



Development of Subsurface Geological Cross-Section from Limited Site-Specific Boreholes and Prior Geological Knowledge Using Iterative Convolution XGBoost

Chao Shi, P.E.¹; and Yu Wang, Ph.D., P.E., F.ASCE²

Abstract: The delineation of vertical geological cross-sections is an essential task in geotechnical site characterization and has a profound impact on subsequent geotechnical designs and analyses. It is a long-lasting challenge, particularly for complex geological settings, to properly produce a subsurface geological cross-section from limited boreholes that are usually encountered in engineering practice. Emerging machine learning methods, such as the convolutional neural network (CNN), provide a fresh perspective of this challenge and effective alternatives for exploiting the complex stratigraphic relationships between different soil deposits. In this study, a novel iterative convolution eXtreme Gradient Boosting model (IC-XGBoost) is proposed. This model interpolates a subsurface geological cross-section from limited site-specific boreholes and a training geological cross-section obtained from previous projects with similar geological settings. This direct application of previous geological cross-sections for training is based on the assumption of similar local spatial connectivity or stratigraphic relationships between soils in areas with similar geological settings. The proposed method can learn stratigraphic patterns from a training image in an automatic manner. In addition, the proposed method is purely data-driven and does not require the specification of any parametric function form. The model performance is illustrated using both a simulated example and real data from a tunnel project in Australia. The proposed method not only infers the most probable geological cross-section but also quantifies the associated interpolation uncertainty from multiple realizations. The effect of the borehole number on the interpolation performance is also explicitly investigated. DOI: [10.1061/\(ASCE\)GT.1943-5606.0002583](https://doi.org/10.1061/(ASCE)GT.1943-5606.0002583). © 2021 American Society of Civil Engineers.

Author keywords: Data-driven; Machine learning; Convolutional neural network (CNN); Spatial interpolation; Sparse data.

Introduction

Subsurface stratigraphy is indispensable for geotechnical site characterization. A good understanding of the subsurface stratigraphic distribution is essential to the development of vertical geological cross-sections for subsequent geotechnical design and analysis. An inaccurate interpretation of subsurface geological profiling can increase financial risks and significantly increase the likelihood of change orders, claims, costs, and schedule overruns during construction (e.g., Boeckmann and Loehr 2016). Following an investigation of more than 300 construction contracts in the US, Prezzi et al. (2011) concluded that geotechnical change orders were mainly caused by the insufficient geotechnical investigation. A survey by Clayton (2001) revealed that more than 40% of geotechnical problems arising from underground construction in the UK were attributed to inherent variability and uncertainties in the ground and its stratigraphy. Such stratigraphic uncertainties are difficult to alleviate using current engineering practices, which rely heavily on limited boreholes in a specific site and the engineer's knowledge of the local geology for developing geological cross-sections.

In current engineering practice, a subsurface geological cross-section is developed by drawing straight lines to connect the boundaries of the same soil types between any two adjacent boreholes. This usual practice is acceptable when a site has relatively simple geology or when extensive ground investigations are available. However, complex geological processes, such as erosion and deposition, can lead to inconsistencies in the soil types and strata thicknesses revealed by adjacent boreholes. In such situations, it is difficult to infer geological cross-sections solely from these limited boreholes. Recently, Markov chain-based probabilistic models, such as the Markov random field (Li et al. 2016b) and coupled Markov chain (Elfeki and Dekking 2001; Qi et al. 2016; Li et al. 2016a), have been applied to interpolate geological profiles from boreholes. These approaches rely on the stationary transition probability for a stochastic simulation. More recently, Deng et al. (2020) used a generalized coupled Markov chain and nonstationary transition probabilities to interpolate a subsurface geological profile. However, it may still be difficult to estimate the transition probability in a practical application. Therefore, it is imperative to develop a handy and reliable tool for interpolating subsurface stratigraphic cross-sections (e.g., Li et al. 2016a; Juang et al. 2019; Wang et al. 2020).

As an example, a real tunnel project in Australia is illustrated in Fig. 1. Four vertical boreholes, labeled BH1–BH4, were collected at equal horizontal intervals of 100 m, and four different soils were revealed by these boreholes. A direct visual examination indicates that most of the soil horizons can be determined by connecting the stratigraphic boundaries separating the different soil horizons via straight lines. However, the interpretation of the Gngangara sand (GS)/Ascot formation (AF) boundaries between BH2/BH3 and BH3/BH4 is challenging because GS is revealed in both BH2

¹Ph.D. Student, Dept. of Architecture and Civil Engineering, City Univ. of Hong Kong, Tat Chee Ave., Kowloon, Hong Kong. Email: chaoshi6-c@my.cityu.edu.hk

²Professor, Dept. of Architecture and Civil Engineering, City Univ. of Hong Kong, Tat Chee Ave., Kowloon, Hong Kong (corresponding author). ORCID: <https://orcid.org/0000-0003-4635-7059>. Email: yuwang@cityu.edu.hk

Note. This manuscript was submitted on July 18, 2020; approved on April 5, 2021; published online on June 17, 2021. Discussion period open until November 17, 2021; separate discussions must be submitted for individual papers. This paper is part of the *Journal of Geotechnical and Geoenvironmental Engineering*, © ASCE, ISSN 1090-0241.

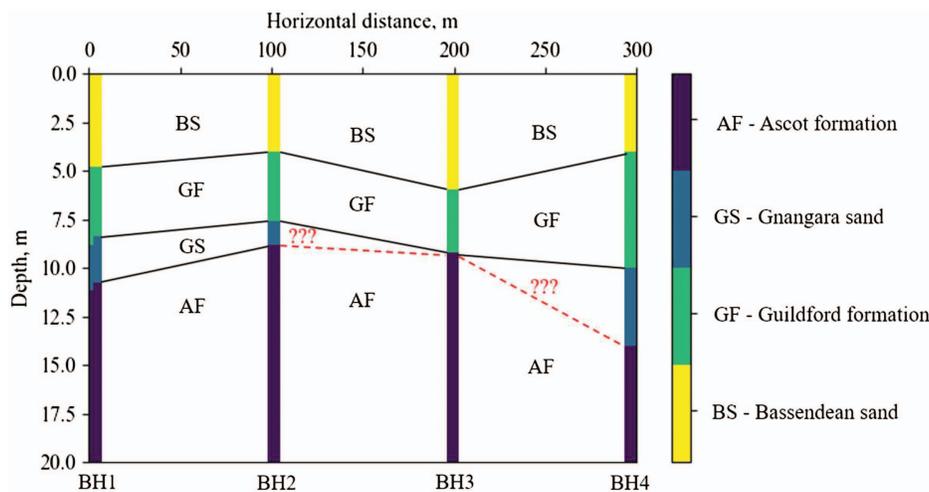


Fig. 1. (Color) Illustration of the challenge encountered in the delineation of the vertical geological cross-section for a tunnel project in Australia.

and BH4 but is missing from BH3. This inconsistency poses a challenge for determining the stratigraphic boundaries of GS. To solve this dilemma, engineers normally judge and define a typical geological cross-section by combining information from the limited site-specific boreholes and prior geological knowledge, which includes, but is not limited to, full geological cross-sections obtained from previous project sites with similar geological settings.

The previously described use of engineering judgment is essentially consistent with emerging machine learning methods, which adaptively learn intrinsic relationships from training datasets and improve automatically through experience (Mitchell 1997). Therefore, it is beneficial to consider this problem from the perspective of machine learning [e.g., the convolutional neural network (CNN)], with the dual aims of (1) improving the accuracy and efficiency of the interpretation and (2) automatically quantifying the uncertainties associated with the interpretation. Using this approach, prior geological knowledge is learned automatically from previous engineering experience (e.g., geological cross-sections from previous projects with similar geological settings), and the application of engineering judgment to the delineation of vertical geological cross-sections from limited boreholes is equivalent to the use of a well-trained machine learning model for conditional predictions.

The CNN model is a popular machine learning method and has been used successfully for image restoration or spatial interpolation when a substantial part of the image (e.g., > 90%) is available. Normally, a large (i.e., millions) set of training images is needed to train a well-performing CNN model. However, geotechnical training images, namely, geological cross-sections from previous projects with similar geological settings, are normally quite limited in quantity. Site-specific boreholes are also limited, accounting for less than 10% of a target cross-section. These limitations pose great challenges to the direct application of CNN models to geotechnical site characterization.

To address the challenges associated with limited boreholes and training images, a novel iterative convolution model for subsurface geological cross-section delineation is proposed in this study. Prior knowledge of local geology and typical stratigraphic connectivity is represented concisely in a single training geological cross-section (i.e., the proposed method only requires a single training image). Spatial stratigraphic patterns of different scales are extracted iteratively from the training cross-section and combined with the limited boreholes to infer the subsurface geological cross-sections. In this study, the training image is not simulated but is obtained

directly from previous project sites with similar geological settings. This direct application of a previous geological cross-section for training purposes is based on the assumption of a similar local spatial connectivity or stratigraphic relationships between soils in areas with similar geological settings. Training images can also be constructed using process-based models (Koltermann and Gorelick 1992), process-mimicking models (Koltermann and Gorelick 1996), or unconditional object-based simulations (e.g., Deutsch and Wang 1996). As the accuracy of a prediction relies heavily on the quality of the training cross-section, a training image that properly represents the potential stratigraphic relationships at the target site will improve the accuracy of interpolation. In contrast, a training image that is not representative of the target site might compromise the performance of the proposed method. When no detailed information about a target site is available, the training image should be selected in accordance with the engineers' knowledge of local geological formation processes.

The remainder of this study is organized as follows. In the next section, the basic architecture of the CNN model is reviewed. In the "Proposed Iterative Convolution Model" section, details of the proposed iterative convolution model for developing a geological cross-section are given. Subsequently, the process used to simulate an illustrative example is described. In the two sections that follow, the interpolation results from both the simulated example and real data from a tunnel project in Australia are discussed.

Brief Review of the Convolutional Neural Network (CNN)

The CNN is considered one of the most impressive neural network types and has been applied mainly to image-driven pattern classification and recognition (e.g., LeCun et al. 1989, 1997; Boureau et al. 2010a). A classical artificial neural network (ANN) (e.g., Kumar et al. 2000; Juang et al. 2001) directly approximates input-output relationships by training multiple layers of hidden neurons. In this context, a neuron is a mathematical function that collects and transforms input information, and multiple neurons are aligned and stacked to form an ANN. A conventional ANN involves a large number of coefficients and does not explicitly consider the intrinsic input data structure. In comparison, a CNN exploits the intrinsic shift-invariant properties of input images (Mairal et al. 2014) via multistage feature extraction and subsampling before feeding into a conventional multilayer perceptron (MLP)

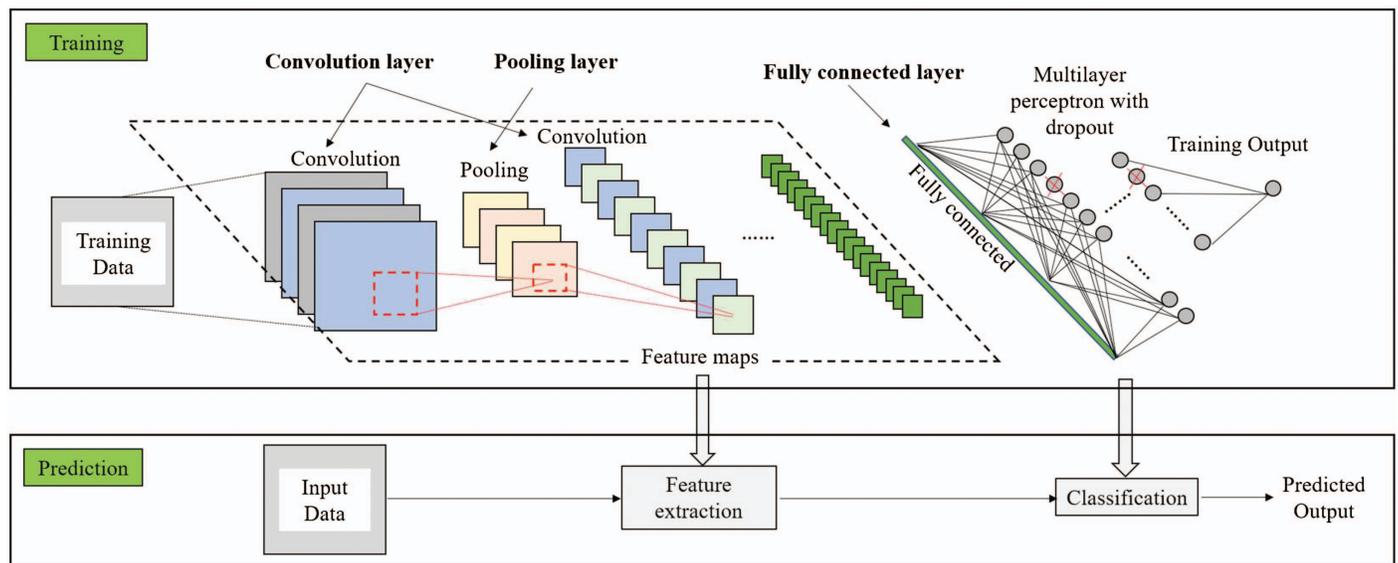


Fig. 2. (Color) Basic architecture of the convolutional neural network.

neural network for classification. In this study, shift invariance denotes the fact that intrinsic connectivity patterns between objects within an image can be successfully recognized and converted into invariant patterns in feature maps after convolution with proper filters. In other words, the local ordering of objects and patterns within an image is preserved, and nonimportant features are filtered out during subsequent sampling and pooling stages. This approach can significantly reduce the number of training parameters. CNN-based models have also been widely applied to image restoration (e.g., Pathak et al. 2016; Steffens et al. 2020).

