ORIGINAL PAPER



Refined implicit characterization of engineering geology with uncertainties: a divide-and-conquer tactic-based approach

Mingchao Li¹ · Chuangwei Chen¹ · Hui Liang¹ · Shuai Han^{1,2} · Qiubing Ren¹ · Heng Li²

Received: 27 November 2023 / Accepted: 22 May 2024 / Published online: 18 June 2024 © The Author(s) 2024

Abstract

In engineering geology, a reasonable assessment of the spatial distribution of uncertainty in a region is vital in guiding research, saving money, and shortening the period. However, the traditional modeling process requires a lot of manual interaction, and the uncertainty of the geological model cannot be accurately quantified and utilized. This paper proposes a novel implicit geological modeling and uncertainty analysis approach based on the triangular prism blocks, which is divided into data point acquisition, ensemble model with divide-and-conquer tactic (EMDCT), uncertainty analysis, and post-processing. By employing machine learning algorithms, the EMDCT gives superior results for implicit modeling. The sensitivity analysis of the prediction results is further evaluated via information entropy. According to the distribution of uncertainty, supplementary boreholes are selected as additional knowledge to retrain the local components of the model to enhance their performances. The implicit modeling method is applied to real hydraulic engineering problems by employing the EMDCT, and the proposed model has obvious advantages in the implicit geological characterization. The overall accuracy in the working area with sparse boreholes reaches 0.922, which is 0.013 higher than the traditional method. By evaluating the distribution of uncertainty, an accuracy of 0.962 can be achieved, which is equivalent to reducing 10 boreholes.

Keywords Geological modeling · Borehole · Implicit modeling · Uncertainty · Ensemble model · Divide-and-conquer

Introduction

3D geological modeling, which essentially relies on boreholes, shallow sections, *in-situ* test data, and geological outcrops, is vital for the interpretation of geological structures (Wang et al. 2020; Zhou et al. 2019), serving as a

Highlights

- A geology model with sparse boreholes is constructed by the method of implicit.
- The accuracy of formation prediction is improved by the EMDCT approach.
- Uncertainty analysis can guide the supplement of boreholes and model updating.
- The overall accuracy of the working area is 0.962 and the accuracy of the dam site is 0.974.

Shuai Han shuaihan@polyu.edu.hk

- ¹ State Key Laboratory of Hydraulic Engineering Intelligent Construction and Operation, Tianjin University, Tianjin 300350, China
- ² Department of Building and Real Estate, The Hong Kong Polytechnic University, Hong Kong 999077, China

basis for subsequent mechanical analysis (Chen et al. 2023), safety assessment (Chen et al. 2021), and construction planning of engineering construction (Aghamolaie et al. 2019; Schaaf and Bond 2019). However, due to the complexity of geological conditions in most projects (Chen et al. 2022a), especially large-scale hydraulic engineering projects, and the limitation of the investigation, the process of geological interpretation has great uncertainty. The modeling of ground and underground conditions is a challenging issue, especially in the case of complex geometries and spatially variable geotechnical properties (Petrone 2023; Antonielli 2023) Hence, it is necessary to use the limited primary data to predict the underground space and provide an immediate preliminary geological model, especially for the dam site. The exploration drilling plan is further guided by the uncertainty analysis and then the refined implicit characteristics are reconstructed. Explicit modeling mainly deals with deterministic methods for large-scale structures (Tian et al. 2023), and heavily relies on manual interpretation which is prone to subjective errors. Although scholars have proposed the semi-automated method (Song et al. 2019), it also has difficulties in assessing uncertainty (Ouyang et al.

2023), which is essential in guiding the investigation. By contrast, implicit modeling has the characteristics of fast modeling speed, less subjective interference, and a strong mathematical foundation. Especially in the current prosperity of data science, implicit modeling has gradually received more and more attention and has played an important role in uncertainty assessment investigation, survey work guidance, and decision-making (Guo et al. 2022).

Implicit modeling essentially employs the spatial interpolation method (Lorensen and Cline 1987), fitting the space surface function to the sampled data, resulting in visualization only in the final output. Implicit modeling generates a scalar field for the entire geological study area (Guo et al. 2021; Lipus and Guid 2005; Scalzo et al. 2022; Zhong et al. 2021), and common reconstruction techniques include, radial basis function (RBF) method (Macêdo et al. 2011), moving least squares (MLS) (Wan and Li 2022), and surface reconstruction approach based on the stochastic velocity fields (Yang et al. 2019). Recently, machine learning has been integrated into implicit modeling to quickly generate more accurate models. Guo et al. (2021) proposed a borehole-based implicit modeling approach that offers new possibilities for controllable 3D automated modeling. Shi et al. (2021) employed the marching cubes (MC) method to reconstruct the interpolating implicit function surface. A combination of implicit modeling and machine learning is capable of effectively enhancing the overall modeling accuracy by leaving the implicit surface-solving process to the algorithm (Hou et al. 2023; Wang et al. 2023; Yang et al. 2022). However, there are still several challenges associated with implicit modeling: 1) Assessing the limited model uncertainty; 2) Difficulty in providing accurate implicit specifications for non-renewable models. Commonly, geological conditions with strong anisotropy and sparse survey data make it difficult for experienced geological experts to infer definitive geological structure, uncertainty being the most acute challenge (Lindsay et al. 2012; Madsen et al. 2022; Ouyang et al. 2023; Zhao et al. 2023a). Due to formation complexity, measurement uncertainty, data scatter and heterogeneity (Shi et al. 2021), surface construction uncertainty (Schaaf and Bond 2019) and random error of sampling, uncertainty exists in most geologic model construction areas. Conventional experience is able to solve the formation and judgment of the surface geological structure, but not the deep formation and large-scale evaluation (Wang et al. 2022). Commonly, deterministic models are not appropriate for examining the uncertainty of geological structure and hinder model renewal evaluations. By making use of the implicit modeling method, the uncertainty expression (Grose et al. 2018; Røe et al. 2014) is appropriate based on geographic statistics. With the aid of quantitative simulations, it is possible to precisely locate the formation zone with a high degree of certainty. Our assessment of the level of uncertainty enables us to recommend supplementary boreholes, ensuring the most comprehensive and accurate results.

Many machine learning algorithms, such as neural networks (Shi and Wang 2022, 2021b; Titus et al. 2022), mixed density networks (Han et al. 2022), random forest, deep forest, and SVM are utilized to enhance accuracy in all types of lithology identification, geological tectonic prediction, and geological modeling. The formation, mineral prediction, and identification have been extensively implemented in geological disciplines (Deng et al. 2022; Harris et al. 2023; Ren et al. 2022; Zhao et al. 2022; Difficulty in generalizing is capable of precisely identifying stratum distributions, there are several challenges that must be addressed: 1) Dependency on large training datasets; 2) Difficulty in generalizing a large area via a single model. This means that the machine learning algorithm used in the implicit model requires more processing of the data to enhance the modeling accuracy, especially the refined features.

Considering the above engineering problems, such as frequent manual interaction, difficult model updating, sparse data, incomplete detail refinement, and difficult uncertainty assessment in geological modeling, an implicit modeling and uncertainty analysis method based on ensemble modeling is proposed and applied to the engineering geological exploration, particularly for dam construction. The chief structure of the present paper is as follows: Section 2 introduces the detailed workflow of the implicit modeling approach and the main utilized research methods. Section 3 introduces the real case study of Yunnan Province, China. Section 4 outlines the prediction of the model and discusses the implicit modeling rules, triangular prism block method, and uncertainty analysis in this process. Section 5 presents the main obtained results.

Methodology

The detailed workflow

An implicit modeling method is proposed. The workflow of the approach is mainly divided into three parts (Fig. 1): preprocessing, main process, and post-processing. Delaunay-triangular prism, implicit modeling, and uncertainty quantization methods mentioned in these processes are introduced in the following sections.

