

Prediction of Shear-Wave Velocity from CPTu Data Using Robust Tanh Normalization and Quadratic Vector Regression

Suho Cho ^a, Han-Saem Kim^{a*}

^a Department of Civil and Environmental Engineering, Dongguk University, Seoul 04620, South Korea

* Corresponding author: hansaem@dgu.ac.kr

Abstract

Shear-wave velocity (V_s) is widely used for dynamic ground characterization, but existing CPT/CPTu-based correlations often exhibit limited transferability across soil types and geological settings. This study proposes a transparent prediction framework that combines robust tanh normalization with an explicit Quadratic Vector Regression (QVR) formulation for direct V_s estimation from four primary CPTu inputs: corrected cone resistance (q_t), sleeve friction (f_s), pore pressure (u_2), and depth (z). The inputs are robustly scaled using the training median and interquartile range and mapped into a bounded $(-1,1)$ feature space before being represented by first-order and feature-wise quadratic terms. The framework was evaluated using the North Sea database of Stuyts et al. (2024) and the multi-regional D1 database of Zhang et al. (2026). On the North Sea hold-out test set, QVR improved R^2 from 0.105 to 0.257 and reduced MAE from 39.13 to 35.89 m/s relative to the stress-dependent correlation of Stuyts et al. (2024). Comparable or improved performance was also obtained for additional held-out North Sea profiles. In cross-dataset evaluation from the North Sea training set to the D1 subsets, QVR achieved $R^2 = 0.221\text{--}0.259$, with performance comparable to SVR, DNN, and XGBoost despite its substantially simpler closed-form structure. Conversely, models trained on D1 showed degraded performance when transferred to the small North Sea hold-out set, indicating persistent dataset-dependent transfer limitations. Overall, the results show that a meaningful component of the multivariate CPTu– V_s relationship can be represented by a compact explicit quadratic formulation after robust bounded transformation, while broader transferability requires strictly site-independent validation.

Keywords Shear-wave velocity; Piezocone penetration test (CPTu); Robust tanh normalization; Quadratic vector regression; Machine learning; Cross-dataset transferability;

1. Introduction

Shear-wave velocity (V_s) is a fundamental property characterizing the dynamic and elastic response of geomaterials and, together with mass density (ρ), is used to determine the small-strain shear modulus through Eq. (1).

$$G_{max} = \rho V_s^2 \quad (1)$$

Eq. (1) is widely used in dynamic site characterization and liquefaction assessment (Tonni and Simonini 2013; Entezari et al. 2022). It is also relevant to serviceability and foundation analyses, including offshore wind-turbine foundations (Zhang et al. 2026). Because direct seismic measurements require additional equipment and are commonly available only at discrete locations, estimating V_s from continuous CPTu profiles can support preliminary design and spatial ground characterization (Entezari et al. 2022; Zhang et al. 2026).

Empirical CPT/CPTu-based relationships for V_s or G_{max} have been developed for different soil types and regions. Rix and Stokoe (1991) examined G_{max} in relation to q_c and σ'_{vo} using calibration-chamber tests on washed mortar sand. Mayne and Rix (1993, 1995) developed clay correlations, with the 1995 V_s model based on 31 natural-clay sites and improved by including void ratio. Later studies introduced stress-normalized sand relationships (Wride et al. 2000), a global all-soils formulation involving q_t , σ'_{vo} , and I_c (Hegazy and Mayne 2006), and models that additionally considered depth and geologic age (Andrus et al. 2007). Robertson (2009) related V_s to normalized net cone resistance and I_c , whereas Tonni and Simonini (2013) reported alternative correlations involving q_t , total and effective overburden stress, I_c , and Q_{tn} for sand-silt mixtures. Region-specific models were subsequently developed for Banks Peninsula loess (McGann et al. 2018) and North Sea soils (Peuchen et al. 2022).

Nevertheless, many empirical relationships remain sensitive to the soil type and regional conditions represented in their calibration datasets. McGann et al. (2018), for example, showed that a generalized correlation underestimated V_s in regional loess, while Stuyts et al. (2024) reported substantial bias when several established relationships were transferred to North Sea soils. In the North Sea benchmark, many evaluated correlations produced low or negative R^2 values, including some developed for mixed soils (Stuyts et al. 2024). In the North Sea benchmark of Stuyts et al. (2024), similarity in soil-behavior classification did not by itself result in consistent predictive performance across the evaluated correlations.

The reliability of u_2 measurements is important because pore pressure affects the area

correction of q_t and several normalized CPTu parameters. Although offshore testing can provide favorable saturation conditions, incomplete saturation, sensor response, and data-processing errors can still affect the measured pore pressure (Robertson and Campanella 1983; Robertson 2009). The CPTu records were therefore screened for algebraic consistency before model development.

Machine-learning (ML) methods have recently been applied to predict V_s directly from CPTu measurements. Entezari et al. (2022) demonstrated improved prediction using random forests, while Zhang et al. (2026) evaluated several ML approaches and recommended a DNN based on depth, f_s , q_t and u_2 . However, Zhang et al. also showed that predictive performance deteriorated for external sites and soil types that were poorly represented in the training data. Although the RF and DNN models used in these studies can represent nonlinear relationships, they do not yield a single closed-form equation that can be directly inspected or implemented without the fitted model object. These studies therefore motivate the evaluation of simpler explicit formulations alongside nonlinear ML models, particularly when model interpretability and cross-dataset performance are of interest.