Fig. 2 depicts the architecture of a basic CNN, which can be split into two separate stages: training and prediction. The training stage mainly comprises the convolution, pooling, and fully connected layers. All of the intermediate convolution and pooling layers are organized in a hierarchical manner to produce feature maps. Primarily, these convolution and subsampling layers are used to develop image representations that are invariant to a particular convolution (Mairal et al. 2014). The extracted features are then flattened into a fully connected layer for subsequent classification. A classifying mechanism based on a CNN model is set up when all parameters are determined using training data (i.e., a known input and known output). Any new input data can then be classified (i.e., prediction) using the well-trained feature extraction and classification components from the training stage.

Convolution Layer

Input training data in the form of images or sound clips produce a multidimensional matrix, with each entry representing a pixel intensity. For example, a single feature map in the first convolution layer can be obtained by sliding a filter, such as a 3×3 matrix of weights, across all of the input data and performing a linear transformation (i.e., convolution) at each visited pixel position. As shown in Fig. 3, convolution is a mathematical operation that combines point-wise multiplication between a filter (e.g., weight matrix) and an underlying subimage called the receptive field. For example, as shown in Fig. 3, the 3×3 filter, K , convolves with a 4×4 input data matrix, $f(\mathbf{x})$, to produce a 2×2 output feature map, y . Mathematically, the convolution operation can be expressed as follows

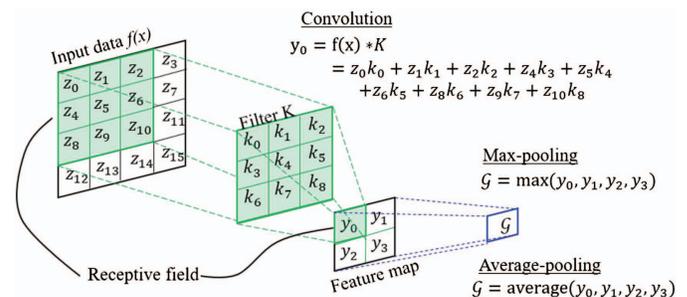


Fig. 3. (Color) Example of a discrete convolution for a 3×3 filter applied to a 4×4 input data followed by a 2×2 feature pooling operation.

$$y = f(\mathbf{x}) * K, \quad \mathbf{x} \in \mathbb{R}^2 \quad (1)$$

where $*$ represents the convolution operation; and \mathbb{R}^2 is a two-dimensional real space. A single output feature map in a convolution layer is produced by applying an identical filter to multiple receptive fields in the input data. This filter sharing across the same feature map preserves the local ordering of input data (Dumoulin and Visin 2016). Multiple feature maps corresponding to various filter effects, such as blurring, sharpening, and enhancing, can be generated by applying different filters (Ludwig 2013). For example, the boundaries or edges of soil layers in geological cross-sections can be located by convolution with commonly adopted edge detection filters, such as Prewitt, Sobel, and Laplacian (Katiyar and Arun 2014).

Feature Pooling Layer

A convolution layer is usually followed by a down-sampling operation called feature pooling, which aggregates several neighbor points in the previous convolution feature map into a local or global bag of features by using an averaging or maximizing operation (i.e., average or max pooling, respectively). As shown in Fig. 3, for example, four receptive fields in the original data are

represented concisely by a single pixel in the pooled feature map. A pooling operation can remove irrelevant details and extract more compact representations of feature maps (Boureau et al. 2010b), thus reducing the dimensions of the input data significantly. Average pooling tends to capture the average intensity of the input data, while max pooling is more suitable for detecting prominent features, such as edges. Shift invariance is realized mainly by weight sharing during the convolution operation and the max-pooling stage because only the max values and not the detailed spatial information are retained in the pooled feature map. More importantly, the pooling results maintain the shift-invariance of the input data patterns and are more robust to perturbations of the image (Boureau et al. 2010b; Dumoulin and Visin 2016; Zhang et al. 2020). Multiple convolutions and pooling layers may be stacked in a hierarchical manner to strengthen the shift-invariance (Zhang et al. 2020) and increase the discriminative power of the learned feature maps for the final classification. Note that only basic object features (e.g., edges) from the input data are extracted from shallow convolution and/or pooling layers.

Fully Connected Layer

More abstract feature maps of input data, recognizable for computers only, are obtained through multiple hierarchical convolutions and feature pooling operations. A fully connected layer is constructed by flattening the output from multiple convolutions or pooling layers into a one-dimensional (1D) vector for the final classification. Flattening is defined as the process by which a series of two-dimensional (2D) feature maps is transformed into a 1D vector. Because the local space-dependent properties are retained during the convolution and pooling operations, the spatial connectivity and stratigraphic relationships are correctly preserved after flattening (Zhang et al. 1990). The resulting 1D vector serves as the input for MLP network learning. To minimize the differences between the prediction and true training labels, the weights of the final multiple neurons and convolution filter are varied by using algorithms, such as the gradient descent algorithm (Ruder 2016), which dynamically adjusts coefficients in the direction of the deepest descent until the calculated difference is below a specified limit. To prevent overfitting of the network learning, a regularization technique called dropout (Srivastava et al. 2014) may be applied to remove neurons randomly from the neural network. Upon completing CNN training, predictions based on new input data can be inferred via the trained feature extraction and classification.

eXtreme Gradient Boosting (XGBoost)

To improve the computational efficiency, the MLP network within a conventional CNN algorithm is often replaced by state-of-the-art classification and regression algorithms. One such algorithm, eXtreme Gradient Boosting (XGBoost) (Chen and Guestrin 2016), is a boosting tree model that sequentially improves the prediction accuracy by combining the outputs of many base decision trees. In this study, a single tree operates by partitioning a feature space into a set of consecutive subspaces (Fig. 4). Fig. 4(a) depicts a typical partition of a 2D feature space with two input variables, x_1 and x_2 , and Fig. 4(b) shows the corresponding binary tree structure. The feature space is partitioned into five regions by using four binary splitting nodes determined from the training data. After establishing the binary tree from these training data, a new input can be assigned to a corresponding region according to the developed binary tree. The tree is called a regression tree when each leaf in each partitioned region is represented by a constant value corresponding to the average of all of the training samples in the region

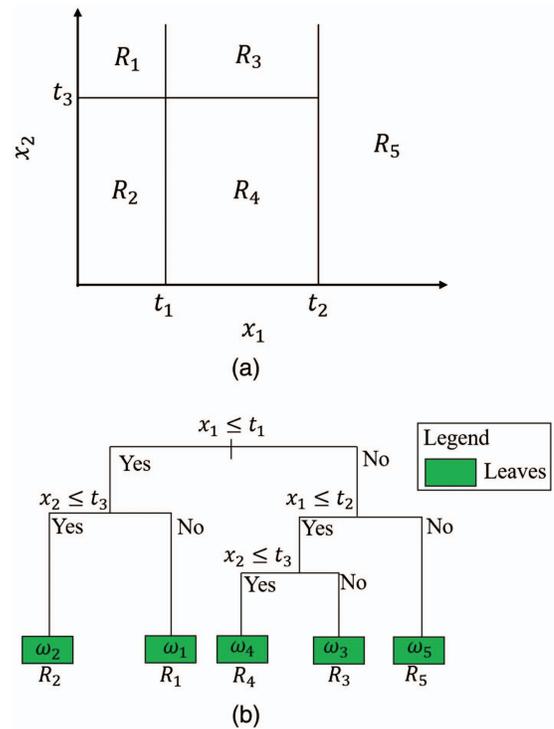


Fig. 4. (Color) Prediction using a single decision tree: (a) partition of a two-dimensional feature space; and (b) corresponding tree structure.

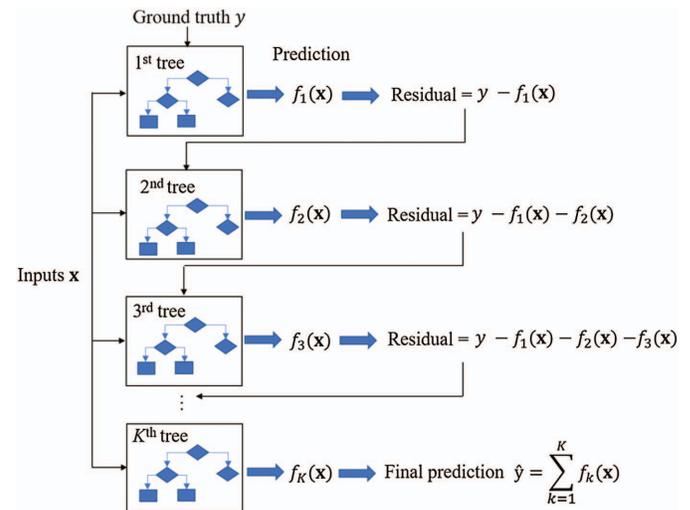


Fig. 5. (Color) Schematic diagram of a boosting tree.

[Fig. 4(b)]. For example, ω_1 equals the average values of all samples assigned to region R_1 . Otherwise, the tree is called a classification tree when each partitioned region denotes a category; in this study, the category equals the majority category of all samples assigned to the partitioned region. All of the splitting intervals (e.g., t_1 , t_2 , and t_3) shown in Fig. 4(b) are determined from training data using a suitable loss function. For example, the mean square error is a commonly used loss function for a regression tree, whereas the misclassification error or cross-entropy loss is often used when training a classification tree.

A boosting tree is a sequential ensemble technique in which the prediction of an individual regression or classification tree f_k is

combined successively to enlarge the model capacity until the expected accuracy or specified maximum number of trees is achieved. Fig. 5 shows a schematic diagram of a boosting tree. The k -th tree, f_k , was trained to fit the residual between the ground truth y and the prediction of a previous tree, f_{k-1} . The final prediction is expressed as follows

$$\hat{y} = \sum_{k=1}^K f_k(\mathbf{x}_i), \quad \mathbf{x}_i \in \mathbb{R}^m, \quad f_k \in \mathcal{F} \quad (2)$$

where \hat{y} = prediction from the boosting process; K = total number of trees; \mathbf{x}_i = i -th training vector with m features; \mathbb{R}^m = m -dimensional real space; and \mathcal{F} = space of the decision trees. For a multiclass classification, the prediction for each category \hat{y}_z is calculated separately according to Eq. (2), and the final class probability p_z is obtained by converting \hat{y}_z through the softmax function (Hastie et al. 2009), as shown subsequently

$$p_z = \frac{e^{\hat{y}_z}}{\sum_{z=z_1}^{z_{N_c}} e^{\hat{y}_z}}, \quad z = z_1, z_2, \dots, z_{N_c} \quad (3)$$

where Z_{N_c} = total number of categories.

XGBoost, a scalable tree-boosting system, penalizes the complexity of an individual tree by imposing a regularization term on the traditional loss function \mathcal{L} . As shown subsequently

$$\mathcal{L} = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k), \quad \text{where } \Omega(f) = \gamma T + \frac{1}{2} \lambda \|\omega\|^2 \quad (4)$$

where $l(\hat{y}_i, y_i)$ = total training loss between the prediction and training targets; $\Omega(f_k)$ = regularization term used to control the complexity of the k -th decision tree; T = number of leaves; λ = coefficient with a default value of 1.0; and γ = tuning parameter. Refer to the study by Chen and Guestrin (2016) for details. A Taylor expansion facilitates the iterative training process. The loss associated with the t -th iteration ($t \leq K$) is calculated as follows

$$\begin{aligned} \mathcal{L}^t &= \sum_{i=1}^n [l(y_i, \hat{y}_i^{(t-1)} + f_t(\mathbf{x}_i))] + \Omega(f_t) \\ &\approx \sum_{i=1}^n [l(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(\mathbf{x}_i) + \frac{1}{2} h_i f_t^2(\mathbf{x}_i)] + \Omega(f_t) \end{aligned} \quad (5)$$

where $g_i = \partial_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)})$ and $h_i = \partial_{\hat{y}_i^{(t-1)}}^2 l(y_i, \hat{y}_i^{(t-1)})$ are the first- and second-order derivatives, respectively, pertaining to the loss function (i.e., l). The loss function can be customized, and a categorical cross-entropy loss (Hastie et al. 2009) can be used for multiclass classification problems. The combination of a conventional neural network with the XGBoost algorithm was shown to outperform a conventional CNN in classification tasks (Thongsuwan et al. 2021).