Pre-processing on investigation data

Sample points of geological data

Assumptions and supplements to existing borehole data are needed to contain more formation information. Generally, there are two methods: (i) The profile is assumed by connecting the boreholes (Fig. 2(a)). The formation can be considered fixed at a certain distance, so two boreholes at this distance can be simply connected linearly to simulate the formation structure between them, that is, a plane is formed in space



Fig. 1 Main flowchart of the implicit modeling method

Fig. 2 Schematic representation of the borehole data sampling: **a** Sampling on the line connecting two adjacent boreholes, **b** Sampling around the borehole and labeling the sample points



between two vertical lines, and the samples are randomly collected on this plane; (ii) Borehole information radiates around (Fig. 2(b)). According to the central limit theorem (Ash and Doleans-Dade 1999; Resnick 2013), a substantial proportion of measurement errors adhere to the normal distribution (Sukhov 1999). Similarly, it can be assumed that the geological structure does not mutate on a small-scale region relative to the understudy area; in other words, the lithology at the central point of the borehole radiates to the surrounding area with a certain probability. The closer to the center of the borehole, the greater the probability. The normal distribution is characterized by concentration near the mean, symmetry, and uniform decline. To meet the engineering requirements in this paper, we use two-dimensional normal distribution here. So, normal distribution tries to reasonably disperse the data amount in space, so that the sample points carry weak information near the borehole locations. The existing boreholes are regarded as cylinders with radius $N(0, \sigma)$ of normal distribution. In this spatial region, the vertical directions are divided into 30 parts on average. In each sub-cylinder space, 20 sampling points are randomly generated, and the strata of borehole sampling points are appropriately labeled according to the relationship between the elevations of the random sampling points and the borehole formation (see Fig. 2(b)). According to the normal distribution law, almost all sampling points are in a circle centered on the borehole and radius 3σ . This means that compared to the center of the borehole, the outside stratum category possesses a probability of being the same as that obtained at the same elevation of the borehole, but there is uncertainty. The sampling point p(x, y, z) was employed as the training input, and the formation number S as the label to complete the labeling of all datasets. In particular, in order to reflect the ground surface, it is necessary to mark the cylindrical space of the part above the ground elevation of the borehole as S=0.

$$f(p) = f(x, y, z) = S \tag{1}$$

In addition to the sampling points, we also need to condense the scattered points of the mesh arrangement as much as possible in the space to be retrieved to form the test set of the classifier. In actual modeling, these scattered points are considered as points of unknown strata, and the classifier must provide its own strata to complete the construction of the entire geological space. Geological exploration research areas usually have ground surface contours that can be integrated into the modeling process. When sampling points and grid scattered points are taken in space, it must be determined whether the location of the point is above the surface or not, and if so, the sampling point is removed (Fig. 2(a)). After filtering with prior knowledge, all the sampling points and grid scatter points are below the surface.

Delaunay-triangular prism blocks

Due to the large region of engineering geology and the significant regional characteristics of the study area (such as local small strata, intrusions, or special local geological structures). The nonlinear fitting ability of a single machine learning algorithm is not enough to accurately characterize all the characteristics of the geological structure, and local special structures cannot be well fitted. Therefore, it is needed to divide the research area into different sections and let each algorithm learn them independently. To improve the training effect of classifiers for details, the understudy region is divided into several parts, and then several classifiers are appropriately trained (divide-andconquer), whose results are combined to form prediction results for the final study area. Divide-and-conquer possesses some special features (Fig. 3(c)): (i) Simple classifiers predict simple regions: The ensemble model is a process of repeated calculation that includes a small range for a small triangular prism. So, we can avoid using overly complex classifiers and reuse many simple classifiers; (ii) Avoiding excessive data imbalance: Ensemble models include local geologic regions, so global space data imbalance has little effect on the machine learning effect of a single triangular prism; (iii) The preferred scheme is selected: Different classifiers can be selected according to the characteristics of different triangular prisms to achieve optimal prediction results, and the overall modeling accuracy is improved by "divide-and-conquer". Divide-andconquer improves accuracy because the geology is complicated, which is an "ensemble learning" strategy that we put forward to solve this geological problem. Common methods of dividing the region are the Delaunay triangle, general triangle, Tyson polygon (also known as the Voronoi diagram), and ordinary quadrilateral mesh. Different algorithms are applied to geological regions after region division (see Table 1, Fig. 3(a)). The region division mode must meet the project requirements: (i) The particularity of its engineering lies in the uneven distribution of boreholes. (ii) The area to be divided should be related to the distribution of boreholes, ensuring that there are sampling points in each area. (iii) Partitions need not be too dense or too sparse in one subblock. (iv) The time complexity of generating partitions is small, which is convenient for drilling and updating. For the above four reasons, Delaunay approach is chosen to divide the region with appropriate triangles horizontally and then stretch them vertically to obtain a number of Triangular prisms, the so-called Delaunay-triangular prism. Suppose V is a finite set of points on a two-dimensional field of real numbers, edge e represents a closed line segment consisting of points in the point set as endpoints, E is the set of e, and the vector stretched vertically is denoted by \mathbf{Z} . The triangular prism set G can be obtained from the point set V (see Fig. 3(b)). In this paper, the triangular prism set G is generated based on the (X, Y)coordinates of the existing borehole positions as point set V. Include all sampling points in the cylinder contained in the outer circle stretched on the base of the triangular prism into the training set (this circle can be enlarged appropriately to allow the training set of different prisms to intersect). All scattered grid points contained in each prism are then included in the test set.

All scattered grid points contained in each prism are then included in the test set. Delaunay-triangular prisms (Liu et al. 2023) possess the same properties as Delaunay triangles: (i) A triangular prism is constructed with the locations of



Fig. 3 (a) Region division method; (b)Schematic representation of the Delaunay-triangular prism; (c) The special features of the Delaunay-triangular prism block in divide-and-conquer.

Methods of dividing the region	Delaunay triangle	Tyson polygon	Ordinary quadrilateral mesh
Time complexity	O(n*logn)	<i>O</i> (n ²)	-
Basis of dividing	Borehole location and boundary location	Delaunay triangle	Boundary only
Sub-prism	Triangular prism with the smallest radius of the outer circle	Polygon prism. difficult to manage the data.	Quadrangle
Training set data	Several boreholes near a sub-prism, which can represent the sub-prism	A sub-prism has only one borehole. The selection of boreholes used for modeling each prism is difficult.	Some sub-prisms may have many boreholes, and some sub-prisms may have no boreholes.
Stratigraphic characteristics in sub-prism	The stratum is simple due to the small partition	The stratum is relatively simple.	The stratum is simple due to the small partition

Table 1 The characteristics of different methods of dividing the region

boreholes, which guarantees that every point in the region belongs to a triangular prism. (ii) No two surfaces or edges in the triangular prism set G will ever intersect, ensuring that the triangular prism to which each sample point belongs is unique. (iii) Adding, omitting, or moving a vertex affects only the adjacent triangular prism, so that only a small part of the classifiers corresponding to the prism should be updated.

Main process

Machine learning

Herein, the problem of implicit modeling represents a multiclassification problem. For this purpose, the sampling points in the previous section are utilized as training sets to predict the stratigraphic classification of scattered grid points (i.e., test sets), restoring the strata of the entire geological space. The commonly used algorithms for multiple classifications are: deep forest (DF) (Han et al. 2023; Zhu et al. 2023), random forest (RF) (Chen et al. 2022b), decision tree (DT), extreme gradient boosting (XGB) (Shi and Wang 2021a; Zhang et al. 2023), categorical boosting (CB) (Prokhorenkova et al. 2017), support vector machines (SVM) (Abedi et al. 2012; Fan et al. 2023), and K-nearest neighbor (Bullejos et al. 2022; Wang et al. 2018). In this paper, seven classification algorithms are aimed to be employed for implicit modeling. Compared with the ensemble model, an approach is selected from the aspects of overall accuracy, robustness, borehole data requirements, and uncertainty asseement. The selection of classifier hyperparameters is carried out in the form of a grid search. The optimal combination of hyperparameters selected and the prediction effect of boreholes in the test set are presented in Table 2.

Implicit modeling

Implicit modeling is the reconstruction of potential function. It includes Euclidean space and implicit space (Adzhiev et al. 2000): The real-valued function is given on the whole Euclidean space, and the volume object is defined by the inequality f(p) < 0, where f(p) = 0 represents the equipotential surface (i.e., the geological surface). As seen in Eq. (2), point $p_i(x_i, y_i, z_i)$ can be written as a 3D potential field in Euclidean space \mathbb{R}^3 . In addition, in a twodimensional implicit space, an operator G is defined to unify different real-valued functions f(p) into a model. As given in Eq. (3), the binary operator G is applied to the potential functions f_1 and f_2 , and the two-dimensional potential function G(X, Y) can be written as a combination of the 3D potential functions f_1 and f_2 in the two-dimensional implicit space, which represents the intersection of two geological surfaces. The point $P(f_1, f_2)$ defined in \mathbb{F}^2 is expressed as the intersection line of surface f_1 and f_2 in \mathbb{R}^3 , and the equipotential line G(X, Y) = 0 defined in \mathbb{F}^2 is expressed as the part of the geological surface that is extruded along the intersection lines of f_1 and f_2 in \mathbb{R}^3 . (see Fig. 4).

$$f_i : \mathbb{R}^3 \to \mathbb{R}$$

$$p(x, y, z) \to f_i(p) = f_i(x, y, z)$$
(2)

$$G : \mathbb{F}^2 \to \mathbb{F}$$

$$P(f_1, f_2) \to G(P) = G(f_1, f_2)$$
(3)

We combine implicit modeling with machine learning to reconstruct implicit functions based on borehole data. The label information of the sampling points is appropriately learned by the machine, and then the relationship is mapped and the scattered points are predicted. The difference with general implicit modeling is that a specific correspondence rule (i.e., an implicit function equivalent) is configured inside the classifier, and such correspondence rules are not analytically expressed but exist in the classifier at any time to call. The spatial surface function f(x,y,z) = 0 is fitted based on the difference in the sampling data, and the formation of the spatial voxel is appropriately predicted. Finally, the 3D visualization model is generated by computer visualization technology. 3D geological implicit modeling is based on borehole data Table 2Optimalhyperparameter combination

Algorithms	Hyperparameter combination
SVM	C=1.0, kernel="rbf", gamma=0.33333; max_iter=4000
CB	iterations = 1000; learning_rate = 0.03; l2_leaf_reg = 3; depth = 6
DF	n_estimators = 4, n_trees = 500, max_layers = 20
RF	n_trees = 100, criterion = "gini", min_samples_split = 2, min_samples_leaf = 1
KNN	$leaf_size = 30$, n_neighbors = 5
XGB	learning_rate = 0.1, n_estimators = 450, max_depth = 5, min_child_weight = 6
DT	criterion = "gini", min_samples_split = 2, min_samples_leaf = 1



Fig. 4 The correspondence of the geometric elements between the implicit space and the Euclidean space

in the understudy area whose processes can be derived (see Fig. 5).

