Proceedings of the 21st ICSMGE, Vienna, Austria, 14 – 19 June 2026. Pistrol, Adam & Schweiger (eds.)
Published by: ÖGG, Austrian Society for Geomechanics, Salzburg, Austria, ISBN 978-3-9503898-4-5
https://doi.org/10.53243/ICSMGE2026-1053

A data-driven framework for constructing engineering soil profile

Mingzhou Li
Shanghai Engineering Research Center of Geotechnical Test for Underground Space, SGIDI Engineering Consulting (Group) Co., Ltd., Shanghai, China, limingzhou@sgidi.com

Shifei Yang
Shanghai Engineering Research Center of Geotechnical Test for Underground Space, SGIDI Engineering Consulting (Group) Co., Ltd., Shanghai, China

Jeenshang Lin
Department of Civil and Environmental Engineering, University of Pittsburgh, Pittsburgh, USA

ABSTRACT: Defining the engineering soil profile is a crucial component of site characterization. A three-stage data-driven framework is proposed for deriving stratification using multiple data sources including Cone Penetration Tests (CPT) and other soil properties. We train the model with only the assigned stratification label without resorting to CPT-based empirical classification relationships. At the core of our framework is a gradient-aware attention mechanism implemented in a TabTransformer-based neural network. In the first stage, sparse soil properties are interpolated along the depth of each borehole using a Histogram Gradient Boosting Regressor (HGB) coupled with Gaussian Process (GP) via a composite objective that balances fit and restraint. In the second stage, the spatial dependencies of soil properties across large geographic areas are captured and modeled by integrating HGB via a 3D Approximate GP with Automatic Relevance Determination (ARD) to adapt to anisotropic smoothness. In the third stage, the framework uses a gradient-aware TabTransformer-based classification algorithm with positional encoding to identify lithological transitions in CPT soundings with subsurface constraints. Each fused sounding is treated as a sequential row, allowing the Transformer's multi-head attention mechanism to capture depth-wise patterns. We applied the approach to geotechnical data collected from 1,098 geotechnical investigation projects in Shanghai city over the last 25 years. This dataset we considered includes 14,340 CPT and soil properties from 10,720 boreholes. Comparisons with manual interpretation confirm model's high accuracy, with overall precision, recall, and f1-score around 0.88 on the evaluation projects. Among our findings is that our framework is able to use multiple data sources to produce stratification results accurately and identify soil layer boundaries closely aligned with manual interpretations. Among the challenges we identified is that the nature of transition zones resulting in over-segmentation, which is partially solved by a gated-gradient module that enhances the attention mechanism.

KEYWORDS: Site characterization, soil stratification, Gaussian Process, gated-gradient module, TabTransformer.

1 INTRODUCTION

Stratifying the subsurface soil is fundamental to geotechnical site characterization for safe and economical design of underground structures. In practice, cone penetration test (CPT) tip resistance and sleeve friction are mapped to lithological classes by soil behavior type (SBT) charts. Since those charts are calibrated globally, misclassification is common when applied to sites with different stress histories (Mayne, 2014). SBT methodology relies only on CPT data, and laboratory-derived properties are patched with the CPT-based stratification manually in practice, introducing subjectivity and consuming substantial manual effort (Rauter & Tschuchnigg 2021). Recent works address part of this gap with machine-learning and Gaussian processes (GPs) for probabilistic site characterization (Wang et al. 2019; Zinas et al. 2025). Although GPs capture local spatial variability (Williams, 2006), inference scales cubically with the number of points, making 3-D interpolation over dense CPT grids computationally expensive (Wilson, 2015). Classical ML models efficiently learn large-scale trends but tend to over-smooth sharp interfaces, interpreting true lithologic changes as noise (Deng et al. 2018). Neither approach alone simultaneously captures the local jump as well as remains computationally tractable. Moreover, soils are natural geomaterials whose properties reflect their depositional and post-depositional history, stress path, and environmental changes; consequently, they commonly exhibit pronounced vertical heterogeneity at the site scale. CPT-only classifications frequently display over-segmentation so that post-editing is required before engineering use. Few works address an end-to-end mapping between CPT data fused with laboratory information nearby and the final stratification results involving the local experience of post-editing.

To address these limitations, this study first interpolates soil properties in a way that captures both global trends and local fluctuations, and then learns from regional stratification practice to produce per-hole stratifications at site scale. The authors collected a dataset of 1,098 geotechnical investigation sites in Shanghai, comprising 14,340 boreholes and 10,720 CPT soundings (Figure 1, left). Figure 1 (right) shows the layout of CPT locations and boreholes at a typical geotechnical investigation site. The city-scale dataset covers almost all geomorphological categories in Shanghai as shown in Figure 2. Figure 3 (a) illustrates a typical single-bridge CPT widely used in Shanghai with the specific penetration resistance Ps sampled every 0.1 m along depth. Figure 3 (b) shows the sparse laboratory measurements used in this study, including water content (w), compression modulus (E), density (γ), direct shear cohesion (c) and friction angle (φ). The colored zones in Figure 3 (a) represent the human-interpreted stratification and serve as the supervision labels for training.