This study aims to identify systematic relationships between the multivariate behavior of offshore CPTu measurements and variations in V_s , and to represent these relationships using a simple and explicit predictive framework. The primary CPTu inputs, which differ in physical scale and outlier characteristics, are first transformed into a common bounded feature space using robust scaling followed by an element-wise hyperbolic tangent transformation. The transformed inputs are assembled into a bounded feature vector and used in a study-specific additive quadratic regression, termed QVR, for direct prediction of V_s . The same inputs are also supplied to SVR, DNN, and XGBoost models to compare predictive performance under the two training and transfer settings. Through cross-dataset evaluations between these independent regional and multi-regional compilations, this study demonstrates the practical viability and transferability of the explicit quadratic formulation alongside nonlinear ML models.

2. Data and Outlier Screening

This study utilized the publicly available seismic piezocone penetration test (S-PCPT) database compiled from offshore wind farm investigations across the southern and central North Sea by Stuyts et al. (2024). The study region has undergone a complex geological history, including Paleogene marine clay deposition, extensive Mid-Miocene to Pleistocene deltaic sedimentation, and Quaternary glacial–interglacial sea-level fluctuations accompanied by sub-glacial channel incision and infilling. Consequently, the subsurface is characterized by overconsolidated sediments resulting from historical erosion and highly variable silica-dominated deposits, consisting predominantly of cohesionless clean to silty sands ($I_c < 2.5$) with limited cohesive intervals. To avoid the higher signal-to-noise ratios and depth uncertainties associated with downhole CPT systems, Stuyts et al. (2024) retained only measurements acquired using seafloor CPT systems, discarded unreliable near-surface data at depths shallower than 5m below the seafloor in accordance with ISO 19901-8, and averaged the CPT parameters over the 0.5 m interval between the dual geophone arrays.

The public-domain dataset covering the Dutch and German sectors comprises 143 boreholes across seven project zones: Borssele, Hollandse Kust Noord (HKN), Hollandse Kust West (HKW), Hollandse Kust Zuid (HKZ), Ten Noorden van Waddeneilanden (TNW), IJmuiden Ver (IJV), and the German Zone. In this study, eight boreholes from the HKZ project zone were designated as the hold-out test dataset, and two boreholes from the IJmuiden Ver zone—IJV162-SCPT, representing a predominantly sandy profile, and IJV038-SCPT, representing a layered profile—were reserved as additional held-out profiles, while all remaining boreholes were assigned to the training dataset.

Securing high-quality data and verifying its physical consistency are essential prerequisites for reliable data-driven modeling. As a data-quality check, the reported derived parameters were algebraically reconstructed from the available CPTu variables. The reconstruction implied equivalent saturated densities (ρ_{sat}) of 1.98-2.04 Mg/m³ and a seawater unit weight (γ_w) of 10.25 kN/m³ under the assumptions embedded in the published dataset. Specifically, based on the fundamental relationship among corrected cone tip resistance (q_t), measured cone tip resistance (q_c), and porewater pressure at the cone shoulder (u_2) in Eq. (2), the net area ratio (A_{net}) at each measurement depth was back-calculated using Eq. (3):

$$q_t = q_c + u_2(1 - A_{net}) \quad (2)$$

$$A_{net} = 1 - \left(\frac{q_t - q_c}{u_2} \right) \quad (3)$$

Because A_{net} is a geometric calibration constant specific to the cone penetrometer, it should remain constant across the entire depth profile of a given test. However, inspection of the back-calculated A_{net} profiles revealed that measurement errors—particularly in porewater pressure (u_2)—frequently occurred at the final penetration depth at the end of offshore CPTu tests (Fig. 1).

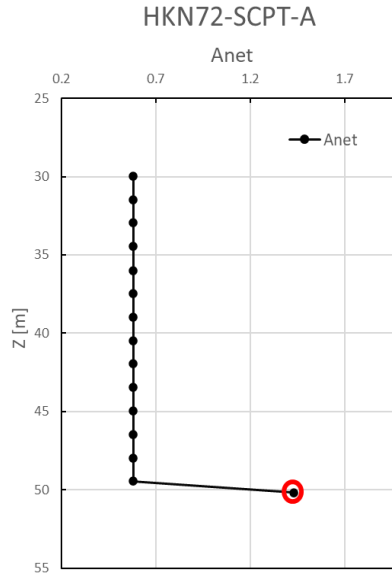


Figure 1. Back-calculated net area ratio (A_{net}) profile with depth for borehole HKN72-SCPT-A, illustrating a terminal porewater pressure measurement error at the end of the CPTu test (indicated by the red circle).

As illustrated for borehole HKN72-SCPT-A in Fig. 1, the back-calculated A_{net} remains constant from 30 m to 49 m depth, but exhibits an abrupt anomalous spike at the terminal measurement point near 50 m (indicated by the red circle). All such terminal measurement anomalies identified through A_{net} back-calculation were classified as outliers and removed prior to model calibration and evaluation.