Uncertainty analysis

The expression method of geological model uncertainty essentially includes the geological statistical-based model (Dell'Arciprete et al. 2012; Madsen et al. 2022), probability-based model (Wang et al. 2017; Wang 2020), and weighting model(Fatehi et al. 2020). Information entropy

is frequently implemented as a quantitative indicator of the information content of a probabilistic system. After predicting the formation result, the information entropy of each voxel is suitably provided to measure its uncertainty quantitatively. The definition of information entropy (Shannon 1948) can be interpreted here as: the weight of the product of the natural logarithm in the predicted probabilities for each layer, and the corresponding formula reads:

$$H(x) = -\sum_{x \in X}^{n} p(x_i) \ln p(x_i)$$
(4)

where *x* represents the probability distribution of any random event, *X* is the set of possible values of *x*, *n* denotes the number of layers, and $p(x_i)$ signifies the probability value of the *i*-th value type. The more average the probability distribution of each value is, the greater the entropy will be, and the greater the uncertainty of the random variable is. The entropy ranges in the following form:

$$0 \le H(x) \le \ln n \tag{5}$$



implicit geological modeling

Fig. 5 The process of 3D

According to Eq. (4), we see that information entropy possesses the following good properties: (i) Non-negative; (ii) Monotonicity: the higher the probability of an event, the lower the information it carries; (iii) Accumulative: the measure of total uncertainty of multiple random events occurring at the same time can be expressed as the sum of measures of uncertainty of each event. The number of strata *n* in the geological layer is certain; moreover, to unify the uncertainty measure in the range of 0 to 1, multiply the coefficient of $\frac{1}{\ln n}$ in front of the formula (Eq. 6):

$$H_1(x) = -\frac{1}{\ln n} \sum_{x \in X}^n p(x_i) \ln p(x_i)$$
(6)

where $H_1(x)$ denotes the system information entropy after standardization.

The intuitive sense of the information entropy value is fuzzy, and it cannot be employed to sense the uncertainty of geological models like confidence. In the probability representation of implicit modeling, the probability of a voxel is mostly placed between two adjacent strata, and the probability of other strata outside the two strata is almost zero. Therefore, we assume that each voxel is at most between two types of strata, and the result of Eq. (6) can be appropriately approximated as standardized binary information entropy (Wellmann 2013; Wellmann and Regenauer-Lieb 2012), so as to conveniently relate information entropy and confidence per the following relation:

$$H_1(x) = -\frac{1}{\ln n} \sum_{x \in X}^n p(x_i) \ln p(x_i) \approx -\frac{1}{\ln 2} (\beta \ln \beta + (1 - \beta) \ln(1 - \beta))$$
(7)

where β represents the probability of the binary information entropy system, that is, the confidence.

Post-processing

Boreholes supplement

The post-processing process is mainly to improve the prediction of the classifier and form the final result. In the geological exploration process, relatively sparse data can be used to quickly obtain a preview of the spatial geological structure and analyze the uncertainty to guide the planning of subsequent boreholes. When there is sparse data, the intuitive reflection is that the information entropy in this region is large, and the place with large information entropy may not predict accurately, expecting more borehole information. The accuracy of geological space can be effectively improved by adding artificial boreholes in these areas. The investigation of the effect of the supplementary borehole has been provided in Sect. 3.5.3.

Detail refinement

A local interpolation of the scattered grid points can be performed at the predicted formation interface, which can iteratively approximate the actual interface. As illustrated in Fig. 6, the length of the first grid scatter interval is set as 5 m, and the sudden change of formation between the scattering points A and B proves that the interface passes through here. The grid was interpolated between A and B (the grid is divided into 10 parts, and the interval length is set as 0.5 m). After re-prediction, the formation jump occurred between C and D, and it can be considered that the interface appeared at a random location within 0.5 m. The interpolation points can be randomly generated at the adjacent scattered grid points with sudden changes in layer types, and the scattered points located at the interface can be extracted (see Eq. (8)). The point cloud formed by each interface



Fig. 6 Grid points adjacent to the interface

smoothly simulates a surface, and the position of this surface represents the interface position of the two soil types. Therefore,

$$\boldsymbol{P}_{in} = \boldsymbol{r} \cdot \boldsymbol{P}_0 + (1 - r) \cdot \boldsymbol{P}_1 \tag{8}$$

where P_{in} represents random interpolation points; P_0 , P_1 represents the vectorial form of two adjacent scattered grid points of abrupt change in stratigraphic types, and r can be set as 0.5. After the surface generation by nonuniform rational B-spline (NURBS), solid cutting is performed to configure an appropriately complete geological model.

Case study

General description of the project

The geological model of the understudied area has been reconstructed using measured boreholes. It is located downstream of the Jinsha River on the border of the Shuifu and Sichuan Province in Yunnan, China. The dam area can be considered as a cube with dimensions of $1,300 \text{ m} \times 1,000 \text{ m} \times 800 \text{ m}$, with compact boreholes. Being located in a mountainous area and a river valley, the land surface has many fluctuations, so its geological conditions affect the construction of the dam. The interlayer shape has been distorted, some layers are relatively thin, and flow erosion destroys many layers, which makes the modeling of the problem somewhat difficult. The understudied area exhibits an obvious tectonic uplift, particularly the bottom layer, which is a favorable structural zone for dam construction and river storage. The dam area is surrounded by river valleys, mountains, and mountains (see Fig. 7). There are a total of 131 boreholes in the area of the dam site (1,000 m long and 800 m wide, averaging 163.75 borehole / km²). There are only 63 boreholes in the entire outer enclosure area $(2,353 \text{ m} \log \text{ and} 2,380 \text{ m} \text{ wide}$, averaging 34.64 borehole / km², 13.21 borehole / km² in the overall area). On average, the less boreholes per unit area, the more scattered the borehole distribution, so this part of the borehole area is more scattered. The distribution of the boreholes is very limited and uneven due to the ground and surface conditions. In this research, the totally 194 boreholes have been used to finely reestablish the understudied area near the dam site. A model developed by explicit method based on boreholes, shallow sections, and geological outcrops, which is considered perfectly correct, is for the approach validation.

Data set preparation

The data of 149 randomly selected boreholes in the working area were sampled and 89,956 sampling points were obtained, which are capable of predicting the state of formation of the dam site with compact boreholes. At the same time, 131 boreholes were sampled, and 86,843 sampling points were obtained, which are aimed at reasonably predicting the state of the entire layer of the area with sparse boreholes. Obviously, it is more difficult to model with sparse drilling data in the overall area, which are basically modeled using interpolation and estimation methods. The trained classifier uses the test set (test set: $100 \times 130 \times 80$ grid points in the research area of the dam site $+80 \times 80 \times 40$ grid points at the edge of the research area, adding previous conditions and removing points above the surface, a total of 876,593 grid points) for the validity of the accuracy of the proposed method in this paper. The coordinates of various boreholes often possess different numerical size scales, which could have a special effect on the training of models, especially those constructed based on distance and similarity. Therefore,



Fig. 7 Map of the understudied area adjacent to the dam site

standardization is a vital step to remove the effect of numerical size. After standardization, the input training dimensions X, Y, and Z are appropriately scaled to be close to zero, which is more suitable for comprehensive comparative evaluation. The standardized formula reads:

$$X_n = \frac{x - \bar{x}}{\sigma} \tag{9}$$

The research area is divided into 7 strata, from top to bottom, respectively numbered as 1–7. There is an obvious imbalance in the distribution of data among classes: Class 1, 2, and 3 strata represent small strata that make up only less than 10%, Class 1 strata even account for only about 2%, and the number of samples of a few classes is small. The lithological imbalance between the classes greatly affects the judgment of the machine learning model. Since the borehole-based sampling points are essentially concentrated in a cylindrical space, it is easy to risk overfitting if synthetic minority over-sampling technique (SMOTE) is applied to the relatively small formation (calculate the Euclidean distance and select a point among K nearest neighbors as the endpoint of the interpolation, as given by Eq. (10)).

$$X_n = x + random(0,1) \cdot \sqrt{(x - x_n)^2}$$
(10)

The subtriangular prism block does not involve all layers, and therefore, the effect of disequilibrium between lithological classes is noticeably reduced. The the dataset has been illustrated Fig. 8.