Figure 4 summarizes the workflow including three stages. First, sparse laboratory measurements are enriched vertically at boreholes. HGB regressor fits the deterministic trend, coupled with an exact GP that models the residual using a composite objective of the optimization. Second, a HGB coupled with a sparse variational GP (SVGP) generalizes these soil properties to CPT positions. Third, a gradient-aware TabTransformer with sinusoidal positional encoding and gated-gradient module is trained to stratify the underground soil. A window-based depth accuracy metric confirms its superior boundary localization. GP variances are propagated to stratification inference by Monte

Carlo, forming a depth-wise label distribution and providing both the predictions and uncertainty.

Figure 1. Plan view of all the in-situ investigation holes in Shanghai (left) used and CPT (green) - borehole (yellow) layout at a typical geotechnical investigation site (right).

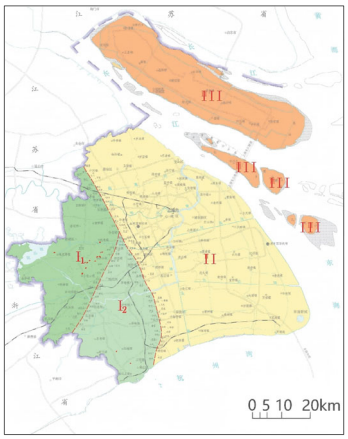

Figure 2. Geomorphological zones of Shanghai (I_1: Lake-Marshy Plain I_1; I_2: Lake-Marshy Plain I_2; II_1: Coastal Plain; II_2: New Coastal Plain; III: Estuarine-Spit Sandy Island).

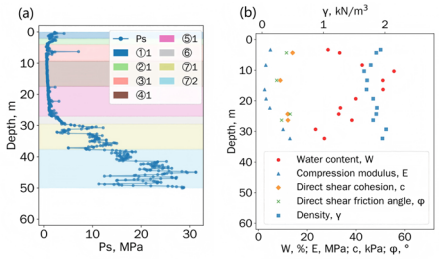

Figure 3. An example of (a) CPT sounding with colored zones represents manual interpretation of stratifications and (b) soil property measurement from a borehole.

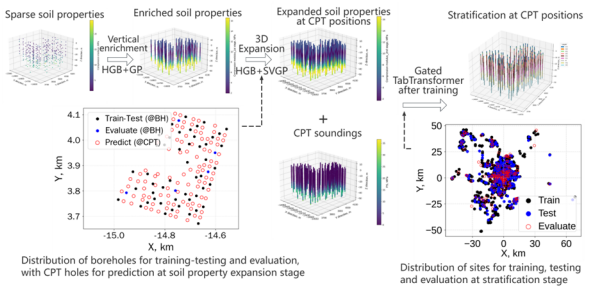

Figure 4. Diagram of the proposed workflow for stratification prediction.

2 APPROACH

The first stage models each soil property (w, E, γ, c, φ) as a deterministic trend plus a correlated residual and fills every borehole on a 0.1 m depth grid. The large-scale trend is fit using HGB, which is tuned by Bayesian optimization on a composite objective $\mathcal{L}_{\text{trend}} = -J$, where $J = \mathrm{R}^2_{\text{trend}} \cdot \frac{\text{Var(residual)}}{\text{Var(trend)}}$. The first term measures how well the trend predicts and the second term measures how much variation is left for the GP. Maximizing J (equivalently minimizing $\mathcal{L}_{\text{trend}} = -J$) simultaneously encourages accuracy and penalizes overfitting on sparse data, leaving meaningful variability for the stochastic residual model. After removing the trend given by HGB, residual is modeled using Gaussian Process (GP). Given inputs $\mathbf{X}$, GP prior $p(\mathbf{f}|\mathbf{X}) = \mathcal{N}(\mathbf{f}; \boldsymbol{\mu}, \mathbf{K})$ places a multivariate normal distribution over latent function values $\mathbf{f}$ with mean $\boldsymbol{\mu}$ and covariance $\mathbf{K}$. We use a scaled Matérn kernel with its length-scale l sets the spatial correlation range and smoothness ν controls the field roughness. With i.i.d. Gaussian observation noise, observations differ from latens as $\mathbf{y} = \mathbf{f} + \boldsymbol{\varepsilon}$ where $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma_n^2\mathbf{I})$. Integrating out $\mathbf{f}$ in the likelihood $p(\mathbf{y}|\mathbf{f}) = \mathcal{N}(\mathbf{y}; \mathbf{f}, \sigma_n^2\mathbf{I})$ yields the marginal log likelihood (MLL) $\log p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}) = \log \int p(\mathbf{y}|\mathbf{f}) p(\mathbf{f}|\mathbf{X}) df$, which measures how well $\mathbf{f}$ explain $\mathbf{y}$ under the chosen hyperparameters, $\boldsymbol{\theta}$ (l, ν, σ_n^2). In practice, it is implemented by equivalently minimizing the closed-form negative MLL $\log p(\mathbf{y}|\mathbf{X}) = -\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu})^T(\mathbf{K} + \sigma_n^2\mathbf{I})^{-1}(\mathbf{y} - \boldsymbol{\mu}) - \frac{1}{2}\log|\mathbf{K} + \sigma_n^2\mathbf{I}| - \frac{n}{2}\log(2\pi)$, where the quadratic term penalizes low probability on $\mathbf{y}$ under the model and the log-determinant penalizes complexity. Combining HGB (Figure 5 a) and GP (Figure 5 b) gives final predicted mean and the uncertainty (Figure 5, c). Metrics (Table 1) on the held-out 10% set validate the model in delivering the vertical distributions of soil properties close to the laboratory measurements.