3. Methodology

3.1 Exploratory Analysis of Vs and CPTu Parameter Variations

Campanella and Stewart (1992) noted that cross-correlation of signals recorded at adjacent 1 m depth intervals is generally reliable when no soil-layer boundary is intercepted; near boundaries, waveform distortion can increase discrepancies between signal-processing methods. Accordingly, abrupt point-to-point changes were treated as potential combinations of real stratigraphic variability and measurement or processing effects. Building on this observation, the present study analyzed the co-variation characteristics between q_t , f_s , u_2 , and Vs along depth(z). In actual field measurements, CPTu parameters and Vs do not follow a simple linear trend with depth; instead, they exhibit nonlinear, multi-directional fluctuations across thin interbedded layers and transition zones. To visually compare the depth-wise behavioral patterns of variables with different physical units and scales, each parameter was first scaled between 0 and 1 via location-wise Min–Max normalization, and its variation pattern along depth was examined (Fig. 2).

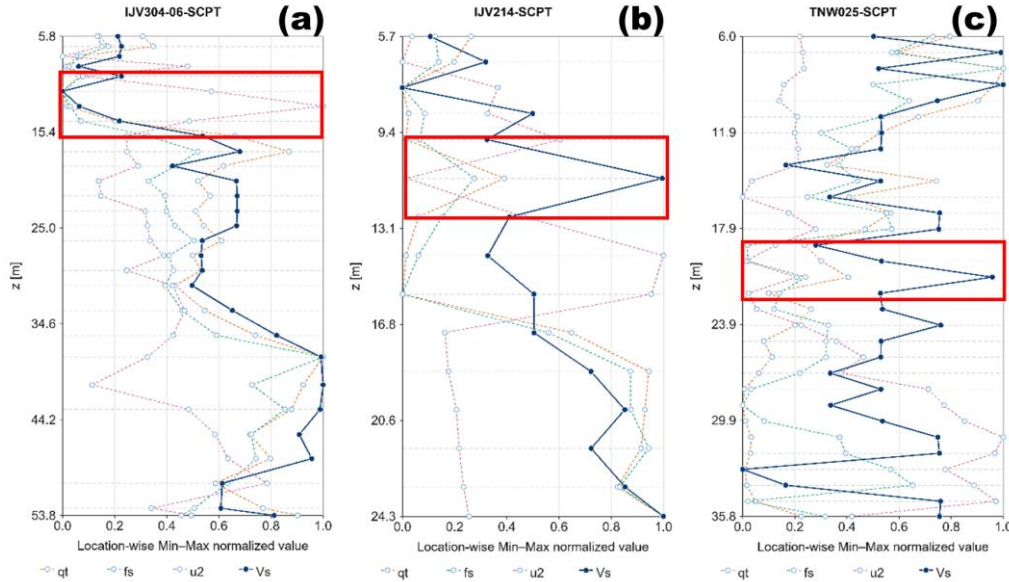


Figure 2. Representative examples of local co-directional and inverse variations between location-wise Min–Max normalized CPTu parameters (q_t , f_s , u_2) and Vs along depth (highlighted by red boxes)

Visual inspection revealed various local patterns, including intervals where Vs fluctuates in the same direction as f_s and q_t as seen in borehole IJV304-06-SCPT in Fig. 2a, intervals

where V_s moves in the same direction as f_s and q_t but inversely to u_2 as shown in borehole IJV214-SCPT in Fig. 2b, and intervals where V_s varies in the same direction as all three CPTu parameters as observed in borehole TNW025-SCPT in Fig. 2c. However, these visual patterns reflect only qualitative tendencies and exhibit excessive variability across different stratigraphic conditions to be generalized into a single consistent rule. Furthermore, applying location-wise Min–Max normalization individually to each borehole makes the scale dependent on the local minimum and maximum values within that borehole, thereby losing the physical meaning of the absolute magnitudes of q_t , f_s , and u_2 across sites.

Although local co-variation between the CPTu measurements and V_s was repeatedly observed, the relative behavior of the individual parameters varied considerably between depth intervals and boreholes. In addition, location-wise Min–Max normalization depends on the local extrema of each sounding and therefore does not preserve absolute parameter magnitudes across different locations. These observations motivated the use of a common normalization scheme that places the CPTu inputs in a bounded feature space while retaining consistent scaling across the training and evaluation datasets.

3.2 Robust tanh normalization

Raw CPTu measurements exhibit substantially different physical scales and statistical distributions, while their relationships with V_s may be nonlinear. Therefore, the input parameters were mapped into a common bounded feature space without imposing a priori physical weighting on individual features. Large deviations from the central range were preserved in direction and relative ordering, while their influence was progressively limited through nonlinear compression.

In this study, each CPTu input parameter was first robustly scaled using the median and interquartile range (IQR) calculated from the training data, followed by an element-wise hyperbolic tangent (tanh) transformation. The transformed parameters were subsequently assembled into a bounded feature vector $x^* = [q_t^*, f_s^*, u_2^*, z^*]^T$. This preprocessing procedure is hereafter referred to as robust tanh normalization. Similar robust tanh normalization approaches have previously been investigated in statistical learning, including feature-wise tanh estimators (Guimerà Cuevas and Schmid, 2024) and median/IQR-based tanh scaling approaches (Lin and Lyu, 2025).

The robust tanh normalization is defined as

$$x_{i,j}^* = \tanh \left(\alpha \frac{x_{ij} - \text{median}_j}{IQR_j} \right) \quad (4)$$

Where, $x_{i,j}^*$ denotes the transformed value, where i indexes the observation and j indexes the input feature. Robust scaling using the training median and IQR is followed by the tanh transformation to bound the input to $(-1, 1)$. The coefficient α controls the strength of compression and the rate at which the transformed values approach saturation. The resulting distributions for different values of α are shown in Figure 3a, 3b.