Accuracy evaluation

Evaluating indicators

In the problem of multi-classification of strata, especially in the case of complex geological conditions, multiple interfaces, and unbalanced volume ratio between layers, the prediction result of unbalanced classification should be appropriately highlighted. To this end, herein, the following classification evaluation indices are utilized: accuracy, precision, recall, and F1, and the corresponding relations read the following formulations:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$
(11)

$$precision = \frac{TP}{TP + FP}$$
(12)

$$recall = \frac{TP}{TP + FN}$$
(13)

$$F_1 = \frac{2 \cdot precision \cdot recall}{precision + recall}$$
(14)

where TP (True Positive) indicates the number of positive classes predicted correctly; FP (False Positive) indicates the number of negative classes predicted as positive; TN (True Negative) indicates the number of negative classes predicted correctly; FN (False Negative) predicts the number of positive classes as negative classes.

When evaluating the detail prediction, the accuracy of the interface trend should be measured and the degree of correlation and error between the actual surface and the predicted surface should be quantified. The root mean square error (*RMSE*) and the correlation coefficient (R^2) were chosen to configure the evaluation system of the detailed prediction results. The formula is as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (O_i - T_i)^2}{n}}$$
(15)

$$R^{2} = 1 - \frac{\sum_{i=1}^{n} (O_{i} - T_{i})^{2}}{\sum_{i=1}^{n} (T - T_{i})^{2}}$$
(16)

where O_i represents the prediction value of z coordinates at the *i*-th interface, T_i denotes the real value of z coordinates of the corresponding scatter point at the *i*-th interface, T is the mean value of the true z coordinates of the entire interface, and n signifies the number of point pairs.

Overall degree of accuracy comparison

The geological regions are divided into 62 Delaunay-triangular prism blocks, and seven classifiers are employed to predict each block. The maximum accuracy of each block is combined to configure an ensemble model, the so-called ensemble model with divide-and-conquer tactic (EMDCT), which is compared with that of the other seven separate classifiers. The grid points in the understudied area are able to represent the global prediction of the algorithm, and the results obtained are given in Table 3, the confusion matrix is illustrated in Fig. 9, and the recognition effect of each stratum is demonstrated in Fig. 10 on the bar chart (where the corresponding data are presented Table 4).

As can be seen from Table 3, most of the classifiers predict the dam site area with an accuracy above 0.9, and the prediction result is good. Meanwhile, the EMDCT exhibits the highest accuracy of the scattered point of the dam site, which reached 0.973, followed by SVM and CB with an accuracy of 0.962 and 0.953. Similarly, the EMDCT possesses the highest prediction accuracy for the whole area with scattered boreholes, reaching Fig. 8 Class distribution of samples: a Training set, b Test set





0.922. This was followed by the SVM and CB with accuracy rates of 0.909 and 0.894, respectively. Obviously, the EMDCT has the best prediction under the conditions of various sparsity of boreholes, different regional sizes, and different geological structures. In the case of scattered boreholes, the accuracy is improved, which is 0.013 higher than 0.909. In compact drilling conditions, the accuracy improvement is limited, but it can still be improved by 0.011 with a higher accuracy of 0.962.

In Fig. 10 and Table 4, the predicted results of each stratigraphic class are visually displayed. Transversely, each stratum is regenerated to varying degrees. The performance of large strata (strata 5, 6, and 7) is generally good. Especially in strata 7, the indices of all algorithms are basically around 0.95, and the F1 value of each algorithm is above 0.9. Such a fact reveals that all the algorithms exhibit good prediction on massive and large strata. The performance of small strata

Table 3 Accuracy of each algorithm

Algorithm	Accuracy of dam site area (149 boreholes)	Accuracy of the overall area (131 boreholes)
EMDCT	0.973	0.922
SVM	0.962	0.909
CB	0.953	0.894
DT	0.931	0.879
RF	0.919	0.866
KNN	0.917	0.868
XGB	0.901	0.858
DT	0.887	0.833

(strata 1, 2, 3, and 4) is not satisfactory: Strata 1, 2, and 3 are located in the valley cover, erosion, more complex structure, and have a border with the surface, and the DT and XGB algorithms are relatively poor in predicting its effect. Strata 1 and 4 in order are a thin cover layer and an interlayer with a small volume, and the prediction effect is average. Commonly, the prediction of some types of samples is low, and the imbalance between layers is a factor that limits the prediction in common classifiers. Longitudinally, the prediction performance of each algorithm is different. DF, RF, DT, XGB, and KNN have different weaknesses for different formations with various shapes and topological relationships, and the recall or accuracy is low. On the contrary, the EMDCT performance is relatively stable in each layer, the *recall* and the *precision* are similar, and the F1 value is stable and maintains a high level, even in small layers that are considered to be of good robustness followed by the SVM and CB.

Refined implicit characterization of the interface

We found that the EMDCT exhibits the best effect on the formation prediction, and the combination of various algorithms suitably cooperates with the triangular prism segmentation approach. The predicted scattered points were visualized and the 3D geological structure was reproduced. The cross-section of the dam axis location (i.e., X = 40725.513 m) is calculated via the grid point mapping, and the corresponding calculations are performed. The results of the root mean square error (*RMSE*) and the correlation coefficient (R^2) data are presented in Tables 5 and 6. The cross-sectional details are illustrated in Fig. 11. The bar graph and line graph for the comparison between R^2 and *RMSE* of the EMDCT are presented in Fig. 12. Compared to the ground truth (black curve), the EMDCT is able to reproduce the details better, with 6 interfaces correlation coefficients above 0.995 and the RMSE averaging of 5.0 m, which is remarkably superior to other algorithms without triangular prism blocks (see Fig. 12). Compared to the ground truth, the EMDCT is visually close, which is closer to reality than the interface that is not used (blue curve). As in the regional details of Fig. 11(a), the green curve most likely represents the trend of formation. Therefore, the EMDCT is capable of refining the implicit characteristics of the engineering geological entity in detail, improving the interface correlation coefficient, and reducing the interface error. We use existing data to model explicitly, resulting in detailed geological errors of the model. The average elevation error of the six main formations is 11.823 m, 6.8 m more than that of EMDCT. The average correlation coefficient of them is 0.9806, 0.016 less than that of EMDCT. These shows that EMDCT is suitable for local detail characterization under relatively sparse boreholes, and is an effective method for refine modeling.

Sensitivity analysis of prediction

To examine the case of more sparse boreholes and uneven distribution, sensitivity analysis of the detection effect is required. Herein, the virtual drilling training sets are combined with the EMDCT for training (see Table 7), and the performance in dam site of each classifier on various borehole numbers and distributions is compared.

Sensitivity to number of boreholes

Training sets T1, T2, T3, and T4 represent boreholes with uniform grid distribution and decreasing grid rows and columns, and the corresponding results are illustrated in Fig. 13(a). As presented in the line graph, the predicted results by the EMDCT in this region are relatively stable and good, and its accuracy does not substantially reduce. In addition, this approach is not sensitive to the data volume, and the accuracy rate remains high (i.e., 0.969, 0.967, 0.966, and 0.965 for T1-T4). Under different distributions of borehole numbers, the formation prediction effect of the divide-and-conquer method (EMDCT) shows significantly better results, compared with other single algorithms. There is almost the same trend between them, but the declining trend of EMDCT is significantly slower than that of other single algorithms, indicating that it is robust to the number of boreholes.

Sensitivity to distribution of boreholes

Training sets T1, T5, and T6 signify the same number of boreholes, but with dissimilar distributions. The achieved results are presented in Fig. 13(b). The uniformity of borehole distribution is another factor in the prediction of any algorithm. The more chaotic the distribution, the less effective the prediction results. The EMDCT exhibits a relatively stable prediction and maintains high accuracy (i.e., 0.969, 0.968, 0.965 for the above training sets) in the case of uneven borehole distribution, which can be applied to the Fig. 9 Confusion matrices of 8 classifiers: a DF, b RF, c DT, d XGB, e CB, f SVM, g KNN, h EMDCT

Frequency

0.8

0.6

0.4

0.2

0

0.0238 0.951

0.928 0.0494

0 0

0 0

0 0

#6 #7

(b)

0

0

0 0.029

0 0

0 0 0 0











Fig. 10 Comparison of 8 classifiers on the identification of each class in histograms, and the seven bar charts from top to bottom in order represent strata 1 to 7

Table 4 Classification evaluation metrics for various classifiers, dam site area with 149 boreholes