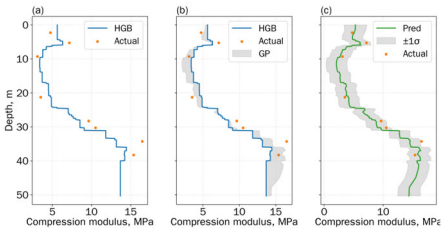

Figure 5. Vertical enrichment at a representative borehole. (a) HGB trend versus observations; (b) HGB + GP posterior mean versus observations; (c) final posterior mean with $\pm 1\sigma$ uncertainty versus observations.

Table 1. Evaluation metrics for vertical enrichment of laboratory measurements.

Laboratory measurement	R^2	RMSE	MAE
Water content	0.87	3.26	2.46
Density	0.89	0.04	0.03
Compression modulus	0.88	1.76	1.16
Direct shear cohesion	0.89	2.54	1.58
Direct shear friction angle	0.87	2.21	1.55

After vertical interpolation within boreholes, the second stage expands them to CPT positions via a similar trend-residual framework. Exact GP inference scales as $O(N^3)$ in time and $O(N^2)$ in memory. Therefore, we use SVGP to employ a set of $m \ll n$ inducing points $\mathbf{X}_m = \{x_j\}_{j=1}^m$ with inducing values $\mathbf{f}_m = [f(x_1), \ldots, f(x_m)]^{\mathrm{T}}$ to summarize the full dataset. K-means clustering yields m centroids that serve as inducing inputs. A GP prior is placed over $\mathbf{f}_m$ gives $p(\mathbf{f}_m) = \mathcal{N}(0, \mathbf{K}_{mm})$, where $\mathbf{K}_{mm} \in \mathrm{R}^{\mathrm{m}\times\mathrm{m}}$. Conditioning GP prior over training latents $\mathbf{f}$ on $\mathbf{f}_m$ gives the augmented conditional prior $p(\mathbf{f}|\mathbf{f}_m) = \mathcal{N}(\mathbf{K}_{nm}\mathbf{K}_{mm}^{-1}\mathbf{f}_m, \mathbf{K}_{nn} - \mathbf{Q}_{\mathrm{nn}})$, where $\mathbf{Q}_{\mathrm{nn}} = \mathbf{K}_{nm}\mathbf{K}_{mm}^{-1}\mathbf{K}_{mn}$ represents a Nyström approximation, i.e., the full covariance $\mathbf{K}_{nn}$ is replaced by a low-rank projection. Instead of inferring the posterior of $\mathbf{f}$ at all inputs, SVGP uses

an augmented variational posterior $q(\mathbf{f}, \mathbf{f}_m) \triangleq p(\mathbf{f}|\mathbf{f}_m)\, q(\mathbf{f}_m)$ to approximate the augmented true posterior $p(\mathbf{f}, \mathbf{f}_\mathrm{m}|\mathbf{y})$, where $q(\mathbf{f}_m)$ is a variational Gaussian distribution with its own mean $\boldsymbol{\mu} \in \mathbb{R}^\mathrm{m}$ and covariance matrix $\mathbf{A} \in \mathbb{R}^{\mathrm{m}\times\mathrm{m}}$. Variational parameters $(\boldsymbol{\mu}, \mathbf{A}, \mathbf{X}_m)$ are optimized via minimizing Kullback-Lerbler (KL) divergence between the augmented true posterior $p(\mathbf{f}, \mathbf{f}_\mathrm{m}|\mathbf{y})$ and the variational posterior $q(\mathbf{f}, \mathbf{f}_\mathrm{m})$, which is equivalent to maximizing the Evidence Lower Bound (ELBO), the objective of SVGP training

$$\mathcal{L}_{SVGP} = \log[\mathcal{N}(\mathbf{y}|0, \mathbf{Q}_{nn} + \sigma^2\mathbf{I})] - \frac{1}{2\sigma^2}\mathbf{Tr}(\widetilde{\mathbf{K}}_{nn}) \qquad (1)$$

where $\widetilde{\mathbf{K}}_{\mathrm{nn}} = \mathbf{K}_{\mathrm{nn}} - \mathbf{Q}_{\mathrm{nn}}$ represents the portion of covariance that the inducing set fails to explain. The first term evaluates how well the low-rank GP explains the data. The trace penalty discourages leaving covariance unexplained by $\mathbf{X}_m$, ensuring that the model is not overly smooth. Inducing points therefore summarize the dataset. The GP infers values at $\mathbf{X}_m$ and propagates them to all CPT locations. SVGP is implemented in GPyTorch using ApproximateGP with Cholesky Variational Distribution. Parameterizing the covariance as $\mathbf{A} = \mathbf{S}\mathbf{S}^\mathrm{T}$ ensures positive definiteness. A small noise floor and kernel diagonal jitter is imposed for the numerical stability during factorization. A scaled Matérn kernel with ARD is used to adapt to varying spatial smoothness across directions. For large-scale training, we adopt mini-batch optimization that computes only the batch–inducing covariances to save GPU memory. Using GPyTorch's VariationalELBO, the data-fit term is estimated per batch, while the KL term over $\mathbf{f}_\mathrm{m}$ is batch-invariant and can be amortized to improve efficiency. Parameters $(\boldsymbol{\mu}, \mathbf{A}, l, \nu, \sigma_n^2)$ are optimized jointly by back-propagation until the ELBO stabilizes. Figure 6 illustrates the interpolation of a water-content profile at an evaluation hole with the posterior mean close to the measurements. The uncertainty rises at deep depths due to the sampling sparsity as shown in Figure 6 (c). Figure 7 indicates the interpolation along all evaluation holes visually validating the model in expanding the enriched laboratory measurements, indicating that the SVGP brings more horizontal variations.