Note that the successful application of a CNN model for image restoration is based on the assumption that a substantial part of an image (e.g., >90%) is available. More importantly, the establishment of a well-performing CNN model normally requires millions of complete training images to tune the parameters (e.g., filter weights and neural network coefficients). For geotechnical investigation, training images, in the form of geological cross-sections, are normally limited in quantity. Site-specific boreholes are also limited, typically accounting for less than 10% of a target site. This limitation hinders the direct application of CNN models to the development of subsurface geological cross-sections. On the other hand, a subsurface geological cross-section is often a simplified

representation of the real complex geological condition. Local stratigraphic variations are frequently omitted in geological profiling, and therefore, a simplified geological cross-section may be simpler than an image considered in image processing.

Proposed Iterative Convolution Model

The idea of extracting local invariant spatial patterns from input data is appealing to geotechnical practitioners, as subsurface geological profiling depends mainly on the accurate interpolation of stratigraphic relationships between soil deposits. In this section, a two-phase data-driven method based on a CNN framework is proposed for the delineation of subsurface geological cross-sections. Fig. 6 shows the basic concept of the proposed iterative convolution XGBoost model (IC-XGBoost). Note that the spatial statistics from prior geological knowledge are represented implicitly and concisely and are stored in an ensemble training cross-section. Rather than directly relying on a single training geological cross-section for feature extraction and spatial interpolation, the large-scale features amenable to site-specific data are first determined and used to extract large-scale spatial patterns from the ensemble training cross-section. All of the extracted spatial statistics are then included in the MLP, a training classification algorithm. In this study, the conventional MLP is replaced with an XGBoost algorithm to improve computational efficiency. The resulting well-trained model is used only for large-scale spatial interpolation. Next, all of the large-scale interpolation results are combined with site-specific data to determine the features iteratively and adaptively at a gradually reducing scale for the subsequent CNN training and spatial interpolation stages until all of the unknown points have been visited and interpolated.

Although the proposed method deals with a similar topic with a recent publication (Shi and Wang 2021a), which applies multiple-point statistics (MPS) to subsurface stratigraphy, the two studies use entirely different methodologies. In this study, the proposed method modifies the tools available in the CNN algorithm and thus opens a new window for delineation of a subsurface geological cross-section. Multiscale filters are applied to extract spatial statistics successively from large to small scales through a process that is amenable to the intrinsic line structures of geotechnical measurements. In addition, the process of feature extraction does not require explicit specification or tuning of hyperparameters.

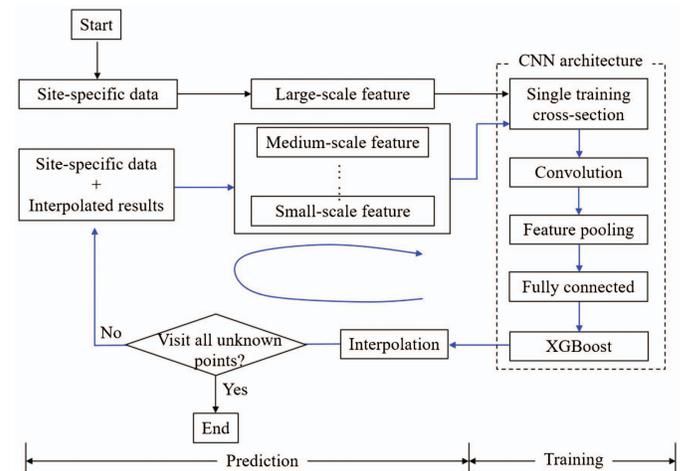


Fig. 6. (Color) Basic idea of the proposed iterative convolution model.

The proposed method also differs fundamentally from the conventional CNN model. A conventional CNN model cannot be used directly for the delineation of subsurface geological cross-sections due to the lack of small-scale measurements. Instead of performing only a single round of training and using two separate stages of training and prediction, as in a conventional CNN model (Fig. 2), an iterative process involving multiple rounds of training and prediction at gradually reduced scales is proposed in this study to address the difficulties associated with the limited training images and measurements. Note that when the number of boreholes is limited, the horizontal spacing between any two adjacent boreholes is relatively large. Therefore, only large-scale spatial patterns, particularly along the horizontal direction, can be revealed from the limited number of boreholes. To accommodate this limited number of boreholes, only the corresponding large-scale spatial features are extracted from the training image via convolution and pooling. These extracted spatial patterns are used to train an XGBoost model (Fig. 6), which is then used only for large-scale predictions (i.e., spatial interpolation). These stages of large-scale convolution and pooling are consistent with the scale of the limited number of boreholes, which resolves the difficulty of performing convolution and pooling based on sparse measurements. In contrast, a conventional CNN model performs both large- and small-scale convolution and pooling simultaneously during the training stage.

Feature extraction at only a large scale is consistent with the scale of the sparse boreholes and thus effectively solves the aforementioned difficulty. However, the trained XGBoost model does not contain small-scale information and can only make large-scale predictions. In other words, when the proposed model is used at a large scale, it can only make predictions at a large scale and cannot develop a complete stratigraphy because the resulting stratigraphy contains many unknown soil/rock types at the unsampled locations. To make a prediction at a relatively small scale, all large-scale prediction results are treated as measurement data. With prediction results being treated as measurement data, the measurements are no longer limited or sparse from the model's perspective. The new measurements are combined with the original measurement data to form a new dataset for the second round of training and prediction at the second scale. The second round of training and prediction at the second scale is performed based on the new dataset and the same training image used on a large scale. Compared with the original dataset, the new measurement dataset contains more data points and has a smaller minimum distance between adjacent measurements and a correspondingly smaller spatial scale. As shown in Fig. 6, spatial features at the second, or medium, scale is extracted again from the training image via convolution and pooling, similar to the large-scale features described previously. A new XGBoost model is then trained using all of the spatial patterns extracted at the second scale and used for predictions (e.g., spatial interpolation) at this scale. Next, all prediction results obtained at the second scale are treated as measurement data, and another round of training and prediction at a gradually reduced scale is repeated until the soil types at all unsampled locations are predicted, and a complete stratigraphy is obtained.

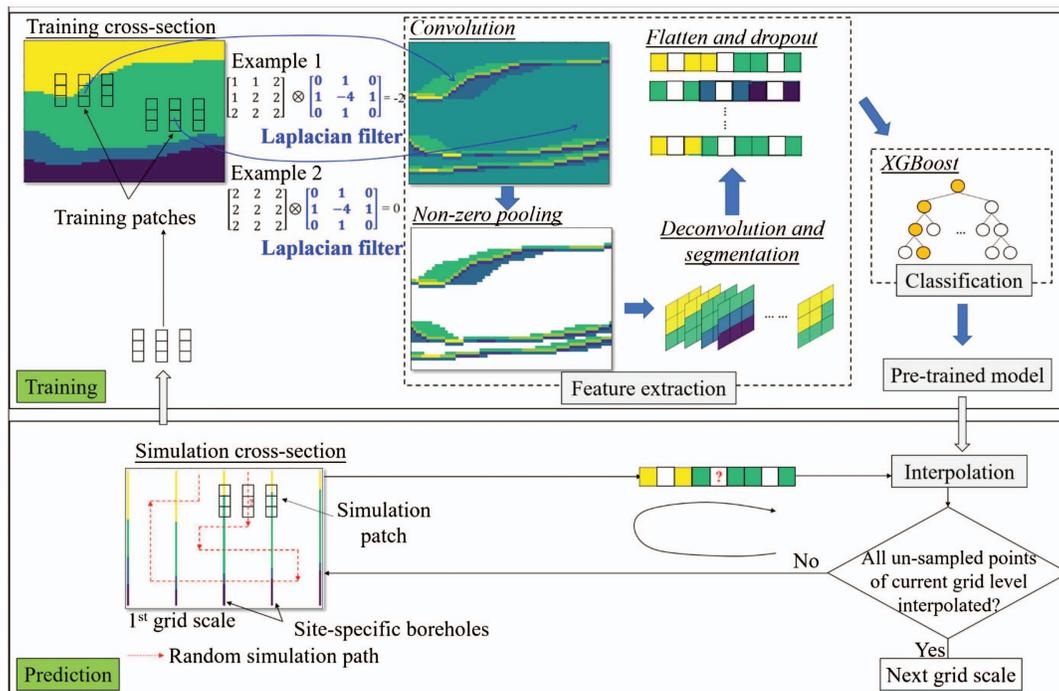
Fig. 7 depicts the detailed architecture of the proposed iterative convolution method. The first stage mainly involves the extraction of features and establishment of the classification model [Fig. 7(a)]. A grid template is established that is amenable to sparse boreholes and in the form of a sandwiched structure, in which each unsampled point is accompanied by two-line measurements. Next, a typical simulation patch is transferred to the training part to extract spatial connectivity and stratigraphic relationships from a single training geological cross-section through convolution with an edge detection filter. The extracted local stratigraphic patterns are

processed and used as the input for XGBoost training. The well-trained XGBoost model is then saved and used for subsequent predictions. In the second stage, all unsampled points with a given grid template are interpolated according to a random simulation path, which is determined using a random number generator with a uniform distribution between 0 and 1. All of the unsampled points are collected at the beginning of a simulation, and a random value is generated and assigned to each unsampled point. Next, all unsampled points are sorted by ascending order of the random values assigned to each point, and the sorted order is used as the random simulation path. Once interpolation is completed, new grid templates are designed conditionally according to the boreholes and previously interpolated results. The preceding two stages of spatial interpolation are then repeated until all unsampled points are interpolated [Fig. 7(b)].

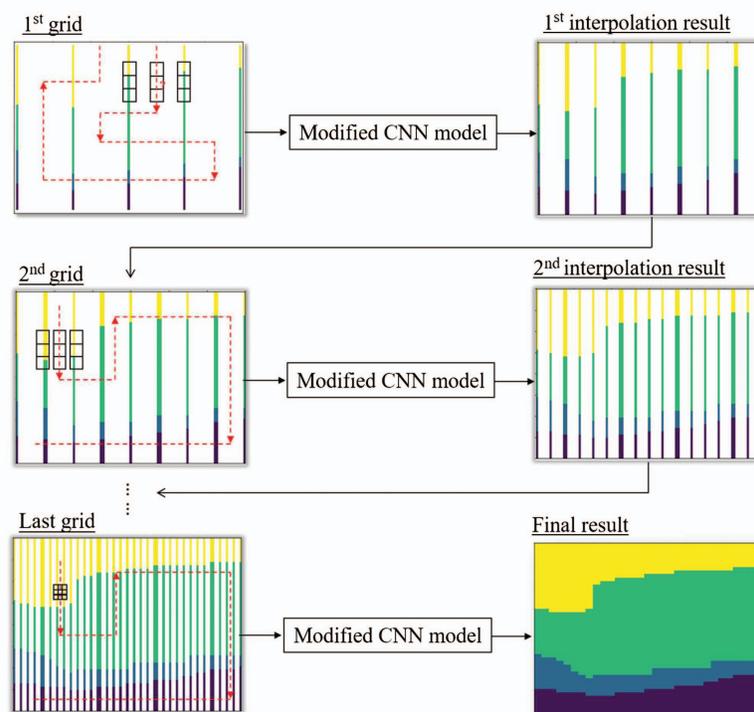
Essentially, each round of spatial interpolation comprises training and prediction. The training part uses the same training image, and it is possible to predetermine all of the potential grid templates to exhaust all of the spatial statistics associated with the single training image. With regard to the prediction part, it has to be completed in an iterative process as the prediction at the current step depends on results from previous steps. Note that the proposed method differs fundamentally from conventional CNN, as the former uses limited training images (i.e., as few as one geological cross-section) for feature extraction and classification model building. More importantly, the target cross-section is interpolated in an iterative manner from large-scale features and limited site-specific boreholes. The proposed method can also be used for boreholes with irregular spacing because the interpolation for locations without measurements in each iteration is performed at the central line between any two closest boreholes, and the spacing between any two closest boreholes is not a parameter used in the proposed method. The current algorithm implementation only considers boreholes with regular spacing. The scenarios of boreholes with irregular spacing will be implemented in a future study.