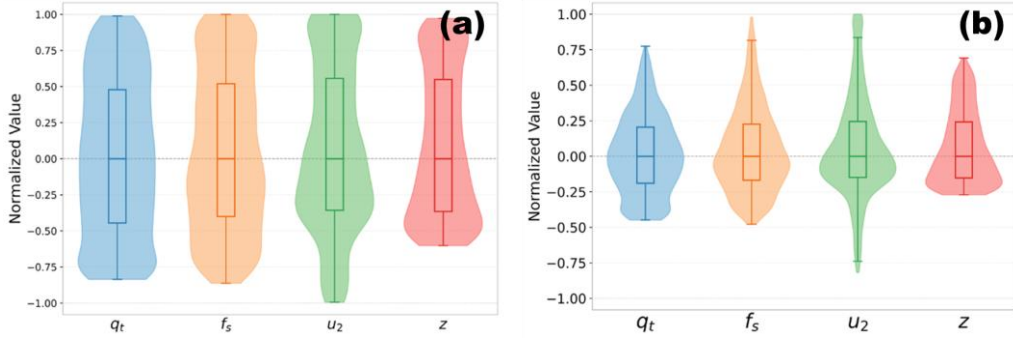


Figure 3. Distributions of robust tanh-normalized CPTu features for different compression parameters: (a) $\alpha = 1$; (b) $\alpha = 0.4$

3.3 Quadratic Vector Regression (QVR)

Using the normalized parameter vector x^* , the proposed Quadratic Vector Regression (QVR) model predicts V_s through vector inner products and element-wise (Hadamard) multiplication as follows:

$$\hat{V}_s = \beta + \omega_1^T x^* + \omega_2^T (x^* \odot x^*) \quad (5)$$

Here, β is a scalar intercept representing the predicted V_s when all transformed features are zero, corresponding to the training medians of the input variables. The vector ω_1 contains the first-order coefficients and retains sign-sensitive information from each transformed feature, whereas ω_2 introduces feature-wise quadratic curvature. Because the squared term $(x^* \odot x^*)$ is invariant to the sign of x^* , it cannot distinguish positive from negative transformed values by itself. The combined first- and second-order terms therefore allow both sign-sensitive linear variation and nonlinear curvature to be represented.

Specifically, for the four normalized CPTu parameters, the formulation expands explicitly as:

$$\begin{aligned} \hat{V}_s = & \beta + \omega_{1,q_t} q_t^* + \omega_{1,f_s} f_s^* + \omega_{1,u_2} u_2^* + \omega_{1,z} z^* \\ & + \omega_{2,q_t} (q_t^*)^2 + \omega_{2,f_s} (f_s^*)^2 + \omega_{2,u_2} (u_2^*)^2 + \omega_{2,z} (z^*)^2 \end{aligned} \quad (6)$$

The formulation in Eq. (6) provides a compact closed-form representation of the first-order and feature-wise quadratic contributions of the normalized CPTu inputs.

4. Predictive Performance and External Validation

4.1. External Dataset and Benchmark Framework

To evaluate the external transferability of the proposed framework and the effect of robust tanh normalization, the public SCPTu database compiled by Zhang et al. (2026) was additionally utilized. This database integrates five public datasets collected from offshore sites in the Netherlands and Germany and onshore sites in Germany, Austria, Taipei, and Nepal. Their experimental design evaluated an extensive matrix of 125 models by combining five input feature groups (Groups A–E, selectively incorporating four raw measurements— z , f_s , q_t , u_2 —and four derived parameters— F_r , Q_t , B_q , I_c), five prediction algorithms (MLR, SVR, RFR, XGBR, and DNN), and five CPTu–Vs pairing methods designed to bridge the resolution gap between 0.02 m CPTu and 0.5–1m Vs measurements. These pairing strategies comprised two down sampling approaches—Point-Point (D1, $n = 5,439$) and Point-Average (D2, $n = 5,990$)—and three data augmentation schemes that expanded the dataset nearly 46-fold ($n = 252,455$ each) by assuming constant (D3), linearly interpolated (D4), or randomly perturbed (D5) Vs profiles within each measurement interval. Each database was randomly partitioned into training, validation, and testing sets at an 80:10:10 ratio, and hyperparameters were tuned via Bayesian optimization over up to 50 iterations.

Because the public database of Zhang et al. (2026) does not provide explicit site or location identifiers for individual records, total vertical stress (σ_{vo}) and effective vertical stress (σ'_{vo}) were back-calculated for all records using the mathematical definitions relating the raw CPTu measurements (f_s , q_t , u_2) to the reported derived parameters (F_r , Q_t , B_q). The total vertical stress was first reconstructed from F_r (Eq. 7).

$$\sigma_{vo} = q_t - \frac{100f_s}{F_r} \quad (7)$$

The hydrostatic pore pressure was then obtained from B_q , and the effective vertical stress was calculated as (Eq. 8, 9).

$$u_0 = u_2 - B_q(q_t - \sigma_{vo}) \quad (8)$$

$$\sigma'_{vo} = \sigma_{vo} - u_0 \quad (9)$$

As an internal consistency check, σ'_{vo} was also independently reconstructed from Q_t as