Strat	a code	DF	RF	DT	XGB	CB	SVM	KNN	EMDCT
#1	Recall	0.886	0.897	0.914	0.902	0.920	0.949	0.865	0.954
	Precision	0.926	0.884	0.797	0.848	0.954	0.976	0.950	0.976
	F1	0.905	0.891	0.851	0.874	0.936	0.963	0.906	0.965
#2	Recall	0.920	0.907	0.848	0.883	0.957	0.964	0.890	0.968
	Precision	0.901	0.879	0.865	0.859	0.924	0.953	0.899	0.966
	F1	0.910	0.893	0.856	0.871	0.940	0.959	0.894	0.967
#3	Recall	0.913	0.888	0.841	0.851	0.930	0.944	0.944	0.963
	Precision	0.884	0.874	0.837	0.830	0.931	0.945	0.945	0.963
	F1	0.898	0.881	0.839	0.840	0.931	0.944	0.944	0.963
#4	Recall	0.884	0.872	0.839	0.819	0.932	0.942	0.864	0.964
	Precision	0.912	0.887	0.826	0.850	0.934	0.929	0.875	0.952
	F1	0.898	0.880	0.832	0.834	0.933	0.935	0.870	0.958
#5	Recall	0.943	0.930	0.858	0.911	0.954	0.960	0.933	0.972
	Precision	0.931	0.919	0.909	0.907	0.958	0.959	0.913	0.971
	F1	0.937	0.924	0.883	0.909	0.956	0.959	0.923	0.972
#6	Recall	0.940	0.928	0.918	0.921	0.957	0.967	0.925	0.977
	Precision	0.938	0.932	0.883	0.924	0.957	0.970	0.933	0.977
	F1	0.939	0.930	0.900	0.923	0.957	0.969	0.929	0.977
#7	Recall	0.957	0.951	0.945	0.952	0.974	0.981	0.958	0.985
	Precision	0.973	0.971	0.957	0.964	0.978	0.988	0.959	0.991
	F1	0.965	0.961	0.951	0.958	0.976	0.985	0.958	0.988

uneven field sampling situation. However, the accuracy of the XGB and DT algorithms sharply lessened (0.026 and 0.021, respectively), indicating that the algorithm is sensitive to the uniform distribution of boreholes. Furthermore, if the distribution of boreholes grows chaotic, the prediction becomes unstable. The uniform equal spacing in this section is a relatively safe distribution method compared to random distribution. In practical engineering, it is valuable to find the location of new boreholes by uncertainty analysis method in existing boreholes. To enhance the accuracy of the overall implicit modeling, placing the most suitable position from a limited number of boreholes. The relationship between the distribution of boreholes and the accuracy of the model provides a basis for the addition and subtraction of boreholes.

Uncertainty analysis

Visualization of uncertainty

Uncertainty is inevitable in a geological model due to measurement error, data scatter, algorithm estimation, and other factors. The geological model can be understood by using the uncertainty quantitative results. One case with the EMDCT is chosen for uncertainty analysis. A 3D information entropy domain can be constructed by collecting the information entropy of each voxel in a 3D space (see Fig. 14(a)). The results of the information entropy isogram provided by the geological model correspond to the actual situation (see Fig. 7): the yellow area is mostly placed in the uncollected boreholes around the dam site. The purple area is mainly in the area of the dam site, which is densely drilled, and its radiation affects the surrounding non-dam area. There is a local maximum of information entropy at the boundary between two formations, which indicates that the spatial prediction results along stratigraphic boundaries exhibit a relatively low confidence level, and the model is expected to increase the sampling points under complex geological conditions and scattered boreholes.

Bulletin of Engineering Geology and the Environment (2024) 83:282

Validation of uncertainty analysis

Uncertainty based on information entropy represents the degree of confidence of the classifiers for the predicted points. This section aims to prove the existence of a correlation between the certainty and the accuracy. According to the plotted results in Fig. 14(b), the decreasing trend of information entropy is almost negative exponential. The points with small information entropy exhibit a large proportion, indicating that the prediction results of most sampling points are relatively reliable. The critical value of information entropy is denoted by h (where, 0 < h < 1), and the value of h can be determined according to certain conditions. The approximate corresponding relationship between information entropy h and confidence β is **Table 5** Detailed comparison of R^2 of various algorithms, damsite area with 149 boreholes

Interface	EMDCT	SVM	DF	RF	DT	XGB	СВ	KNN	Explicit
#1-2	0.9962	0.9923	0.9757	0.9810	0.9449	0.9655	0.9872	0.9758	0.9846
#2-3	0.9979	0.9942	0.9796	0.9759	0.9646	0.9758	0.9907	0.9752	0.9908
#3-4	0.9957	0.9918	0.9829	0.9746	0.9608	0.9622	0.9867	0.9753	0.9812
#4-5	0.9972	0.9947	0.9885	0.9856	0.9809	0.9773	0.9934	0.9779	0.9948
#5-6	0.9957	0.9917	0.9867	0.9765	0.9490	0.9748	0.9909	0.9783	0.9905
#6-7	0.9961	0.9922	0.9657	0.9631	0.8759	0.9642	0.9825	0.9529	0.9414
Mean	0.9965	0.9928	0.9798	0.9761	0.9460	0.9700	0.9886	0.9726	0.9806
Interface	EMDCT	SVM	DF	RF	DT	XGB	СВ	KNN	Explicit
#1-2	5.227	7.442	13.295	11.829	19.400	15.875	9.590	13.197	11.340
#2-3	3.799	6.283	11.742	12.825	15.493	12.859	7.961	13.078	9.103
#3-4	5.645	7.782	11.174	13.589	17.024	16.676	9.923	13.503	13.665
#4-5	5.063	6.925	10.212	11.402	13.203	14.446	7.747	13.948	7.566

13.243

14.311

12.866

Table 6Detailed comparisonof RMSE of various algorithms(unit: m), dam site area with149 boreholes

described in Sect. 2.3.3. According to the scattered boreholes around the dam site, h = 0.6 is taken and the prediction is considered confused (uncertain) if the information entropy becomes greater than 0.6 (corresponding confidence is about $\beta = 0.85$). In this paper, an 85% confidence degree is taken as the boundary value of whether the formation prediction is confident. Then, 0.6 can be determined as the boundary value of whether the prediction is considered to be chaotic. The relationship between the certainty and accuracy of the model predictions is confirmed by the 2×2 contingency table chi-square calculation formula:

#5-6

#6-7

Mean

5.652

4.707

5.016

7.871

6.699

7.167

9.969

13.849

11.707

$$\chi^{2} = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{\left(n_{ij} - \frac{n_{i} \cdot n_{j}}{n}\right)^{2}}{\frac{n_{i} \cdot n_{j}}{n}} \sim \chi^{2}(1)$$
(17)

where *i* and *j* are two related categories, and n_{ij} , n_i , $n_{.j}$, and *n* represent their frequencies. The certainty and the correctness are given in Table 8. By entering the data and after calculating, it is obtainable χ^2 =57,397. Under the condition of significance level α =0.05, the correlation between prediction certainty and correctness can be satisfied. The information entropy represents the confidence of the prediction algorithm in probability, and for the points predicted with confidence, the prediction results are always correct with greater probability. The uncertainty assessment is also always more likely to give a conservative view of points even where the prediction was wrong.

The uncertainty assessment in this section is valid. Therefore, application of the uncertainty to analyze and increase the sampling position is an effective way, and the results of uncertainty can be effectually utilized to express the model.

13.714

14.122

14.615

8.266

9.886

8.896

12.733

16.408

13.811

9.172

20.089

11.823

Boreholes supplement after uncertainty analysis

19.559

26.251

18.488

By mapping points with information entropy greater than 0.95 into space, regions of low confidence can be directly identified (see Fig. 15). The areas of low confidence mainly include: (i) Stratigraphic interface; (ii) Small strata area; (iii) Areas with few boreholes. The region with the highest information entropy is further screened, and the location of new boreholes is artificially selected (see Fig. 15), whose new data set is sampled in new boreholes and added to the training set. The results obtained with 10 times of retraining are presented in Table 9. Adding new boreholes after uncertainty analysis has a certain effect on improving the overall accuracy of the model. But in fact, adding more sampling points always leads to the improvement of modeling. Therefore, by randomly increasing the equal number of boreholes in the understudy area, a new dataset obtained from sampling was added to the training set and trained with the same method for 10 times. Using the T-hypothesis test (Konietschke and Pauly 2014), it was proved that the additional assumptions based on the uncertainty analysis have a substantial effect on improving the overall accuracy of implicit modeling.

Null hypothesis H0: Additional boreholes in uncertainty analysis do not have a substantial effect on enhancing the overall accuracy of implicit modeling. The obtained results after training are provided in Table 9, and the statistical calculation formula reads: Fig. 11 Comparison of various cross-sections on the dam axis: a Cross-sections on the axis of the dam, the green ones are obtained via the EMDCT, whereas the blue ones are extracted in another way, b Implicit surface



$$T = \frac{\overline{Z}}{\overline{S}}\sqrt{n} \sim t(n-1) \tag{18}$$

where Z is the difference between the two samples, $Z_i = X_i - Y_i$, the mean value is defined by $\overline{Z} = \frac{1}{n} \sum_{i=1}^{n} Z_i$, and the standard deviation is given by $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Z_i - \overline{Z})^2$.

Therefore, we obtain T = 7.841. According to the significance level $\alpha = 0.005$, a unilateral test can be conducted to obtain $t_{1-\alpha}(n-1) = t_{0.995}(9) = 3.250 < T$. This indicates that the additional boreholes in the uncertainty analysis of the two algorithms exhibit a more significant effect on



Fig. 12 The degree of detail fit for various interfaces fitted by the classifiers, the orange colored bars are associated with the EMDCT: (a) R^2 , (b) *RMSE*

improving the overall accuracy of implicit modeling, and the difference is statistically significant.

In order to compare uncertainty analysis to guide geological exploration more accurately, we look for as few holes as possible to match the effect of uncertainty analysis when adding random holes. We find that the effect of randomly adding 30 holes (the last column of Table 9) is comparable to that of adding 20 holes after uncertainty analysis. Therefore, it can be figured that in the case of this paper, uncertainty analysis can reduce the engineering quantity of about 10 boreholes. After uncertainty analysis, the number of boreholes increased, and the information entropy obtained again is shown in Fig. 16. It can be seen that the overall information entropy decreased significantly.