Multiscale Interpolation Template

Borehole measurements are considered structured data because they are frequently sampled at a given horizontal spacing and have multiple measurements along the vertical direction. In this study, a multiscale interpolation template that begins at a large scale and is amenable to the intrinsic data structure is adopted. The first interpolation scale, g_1 , comprises all boreholes and unsampled profiles between any two adjacent boreholes. For example, five boreholes within a simulated cross-section and the four central unsampled profiles comprise the first interpolation template shown in Fig. 7(a). Each point along the unsampled profiles is visited using a random simulation path, which is represented by the red dashed line. An associated simulation patch at a 3×3 matrix containing both sampled and unsampled soil types is extracted for the subsequent conditional interpolation of the central unsampled point, which is represented by the cell with a question mark in the simulation patch in Fig. 7(a). A 3×3 matrix is commonly used in image processing for feature extraction. This approach is adopted in this study because the soil type at any point is most relevant to the soil type of the nearest neighbors. The soil type at the central point within a simulation patch is interpolated using the XGBoost algorithm, which is discussed in detail in the "Gradient Boosting" section. The inferred soil type is assigned to the unsampled point and used as the input for a subsequent conditional interpolation. When all points along the unsampled line profiles within the g_1 -th scale of the interpolation template are interpolated, a reduced scale g_2 with nine line measurements corresponding to the five original



(a)



(b)

Fig. 7. (Color) Architecture of proposed iterative convolution model for delineation of a subsurface geological cross-section: (a) extraction of large-scale features and corresponding spatial interpolation of the first grid template; and (b) process of iterative interpolation.

boreholes plus four simulated profiles is obtained from the first scale of the interpolation template, and eight new intermediate un-sampled profiles [Fig. 7(b)] are constructed for spatial interpolation. The preceding procedures are repeated until all un-sampled profiles at different scales have been visited and interpolated in a hierarchical manner. Note that different scales of the interpolation

template correspond to the spatial features of different scales. At a large scale, such as the first interpolation template, the spacing between adjacent measurements is large, and conditional interpolation tends to capture long-range stratigraphic connectivity. As the scale decreases in size, local spatial patterns are added gradually. Note that there is a possibility of error propagation, as the

interpolation of small-scale simulation patches depends on the interpolated results of large-scale patches. The error from small-scale spatial interpolation can possibly be mitigated by assigning more weights to large-scale patches containing original measurements.

The multiscale interpolation templates used to delineate the vertical geological cross-section in this study use the soil types measured at surrounding points to interpolate the soil type at an unsampled point. Initially, all three cells along the central line are unsampled, and only the central single cell within a 3×3 simulation patch is interpolated in each round of the process. The interpolated cell is then used as the measurement for subsequent conditional simulations. At subsequent stages, the total number of conditional cells (including the six true measurements and previously interpolated cells) of a simulation patch can be seven or eight. When the total number of conditioning cells equals seven, the distribution of conditioning cells within the simulation patch can have two different forms. The seven conditioning cells comprise the six true measurements on both sides and another one in either the central upper cell or the central bottom cell. In total, there are four different forms of simulation patches when the conditioning number varies between 6 and 8. Therefore, four sets of XGBoost models are separately trained for the four simulation patches, which is discussed in detail in the “Gradient Boosting” section. The proposed method ensures that each unsampled point is surrounded by at least six measurements for the conditional simulation. This constraint can be relaxed for cases with limited measurements (e.g., one borehole), although this causes a large interpolation uncertainty.

Filter Selection

Spatial stratigraphic relationships between the points within each simulation patch on a 3×3 matrix containing both sampled and unsampled soil types are learned directly from the training data. These data may be a single conceptual geological cross-section drawn by engineers or geological cross-sections borrowed from previous project sites with similar geological settings (Shi and Wang 2021a, b). The direct application of existing geological cross-sections as training datasets is based on the assumption of a similar local spatial connectivity or stratigraphic relationships between soils in areas with similar geological settings. Because geological cross-sections are often simplified representations of real complex geological conditions, and each simplified geological cross-section contains a limited number of soil types, only points lying along the stratigraphic boundaries that separate different soil layers carry important stratigraphic information. Therefore, filters capable of edge detection are considered appropriate for identifying the stratigraphic relationships between soil deposits. In this study, a 3×3 spatial derivative filter is used to extract training patches with useful stratigraphic information from the training data via a convolution operation. This type of filter is commonly used to detect transitions of pixel intensity during image processing. Note that the filter is learned automatically from training images in a conventional CNN model. However, in the proposed method, the filter is specified directly as a spatial derivative filter because only a single training image is available, and the extracted training patches are insufficient for training a well-performing filter. Consistent with the multiscale interpolation templates, the typical spatial features around the simulation patches from the training cross-section are extracted iteratively using a filter adapted to a given simulation patch within each interpolation scale. The spatial features are extracted using a discrete Laplacian convolution filter, which is a 2D isotropic measure of the second spatial derivatives (Van Vliet et al. 1989) that can be used to highlight regions of rapid changes in

intensity (i.e., transitions or boundaries between different objects). Mathematically, a Laplacian $L(h, v)$ of an image with a pixel intensity value I can be estimated using the following equation

$$L(h, v) = \frac{\partial^2 I}{\partial h^2} + \frac{\partial^2 I}{\partial v^2} \quad (6)$$

where h and v = horizontal distance and vertical depth, respectively. The pixel intensity I represents the soil type, which is indexed as an integer (e.g., 1 and 4 refer to Soil1 and Soil4, respectively). For discrete fields, the commonly used Laplacian filter is simplified as follows

$$\begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

A nonzero convolution value resulting from the Laplacian filter indicates the locations of the edges between different stratigraphic boundaries. For example [Fig. 7(a)], all points within the receptive field in Example 2 fall entirely within a single soil type (i.e., Index 2). After convolution with the Laplacian filter, the corresponding value in the feature map is zero. The receptive field in Example 1 is occupied by two different soil types, which are indexed as 1 and 2. The calculated convolution value is -2 , indicating the detection of a stratigraphic boundary.

Nonzero Pooling and Full Connection

A single filter is convoluted with the training cross-section to construct a feature map corresponding to a given grid scale. Similar to the conventional CNN algorithm, a feature pooling layer is applied to filter out the unimportant stratigraphic relationships. A value of zero within the convoluted feature map essentially refers to training patches located far from the stratigraphic boundaries, as the corresponding points within the original receptive fields are filled with the same soil type. In this study, a nonzero pooling layer is adopted to retain only nonzero values in the convoluted feature map. Note that only one convolution layer and one pooling layer are used in this study because geotechnical cross-sections are often simplified representations of real complex geological conditions, and local stratigraphic variations are frequently simplified.

After identifying all of the nonzero convoluted values within the feature maps, a deconvolution operation is applied to recover the corresponding training spatial patches in the original receptive fields. Note that in the conventional CNN model, the feature maps are used directly in subsequent classification tasks. Feature maps are abstract and can be successfully recognized by computers. For improving the interpretability of the proposed algorithm, the feature maps generated in this study have been deconvoluted inversely to recover the original 3×3 training patches for classification. The collected training patches within each interpolation scale are then flattened and stacked together for subsequent supervised learning via the XGBoost algorithm, which is discussed in the “Gradient Boosting” section. Note that the simulation patch in the simulation cross-section contains both sampled and unsampled soil types, which are treated differently. The circumferential points with the unsampled soil category do not impart useful stratigraphic information and are dropped from both the simulation patch and stacked training patches. Note also that if a simulation patch lies entirely within a single soil layer, all three central points can be interpolated simultaneously. However, when a simulation patch lies along a stratigraphic boundary, the interpolation sequence of the

central three points can yield different results. Therefore, only the central single unknown cell within the simulation patch is interpolated every time in this study.

Gradient Boosting

The interpolation of a central unknown soil type within a simulation patch can be considered a multiclass classification problem. Training patches extracted from a given interpolation scale are flattened in a 1D vector and serve as an input variable. The corresponding soil type at the central point is labeled as the output variable. For example, the training patch $\begin{bmatrix} 1 & 1 & 1 \\ 1 & \underline{2} & 2 \\ 2 & 2 & 2 \end{bmatrix}$ has a target label of 2; in this study, the underlined integer represents Soil2, and all other integers are taken as the input. As correlations between the different soil types may affect the results of spatial interpolation, natural ordering relationships between different integers are removed using a one-hot encoder (Rodríguez et al. 2018) to avoid any potential privileges associated with the numbers applied

to represent different soil types. The one-hot encoder can transform each indexed soil category into a vector with a value of 0 or 1. For example, Soil1 and Soil2 can be encoded as $[1, 0, 0, 0]$ and $[0, 1, 0, 0]$, respectively, and a vector length of 4 equals the total number of soil types within the simulation cross-section. In this study, all indexed soil integers within a single training patch are converted into a 1D vector. For example, a training patch occupied by Soil1 and Soil2, which are coded as Integer 1 and Integer 2, respectively, is flattened into a 1D vector, as shown in Eq. (7)

$$\begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 2 \\ 2 & 2 & 2 \end{bmatrix} \xrightarrow{\text{flattening}} [1, 1, 1, 1, 2, 2, 2, 2, 2] \quad (7)$$

Each single soil type within the 1D vector is further expanded using one-hot encoding, which transforms each soil integer into a vector with a value of 0 or 1, as shown subsequently

$$[1, 1, 1, 1, 2, 2, 2, 2, 2] \xrightarrow{\text{one-hot encoding}} [[1, 0, 0, 0], [1, 0, 0, 0], [1, 0, 0, 0], [1, 0, 0, 0], [0, 1, 0, 0], [0, 1, 0, 0], [0, 1, 0, 0], [0, 1, 0, 0], [0, 1, 0, 0]] \quad (8)$$

The multiclass classification problem can be solved using the MLP, a fully connected model used commonly in CNNs. In a fully connected model, every neuron in the previous layer is fully connected with all of the neurons in the next layer. The conventional training process is usually time-consuming because the specified numbers of neurons and hidden layers rely heavily on the modeler's level of experience, and the training of weights for the MLP model requires a large amount of training data. In this study, XGBoost, which uses parallel tree boosting to solve complex classification problems quickly and accurately, is applied to improve computational efficiency. The predictions on the leaves of multiple trees are summed and converted to a class probability p_z , using the softmax function in Eq. (3). The soil type Z at location x is determined by a random draw, according to the calculated class probability p_z . Although the mathematical formulations of XGBoost seem complex, a detailed implementation has been documented in a single open-source GitHub repository, using popular programming languages such as Python, R, and C++. For this study, the Python package (xgboost version 1.3.3/python-package) has been adopted.

Quantification of Interpolation Uncertainty

A realization is completed when all of the unsampled soil types among all of the grid template scales have been interpolated. The spatial interpolation uncertainty associated with the proposed method can be quantified by a statistical analysis of multiple realizations. The primary randomness of the proposed method originates from a random simulation path for each grid scale. The most probable interpolation associated with multiple realizations, $Z_{mp}(\mathbf{x})$, is obtained by setting the interpolated soil type, Z , as the type with the highest occurrence frequency at each point \mathbf{x} . The most probable interpolation is used as the results from the proposed method. It is worthwhile to mention that the proposed method not only provides interpolation results but also quantifies the uncertainty associated with the interpolation. The corresponding interpolation uncertainty at \mathbf{x} can be quantified by the dispersion, $Dp(\mathbf{x})$,

which is defined as the proportion of mismatching soil types at each \mathbf{x} among multiple realizations, N_r , as compared with that of the most probable interpolation (Shi and Wang 2021a)

$$Dp(\mathbf{x}) = \frac{\sum_{r=1}^{N_r} I(Z_r(\mathbf{x}) \neq Z_{mp}(\mathbf{x}))}{N_r} \quad (9)$$

where I = indicator function and has a value of 1 when the predicted soil type $Z_r(\mathbf{x})$ differs from the most probable interpolation. In this study, dispersion measures the deviation of multiple realizations from the most probable interpolation and has theoretical maximum and minimum values of 1 and 0. Note that the dispersion plot is derived from the statistically most probable data rather than the underlying true data. The dispersion only represents the confidence level of the interpolation results obtained using the proposed method. The dispersion at all points within a simulation cross-section is collected and presented as a 2D dispersion map, thus providing a direct visual tool for assessing the interpolation uncertainty originating from random simulation paths. Note that the number of realizations, N_r , is considered sufficient when the percent change associated with the most probable interpolation is virtually zero with additional stochastic realizations. In this study, a relatively small percent change of 0.1% is adopted as the termination criterion.

For the illustrative examples used in the following sections, in which the complete and true testing geological cross-sections are available, the interpolation performance can be assessed by the accuracy, Acc . This is a measure of the deviation of the most probable interpolation, $Z_{mp}(\mathbf{x}_i)$, from the underlying true geological cross-section, $Z_T(\mathbf{x}_i)$

$$Acc = \frac{\sum_{i=1}^{N_h \times N_v} I(Z_T(\mathbf{x}_i) = Z_{mp}(\mathbf{x}_i))}{N_h \times N_v} \quad (10)$$

where N_h and N_v = total node numbers of the simulation profile in the horizontal and vertical directions, respectively. In this study, the

underlying true geological cross-section of a site is normally unavailable in a practical application and is only used in the illustrative examples to validate the proposed method.

Note that this newly proposed method is purely data-driven and does not require the explicit specification of any parametric function form. In practice, an engineer would only need to feed the site-specific measurements and a suitable training image to the trained model, which will return the most probable interpolation results with quantified uncertainty.