(Eq. 10).

$$\sigma'_{vo} = \frac{q_t - \sigma_{vo}}{Q_t} \quad (10)$$

The two estimates of effective vertical stress were found to be numerically identical across the dataset. The equivalent density and unit-weight terms inferred from the reconstructed stresses were effectively constant across all records, indicating that the reconstructed stress variables are consistent with a common unit-weight and hydrostatic pore-pressure assumption across the database (Eq. 11).

$$\sigma'_{vo} = \gamma_{sat} z_{sat} - \gamma_{\omega} z_{\omega} = (\rho_{sat} - \rho_{\omega}) g z \quad (11)$$

Consequently, σ'_{vo} is strictly proportional to depth (z), in this dataset (Figure 4). After robust tanh normalization, the normalized effective stress therefore becomes mathematically identical to the normalized depth, z^* . This identity reflects the construction of the published variables and should not be generalized to datasets with site-specific unit weights or groundwater conditions.

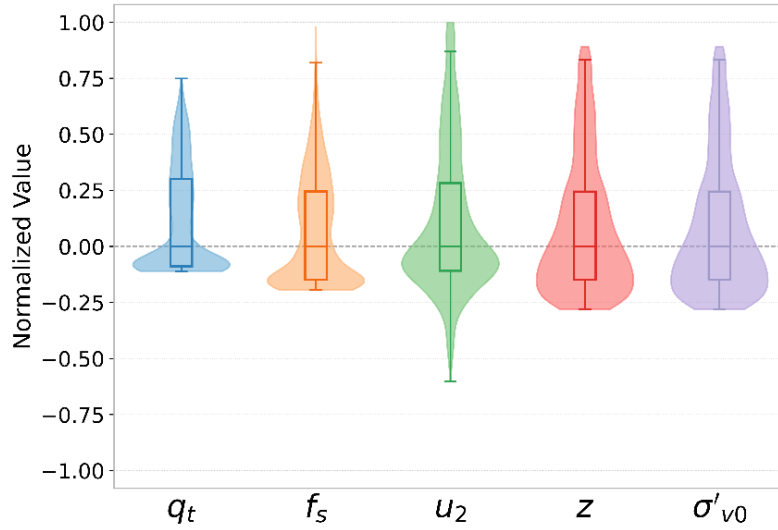


Figure 4. Robust tanh-normalized CPTu feature distributions for Zhang et al. (2026) D1 datasets ($\alpha = 0.4$).

Figure 4 also shows the normalized distributions obtained using the fixed robust tanh normalization parameter. The distributions for the Zhang et al., (2026) dataset exhibit stronger compression of the normalized CPTu parameters. To maintain consistent preprocessing conditions and isolate differences attributable to the datasets, α was kept fixed at 0.4 rather

than re-optimized for each dataset.

4.2. Quadratic Vector Regression (QVR) Prediction Performance

The proposed Quadratic Vector Regression model was calibrated using the North Sea training dataset ($n = 2,754$) compiled by Stuyts et al. (2024), and its predictive accuracy was compared against the stress-dependent correlation proposed in the same study (Table 1). Model performance was evaluated across the training set, the hold-out test set ($n = 145$), and two independent validation profiles: IJV162-SCPT ($n = 20$), consisting entirely of sandy soil, and IJV038-SCPT ($n = 31$), representing a layered mixed-soil profile. In addition to the coefficient of determination (R^2), root mean square error (RMSE), mean absolute error (MAE), ratio-based metrics were computed from the predicted-to-measured Vs ratio, namely the mean ratio (μ) and coefficient of variation (CoV, defined as the sample standard deviation of the ratio divided by μ , following Stuyts et al., 2024).

Table 1. Comparison of prediction performance between the Quadratic Vector Regression (QVR) and Stuyts et al. (2024) on the North Sea dataset.

Data	n	Model	R^2	RMSE	MAE	CoV	μ
Train	2,754	QVR	0.506	44.347	33.629	0.173	1.022
		Stuyts et al. (2024)	0.380	49.669	38.532	0.185	0.979
Test	145	QVR	0.257	50.853	35.89	0.172	1.023
		Stuyts et al. (2024)	0.105	55.803	39.13	0.19	0.986
IJV162-SCPT	20	QVR	0.835	20.852	16.429	0.065	0.997
		Stuyts et al. (2024)	0.825	21.474	17.725	0.058	0.959
IJV038-SCPT	31	QVR	0.466	29.792	26.444	0.060	1.083
		Stuyts et al. (2024)	0.474	29.588	24.264	0.099	1.022

As summarized in Table 1, the QVR outperformed the correlation of Stuyts et al. (2024) on both the training set ($R^2 = 0.506$, MAE = 33.63 m/s vs. $R^2 = 0.380$, MAE = 38.53 m/s) and the test set ($R^2 = 0.257$, MAE = 35.89 m/s vs. $R^2 = 0.105$, MAE = 39.13 m/s). For the sand-dominated IJV162-SCPT profile, the QVR ($R^2 = 0.835$, MAE = 16.43 m/s) also achieved higher accuracy than the Stuyts et al. (2024) model ($R^2 = 0.825$, MAE = 17.73 m/s).