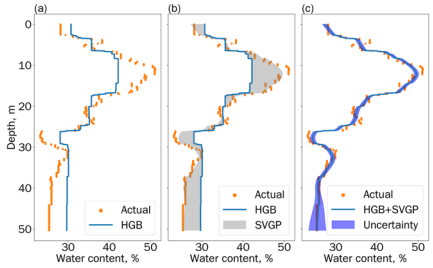

Figure 6. Interpolation results of water content at an evaluation hole. (a) HGB trend versus observations; (b) HGB + SVGP posterior mean versus observations; (c) final posterior mean with $\pm 1\sigma$ uncertainty versus observations.

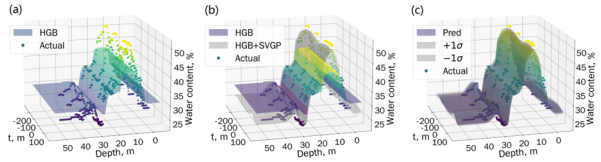

Figure 7. Interpolation results of water content across all evaluation holes. (a) HGB trend versus observations; (b) HGB + SVGP posterior mean versus observations; (c) final posterior mean with $\pm 1\sigma$ uncertainty band versus observations.

For the stratification task in the third stage, we propose a gradient-aware TabTransformer (Figure 8) that improves boundary detection by emphasizing the depth-wise gradients of fused soil parameters. Each depth point is treated as a token here. Let $(\mathbf{X}, \mathbf{y})$ denote a sequence of feature-label pairs, for a fused sounding with l depth points, we write $\mathbf{X} = [\mathbf{x}_1, \ldots \mathbf{x}_l]^\mathrm{T}$, where each token concatenates 6 raw features (CPT tip resistance P_s and five laboratory-derived soil properties) and their first-order depth gradients, i.e., $\mathbf{x}_i = [\mathbf{x}_i^{raw}; \mathbf{x}_i^{raw}] \in \mathbb{R}^{2n}$ with $n = 6$. At depth i, these two are projected via separate linear maps into a shared embedding space to form $\mathbf{e}_i^{raw}, \mathbf{e}_i^{grad} \in \mathbb{R}^d$ $(d = 128)$, which are concatenated to $\mathbf{e}_i \in \mathbb{R}^{2d}$. A gate vector is computed via a linear transform $(\mathbf{w}_{gate} \in \mathbb{R}^{2d\times d}, \mathbf{b}_{gate} \in \mathbb{R}^d)$ as

$$\mathbf{g}_i = \sigma(\mathbf{e}_i\mathbf{w}_{gate} + \mathbf{b}_{gate}) \qquad (2)$$

where σ is the element-wise sigmoid. This generated gate controls the contribution of gradient information when forming the gradient-aware embedding $\mathbf{e}_i = \mathbf{e}_i^{orig} + \mathbf{g}_i \odot \mathbf{e}_i^{grad}$ with $\odot$ denoting the element-wise product. Stacking per-depth embeddings yields the sequence $\mathbf{E} = [\mathbf{e}_1, \ldots, \mathbf{e}_l] \in \mathbb{R}^{L\times 2d}$. A sinusoidal positional encoding encodes the depth order, making the model know token positions along the CPT profile. For token i, the encoded token is $\tilde{\mathbf{e}}_i = \mathbf{e}_i + \mathbf{p}_i$, with $\mathbf{p}_{i,2j} = sin(\frac{i}{10000^{2j/d}})$ and $\mathbf{p}_{i,2j+1} = cos(\frac{i}{10000^{2j/d}})$ for $j = 0, \ldots, d/2 - 1$. This analytic encoding extrapolates naturally to longer sequences (greater depth). Moreover, for any frequency w_k, the positional encoding at $t + \phi$ is a rotation of that at t via $\mathbf{M}_{\phi,k} = \begin{bmatrix} \cos(w_k\phi) & \sin(w_k\phi) \\ -\sin(w_k\phi) & \cos(w_k\phi) \end{bmatrix}$. Therefore, a depth shift ϕ depends only on the offset rather than the absolute index, helping the model recognize recurring boundary patterns at different depths. We alter the number of Transformer encoder layers (N) and self-attention heads (h) to determine the optimal architecture for the Transformer model. Although deeper models with $N = 8$ have a larger theoretical capacity, in our current experiments they do not outperform the 4-layer configuration (Table 2). Deeper post-norm Transformers are harder to optimize and require tailored training strategies including smaller learning rates, stronger regularization and longer training to fully exploit their capacity, which is left for future work. We therefore adopt a moderately deep model with $N = 4$ and $h = 8$ which provides a good balance between stratification accuracy and computational cost on this dataset, using $d_{\mathrm{ff}} = 2048$ for the feed-forward sublayer following the standard practice. With input $\mathbf{H}^{(0)} = \tilde{\mathbf{E}} \in \mathbb{R}^{L\times d}$, the k-th encoder layer produces the contextualized representations $\mathbf{H}^{(k)} = f_{tf,\theta}(\mathbf{H}^{(k-1)})$. In head j, $\mathbf{H}^{(k-1)}$ is linearly projected to queries, keys, values as $\mathbf{Q}_j \in \mathbb{R}^{d\times d_k}$, $\mathbf{K}_j \in \mathbb{R}^{d\times d_k}$, and $\mathbf{V}_j \in \mathbb{R}^{d\times d_v}$ with $d_k = d_v = \frac{d}{h} = 16$. The head output is