Illustrative Example

The flowchart in Fig. 8 describes the procedures used to run stochastic simulations using the proposed method. To implement the proposed method, a Python package was developed such that the only required input is a suitable training geological cross-section (i.e., training data) and sparse borehole measurements from a specific site (i.e., new input data for the prediction). In this section, a pair of geological cross-sections are simulated to illustrate the

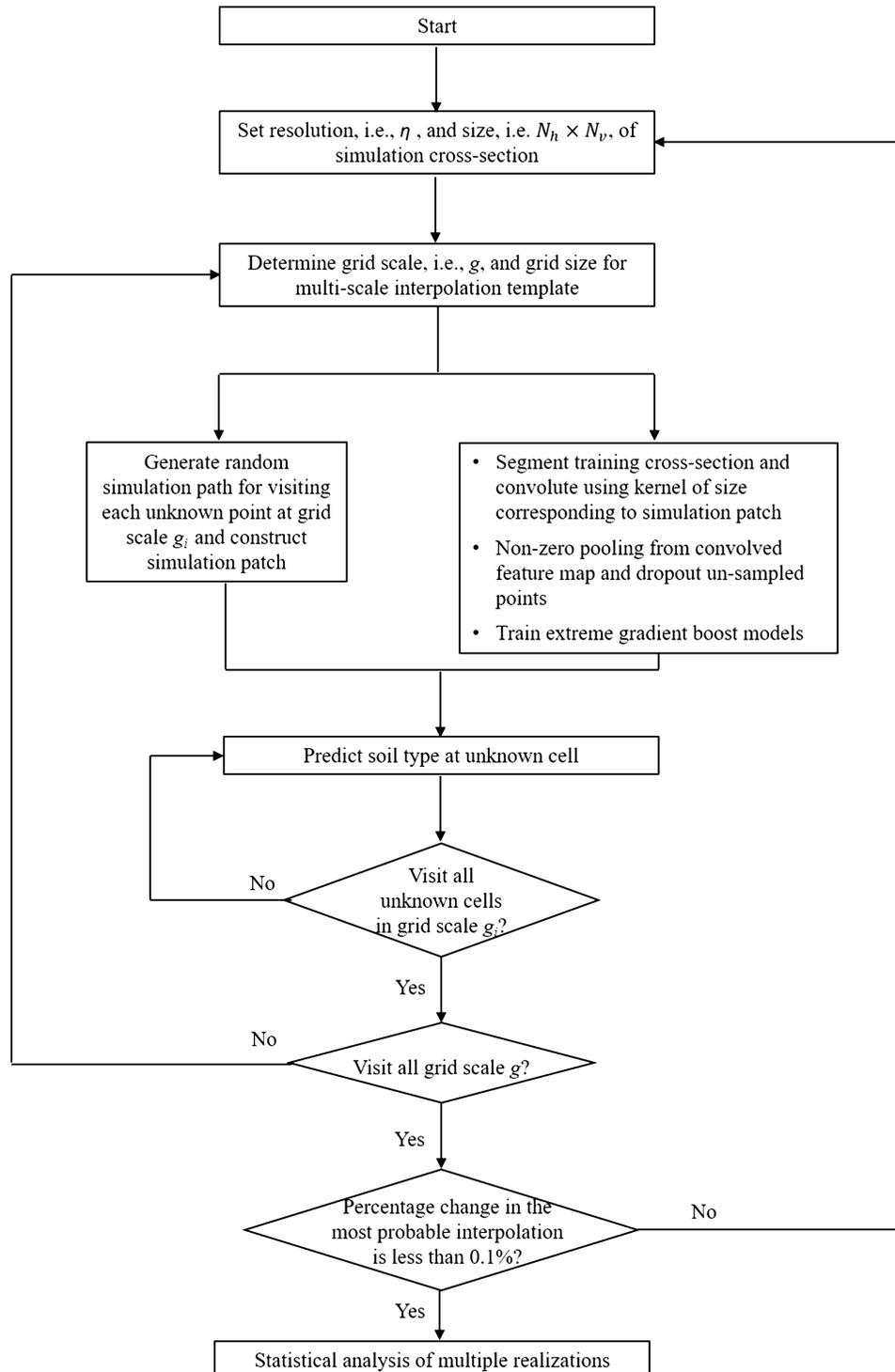


Fig. 8. Flowchart of the proposed method.

proposed method. Both cross-sections comprise four soil types, and each cross-section has a total horizontal length and vertical depth of 30 m and 15 m, respectively. The resolutions in both directions are set at 0.2 and are overly optimistic for most practical purposes. For example, measurements are taken at a common vertical interval of 0.45–0.6 m for the standard penetration test. Note that a small resolution corresponds to a high-resolution cross-section. These high-resolution cross-sections are simulated to demonstrate the capacity of the proposed method for reconstructing the detailed

Table 1. Coefficients for the function $v = A \times h^2 + B \times h + C$

Function ID	A		B		C	
	Mean	Std.	Mean	Std.	Mean	Std.
I	-0.02	0.02	1.5	0.3	2	1
II	-0.04		0.2		5	
III	-0.02		0.5		8.5	
IV	0.04		-3.0		11	
V	-0.02		-0.5		11	
VI	—	—	0.5		9	
VII	-0.01	0.02	-0.5		14	
VIII	-0.02		-0.5		13	

Note: Std. = standard deviation.

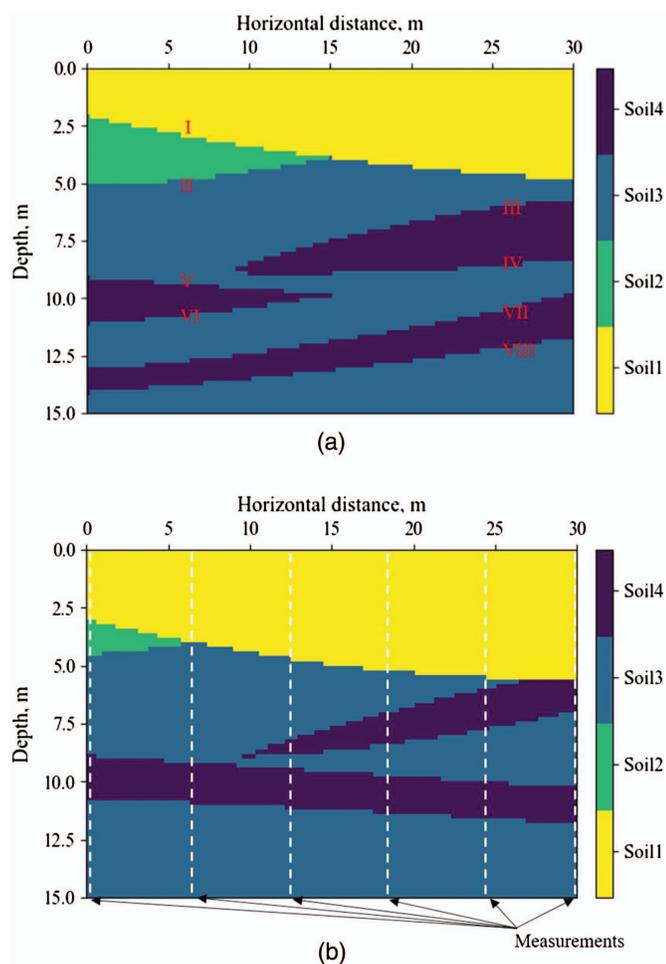


Fig. 9. (Color) Colormaps of training and test cross-sections used in the illustrative example: (a) training cross-section; and (b) test cross-section.

stratigraphic connectivity between different soils. In total, 150 and 75 points were sampled in the horizontal and vertical directions, respectively, with a resolution of 0.2 m in both directions. The stratigraphic boundaries between the different soil deposits are artificially imported as quadratic lines and can be described using the following equation

$$v = A \times h^2 + B \times h + C \quad (11)$$

where h and v = horizontal distance and vertical depth, respectively; coefficients A , B , and C are taken to follow Gaussian distributions, and the associated statistics (i.e., mean and variance) are summarized in Table 1. The quadratic line from Eq. (11) separates the different soil types, such that the sampled points above and below the line are assigned to different soil types. Figs. 9(a and b) show the simulated training and test geological cross-sections. Evidently, both cross-sections share similar spatial stratigraphic relationships. For example, different-sized wedges comprising Soil2 exist in the top left corners of both cross-sections. The largest differences between both cross-sections lie in the extent and orientation of the interbedded layers of Soil3 and Soil4. As an illustration, six vertical line measurements (i.e., boreholes) spaced at an equal interval of 6 m are taken as measurements (i.e., inputs for the proposed method) for subsequent interpolations.

Simulation Results from the Proposed Method

The proposed method adopted recovered subimages for classification, and the classification accuracy varies between 75% and 90% for different spatial interpolation scales. Multiple realizations conditioned on the six-line measurements were generated according to the simulation procedures described in previous sections. The evolution of the percent change in the most probable interpolation is shown in Fig. 10. Evidently, the percent change exhibits a decreasing trend to less than 0.20% as the realization number increases beyond 20. In this study, 100 realizations were performed. Fig. 11(a) presents a color map of the most probable interpolations derived from 100 realizations. Clearly, the spatial stratigraphic relationships between different soil deposits were reproduced correctly by the proposed method, with an overall accuracy of 97.1%. The true soil boundaries are also superimposed in Fig. 11(a) for comparison. Most of the interpolation differences cluster around these true soil boundaries. Encouragingly, the spatial distribution of the bottom seam, i.e., Soil4, can be replicated correctly, even though the seams in the training cross-section shown in Fig. 9(a)

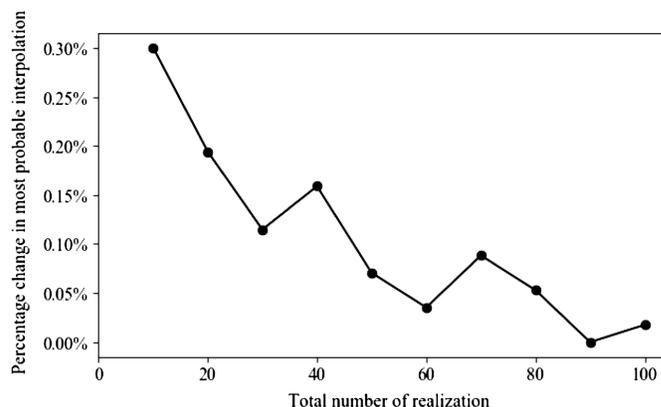


Fig. 10. Percentage change in the most probable interpolation.

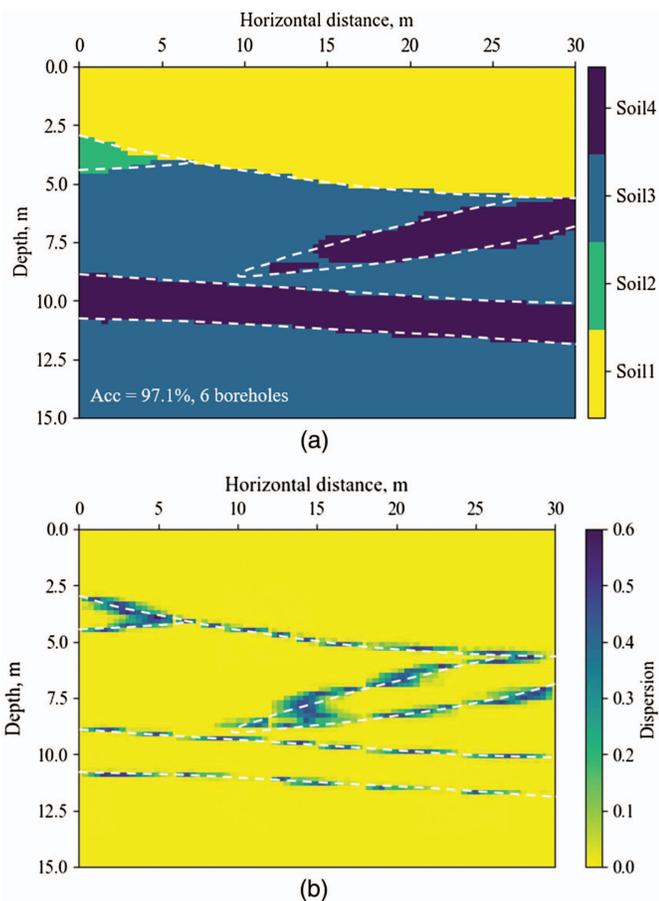


Fig. 11. (Color) Results of the proposed method conditioning on six-line measurements: (a) most probable interpolation; and (b) dispersion.

have different orientations and extents. In this study, a separate analysis was performed to compare the spatial interpolation accuracies of the XGBoost and MLP models. Although the respective spatial interpolation accuracies of 97.1% and 96.3% were comparable, the time for running MLP is approximately fourfold longer than that for XGBoost. The spatial interpolation uncertainty associated with the most probable interpolation was then calculated using Eq. (9), and the obtained dispersion plot is shown in Fig. 11(b). Clearly, the majority of the four soil types were interpolated correctly with zero dispersion, and large dispersion bands cluster mainly around the true stratigraphic boundaries surrounding Soil2 and Soil4. In this study, the dispersion plot provides a handy tool for assessing and visualizing the interpolation uncertainty. The dispersion plot represents the interpolation confidence level associated with the statistically most probable interpolation and should not be taken as the true uncertainty of the ground.