For the layered IJV038-SCPT profile, the Stuyts et al. (2024) correlation yielded slightly higher R^2 and lower RMSE and MAE than QVR ($R^2 = 0.474$ vs. 0.466; MAE = 24.26 vs. 26.44 m/s). QVR nevertheless produced a lower ratio dispersion, with a CoV of 0.060 compared with 0.099. The metrics therefore provide mixed evidence rather than a consistent ranking for this profile.

4.3. ML Performance with Robust Tanh-Normalized Inputs

In this section, the bounded feature vector derived from robust tanh normalization was supplied as input to three widely used nonlinear machine-learning (ML) models—deep neural network (DNN), support vector regression (SVR), and extreme gradient boosting (XGBoost)—to evaluate the effect of the proposed transformation on predictive performance under two experimental cases (Case 1 and Case 2). Hyperparameters were selected by fivefold cross-validation within each training split, without using a separate validation set. Grouped cross-validation by location was used for the North Sea data; because the public D1 data lack site identifiers, D1 tuning used row-wise folds, which may share site-specific profile structure across folds.

In Case 1, the three ML models and the QVR were trained using the robust tanh normalization parameters derived from the Stuyts et al. (2024) training dataset, and subsequently evaluated on the unseen Stuyts test set as well as the external D1 Train ($n = 4,351$) and D1 Test ($n = 544$) datasets of Zhang et al. (2026). Previously, Zhang et al. (2026) reported that conventional ML models (SVR, RFR, and XGBR) suffered from severe prediction divergence and negative R^2 values when applied to unseen external site data (Table 2). Case 1 evaluated how models trained on the North Sea compilation performed on the D1 Train and D1 Test subsets under the robust tanh representation. Because the two compilations may include geographically or geologically related North Sea records, this analysis is treated as a cross-dataset transfer test rather than fully independent external validation. As shown in Table 2, all four models maintained positive R^2 values (0.211–0.261) across the external D1 Train and D1 Test datasets under the robust tanh representation. The explicit QVR achieved R^2 and MAE values comparable to those of the three nonlinear ML models.

Table 2. Prediction performance of the QVR and ML models trained on the Stuyts et al. (2024) training dataset (Case 1).

Data	Model	R ²	RMSE	MAE	CoV	μ
Test	QVR	0.257	50.853	35.890	0.172	1.023
	SVR	0.303	49.234	34.079	0.168	1.013
	DNN	0.238	51.486	36.167	0.173	1.013
	XGBoost	0.280	50.062	36.020	0.171	1.024
D1 Train*	QVR	0.259	81.893	50.888	0.260	1.110
	SVR	0.261	81.783	51.000	0.260	1.103
	DNN	0.242	82.818	51.988	0.262	1.115
	XGBoost	0.225	83.754	54.056	0.267	1.140
D1 Test*	QVR	0.221	99.824	58.098	0.248	1.098
	SVR	0.229	99.275	57.889	0.247	1.090
	DNN	0.218	99.997	59.409	0.251	1.104
	XGBoost	0.211	100.443	60.905	0.254	1.126

* External evaluation datasets from Zhang et al. (2026).

Conversely, in Case 2, the models were trained on the D1 Train dataset of Zhang et al. (2026) and evaluated on the hold-out D1 Test dataset as well as the external Train and Test datasets of Stuyts et al. (2024) (Table 3). On the common D1 Test split, the present implementations yielded modestly lower MAE values than the published Zhang et al. (2026) baselines (SVR: 45.70 to 44.50 m/s; DNN: 46.40 to 44.26 m/s; XGBoost: 43.50 to 43.26 m/s). Because the training, validation, and tuning protocols were not identical, these differences are interpreted as an observed association rather than a causal effect of robust tanh scaling. When transferred to the external Stuyts training dataset, all four models exhibited consistent performance ($R^2 = 0.395\text{--}0.425$; $MAE = 35.62\text{--}37.08\text{m/s}$). On the smaller Stuyts hold-out test set, DNN and XGBoost retained positive R^2 values (0.164 and 0.082) whereas QVR and SVR yielded slightly negative R^2 values (-0.032 and -0.055) while still maintaining stable CoV (0.184–0.197) and μ (0.987–1.021).

Table 3. Prediction performance of the QVR and ML models trained on the D1 Train dataset of Zhang et al. (2026) (Case 2).

Data	Model	R ²	RMSE	MAE	CoV	μ
D1 Test*	QVR	0.414	86.584	51.405	0.235	1.042
	SVR	0.572	74.004	44.500	0.221	1.017
	DNN	0.570	74.116	44.258	0.219	1.028
	XGBoost	0.622	69.517	43.256	0.215	1.032
	SVR*	0.540	-	45.700	-	-
	DNN*	0.570	-	46.400	-	-
	XGBoost *	0.620	-	43.500	-	-
Train	QVR	0.395	49.055	37.076	0.187	1.019
	SVR	0.425	47.835	35.615	0.186	1.000
	DNN	0.407	48.543	36.300	0.187	1.013
	XGBoost	0.414	48.270	36.431	0.192	1.012
Test	QVR	-0.032	59.934	44.518	0.193	1.021
	SVR	-0.055	60.588	43.290	0.197	0.987
	DNN	0.164	53.922	38.677	0.184	0.999
	XGBoost	0.082	56.524	40.991	0.187	1.002

* Benchmark dataset and baseline model results reported by Zhang et al. (2026).

