantification in model predictions, and investigating augmentation techniques with stress-dependent logic and geostatistical realism (using estimated correlation lengths and variability ranges), as well as sequence-to-sequence ML techniques, such as Long Short-Term Memory (LSTM). Finally, this study demonstrated the benefits of integrating local data with global data for constructing site-specific models, and more advanced techniques such as transfer learning [60] could be also investigated in future research.

CRediT authorship contribution statement

Yuting Zhang: Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Conceptualization. **Héctor Marín-Moreno:** Writing – review & editing, Validation, Supervision, Software, Methodology, Conceptualization. **Susan Gourvenec:** Writing – review & editing, Validation, Supervision, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

Susan Gourvenec is supported by the Royal Academy of Engineering under the Chairs in Emerging Technologies scheme. Yuting Zhang is supported through funding provided by the Chair in Emerging Technologies. The work presented forms part of the activities of the Royal Academy of Engineering Chair in Emerging Technologies Centre of Excellence for Intelligent and Resilient Ocean Engineering. We would also acknowledge Christopher R. McGann for kindly providing the dataset collected from Christchurch, New Zealand.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.soildyn.2025.109972.

Data availability

The database used for training, validating, and testing the models, as well as the trained models, can be accessed via: https://github.com/T-Martin111/ML-CPTu-Vs

References

[1] CEN. European standard EN 1998-1: 2005 eurocode 8: design of structures for earthquake resistance. Part 1: general rules, seismic action and rules for buildings. Brussels, Belgium: European Committee for Standardization; 2005.

[2] ASCE. Minimum design loads and associated criteria for buildings and other structures. Reston, UNITED STATES. Am Soc Civil Eng 2021.

[3] BSI. BS EN ISO 22476-1:2023 - TC Geotechnical investigation and testing. Field testing - electrical cone and piezocone penetration test. 2023.

[4] DNV. DNV-RP-C212 offshore soil mechanics and geotechnical engineering. 2021.

[5] ISO. 19901-8:2023. Oil and gas industries including lower carbon energy - offshore structures. Part 8: marine soil investigations. 2023.

[6] Rathje EM, Kottke AR, Trent WL. Influence of input motion and site property variabilities on seismic site response analysis. J Geotech Geoenviron Eng 2010;136:607–19.

[7] Mohamad Nor O, Abbiss CP, Mohd, Raihan T, Khairul Anuar Mohd N. Prediction of long-term settlement on soft clay using shear wave velocity and damping characteristics. Eng Geol 2011;123:259–70.

[8] Gao S, Gong J, Feng Y. Equivalent damping for displacement-based seismic design of pile-supported wharves with soil–pile interaction. Ocean Eng 2016;125:12–25.

[9] Mo R, Cao R, Liu M, Li M. Effect of ground motion directionality on seismic dynamic responses of monopile offshore wind turbines. Renew Energy 2021;175:179–99.

[10] Zhao M, Gao Z, Wang P, Du X. Response spectrum method for seismic analysis of monopile offshore wind turbine. Soil Dynam Earthq Eng 2020:136.

[11] Liu X, Yang J. Shear wave velocity in sand: effect of grain shape. Geotechnique 2018;68:742–8.

[12] Hunter JA, Crow HL. Shear wave velocity measurement guidelines for Canadian seismic site characterization in soil and rock. 2015.

[13] Clayton CRI, Theron M, Best AI. The measurement of vertical shear-wave velocity using side-mounted Bender elements in the triaxial apparatus. Geotechnique 2004;54:495–8.

[14] Wu X. Structure-, stratigraphy- and fault-guided regularization in geophysical inversion. Geophys J Int 2017;210:184–95.

[15] Long M, Trafford A, McGrath T, O'Connor P. Multichannel analysis of surface waves (MASW) for offshore geotechnical investigations. Eng Geol 2020:272.

[16] McGann CR, Bradley BA, Jeong S. Empirical correlation for estimating shear-wave velocity from cone penetration test data for banks peninsula loess soils in Canterbury, New Zealand. J Geotech Geoenviron Eng 2018;144.