In this study, the complete and true test cross-section is available for the simulation example. Therefore, it is possible to compare the interpolation results with the true test cross-section. Fig. 12(a) shows the evolution of accuracy as calculated by Eq. (10) for each soil type. At fewer than 40 realizations, large variations were observed in the calculated accuracies for all of the soil types; beyond this number, the accuracies exhibit stabilizing trends. The worst prediction accuracy was obtained for Soil2, which occupies a small area in Fig. 9(a); a small perturbation in the interpolation results may cause a large change in the calculated accuracy. In addition, the collected interpolation accuracy values at each point from 100

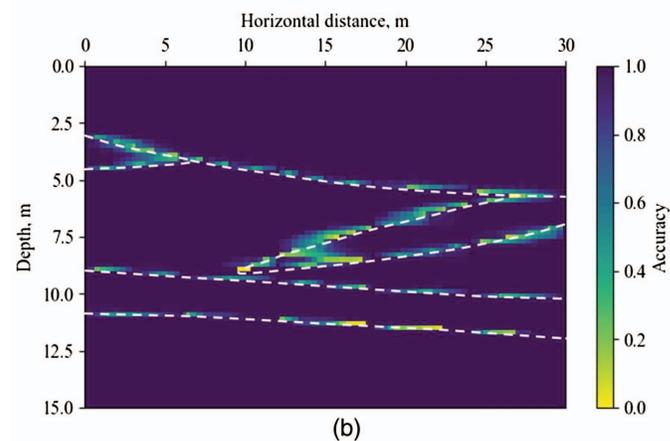
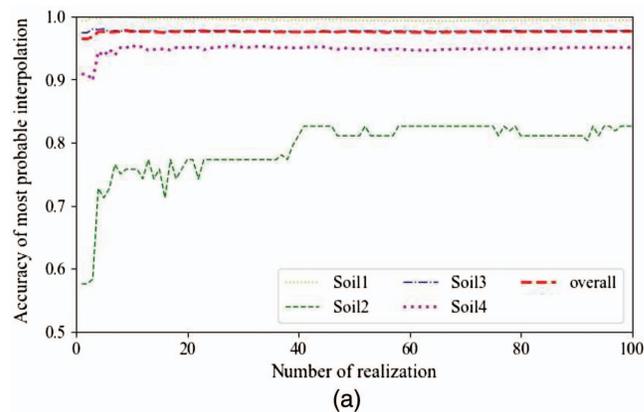


Fig. 12. (Color) Comparison between most probable interpolation and the true test cross-section: (a) variation of interpolation accuracy of each soil type with realization number; and (b) spatial distribution of interpolation accuracy.

realizations are presented in a 2D accuracy color map in Fig. 12(b). Evidently, interpolation errors are clustered mainly at the stratigraphic boundaries that separate different soils, consistent with the dispersion results shown in Fig. 11(b).

Multiscale Representation of Spatial Interpolation

Apart from the final interpolation results, the proposed method can also be used to generate intermediate convolution feature maps and multiscale representations corresponding to different interpolation scales. Fig. 13 shows feature maps generated from a convolution operation involving the Laplacian filter and the corresponding interpolation results. At the first iteration [i.e., large scale; see Fig. 13(a)], long-range training patches are extracted from the training cross-section. As the scale decreases or the iteration number increases, the feature maps corresponding to the detailed stratigraphic relationships are pooled. The stratigraphic boundaries can be identified correctly by the nonzero values within these feature maps, while the zero convolved values mainly lie within the single soil horizons. The intermediate interpolation results corresponding to the multiscale simulation grids of a realization are shown in Figs. 13(c and d). At the first iteration, only the central profiles between any two adjacent boreholes are interpolated, resulting in a spatial interpolation with a coarse resolution. This iteration does not capture the detailed stratigraphic relationships between different soil deposits. As the simulation scale decreases (i.e., iteration number increases from 1 to 4), more detailed spatial

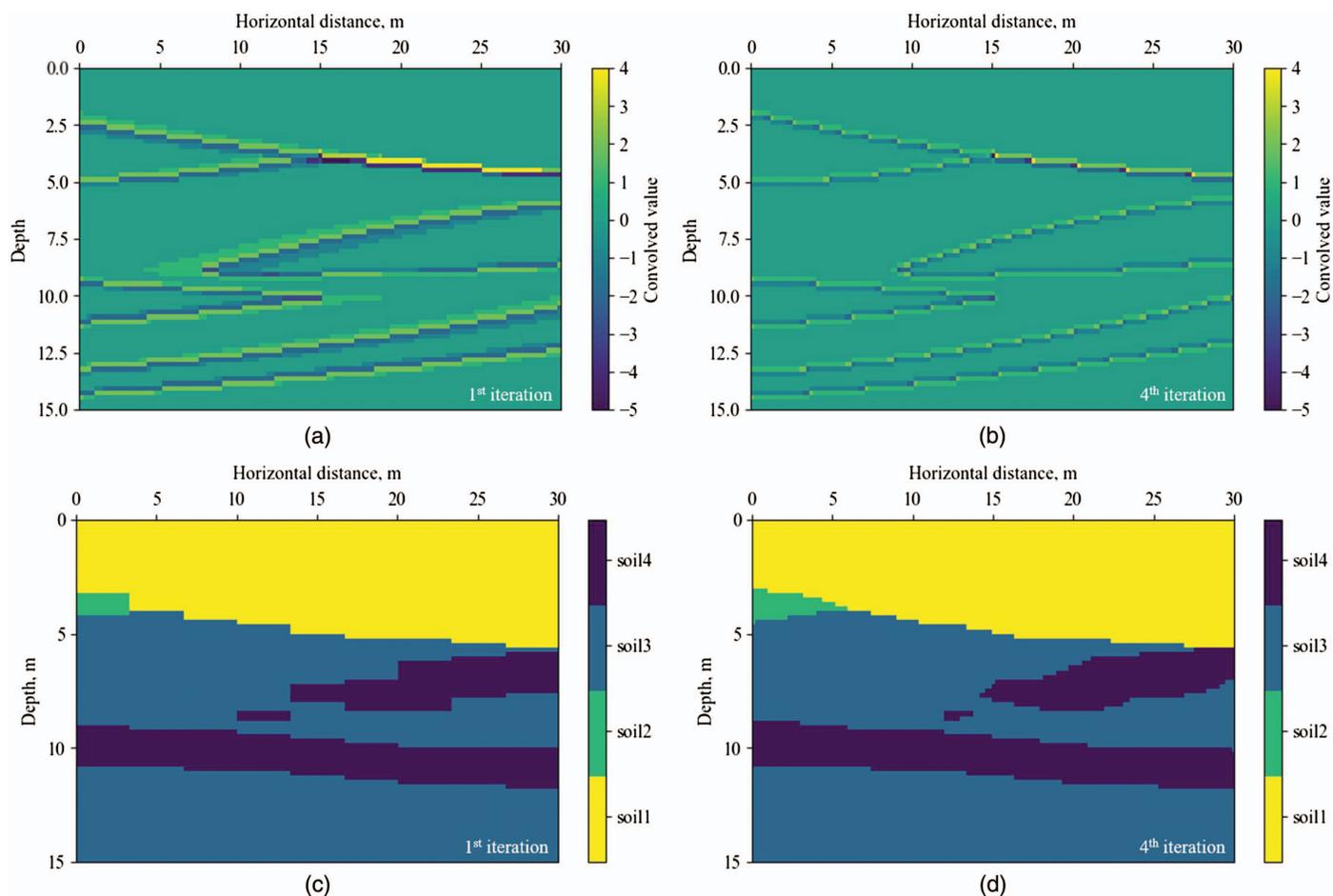


Fig. 13. (Color) Multiscale representation of one realization: (a) 1st iteration convolution feature map; (b) 4th iteration convolution feature map; (c) 1st iteration interpolation result; and (d) 4th iteration interpolation result.

stratigraphic connectivity information becomes available. The proposed method thus enables a user-defined simulation scale.

Effect of the Borehole Number

Figs. 14 and 15 compare the interpolation results generated by the proposed method when the number of boreholes decreases to 3 or increases to 11, respectively. When only three boreholes are used as the input, the most probable interpolation in Fig. 14(a) can reasonably capture the overall stratigraphic connectivity between different soils with an accuracy of approximately 91.2%. The dispersion plot in Fig. 14(b) demonstrates that the interpolation uncertainty largely clustered around the soil boundaries and particularly around the interbedded layers of Soil4 and Soil3, consistent with the accuracy plot shown in Fig. 14(c). In addition, as the number of line measurements increased to 11, the most probable interpolation cross-section derived from 100 realizations [Fig. 15(a)] became increasingly similar to the true test cross-section shown in Fig. 9(b), and the interpolation accuracy increased further to 98.5%. Consistent results are also reflected by the dispersion plot in Fig. 15(b), which depicts the significant shrinkage of the areas of large dispersion values and major stratigraphic uncertainty around Soil2. Similarly, the accuracy plot in Fig. 15(c) indicates that areas of major interpolation error disappear.

The good correlation observed between the dispersion map and the accuracy map demonstrates that the dispersion plot is valuable and can be used to indicate the interpolation accuracy of the

proposed method. Note that the proposed method not only can be used to interpolate the ordered geological patterns but can also be applied to more complex geological features, such as sudden changes in layers and interfolding when a representative training image is available. In practice, it might be possible to collect a group of potential candidate training images. The similarity between potential training images and available site-specific measurements can be evaluated using a distribution of runs (Mood 1940) and the multiple-point density function (MPDF) (Boisvert et al. 2007). In addition, the proposed method may possibly be extended to interpolate geological cross-sections projected to different directions by carrying out three-dimensional (3D) spatial interpolation, which requires further studies.

Real Geological Cross-Section Example

In this section, a 600 (horizontal distance) \times 40 m (depth) geological profile from a tunnel project in Australia is extracted to illustrate the proposed method. The subsurface strata of the cross-sections were interpreted from 14 boreholes by experienced geologists, and the average distance between the adjacent boreholes was approximately 43 m. Fig. 16 shows the interpreted subsurface geological profile (Golder, unpublished data, 2015). The ground consists successively of Bassendean sand (BS), Guildford formation (GF), Gngangara sand (GS), and Ascot formation (AF). Two 300 \times 20 m geological cross-sections were extracted as the training

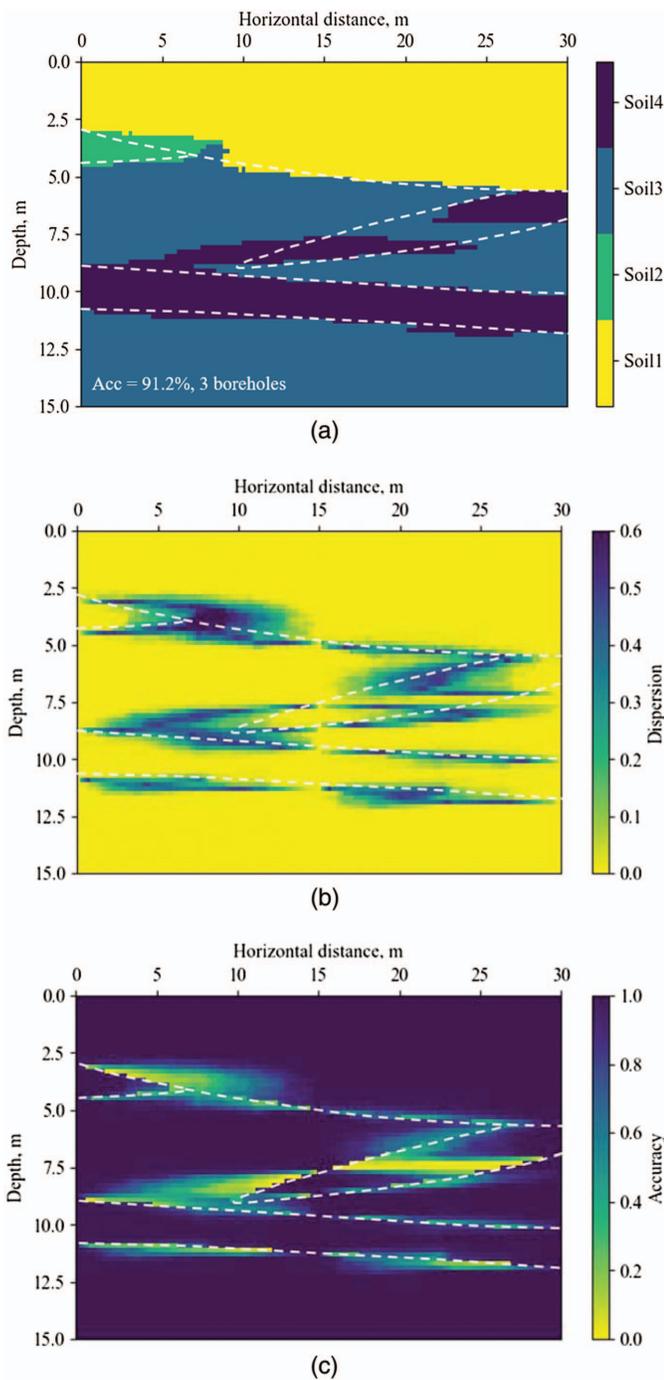


Fig. 14. (Color) Results of the proposed method conditioning on three line measurements: (a) most probable results for three boreholes; (b) dispersion plot for three boreholes; and (c) interpolation accuracy for three boreholes.