5. Discussion and Conclusion

In this study, a median- and interquartile range (IQR)-based robust tanh normalization method and an explicit second-order nonlinear vector formulation, termed the QVR, were proposed to estimate shear wave velocity (V_s) from piezocone penetration test (CPTu) data. The proposed framework was evaluated within and across the North Sea compilation of Stuyts et al. (2024) and the multi-regional D1 compilation of Zhang et al. (2026). Because explicit site identifiers are unavailable in D1 and Dutch/German offshore records may overlap between the two compilations, the cross-dataset results should not be interpreted as strictly site-independent external validation. The main discussions and conclusions are summarized as follows:

First, robust tanh normalization maps four primary input variables—the corrected cone resistance (q_t), sleeve friction (f_s), pore pressure (u_2), and depth (z)—into a common bounded vector space within $(-1, 1)$, without requiring soil-behavior classification indices or stress-normalized derived parameters. Because the proposed transformation does not require a logarithm or a non-integer power of u_2 , it can accept non-positive u_2 values without introducing an arbitrary offset. This advantage applies specifically to formulations that would otherwise transform u_2 directly; it is not a limitation of every power-law or logarithmic CPT correlation. The tanh mapping bounds each transformed feature and therefore limits the leverage of extreme input values by construction. The incremental effect of this compression on prediction error was not isolated in the present experiments and requires a controlled preprocessing ablation.

Second, the explicit QVR yielded higher R^2 and lower RMSE and MAE than the stress-dependent correlation of Stuyts et al. (2024) for the North Sea training set, hold-out test set, and IJV162-SCPT profile, while the IJV038-SCPT profile showed mixed performance across the evaluated metrics. QVR also achieved predictive performance comparable to the nonlinear ML models under several Case 1 and Case 2 evaluation settings. In the North Sea dataset evaluation, the QVR outperformed the correlation of Stuyts et al. (2024) across the training set ($R^2 = 0.506$ vs. 0.380), the hold-out test set ($R^2 = 0.257$ vs. 0.105), and the sand-dominated IJV162-SCPT profile ($R^2 = 0.835$ vs. 0.825), while reducing the coefficient of variation (CoV) of the predicted-to-measured V_s ratio from 0.099 to 0.060 in the layered IJV038-SCPT profile. Compared with the lower predictive performance reported for MLR by Zhang et al. (2026),

QVR incorporates first-order and feature-wise quadratic terms that allow nonlinear curvature in the transformed CPTu-Vs relationship to be represented explicitly. The QVR achieved R^2 and mean absolute error (MAE) values comparable to those of the nonlinear ML models (DNN, SVR, and XGBoost) across the evaluation datasets in Case 1 and the Stuyts training set in Case 2. These results show that the study-specific QVR provides a transparent closed-form predictive formulation with competitive performance relative to the evaluated nonlinear ML models under several test conditions. Its additive quadratic form captures feature-wise curvature, although the mixed performance for IJV038 and the cross-dataset performance differences indicate that important dataset-specific effects remain unresolved.

Third, when transferred from the North Sea training compilation to the D1 subsets in Case 1, all four models using robust tanh inputs maintained positive R^2 values (0.211–0.261), CoV values of 0.168–0.267, and mean prediction ratios of $\mu = 1.013$ –1.140. Whereas Zhang et al. (2026) reported severe prediction divergence and negative R^2 values for conventional ML models at unseen sites, the absence of severe divergence under the bounded representation in the present study is consistent with a potential stabilizing role of robust tanh scaling under distributional shift. However, the present experimental design does not isolate this effect from differences in training data, tuning procedures, or dataset composition.

Moreover, whereas the baseline models in Zhang et al. (2026) used an 80:10:10 split and Bayesian optimization, the Case 2 models in this study were trained solely on the training split via internal cross-validation, omitting a separate validation set. Despite this simplified setup, evaluation on the common D1 Test split showed that DNN ($R^2 = 0.570$) and XGBoost ($R^2 = 0.622$) maintained R^2 values comparable to the published baselines, while SVR yielded a higher R^2 (0.572 vs. 0.540). All three models also produced modestly lower MAE values (SVR: 45.70 to 44.50 m/s; DNN: 46.40 to 44.26 m/s; XGBoost: 43.50 to 43.26 m/s). Because the training and tuning protocols were not identical, these differences are interpreted as observed performance differences rather than an isolated causal effect of robust tanh normalization. The model-specific responses differed, but the present experiments do not isolate the mechanisms responsible for these differences.

Fourth, the results identify performance differences between the additive QVR and the evaluated nonlinear models within the two compiled datasets. On D1 Test, which includes higher-velocity materials, DNN and XGBoost achieved higher predictive R^2 than QVR; however, the present results are insufficient to define general model-selection rules by soil type.

Within the heterogeneous global dataset (D1 Test in Case 2), which includes cemented materials such as chalk, the more complex nonlinear ML models showed higher predictive performance than the QVR ($R^2 = 0.570\text{--}0.622$ vs. 0.414). In addition, the asymmetric performance degradation observed when transferring the D1-trained models to the small North Sea hold-out test set shows that robust tanh normalization does not eliminate dataset-dependent transfer limitations, which may reflect differences in geological setting, stress history, measurement conditions, sampling, and database composition.