Final performance

Through the above comparison, the best prediction results can be obtained by the EMDCT. After the uncertainty analysis, point clouds are processed and surfaces and entities are formed (see Sect. 2.4), where the results of the final modeling have been presented in Fig. 17. The implicit modeling accuracy can be improved at key stages of the entire process (see Table 10 and Fig. 18). The first line segment shows the ascending contribution of the EMDCT, and the second line segment is provided by the uncertainty analysis. The accuracy of the dam site reached 0.974 (the EMDCT contribution increased by 0.011, the uncertainty analysis contribution increased by 0.001, and the total increased by 0.012).

Table 7 Data set setup

Data set	Number of boreholes	Distribution
Training set T1	110	Virtual boreholes are drilled every approximately 210 m in the horizontal range of the understudy area; Evenly distributed according to the shape of squares: $11 \times 10 = 110$
Training set T2	90	Evenly distributed according to the shape of squares: $10 \times 9 = 90$
Training set T3	72	Evenly distributed according to the shape of squares: $9 \times 8 = 72$
Training set T4	56	Evenly distributed according to the shape of squares: $8 \times 7 = 56$
Training set T5	110	Unevenly spaced according to the square shape: $11 \times 10 = 110$
Training set T6	110	Virtual boreholes are drilled randomly in the horizontal range

Additionally, the accuracy of the whole working area around the dam site (sparse boreholes) reached 0.962 (the EMDCT contribution increased by 0.013, the uncertainty analysis contribution increased by 0.040, and the total increase by 0.053). Based on the same data, we built this model using traditional explicit workflows, and the results are shown in Fig. 17. This is non-automated, time-consuming, and the results of the modeling are different from ground truth. Workflows modeled explicitly do not involve machine learning, so the result "accuracy" does not exist. However, we can see from the final profile (Fig. 17) and detailed geological indicators (Fig. 12) presented by the model that its performance is far inferior to EMDCT.

Discussion

Implicit modeling with machine learning

The discrete input of implicit modeling can be suitably combined with machine learning to realize automated modeling. After simple processing, the drilling data can be converted into a training set. The discretized output can be visualized quickly, making it easy for decision-makers to readjust the model. A combination of implicit modeling and machine learning essentially transforms a series of tedious processes of determining topological relationships and surface construction into classification algorithm selection and hyperparameter adjustment. This paper proposes a set of geological implicit modeling and uncertainty analysis methodologies based on the triangular prism blocks method. The accuracies of the seven popular algorithms touch approximately 0.9, with an R^2 value of more than 0.95 between those of the predicted surface and the actual interface. This reveals that the overall model constructed using the implicit method meets the accuracy requirements of general engineering. Additionally, the local model fits well with the surface model of the stratigraphic interface, resulting in smooth surfaces.

Refined characterization of geological entity with divide-and-conquer tactic

Methods of dividing the region

The discrete points' processing is also a major research problem that directly affects the classifier's prediction results. The divide-and-conquer tactic serves as a crucial preprocessing in implicit modeling. Here, we supplement and compare the divide-and-conquer method with different regional segmentation methods. Through the same process as Table 11, we get the results of Table 18. Several classification algorithms show different performance in different segmentation methods, but the Delaunay-triangular prism shows remarkable performance. The Delaunay-triangular prism presented here essentially employs the characteristics of the triangular Delaunay mesh and is coordinated with the corresponding classification algorithms to refine the local details. This strategy allows for capturing detailed formation situations (Ji et al. 2023), and different sub-blocks can be processed via various approaches. The effects of different sampling point patterns are different in classifiers. The divide-and-conquer tactic is capable of reducing the data volume and classification dimension and avoiding the shortcomings of the classification algorithm. Hence, the algorithm is able to readily identify the optimal classification in the feature space, and the simple and repetitive operation ensures both the automaticity and accuracy of the modeling.

The modeling performance of EMDCT

By employing the EMDCT, the accuracy of the algorithm is substantially improved in the region of sparse data (the overall accuracy is improved by 0.013). Breakthroughs can be made based on the effect of better modeling effect in data-intensive areas (the accuracy of the dam site area is enhanced by 0.011, the correlation coefficient of each interface increases by an average of 0.002, and the *RMSE* reduces by approximately 1 m). The average layer error of Fig. 13 Sensitivity analysis of the prediction results: (a) The effect of borehole number on the implicit modeling, (b) The effect of the uniformity of borehole distribution on implicit modeling







(b)

Fig. 14 Information entropy: (a)Isogram of 131 boreholes; (b)Ranking chart of information entropy

Table 8 Contingency table for prediction confidence and accuracy

	Prediction of incorrect points	Prediction of correct points	Total
Prediction of uncertain points (entropy > 0.6)	24,538	49,039	73,577
Prediction of certain points (entropy ≤ 0.6)	54,981	748,035	803,016
Total	79,519	797,074	876,593

EMDCT compared to the general explicit modeling method improved by 6.8 m, \mathbb{R}^2 improved by 0.016 (see Tables 5 and 6). The modeling effect of EMDCT is better than that of traditional methods, from our supplementary experiments. Although not all classifiers are perfect for each data situation and geological structure, using appropriate algorithms in different blocks can result in achieving better results. From the perspective of uncertainty, we calculate the information entropy of two models using SVM alone and EMDCT. The average information entropy of the former model is 0.119

Fig. 15 Uncertainty analysis result



(corresponding to 98.4% confidence), and that of the latter is 0.096 (corresponding to 98.8% confidence), which indicates that the prediction of the model by EMDCT is more stable. EMDCT not only has improved accuracy in the model, but also has higher modeling confidence.

Uncertainties in the geological model

When dealing with sparse data, complex structures, and random sampling, uncertainty quantification is necessary to assess overall geological conditions. The classification of voxel classes is the certainty expression of implicit modeling; in contrast, the probability of each stratum and the entropy information of each voxel are the uncertainty expression of implicit modeling. When we assume that the geological space resembles a binary information entropy system, this expression can be roughly matched to the confidence level. Through cloud maps and confidence levels, people can intuitively grasp the understanding of geological conditions.

Uncertainty analysis provides two advantages for implicit geological modeling: (i) Enabling the geological model to accommodate different possibilities; for instance, in Fig. 14(a), the interface is almost confined to this light blue space at the dam location, but it can occur anywhere in this narrow space, meaning many possibilities for the implicit renewable geological model in the form of scatter; (ii) Guiding geological discoveries; actually, uncertainty

Number of the tests	The accuracy before adding boreholes	The accuracy after adding 20 boreholes according to uncertain analysis	The accuracy after adding 20 random boreholes	The previous two columns of difference	The accuracy after adding 30 random boreholes
1	0.9093	0.9624	0.9479	0.0145	0.9560
2	0.9093	0.9619	0.9449	0.0170	0.9635
3	0.9093	0.9623	0.9552	0.0071	0.9638
4	0.9093	0.9622	0.9467	0.0155	0.9602
5	0.9093	0.9622	0.9577	0.0045	0.9619
6	0.9093	0.9622	0.9489	0.0133	0.9568
7	0.9093	0.9619	0.9497	0.0122	0.9638
8	0.9093	0.9618	0.9526	0.0092	0.9659
9	0.9093	0.9626	0.9591	0.0035	0.9583
10	0.9093	0.9620	0.9492	0.0128	0.9642
Mean	0.9093	0.9622	0.9512	0.0110	0.9621
$Std(\times 10^{-3})$	0.0000	0.2377	4.5276	4.4205	3.2653

Table 9 The results of the boreholes increasing





analysis assistances in locating new boreholes that effectively augment the implicit modeling algorithm, as it reflects the information that is expected to be added in the algorithm. Boreholes around the dam site are very few. Without uncertainty analysis, adding boreholes randomly, and the overall working area accuracy rate is only 0.951. By adding boreholes based on uncertainty analysis, the accuracy of the whole working area reached 0.962 (see Table 10, 0.011 higher), and the visualization effect of the formation structure is significantly closer to the actual situation than without uncertainty analysis (see Fig. 17), which can reduce the quantity of about 10 boreholes, equivalently.

Limitations

The examples throughout this paper are real engineering cases with good results in hydraulic engineering. This case has the following characteristics to fit into this process: (i) The distribution of boreholes is uneven, and in some areas is very sparse. (ii) The layered structure is obvious. (iii) The stratigraphic scale does not cross orders of magnitude.