and test images. Both images were obtained from the same site and share the same geological origin. Fig. 17 depicts the training and test geological cross-sections. Note that the cross-section shown in Fig. 1 is the test cross-section in this example. Both cross-sections are considered to contain similar subsurface stratigraphic information about the site. Each geological cross-section has a total horizontal distance and vertical depth of 300 m and 20 m, respectively. Both sections have been discretized into a grid of 100×50 m, with corresponding horizontal and vertical resolutions of 3 and 0.4 m, respectively. For illustration purposes, four equally spaced vertical

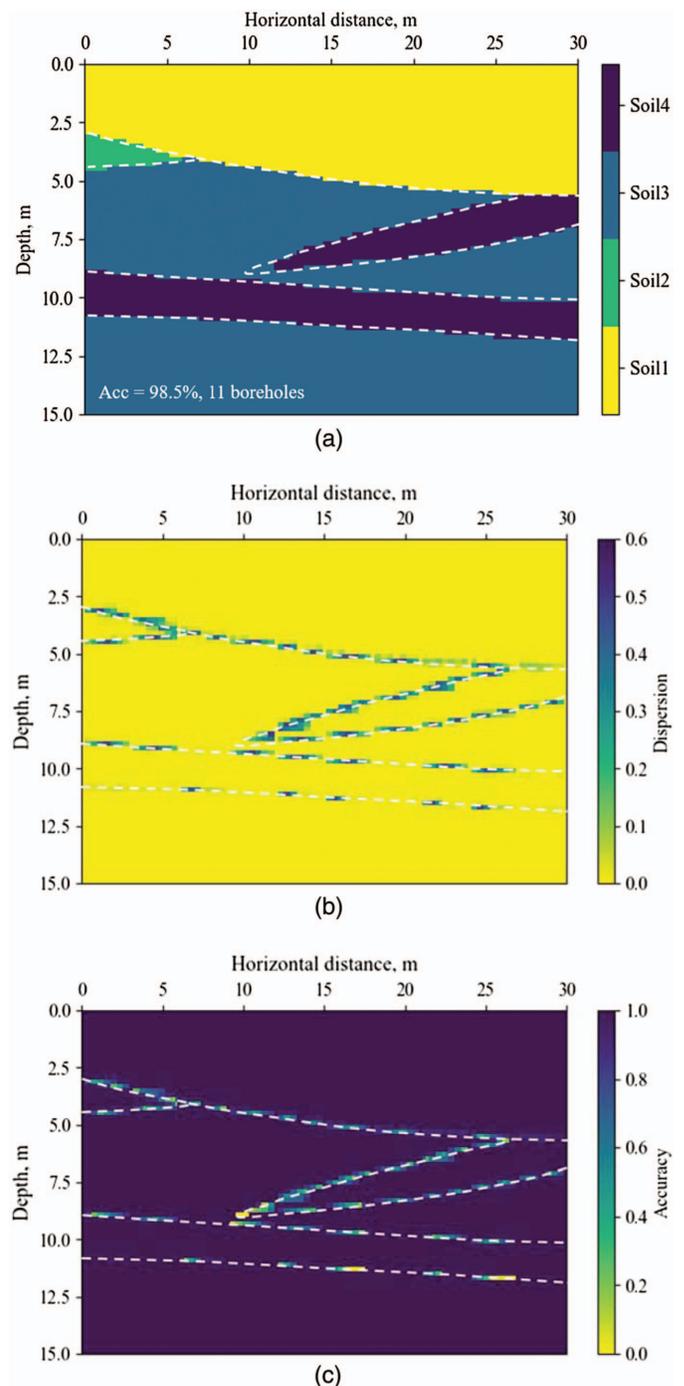


Fig. 15. (Color) Results of the proposed method conditioning on 11 line measurements: (a) most probable results for 11 boreholes; (b) dispersion plot for 11 boreholes; and (c) interpolation accuracy for 11 boreholes.

line profiles have been taken as the borehole measurements, and these represent 4% of all data points. The spacing interval between the adjacent boreholes is 100 m.

A total of 100 realizations were generated following the simulation procedures stated in Fig. 8. Fig. 18(a) shows the most probable interpolation conditioning on four boreholes. The proposed method can reasonably reproduce the spatial stratigraphic relationship between the four soil types at an accuracy of 91.2%. The true soil boundaries are also superimposed for a better comparison. The greatest difference between the most probable interpolation and test

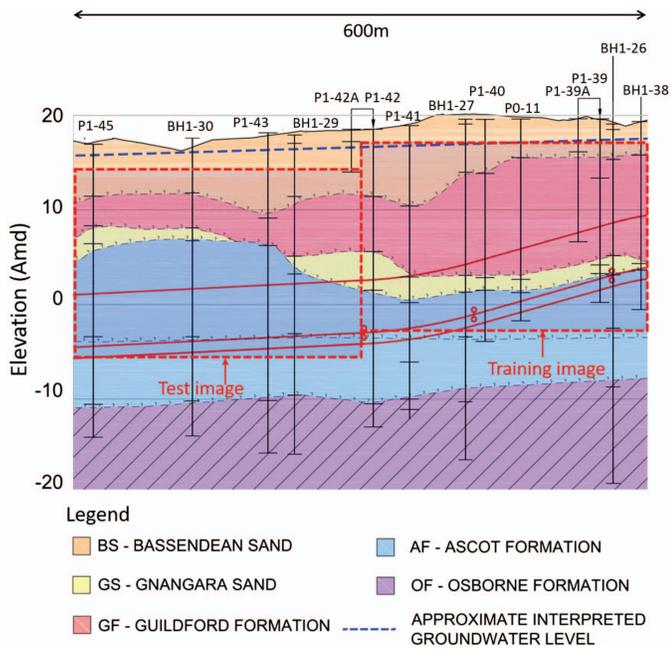


Fig. 16. (Color) Interpreted geological cross-section along a tunnel alignment in Australia. (Adapted from Golder, unpublished data, 2015.)

cross-section is observed at the intersection between GS, GF, and AF. This can be attributed to the lack of a stratigraphic interface between GF and AF within the training cross-section, as shown in Fig. 17(a). Note that the proposed method is purely data-driven, and predictions based on a less representative training image may violate the expected physical rules and knowledge in terms of geological settings. A similar phenomenon is also observed in the dispersion map in Fig. 18(b). Areas with large dispersions are clustered mainly around the stratigraphic boundaries between GS and AF.

Fig. 19 shows the comparison between the most probable interpolation and the underlying true cross-section. The interpolation accuracy associated with each soil type was calculated using Eq. (10) and plotted in Fig. 19(a). The calculated accuracy varies widely at realization numbers between 0 and 20. The calculated accuracy fluctuates most widely within GS, mainly due to the interpolation uncertainties around the intersection between GS and AF. The accuracy tends to stabilize as the realization number increases beyond 20.

Fig. 19(b) presents a color map of calculated accuracy. Similarly, the largest interpolation errors mainly cluster at the soil boundaries, consistent with the results from the dispersion plot in Fig. 18(b). To assess the performance of the proposed method more objectively, a separate analysis was performed by switching the training and test images in Figs. 17(a and b). This analysis yielded similar spatial interpolation results with an overall accuracy of 94.5%, and the interpolation uncertainties are properly reflected

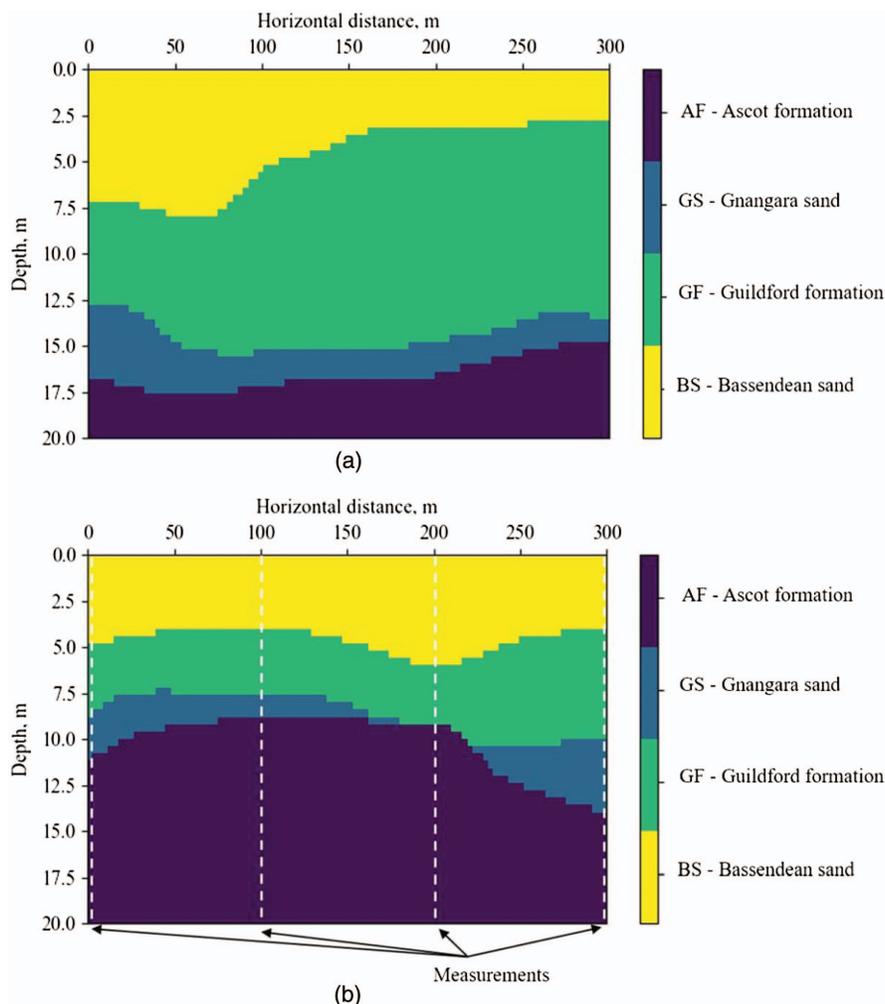
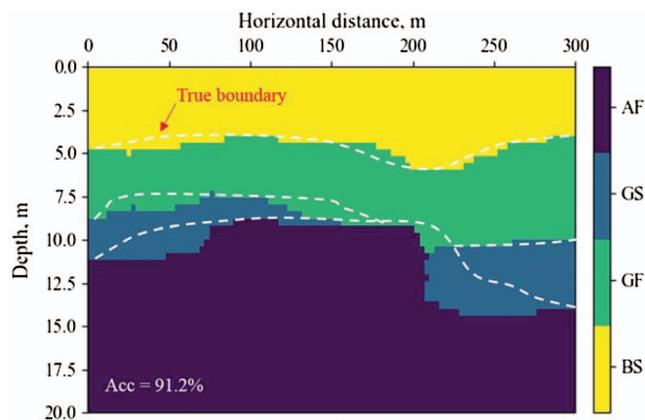
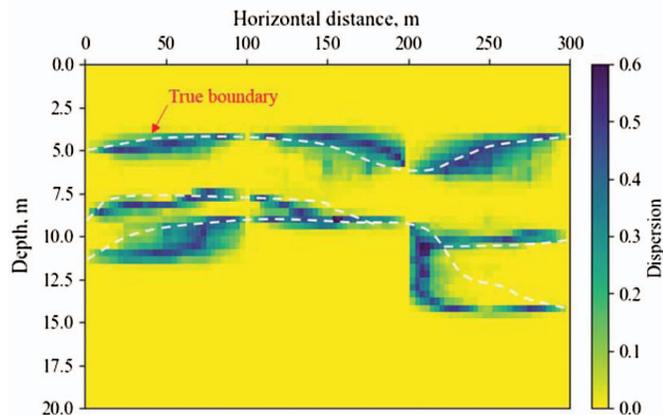


Fig. 17. (Color) Two extracted geological cross-sections from a tunnel project in Australia: (a) training cross-section; and (b) test cross-section.