Consequently, the QVR equation calibrated in this study should not yet be regarded as a universally transferable model. Nevertheless, its competitive performance in the North Sea dataset and its retained predictive capability in the cross-dataset D1 evaluation indicate that the formulation is not limited to within-dataset fitting alone. Broader generalizability remains to be established through strictly site-independent external validation. Overall, the North Sea-calibrated QVR should be regarded as a transparent, low-complexity predictive model with demonstrated regional performance and competitive cross-dataset behavior under the evaluated conditions, while strictly site-independent validation remains necessary before broader or universal transferability can be claimed.

Reference

- Andrus, R. D., Mohanan, N. P., Piratheepan, P., Ellis, B. S., & Holzer, T. L. (2007). Predicting shear-wave velocity from cone penetration resistance. In Proceedings of the 4th International Conference on Earthquake Geotechnical Engineering (Paper No. 1454).
- Campanella, R. G., & Stewart, W. P. (1992). Seismic cone analysis using digital signal processing for dynamic site characterization. *Canadian Geotechnical Journal*, 29(3), 477–486. <https://doi.org/10.1139/t92-052>
- Entezari, I., Sharp, J., & Mayne, P. W. (2022). A data-driven approach to predict shear wave velocity from CPTu measurements. In Proceedings of the 5th International Symposium on Cone Penetration Testing (CPT'22) (pp. 374–380). CRC Press. <https://doi.org/10.1201/9781003308829-51>
- Guimerà Cuevas, F., & Schmid, H. (2024). Robust non-linear normalization of heterogeneous feature distributions with adaptive tanh-estimators. In S. Dasgupta, S. Mandt, & Y. Li (Eds.), Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (Vol. 238, pp. 406–414). PMLR. <https://proceedings.mlr.press/v238/guimera-cuevas24a.html>
- Hegazy, Y. A., & Mayne, P. W. (2006). A global statistical correlation between shear wave velocity and cone penetration data. In Site and geomaterial characterization (Geotechnical Special Publication No. 149, pp. 243–248). American Society of Civil Engineers. [https://doi.org/10.1061/40861\(193\)31](https://doi.org/10.1061/40861(193)31)
- Lin, H.-M., & Lyu, J.-J. (2025). A holistic framework for intradialytic hypotension prediction using generative adversarial networks-based data balancing. *BMC Medical Informatics and Decision Making*, 25, Article 257. <https://doi.org/10.1186/s12911-025-03094-5>
- Mayne, P. W., & Rix, G. J. (1993). $G_{max} - q_c$ relationships for clays. *Geotechnical Testing Journal*, 16(1), 54–60. <https://doi.org/10.1520/GTJ10267J>
- Mayne, P. W., & Rix, G. J. (1995). Correlations between shear wave velocity and cone tip resistance in natural clays. *Soils and Foundations*, 35(2), 107–110. https://doi.org/10.3208/sandf1972.35.2_107
- McGann, C. R., Bradley, B. A., & Jeong, S. (2018). Empirical correlation for estimating shear-wave velocity from cone penetration test data for Banks Peninsula loess soils in Canterbury, New Zealand. *Journal of Geotechnical and Geoenvironmental Engineering*, 144(9), 04018054. [https://doi.org/10.1061/\(ASCE\)GT.1943-5606.0001926](https://doi.org/10.1061/(ASCE)GT.1943-5606.0001926)
- Peuchen, J., Kaltekis, K., Klein, M., Murali, M., van Erp, F. C. W., & Hicks, M. A. (2022). Characteristic values for geotechnical design of offshore monopiles in sandy soils—Case study. In Proceedings of the 4th International Symposium on Frontiers in Offshore Geotechnics (pp. 1892–1901).
- Rix, G. J., & Stokoe, K. H., II. (1991). Correlation of initial tangent modulus and cone penetration resistance. In A.-B. Huang (Ed.), *Calibration chamber testing* (pp. 351–362). Elsevier.
- Robertson, P. K. (2009). Interpretation of cone penetration tests—A unified approach. *Canadian Geotechnical Journal*, 46(11), 1337–1355. <https://doi.org/10.1139/T09-065>

Robertson, P. K., & Campanella, R. G. (1983). Interpretation of cone penetration tests. Part II: Clay. *Canadian Geotechnical Journal*, 20(4), 734–745. <https://doi.org/10.1139/t83-079>

Stuyts, B., Weijtjens, W., Sastre Jurado, C., Devriendt, C., & Kheffache, A. (2024). A critical review of cone penetration test-based correlations for estimating small-strain shear modulus in North Sea soils. *Geotechnics*, 4(2), 604–635. <https://doi.org/10.3390/geotechnics4020033>

Tonni, L., & Simonini, P. (2013). Shear wave velocity as function of cone penetration test measurements in sand and silt mixtures. *Engineering Geology*, 163, 55–67. <https://doi.org/10.1016/j.enggeo.2013.06.005>

Wride, C. E., Robertson, P. K., Biggar, K. W., Campanella, R. G., Hofmann, B. A., Hughes, J. M. O., Küpper, A., & Woeller, D. J. (2000). Interpretation of in situ test results from the CANLEX sites. *Canadian Geotechnical Journal*, 37(3), 505–529. <https://doi.org/10.1139/t00-044>

Zhang, Y., Marín-Moreno, H., & Gourvenec, S. (2026). Advancing Vs estimation from CPTu for engineering practice: A data-driven approach. *Soil Dynamics and Earthquake Engineering*, 201, 109972. <https://doi.org/10.1016/j.soildyn.2025.109972>