Given the scalability and applicability of the method, this section analyzes the limitations of the workflow based on the address characteristics of engineering cases. The method in this paper has limitations in dealing with different practical cases. (i) Data fusion needs to be studied. Borehole is the most commonly used modeling data in geological engineering, which has been proved by most engineering practices. Profiles, geological exposure lines, physical data, etc., can also be used, and many geological modeling methods use the fusion of multisource data. The proposed approach provides valuable insights for data fusion. The study used only borehole data, not profiles or other geological data. According to our proposed method, the profile data can also be included by the sampling method, which can be processed into "location(x, y, z), attribute(S)" sample points for machine learning. After the boundaries on the profile are interpreted based on their spatial positions, they are divided into sampling points and added to the training set. The integration of multi-source data can strengthen the present approach in the near future. (ii) The modeling effect of structural lens needs to be improved. Structural lenses, denoting solids characterized by a thick middle and thin periphery, typically signify compressive stress-induced structural fracture zones. These features, although smaller in size relative to formation structures, possess complex shapes. When the borehole does not contain any lens information totally, the proposed method cannot infer it, even if the lens exists between the boreholes. When the borehole contains less of it, we want to accurately restore it by implicit modeling method, which involves how determining the voxel size of the grid model. In this context, detailed structures require smaller voxels to represent the boundary. If a finer grid is selected, more voxels are needed to form the grid model and uncertainty model for the same area of work, and the amount of computation will increase



Fig. 17 The visualization results obtained by the implicit modeling: (a) Ground truth; (b) Points visualization results before uncertainty analysis; (c) The final geology model; (d) Traditional explicit workflows

Table 10 Performance of each process in this appro	ach
--	-----

Process	Accuracy of dam site area with 149 boreholes	Accuracy of the overall area with 131 boreholes
The SVM algorithm without EMDCT	0.962	0.909
EMDCT	0.973	0.922
Uncertainty analysis	0.974	0.962

exponentially. We try a site-specific hierarchical grid approach in the future.

Conclusions

The 3D geological model provides the basis for the subsequent analysis of the engineering construction of the dam site. In view of current geological modeling problems, such as low local accuracy, difficulty in quantifying uncertainties, and difficulty in guiding model updates, a geological implicit modeling and uncertainty analysis approach is proposed in this study. The contributions include:

- (i). Implicit modeling with machine learning. A combination of implicit modeling and machine learning is classification algorithm selection and hyperparameter adjustment, which is more convenient than reconstructing the potential functions.
- (ii). Application of the EMDCT. EMDCT is able to readily identify the optimal classification in the feature space, which can enhance the reproduction effect of model details.

(iii). Uncertainty analysis. It quantifies the uncertainty of the whole geological model and guides the further development of geological exploration and model updating. The implementation of uncertainty analysis can reduce the number of boreholes by 10 in this study approximately.

The implicit modeling via machine learning techniques is a relatively novel approach in this field. The EMDCT and uncertainty analysis are two critical processes aimed at improving modeling effectiveness. In this paper, data sets consisting of real boreholes and sampling points within the area are collected by a divide-and-conquer tactic. Commonly, the overall predictive accuracy of the EMDCT, which refines implicit features, is higher. Specifically, the overall accuracy of the working area is 0.922 and the accuracy of the dam site is obtained as 0.973. The EMDCT-based analysis indicates less sensitivity to the number and distribution of boreholes and generally exhibits robustness. Uncertainty assessment is appropriately conducted using information entropy, and a correlation test of the contingency table is adopted to establish a vital relationship between the size of uncertainty and the accuracy of formation prediction. The T-test confirms that the given supplementary borehole is capable of noticeably improving the modeling effect according to the uncertainty assessment, and the significance level reaches 0.005. Based on the results of uncertainty analysis, the overall accuracy of the working area is 0.962 and the accuracy of the dam site is 0.974.

The study used only borehole data, not profiles or other geological data. According to our proposed method, in fact, the profile data can also be included by the sampling method. After the boundaries on the profile are interpreted based on their spatial positions, they are equally divided into



Fig. 18 Accuracy of various processes in the proposed approach

Methods of dividing the region	Delaunay triangle	Tyson polygon	Ordinary quadrilateral mesh
Accuracy	0.973	0.906	0.868
R^2	0.9965	0.9535	0.9464
<i>RMSE</i> (m)	5.016	18.301	18.470
Mean information entropy	0.096	0.225	0.256
Modeling confidence	98.8%	96.4%	95.7%

sampling points and added to the training set. The integration of multi-source data is able to strengthen the present approach in the near future. The proposed approach, combining the background of geological implicit modeling with the triangulation ensemble model and uncertainty analysis, provides valuable insights for the rapid generation of models, borehole layout, and uncertainty assessment in the early stages of hydraulic engineering geological explorations. The inspiration to engineering practitioners is as follows. (i) The main significance of this research is that it can help engineers understand geological conditions more deeply and guide geological exploration work. Engineers can clearly locate areas with complex geological structures through uncertainty analysis, so as to find the space where most worthy of investigation, rationally arrange the following investigation work, and improve productivity. For example, in the example given in Section 3.5.3, our method can reduce the effort of 10 boreholes (Table 9). Through our method, the accuracy of the whole working area can be improved from 0.909 to 0.962 (Table 10). The modeling process of the divide-and-conquer method and implicit modeling is beneficial to the uncertainty analysis and the layout of new boreholes. In the preliminary planning stage, the general geological structure of the study area can be obtained with few boreholes. (ii) The method takes the uncertainty information of the model into account. The general relationship established between information entropy and confidence allows engineers to have a deeper understanding of the grid model. The uncertainty analysis method in this paper is not only an evaluation of the confidence level of the modeling effect but also a new expression of geological uncertainty. Furthermore, the reasonable integration of uncertainty is conducive to the quantitative evaluation of the reliability of the calculation results in the numerical simulation of geological safety and stability. (iii) Implicit modeling has an inspiration for geological engineers and is a novel workflow for forming geological models automatically and finely.

Acknowledgements This research was jointly funded by the National Natural Science Foundation of China (Grant No. 42302322), the China Postdoctoral Science Foundation (Grant Nos. 2023M732604 and 2023TQ0239), the Independent Innovation Fund of Tianjin University (Grant No. 2023XJD-0065) and the Open Research Fund of State

Key Laboratory of Hydraulic Engineering Intelligent Construction and Operation (HESS-2046).

Author contributions Mingchao Li: Resources, Writing—Review & Editing, Data Curation, Supervision, Project administration and Funding acquisition;

Chuangwei Chen: Conceptualization, Methodology, Validation, Formal analysis, Visualization, Code, Writing-Original Draft;

Hui Liang: Writing—Review & Editing;

Shuai Han: Conceptualization, Methodology, Code, Validation,

Formal analysis, Writing—Review & Editing and Funding acquisition; Qiubing Ren: Resources, Formal analysis, Visualization, Writing— Review & Editing;

Heng Li: Writing—Review & Editing.

All authors have read and agreed to the published version of the manuscript.

Funding Open access funding provided by The Hong Kong Polytechnic University.

Data availability The authors do not have permission to share data.

Declarations

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

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

- Abedi M, Norouzi G-H, Bahroudi A (2012) Support vector machine for multi-classification of mineral prospectivity areas. Comput Geosci 46:272–283. https://doi.org/10.1016/j.cageo.2011.12. 014
- Adzhiev V, Kazakov M, Pasko A, Savchenko V (2000) Hybrid system architecture for volume modeling. Computers

Graphics-UK 24(1):67–78. https://doi.org/10.1016/S0097-8493(99)00138-7

Aghamolaie I, Lashkaripour GR, Ghafoori M, Hafezi Moghaddas N (2019) 3D geotechnical modeling of subsurface soils in Kerman city, southeast Iran. Bull Eng Geol Env 78(3):1385–1400. https:// doi.org/10.1007/s10064-018-1240-7

Antonielli B, Iannucci R, Ciampi P et al (2023) Engineering-geological modeling for supporting local seismic response studies: insights from the 3D model of the subsoil of Rieti (Italy). Bull Eng Geol Environ 82:235. https://doi.org/10.1007/s10064-023-03259-4