$$\mathbf{head}_j = softmax\left(\frac{\mathbf{Q}_j\mathbf{K}_j^\mathrm{T}}{\sqrt{d_k}}\right)\mathbf{V}_j \in \mathbb{R}^{l\times d_k}, \qquad (3)$$

where $\mathbf{Q}_j\mathbf{K}_j^\mathrm{T}$ gives all pairwise similarities between depths, scaled by $\sqrt{d_k}$ and converted into weights via softmax. For each token, the attention forms a weighted mixture of the most informative depths, yielding a representation enhanced by depth context. Near a boundary it emphasizes contrasting tokens above and below, while within a homogeneous layer it concentrates on similar tokens to denoise. Gating gradients in embedding-space further enhances this by amplifying genuine alternation cues at true boundaries and suppresses gradient noise inside layers. All heads are concatenated and projected back to the model dimension and nonlinearly remapped through a token-wise feed forward network $FFN\left(\mathbf{h}_i^{(k)}\right) = ReLU\left(\mathbf{h}_i^{(k)}\mathbf{W}_1 + \mathbf{b}_1\right)\mathbf{W}_2 + \mathbf{b}_2$ to deliver higher-level

attributes that align with soil classes. Each sublayer is wrapped in the residual connection and layer normalization for stable training. The final token encodings $\mathbf{h}_i^{(N)}$ are fed into an MLP classifier to produce logits $\boldsymbol{\ell}_i = MLP(\mathbf{h}_i^{(N)}) \in \mathbb{R}^C$ and the sequence logits $\boldsymbol{\ell} \in \mathbb{R}^{L\times C}$. Training minimizes cross-entropy over unpadded tokens to find best hyperparameters. At inference, uncertainty is quantified via MC perturbations with 300 stochastic passes, injecting gaussian noise into the soil features. For pass m at depth i, network outputs $\boldsymbol{\ell}_{i,m} \in \mathbb{R}^C$ and the probability for class c as $p_{i,c,m} = \mathrm{softmax}(\boldsymbol{\ell}_{i,m})_c$. After MC perturbations, the model produces for each depth i a class-probability vector $\mathbf{p}_i \in \mathbb{R}^C$ with components $p_{i,c} = \frac{1}{M}\sum_{m=1}^{M} p_{i,c,m}$. The predicted label $\hat{y}_i$ at depth i is the class with the highest probability. To quantify the likelihood of a boundary between adjacent depths i and $i-1$, we define the same-layer probability between depth i and $i-1$ as $p_{same}(i) = \sum_{c=1}^{C} p_{i,c}p_{i-1,c} = \mathbf{p}_i^{\mathrm{T}}\mathbf{p}_{i-1}$. The boundary probability is then computed by taking the complement of the same-layer probability as $p_{\mathrm{bnd}}(i) = 1 - p_{\mathrm{same}}(i) = 1 - \mathbf{p}_i^{\mathrm{T}}\mathbf{p}_{i-1}$, serving as a confidence indicator for a layer transition. Figure 9 visualizes 1,500 tokens using t-SNE plots for the input embeddings (Figure 9, left), final encoder outputs (Figure 9, middle), and MLP logits (Figure 9, right), using Monte Carlo averaged representations. From left to right, the clusters become progressively tighter and better separated, making classes easier to distinguish and visually validating the proposed framework in the classification aspect.

Table 2. Stratification performance evaluations of various structures of the Transformer model.

N	h	Weighted precision	Weighted recall	Weighted f1-score
2	2	0.833	0.833	0.830
	4	0.846	0.851	0.854
	8	0.862	0.858	0.859
4	2	0.863	0.862	0.861
	4	0.873	0.866	0.868
	8	0.879	0.875	0.876
8	2	0.869	0.863	0.864
	4	0.869	0.863	0.865
	8	0.870	0.864	0.865

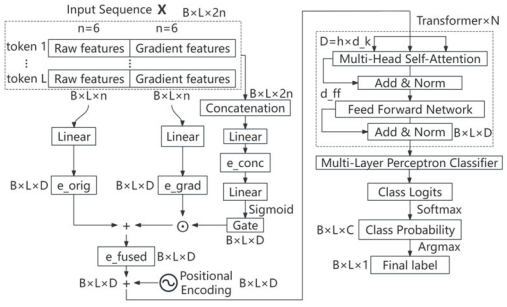

Figure 8. Architecture of the gradient-aware TabTransformer.

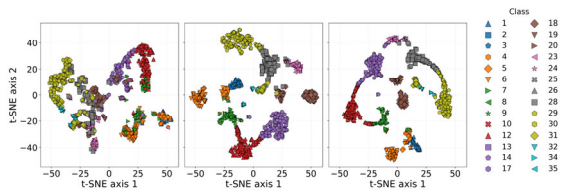

Figure 9. T-SNE plots before TabTransformer (left), after the final Transformer encoder (middle) and MLP (right).