[17] McGann CR, Bradley BA, Taylor ML, Wotherspoon LM, Cubrinovski M. Development of an empirical correlation for predicting shear wave velocity of christchurch soils from cone penetration test data. Soil Dynam Earthq Eng 2015;75:66–75.

[18] Robertson PK. Interpretation of cone penetration tests — a unified approach. Can Geotech J 2009;46:1337–55.

[19] Mayne PW, Rix GJ. Correlations between shear wave velocity and cone tip resistance in natural clays. Soils Found 1995;35:107–10.

[20] Long M, Donohue S. Characterization of Norwegian marine clays with combined shear wave velocity and piezocone cone penetration test (CPTU) data. Can Geotech J 2010;47:709–18.

[21] Tonni L, Simonini P. Shear wave velocity as function of cone penetration test measurements in sand and silt mixtures. Eng Geol 2013;163:55–67.

[22] Tong LY, Che HB, Zhang MF, Pan HS. Determination of shear wave velocity of Yangtze Delta sediments using seismic piezocone tests. Transport Geotech 2018;14:29–40.

[23] McGann CR, Bradley BA, Taylor ML, Wotherspoon LM, Cubrinovski M. Applicability of existing empirical shear wave velocity correlations to seismic cone penetration test data in Christchurch New Zealand. Soil Dynam Earthq Eng 2015;75:76–86.

[24] Stuyts B, Weijtjens W, Jurado CS, Devriendt C, Kheffache A. A critical review of cone penetration test-based correlations for estimating small-strain shear modulus in North Sea soils. Geotechnics 2024;4:604–35.

[25] Salsabili M, Saeidi A, Rouleau A, Nastev M. Development of empirical CPTu-V correlations for post-glacial sediments in Southern Quebec, Canada, in consideration of soil type and geological setting. Soil Dynam Earthq Eng 2022;154.

[26] Chala AT, Ray RP. Machine learning techniques for soil characterization using cone penetration test data. Appl Sci 2023;13.

[27] Entezari I, Sharp J, Mayne PW. A data-driven approach to predict shear wave velocity from CPTu measurements. Cone Penetration Testing. 2022. p. 374–80.

[28] Marin Moreno H, Gourvenec S, Charles J. Application of deep neural network combined with dynamic poroelasticity to define seismic velocities and porosity from cone penetrometer data. In: Proceedings of the 7th International Conference on Geotechnical and Geophysical Site Characterization; 2024. p. 8. https://doi.org/10.23967/isc.2024.248.

[29] Assaf J, Molnar S, El Naggar MH. CPT-Vs correlations for post-glacial sediments in metropolitan Vancouver. Soil Dynam Earthq Eng 2023;165.

[30] BSH. https://pinta.bsh.de/ausschreibungen?lang=en%29%3A; 2024.

[31] Oberhollenzer S, Premstaller M, Marte R, Tschuchnigg F, Erharter GH, Marcher T. Cone penetration test dataset premstaller geotechnik. Data Brief 2021;34:106618.

[32] Rvo. https://offshorewind.rvo.nl/; 2023.

[33] Gilder C, Pokhrel R, De Luca F, Vardanega PJ. Further analysis of CPTu and seismic cone data collected in the kathmandu valley, Nepal. Politehnium Publishing House; 2023. p. 43–50.

[34] Lu CC, Deng YC, Wang JS, Hwang JH, Tseng CC. Datasets of various geotechnical surveys in several arrays in the Taipei basin. Data Brief 2024;53:110195.

[35] ASTM. D7400/D7400M-19 standard test methods for downhole seismic testing. 2019.

[36] Teachavorasinskun S, Lukkunaprasit P. A simple correlation for shear wave velocity of soft Bangkok clays. Geotechnique 2004;54:323–6.

[37] Kaklamanos J, Bradley BA, Moolacattu AN, Picard BM. Physical hypotheses for adjusting coarse profiles and improving 1D site-response estimation assessed at 10 KiK-net sites. Bull Seismol Soc Am 2020;110:1338–58.

[38] Griffiths SC, Cox BR, Rathje EM, Teague DP. Surface-wave dispersion approach for evaluating statistical models that account for shear-wave velocity uncertainty. J Geotech Geoenviron Eng 2016;142.