(a)



(b)

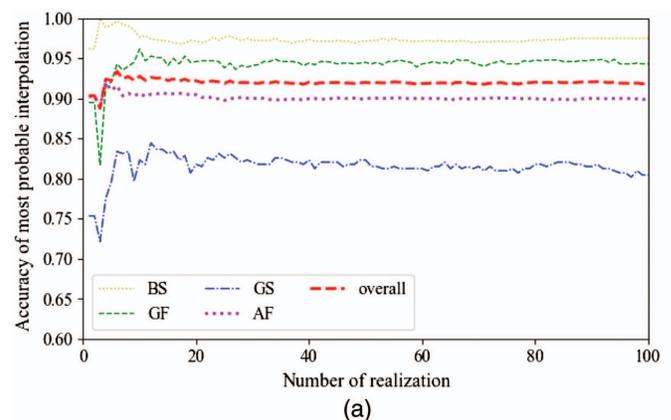
Fig. 18. (Color) Results obtained from the proposed method for the real data example: (a) most probable interpolation; and (b) spatial distribution of dispersion.

in the dispersion plot. Note that this accuracy map is not available in real engineering practice and is used only for illustration and validation in this example. However, the observed good correlation between the dispersion map and the accuracy map indicates that the dispersion plot is valuable and can be used to indicate the interpolation accuracy of the proposed method.

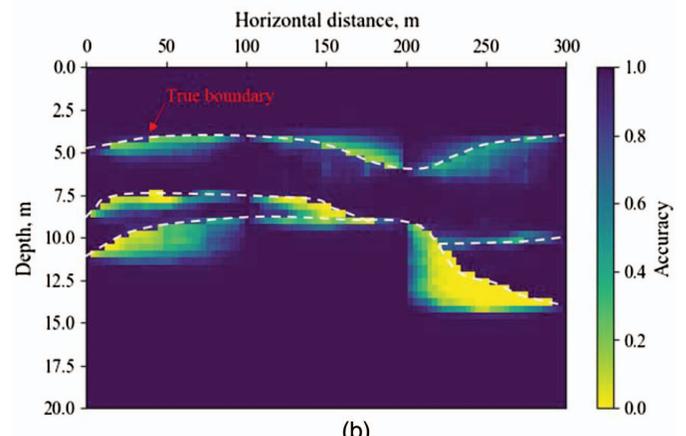
Summary and Conclusions

The delineation of subsurface geological cross-sections is indispensable for geotechnical site characterization and may have a significant impact on the subsequent geotechnical design and analysis. The conventional method used to define a geological cross-section relies heavily on the engineer's experience and judgment, which can be subjective and inconsistent. Emerging machine learning methods, such as CNNs, provide an effective alternative for the spatial interpolation of subsurface stratigraphy in an objective, quantitative, and automated manner.

To explicitly combat the intrinsic difficulty associated with the limited training geological cross-sections and sparse site-specific borehole measurements, a novel iterative convolution model based on a CNN framework was proposed in this study for the development of geological cross-sections for engineering design and analysis. In this model, multiscale grid templates adapted to the available boreholes are determined to extract the shift-invariant stratigraphic relationships of different scales from a single training cross-section



(a)



(b)

Fig. 19. (Color) Comparison between the most probable interpolation and true test cross-section for the real example: (a) variation of interpolation accuracy of each soil type with the realization number; and (b) spatial distribution of interpolation accuracy.

iteratively using a convolution operation and an edge detection filter. The extracted spatial features serve as the input for training an XGBoost classification model, which is used subsequently for the spatial interpolation of the soil type at unsampled locations. Both the simulated example and real data from a tunnel site in Australia were used to illustrate the proposed method, and the results demonstrated that the proposed IC-XGBoost method not only estimates the most probable geological cross-section with a high level of accuracy but also identifies the areas of large interpolation uncertainty. The bands of large interpolation uncertainty mainly clustered around the stratigraphic boundaries, and the extent of uncertainty decreased as more boreholes were included and vice versa.

Data Availability Statement

The developed executable file for the IC-XGBoost algorithm is available at <https://sites.google.com/site/yuwangcityu/ic-xgboost>.

Acknowledgments

The work described in this paper was supported by grants from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project Nos. CityU 11213119 and CityU 11213117). The financial supports are gratefully acknowledged.

References

- Boeckmann, A. Z., and J. E. Loehr. 2016. *Influence of geotechnical investigation and subsurface conditions on claims, change orders, and overruns. National cooperative highway research program synthesis 484*. Washington, DC: Transportation Research Board.
- Boisvert, J. B., M. J. Pycrz, and C. V. Deutsch. 2007. "Multiple-point statistics for training image selection." *Nat. Resour. Res.* 16 (4): 313–321. <https://doi.org/10.1007/s11053-008-9058-9>.
- Boureau, Y., F. Bach, Y. LeCun, and J. Ponce. 2010a. "Learning mid-level features for recognition." In *Proc., 2010 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, 2559–2566. New York: IEEE.
- Boureau, Y., J. Ponce, and Y. LeCun. 2010b. "A theoretical analysis of feature pooling in visual recognition." In *Proc., 27th Int. Conf. on Machine Learning (ICML-10)*, 111–118. Alexandria, VA: National Science Foundation.
- Chen, T., and C. Guestrin. 2016. "Xgboost: A scalable tree boosting system." In *Proc., 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, 785–794. New York: Special Interest Group on Management of Data.
- Clayton, C. R. 2001. *Managing geotechnical risk: Improving productivity in UK building and construction*. London: Thomas Telford Publishing.
- Deng, Z. P., S. H. Jiang, J. T. Niu, M. Pan, and L. L. Liu. 2020. "Stratigraphic uncertainty characterization using generalized coupled Markov chain." *Bull. Eng. Geol. Environ.* 79 (10): 5061–5078. <https://doi.org/10.1007/s10064-020-01883-y>.
- Deutsch, C., and L. Wang. 1996. "Hierarchical object-based stochastic modeling of fluvial reservoirs." *Math. Geol.* 28 (7): 857–880. <https://doi.org/10.1007/BF02066005>.
- Dumoulin, V., and F. Visin. 2016. "A guide to convolution arithmetic for deep learning." Preprint, submitted March 23, 2016. <http://arxiv.org/abs/1603.07285>.
- Elfeki, A., and M. Dekking. 2001. "A Markov chain model for subsurface characterization: Theory and applications." *Math. Geol.* 33 (5): 569–589. <https://doi.org/10.1023/A:1011044812133>.
- Hastie, T., R. Tibshirani, and J. Friedman. 2009. *The elements of statistical learning*. 2nd ed. New York: Springer.
- Juang, C. H., T. Jiang, and R. A. Christopher. 2001. "Three-dimensional site characterisation: Neural network approach." *Géotechnique* 51 (9): 799–809. <https://doi.org/10.1680/geot.2001.51.9.799>.
- Juang, C. H., J. Zhang, M. Shen, and J. Hu. 2019. "Probabilistic methods for unified treatment of geotechnical and geological uncertainties in a geotechnical analysis." *Eng. Geol.* 249 (Jan): 148–161. <https://doi.org/10.1016/j.enggeo.2018.12.010>.
- Katiyar, S. K., and P. V. Arun. 2014. "Comparative analysis of common edge detection techniques in context of object extraction." Accessed April 1, 2021. <https://arxiv.org/ftp/arxiv/papers/1405/1405.6132.pdf>.
- Koltermann, C. E., and S. M. Gorelick. 1992. "Paleoclimatic signature in terrestrial flood deposits." *Science* 256 (5065): 1775–1782. <https://doi.org/10.1126/science.256.5065.1775>.
- Koltermann, C. E., and S. M. Gorelick. 1996. "Heterogeneity in sedimentary deposits: A review of structure-imitating, process-imitating, and descriptive approaches." *Water Resour. Res.* 32 (9): 2617–2658. <https://doi.org/10.1029/96WR00025>.
- Kumar, J. K., M. Konno, and N. Yasuda. 2000. "Subsurface soil-geology interpolation using fuzzy neural network." *J. Geotech. Geoenviron. Eng.* 127 (7): 632–639. [https://doi.org/10.1061/\(ASCE\)1090-0241\(2000\)126:7\(632\)](https://doi.org/10.1061/(ASCE)1090-0241(2000)126:7(632)).
- LeCun, Y., B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. 1989. "Backpropagation applied to handwritten zip code recognition." *Neural Comput.* 1 (4): 541–551. <https://doi.org/10.1162/neco.1989.1.4.541>.
- LeCun, Y., L. Bottou, and Y. Bengio. 1997. "Reading checks with multi-layer graph transformer networks." In Vol. 1 of *Proc., IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, 151–154. New York: IEEE.
- Li, D., X. Qi, Z. Cao, X. Tang, K. Phoon, and C. Zhou. 2016a. "Evaluating slope stability uncertainty using coupled Markov chain." *Comput. Geotech.* 73 (Mar): 72–82. <https://doi.org/10.1016/j.compgeo.2015.11.021>.
- Li, Z., X. Wang, H. Wang, and R. Y. Liang. 2016b. "Quantifying stratigraphic uncertainties by stochastic simulation techniques based on Markov random field." *Eng. Geol.* 201 (Feb): 106–122. <https://doi.org/10.1016/j.enggeo.2015.12.017>.
- Ludwig, J. 2013. "Image convolution, Portland State University." Accessed April 1, 2020. http://web.pdx.edu/~jduh/courses/Archive/geog481w07/Students/Ludwig_ImageConvolution.pdf.
- Mairal, J., P. Koniusz, Z. Harchaoui, and C. Schmid. 2014. "Convolutional kernel networks." In *Advances in neural information processing systems*, 2627–2635. Cambridge, MA: MIT Press.
- Mitchell, T. M. 1997. *Machine learning*. Burr Ridge, IL: McGraw Hill.
- Mood, A. M. 1940. "The distribution theory of runs." *Ann. Math. Stat.* 11 (4): 367–392. <https://doi.org/10.1214/aoms/1177731825>.
- Pathak, D., P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros. 2016. "Context encoders: Feature learning by inpainting." In *Proc., IEEE Conf. on Computer Vision and Pattern Recognition*, 2536–2544. New York: IEEE.
- Prezzi, M., B. McCullough, and V. K. D. Mohan. 2011. *Analysis of change orders in geotechnical engineering work at INDOT*. Publication FHWA/IN/JTRP-2011/10. West Lafayette, IN: Joint Transportation Research Program, Indiana DOT and Purdue Univ.
- Qi, X., D. Li, K. Phoon, Z. Cao, and X. Tang. 2016. "Simulation of geologic uncertainty using coupled Markov chain." *Eng. Geol.* 207 (Jun): 129–140. <https://doi.org/10.1016/j.enggeo.2016.04.017>.
- Rodríguez, P., M. A. Bautista, J. Gonzalez, and S. Escalera. 2018. "Beyond one-hot encoding: Lower dimensional target embedding." *Image Vision Comput.* 75 (Jul): 21–31. <https://doi.org/10.1016/j.imavis.2018.04.004>.
- Ruder, S. 2016. "An overview of gradient descent optimization algorithms." Preprints, submitted September 15, 2016. <http://arxiv.org/abs/1609.04747>.
- Shi, C., and Y. Wang. 2021a. "Nonparametric and data-driven interpolation of subsurface soil stratigraphy from limited data using multiple point statistics." *Can. Geotech. J.* 58 (2): 261–280. <https://doi.org/10.1139/cgj-2019-0843>.
- Shi, C., and Y. Wang. 2021b. "Smart determination of borehole number and locations for stability analysis of multi-layered slopes using multiple point statistics and information entropy." *Can. Geotech. J.* <https://doi.org/10.1139/cgj-2020-0327>.
- Srivastava, N., G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. 2014. "Dropout: A simple way to prevent neural networks from overfitting." *J. Mach. Learn. Res.* 15 (1): 1929–1958.
- Steffens, C. R., L. R. Messias, P. J. Drews Jr., and S. S. D. C. Botelho. 2020. "CNN based image restoration." *J. Intell. Rob. Syst.* 99: 609–627. <https://doi.org/10.1007/s10846-019-01124-9>.
- Thongsuwan, S., S. Jaiyen, A. Padcharoen, and P. Agarwal. 2021. "ConvXGB: A new deep learning model for classification problems based on CNN and XGBoost." *Nucl. Eng. Technol.* 53 (2): 522–531. <https://doi.org/10.1016/j.net.2020.04.008>.
- Van Vliet, L. J., I. T. Young, and G. L. Beckers. 1989. "A nonlinear laplace operator as edge detector in noisy images." *Comput. Vision Graphics Image Process.* 45 (2): 167–195. [https://doi.org/10.1016/0734-189X\(89\)90131-X](https://doi.org/10.1016/0734-189X(89)90131-X).
- Wang, Y., Y. Hu, and T. Zhao. 2020. "CPT-based subsurface soil classification and zonation in a 2D vertical cross-section using Bayesian compressive sampling." *Can. Geotech. J.* 57 (7): 947–958. <https://doi.org/10.1139/cgj-2019-0131>.
- Zhang, W., K. Itoh, J. Tanida, and Y. Ichioka. 1990. "Parallel distributed processing model with local space-invariant interconnections and its optical architecture." *Opt. Soc. Am.* 29 (32): 4790–4797. <https://doi.org/10.1364/AO.29.004790>.
- Zhang, W., R. Zhang, C. Wu, A. T. Goh, and L. Wang. 2020. "Assessment of basal heave stability for braced excavations in anisotropic clay using extreme gradient boosting and random forest regression." *Underground Space.* <https://doi.org/10.1016/j.undsp.2020.03.001>.