3 RESULTS

3.1 *Influence of the number of inducing points on SVGP interpolation accuracy (experiment 1)*

The effect of the number of inducing points M on the SVGP performance is investigated by fitting the model on a same dataset with $M = \{400, 600, 800, 1000, 1200\}$. The data-splitting strategy is aligned with the intended application of the model in the current stage, which is to expand borehole soil property to CPT positions. To be consistent with the intended workflow, the dataset is split at the borehole level into training, testing, and evaluation subsets with a 70–20–10% ratio. Table 3 summarizes the depth-averaged metrics in this stage. Negative log-likelihood (NLL) of laboratory measurement y_i under the predictive posterior (with mean $\hat{y}_i$ and standard deviation σ_i) at each depth point i is computed as $\mathrm{nll}_i = -\log p\left(y_i \middle| \hat{y}_i, \sigma_i^2\right) = \frac{1}{2}\left[\log\left(2\pi\sigma_i^2\right) + \frac{(y_i-\hat{y}_i)^2}{\sigma_i^2}\right]$. Lower NLL indicates better SVGP predictive performance. Whereas RMSE and MAE assess only the central prediction and ignore uncertainty, NLL penalizes both the large residuals $|y_i - \hat{y}_i|$ and the over-confident variances σ_i^2, aligning with SVGP's ELBO-based training objective. To further illustrate how SVGP fits soil-property profiles, we use box-plots to better analyze NLL (Figure 10) and show interpolation results in an evaluation borehole using different M (Figure 11). For water content which varies gradually with depth, the predicted curves in Figure 11 (a) converge towards the true curve as M grows, which is accordant with Equation (1), where increasing M reduces the Nyström approximation residual $\tilde{\mathbf{K}}_{\mathrm{nn}} = \mathbf{K}_{\mathrm{nn}} - \mathbf{Q}_{\mathrm{nn}}$, improving the ELBO. The NLL distributions in Figure 10 (a) are consistent with these interpolations. However, more inducing points are not always better for cohesion, which is mostly smooth but exhibits a sharp jump with local variation in the clay–sand transition zone at depth 25-27 m. Model with $M = 1000$ achieves the lowest median NLL (Figure 10 b) and captures the local variation (Figure 11 b), which better serves stratification needs. In this case, inducing centroids fall on both sides of the discontinuity, allowing the Nyström approximation $\mathbf{Q}_{\mathrm{nn}} = \mathbf{K}_{\mathrm{nm}}\mathbf{K}_{\mathrm{mm}}^{-1}\mathbf{K}_{\mathrm{mn}}$ to capture a steep gradient. $\mathbf{Tr}\left(\tilde{\mathbf{K}}_{nn}\right)$ remains non-negligible and continues to penalize over-fit. If M is too small, the local variation is missed and NLL increases. If M is too large, $\tilde{\mathbf{K}}_{\mathrm{nn}}$ becomes close to zero and KL penalty vanishes. The posterior is dominated by the smooth kernel fit, flattening the local fluctuation which is crucial to stratification. This reflects a trade-off, while more inducing points reduce predictive variance, they can also over-smooth abrupt local variations. Therefore, the most appropriate M should be determined by jointly examining the NLL distribution and the prediction-vs-measurement profiles, so that the model captures the background trend while retaining key local variations. With the selected inducing points, the framework is applied to generate the spatial distribution of soil properties at CPT positions (Figure 12, right) from borehole data (Figure 12, left & middle), providing the CPT-aligned soil property features for the stratification.

Table 3. Evaluation results of five models using different numbers of inducing points.

Laboratory measurement	Metric	600	800	1000	1200	1400
Compression modulus	NLL	0.95	0.35	0.55	0.30	0.60
	RMSE	0.54	0.36	0.47	0.35	0.37
	MAE	0.55	0.25	0.38	0.20	0.26
Density	NLL	0.18	0.48	0.05	0.56	0.04
	RMSE	0.70	0.95	0.63	0.65	0.52
	MAE	0.40	0.60	0.34	0.36	0.23

Laboratory measurement	Metric	600	800	1000	1200	1400
Direct shear cohesion	NLL	0.42	0.73	0.64	0.95	0.59
	RMSE	2.05	2.45	2.15	2.30	2.25
	MAE	0.16	0.26	0.17	0.22	0.23
Direct shear friction angle	NLL	0.66	0.95	0.63	0.62	0.75
	RMSE	0.68	0.95	0.60	0.59	0.70
	MAE	0.28	0.42	0.21	0.19	0.23
Water content	NLL	0.82	0.70	0.95	0.75	0.55
	RMSE	1.12	1.04	1.20	0.99	0.89
	MAE	0.51	0.40	0.60	0.34	0.27

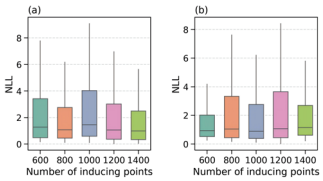

Figure 10. NLL box plots for (a) water content and (b) direct shear cohesion interpolated using different numbers of inducing points.

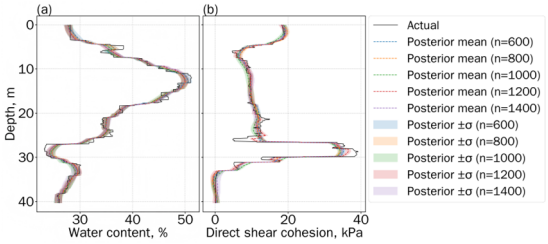

Figure 11. SVGP interpolation (dashed lines) of (a) water content and (b) direct shear cohesion using different numbers of inducing points on an evaluation hole

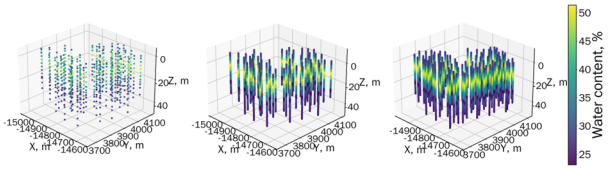

Figure 12. Spatial distribution of soil property (water content as an example) from the sparse borehole sampling (left) vertically enriched (middle) and expanded to every depth of CPT holes (right).