[39] Toro GR. Probabilistic models of site velocity profiles for generic and site-specific ground-motion amplification studies. Technical Rep 1995.

[40] Ramboll. Geotechnical data report of the preliminary investigation of FEP-site O-1.3. Hamburg 2021.

[41] Singh R, Umrao RK, Ahmad M, Ansari MK, Sharma LK, Singh TN. Prediction of geomechanical parameters using soft computing and multiple regression approach. Measurement 2017;99:108–19.

[42] Goh ATC, Zhang WG. An improvement to MLR model for predicting liquefaction-induced lateral spread using multivariate adaptive regression splines. Eng Geol 2014;170:1–10.

[43] Youd TL, Hansen CM, Bartlett SF. Revised multilinear regression equations for prediction of lateral spread displacement. J Geotech Geoenviron Eng 2002;128:1007–17.

[44] Vapnik V, Golowich S, Smola A. Support vector method for function approximation, regression estimation and signal processing. Adv Neural Inf Process Syst 1996;9.

[45] Samui P. Support vector machine applied to settlement of shallow foundations on cohesionless soils. Comput Geotech 2008;35:419–27.

[46] Cipolla S, Gondzio J. Training very large scale nonlinear SVMs using alternating direction method of multipliers coupled with the hierarchically semi-separable kernel approximations. EURO J Comput Optimiz 2022;10:100046.

[47] Samui P, Sitharam TG, Kurup PU. OCR prediction using support vector machine based on piezocone data. J Geotech Geoenviron Eng 2008;134:894–8.

[48] Debnath P, Dey AK. Prediction of bearing capacity of geogrid-reinforced stone columns using support vector regression. Int J GeoMech 2018;18.

[49] Mahmoodzadeh A, Nejati HR, Mohammadi M, Ibrahim HH, Rashidi S, Ibrahim BF. Forecasting face support pressure during EPB shield tunneling in soft ground formations using support vector regression and meta-heuristic optimization algorithms. Rock Mech Rock Eng 2022;55:6367–86.

[50] Bherde V, Kudlur Mallikarjunappa L, Baadiga R, Balunaini U. Application of machine-learning algorithms for predicting California bearing ratio of soil. J Transport Eng 2023:149. Part B: Pavements.

[51] Zhang WG, Wu CZ, Li YQ, Wang L, Samui P. Assessment of pile drivability using random forest regression and multivariate adaptive regression splines. Georisk 2021;15:27–40.

[52] Chen T, Guestrin C. XGBoost: a scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016. p. 785–94.

[53] Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat 2001:1189–232.

[54] Zhang W, Zhang R, Wu C, Goh ATC, Wang L. Assessment of basal heave stability for braced excavations in anisotropic clay using extreme gradient boosting and random forest regression. Undergr Space 2022;7:233–41.

[55] Wang M-X, Huang D, Wang G, Li D-Q. SS-XGBoost: a machine learning framework for predicting newmark sliding displacements of slopes. J Geotech Geoenviron Eng 2020;146.

[56] Vali R, Alinezhad E, Fallahi M, Beygi M, Saberian M, Li J. Developing a novel big dataset and a deep neural network to predict the bearing capacity of a ring footing. J Rock Mech Geotech Eng 2024;16:4798–813.

[57] Wang L, Wu C, Yang Z, Wang L. Deep learning methods for time-dependent reliability analysis of reservoir slopes in spatially variable soils. Comput Geotech 2023;159.

[58] Shahriari B, Swersky K, Wang Z, Adams RP, de Freitas N. Taking the human out of the loop: a review of bayesian optimization. Proc IEEE 2016;104:148–75.

[59] Xie J, Huang J, Zhang F, He J, Kang K, Sun Y. Enhancing the resolution of sparse rock property measurements using machine learning and random field theory. J Rock Mech Geotech Eng 2024;16:3924–36.

[60] Xie J, Chen B, Jiang S-H, Guo H, Xie S, Huang J. Enhancing data reuse in tunnelling site investigation through transfer learning-based historical data mining. Undergr Space 2025;23:161–74.