3.2 *Influence of gated gradient on TabTransformer stratification accuracy (experiment 2)*

Four models using different strategies of raw-gradient embedding (Figure 13) are compared with Model-0 using 6 raw features. Leave-project-out split is adopted in training process of the stratification stage, where the model seeks to generalize from the annotated projects to a new one. Evaluation in this stage is conducted based on the leaved-out 10% projects. Model-1 concatenates raw and gradients without gate. Model-2 applies the gate on the embedding space, concatenating raw and gated gradients. Model-3 applies gate on the original feature space prior to embedding. Model-4 computes the gate from raw features only. On the same held-out projects, Model-2 performs best, improving 1.0% over Model-1 (Table 4). Simply appending gradients without gating (Model-1) brings no gain. Gating outside the embedding space (Model-3) or ignoring gradient when computing the gate (Model-4) degrades the performance. For any depth d, a window-based depth accuracy (WDA) is calculated by dividing the number of label matches ($\hat{y}_i = y_i$) from all holes within the sliding window of width $w = 0.30$ m (Figure 15) by the total number of points N_d in the window as $\mathrm{WDA}(d) = \frac{1}{N_d}\sum_{i:\, z_i \in [d-w/2,\, d+w/2]} 1(\hat{y}_i = y_i)$. WDA penalizes misplaces of boundaries and assesses the localization performance. The boxplots across the held-out set (Figure 15) show the differences between models in boundary localization accuracy. Model-2 behaves better than Model-1 and achieves the highest median WDA and the tightest interquartile range, indicating the contribution of gated gradient to generating a more accurate boundary placement. Model-3 and Model-4 underperform, showing that post-projection gating captures richer cross-feature interactions and conditioning the gate on gradient embeddings is critical for precise boundary localization.

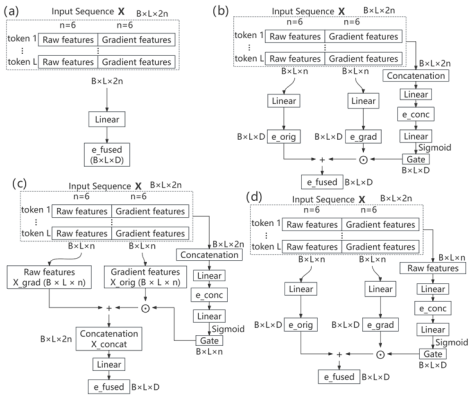

Figure 13. Diagrams of embedding via different gate strategies in (a) Model-1, (b) Model-2, (c) Model-3, and (d) Model-4.

Table 4. Evaluations of different embedding strategies.

Model	Weighted precision	Weighted recall	Weighted f1-Score
0(raw)	0.871	0.865	0.868
1(concatenation without gate)	0.871	0.866	0.867
2(gate-embedding-space)	0.879	0.875	0.876
3(gate-original-space)	0.866	0.861	0.863
4(gate-embedding-space-raw)	0.869	0.864	0.866

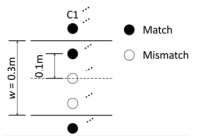

Figure 14. Sliding window for WDA computation at a CPT depth.

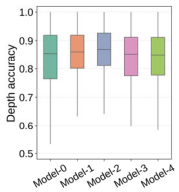

Figure 15. WDA box plots across the evaluation dataset using different models.

Predicted stratifications using Model-1 and Model-2 are plotted in evaluation CPT holes. The over-segmentation around 16 m and 34 m observed with Model-1 (Figure 16 a, left) is reduced by Model-2 (Figure 16 a, right) at an evaluation CPT hole. The remaining spurious splits may correspond to natural heterogeneity in transition zones of sandy silt and silty sand. The second example shown in Figure 16 (b) illustrates a case with layer ②2, silty clay with interbeds of clayey silt (Table 5). Although human annotations do not split these clayey silt interbeds explicitly and subsume them into ②2, the model does not follow and reveals them during prediction. Because the label set contains no dedicated class for this sublayer, the model assigns the closest learned class, ②3 as a surrogate. Compared with the ungated Model-1 (Figure 16 b, left), Model-2 exposes more of these lithologic interbeds (Figure 16 b, right). Overall,

the results indicate that the gradient-aware TabTransformer effectively suppress spurious splits while enhancing geologically supported interbeds.

Table 5. Layer codes and names involved in this experiment.

Layer code	Layer name
①1	Fill
②1	Silty clay
②2	Silty clay with interbeds of clayey silt
②3	Sandy silt
③1	Muddy silty clay
③2	Sandy silt
④1	Muddy clay
⑤1	Clay
⑤2	Sandy silt
⑥	Silty clay
⑦1	Sandy silt
⑦2	Silt

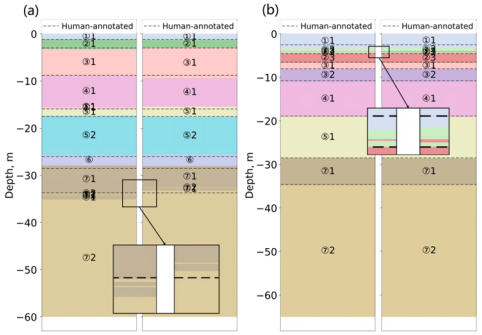

Figure 16. Influence of gated gradients on (a) reducing over-segmentation and (b) revealing lithologic interbeds.

In Figure 17, Model-1 (left three) misplaces the ④1 - ⑤1 interface by around 1.2 m, whereas model-2 (right three) reduces the offset to around 0.5 m, indicating that the gated network shifts boundary localization closer to manual interpretations. Figure 17 also compares layer probability distributions and boundary-probability envelopes via model-1 and Model-2. The gated model assigns high boundary probability closer to the manual interpretation, giving a more confident, better-localized delineation. Figure 18 compares Model-2 stratifications (middle) with human annotations (left) at an evaluation site. In this example, the predicted layers generally agree with the manual interpretation, with noticeable differences mainly occurring in layers ②1 and ②2, as shown in their WDA violin plots (Figure 18, right). Overall, the results in this chapter indicate that concatenating raw embeddings with their gated gradients sharpens the attention mechanism, linking lithologic gradient signals to layer boundaries while suppressing noise, and thus produces site-scale stratifications close to expert interpretations, with some local discrepancies.

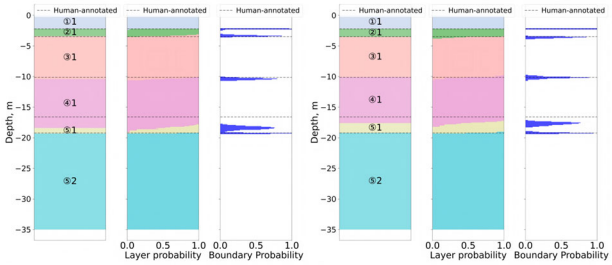

Figure 17. Comparison between Model-1 (left three) and Model-2 (right three) in the boundary localization, layer and boundary probability.

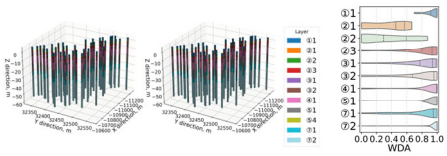

Figure 18. Site-scale comparison of manual-interpreted (left) and machine-predicted (middle) stratifications at CPT holes, with WDA violin plots by layer (right) for an evaluation project.

4 CONCLUSIONS

Our proposed framework integrates multiple data sources to deliver accurate stratifications and localize boundaries close to manual interpretations, achieving around 0.88 on both classification and boundary metrics. The algorithm thus delivers reliable site-scale stratifications at CPT holes in Shanghai using CPT sounding fused with borehole data. The framework couples HGB and SVGP via a composite optimization objective, fitting the trend and local variation respectively. It captures both smooth and genuine local fluctuations at modest computational cost. GP uncertainty is propagated via Monte Carlo to per-depth layer and boundary probability. We observe challenges in transition zones, where natural heterogeneity causes spurious splits. Concatenating raw embeddings with gated-gradient enhances attention mechanism, mitigating noise-driven over-segmentation yet revealing more lithologic interbeds that are implicit in the layer nomenclature. It also improves boundary localization assessed via WDA. Future work will develop more efficient ways of propagating GP uncertainty and optimize training strategies for deeper Transformer to better exploit their capacity. Beyond assisting stratification, coupling HGB and GP offers a potential framework for multi-scale spatial analysis of soil properties.

5 ACKNOWLEDGEMENTS

This work is funded by Shanghai Municipal Science and Technology Commission through the program, Research and Demonstration of Technologies for Digital Construction and Risk Prevention in Underground Engineering (21DZ1204303). The authors gratefully acknowledge this financial support.

6 REFERENCES

Deng, R., Shen, C., Liu, S., Wang, H. and Liu, X., 2018. Learning to predict crisp boundaries. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 562-578).

Mayne, P.W., 2014, May. Interpretation of geotechnical parameters from seismic piezocone tests. In *Proceedings, 3rd international symposium on cone penetration testing* (Vol. 102, pp. 47-73).

Williams, C.K. and Rasmussen, C.E., 2006. *Gaussian processes for machine learning* (Vol. 2, No. 3, p. 4). Cambridge, MA: MIT press.

Rauter, S. and Tschuchnigg, F., 2021. CPT data interpretation employing different machine learning techniques. *Geosciences*, *11*(7), p.265.

Wang, H., Wang, X., Wellmann, J.F. and Liang, R.Y., 2019. A Bayesian unsupervised learning approach for identifying soil stratification using cone penetration data. *Canadian Geotechnical Journal*, *56*(8), pp.1184-1205.

Wilson, A. and Nickisch, H., 2015, June. Kernel interpolation for scalable structured Gaussian processes (KISS-GP). In *International conference on machine learning* (pp. 1775-1784). PMLR.

Zinas, O., Papaioannou, I., Schneider, R., Cuéllar, P. and Baeßler, M., 2025. 3D spatial modelling of CPT data for probabilistic preliminary assessment of potential pile tip damage upon collision with boulders. In *Proceedings of the fifth International Symposium on Frontiers in Offshore Geotechnics (ISFOG 2025)* (pp. 505-510). International Society for Soil Mechanics and Geotechnical Engineering.