- Ash RB, Doleans-Dade CA (1999) Probability and measure theory, edn 2. Elsevier
- Bullejos M, Cabezas D, Martín-Martín M, Alcalá FJ (2022) A K-nearest neighbors algorithm in Python for visualizing the 3D stratigraphic architecture of the Llobregat River Delta in NE Spain. J Mar Sci Eng 10(7):986. https://doi.org/10.3390/jmse10070986
- Chen P, Liu Y, Li Y (2022a) A large deformation finite element analysis of uplift behaviour for helical anchor in spatially variable clay. Comput Geotech 141:104542
- Chen Z, Wu Q, Han S, Zhang J, Yang P, Liu X (2022b) A study on geological structure prediction based on random forest method. Artif Intell Geosci 3:226–236. https://doi.org/10.1016/j.aiig.2023. 01.004
- Chen P, Hu Y, Yao K (2023) Probabilistic investigations on the elastic stiffness coefficients for suction caisson considering spatially varying soils. Ocean Eng 289:116273
- Chen X, Li D, Tang X (2021) A three-dimensional large-deformation random finite-element study of landslide runout considering spatially varying soil. Landslides 18(9):3149–3162
- Dell'Arciprete D, Bersezio R, Felletti F, Giudici M, Comunian A, Renard P (2012) Comparison of three geostatistical methods for hydrofacies simulation: a test on alluvial sediments. Hydrogeol J 20(2):299–311. https://doi.org/10.1007/s10040-011-0808-0
- Deng H, Zheng Y, Chen J, Yu S, Xiao K, Mao X (2022) Learning 3D mineral prospectivity from 3D geological models using convolutional neural networks: Application to a structure-controlled hydrothermal gold deposit. Comput Geosci 161:105074. https:// doi.org/10.1016/j.cageo.2022.105074
- Fan M, Xiao K, Sun L, Xu Y (2023) Metallogenic prediction based on geological-model driven and data-driven multisource information fusion: A case study of gold deposits in Xiong'ershan area, Henan Province, China. Ore Geol Rev 156:105390. https://doi.org/10. 1016/j.oregeorev.2023.105390
- Fatehi M, Asadi HH, Hossein Morshedy A (2020) 3D design of optimum complementary boreholes by integrated analysis of various exploratory data using a sequential-MADM approach. Nat Resour Res 29(2):1041–1061. https://doi.org/10.1007/ s11053-019-09484-7
- Grose L, Laurent G, Aillères L, Armit R, Jessell M, Cousin-Dechenaud T (2018) Inversion of structural geology data for fold geometry. J Geophys Res: Solid Earth 123(8):6318–6333. https://doi.org/10. 1029/2017JB015177
- Guo J, Wang X, Wang J, Dai X, Wu L, Li C, Li F, Liu S, Jessell MW (2021) Three-dimensional geological modeling and spatial analysis from geotechnical borehole data using an implicit surface and marching tetrahedra algorithm. Eng Geol 284:106047. https://doi. org/10.1016/j.enggeo.2021.106047
- Guo J, Wang Z, Li C, Li F, Jessell MW, Wu L, Wang J (2022) Multiple-point geostatistics-based three-dimensional automatic geological modeling and uncertainty analysis for borehole data. Nat Resour Res 31(5):2347–2367. https://doi.org/10.1007/ s11053-022-10071-6
- Han S, Li H, Li M, Zhang J, Guo R, Ma J, Zhao W (2022) Deep learning-based stochastic modelling and uncertainty analysis of fault

networks. Bull Eng Geol Env 81(6):242. https://doi.org/10.1007/s10064-022-02735-7

- Han R, Wang Z, Wang W, Xu F, Qi X, Cui Y, Zhang Z (2023) Igneous rocks lithology identification with deep forest: Case study from eastern sag, Liaohe basin. J Appl Geophys 208:104892. https://doi.org/10.1016/j. jappgeo.2022.104892
- Harris JR, Ayer J, Naghizadeh M, Smith R, Snyder D, Behnia P, Parsa M, Sherlock R, Trivedi M (2023) A study of faults in the Superior province of Ontario and Quebec using the random forest machine learning algorithm: Spatial relationship to gold mines. Ore Geol Rev 157:105403. https://doi.org/10.1016/j.oregeorev.2023.105403
- Hou W, Chen Y, Liu H, Xiao F, Liu C, Wang D (2023) Reconstructing Three-dimensional geological structures by the Multiple-point statistics method coupled with a deep neural network: A case study of a metro station in Guangzhou, China. Tunn Underground Space Technol 136. https://doi.org/10.1016/j.tust.2023.105089
- Ji G, Wang Q, Zhou X, Cai Z, Zhu J, Lu Y (2023) An automated method to build 3D multi-scale geological models for engineering sedimentary layers with stratum lenses. Eng Geol 317:107077. https://doi.org/10.1016/j.enggeo.2023.107077
- Konietschke F, Pauly M (2014) Bootstrapping and permuting paired t-test type statistics. Stat Comput 24(3):283–296. https://doi.org/ 10.1007/s11222-012-9370-4
- Lindsay MD, Aillères L, Jessell MW, de Kemp EA, Betts PG (2012) Locating and quantifying geological uncertainty in three-dimensional models: Analysis of the Gippsland Basin, southeastern Australia. Tectonophysics 546–547:10–27. https://doi.org/10.1016/j. tecto.2012.04.007
- Lipus B, Guid N (2005) A new implicit blending technique for volumetric modelling. Visual Computer 21(1–2):83–91. https://doi. org/10.1007/s00371-004-0272-0
- Liu H, Li W, Gu S, Cheng L, Wang Y, Xu J (2023) Three-dimensional modeling of fault geological structure using generalized triangular prism element reconstruction. Bull Eng Geol Env 82(4):118. https:// doi.org/10.1007/s10064-023-03166-8
- Lorensen WE, Cline HE (1987) Marching cubes: A high resolution 3D surface construction algorithm. Computer Graphics 21(4):163– 169. https://doi.org/10.1145/37402.37422
- Macêdo I, Gois JP, Velho L (2011) Hermite radial basis functions implicits. Computer Graphics Forum 30(1):27–42. https://doi. org/10.1111/j.1467-8659.2010.01785.x
- Madsen RB, Høyer A-S, Andersen LT, Møller I, Hansen TM (2022) Geology-driven modeling: A new probabilistic approach for incorporating uncertain geological interpretations in 3D geological modeling. Eng Geol 309:106833. https://doi.org/10.1016/j. enggeo.2022.106833
- Ouyang J, Zhou C, Liu Z, Zhang G (2023) Triangulated irregular network-based probabilistic 3D geological modelling using Markov Chain and Monte Carlo simulation. Eng Geol 320:107131. https:// doi.org/10.1016/j.enggeo.2023.107131
- Petrone P, Allocca V, Fusco F et al (2023) Engineering geological 3D modeling and geotechnical characterization in the framework of technical rules for geotechnical design: the case study of the Nola's logistic plant (southern Italy). Bull Eng Geol Environ 82:12. https://doi.org/10.1007/s10064-022-03017-y
- Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A (2017) CatBoost: unbiased boosting with categorical features. Adv Neural Inf Process Syst (31):23. https://doi.org/10.48550/arXiv.1706.09516
- Ren Q, Zhang H, Zhang D, Zhao X, Yan L, Rui J (2022) A novel hybrid method of lithology identification based on k-means++ algorithm and fuzzy decision tree. J Petrol Sci Eng 208:109681. https://doi.org/ 10.1016/j.petrol.2021.109681
- Resnick SI (2013) A probability path. Springer Science & Business Media

- Røe P, Georgsen F, Abrahamsen P (2014) An uncertainty model for fault shape and location. Math Geosci 46(8):957–969. https://doi. org/10.1007/s11004-014-9536-z
- Scalzo R, Lindsay MD, Jessell MW, Pirot G, Giraud J, Cripps E, Cripps S (2022) Blockworlds 0.1.0: a demonstration of anti-aliased geophysics for probabilistic inversions of implicit and kinematic geological models. Geosci Model Dev 15(9):3641–3662. https://doi. org/10.5194/gmd-15-3641-2022
- Schaaf A, Bond CE (2019) Quantification of uncertainty in 3-D seismic interpretation: implications for deterministic and stochastic geomodeling and machine learning. Solid Earth 10(4):1049–1061. https://doi.org/10.5194/se-10-1049-2019
- Shannon (1948) A mathematical theory of communication. Bell Syst Tech J 27(3):379–423. https://doi.org/10.1002/j.1538-7305.1948. tb01338.x
- Shi C, Wang Y (2021a) Development of subsurface geological crosssection from limited site-specific boreholes and prior geological knowledge using iterative convolution XGBoost. J Geotech Geoenviron Eng 147(9):04021082. https://doi.org/10.1061/ (ASCE)GT.1943-5606.0002583
- Shi C, Wang Y (2021b) Training image selection for development of subsurface geological cross-section by conditional simulations. Eng Geol 295:106415. https://doi.org/10.1016/j.enggeo.2021.106415
- Shi C, Wang Y (2022) Data-driven construction of three-dimensional subsurface geological models from limited site-specific boreholes and prior geological knowledge for underground digital twin. Tunn Undergr Space Technol 126:104493. https://doi.org/10. 1016/j.tust.2022.104493
- Shi T, Zhong D, Wang L (2021) Geological modeling method based on the normal dynamic estimation of sparse point clouds. Mathematics 9(15):1819. https://doi.org/10.3390/math9151819
- Song R, Qin X, Tao Y, Wang X, Yin B, Wang Y, Li W (2019) A semi-automatic method for 3D modeling and visualizing complex geological bodies. Bull Eng Geol Env 78(3):1371–1383. https:// doi.org/10.1007/s10064-018-1244-3
- Sukhov AN (1999) A mathematical model for the normal law of probability distribution of random errors. Meas Technol 42(3):205–214. https://doi.org/10.1007/BF02505141
- Tian M, Ma K, Liu Z, Qiu Q, Tan Y, Xie Z (2023) Recognition of geological legends on a geological profile via an improved deep learning method with augmented data using transfer learning strategies. Ore Geol Rev 153:105270. https://doi.org/10.1016/j.orege orev.2022.105270
- Titus Z, Heaney C, Jacquemyn C, Salinas P, Jackson MD, Pain C (2022) Conditioning surface-based geological models to well data using artificial neural networks. Comput Geosci 26(4):779–802. https://doi.org/10.1007/s10596-021-10088-5
- Wan J, Li X (2022) Analysis of a superconvergent recursive moving least squares approximation. Appl Math Lett 133:108223. https:// doi.org/10.1016/j.aml.2022.108223
- Wang X (2020) Uncertainty quantification and reduction in the characterization of subsurface stratigraphy using limited geotechnical investigation data. Underground Space 5(2):125–143. https://doi. org/10.1016/j.undsp.2018.10.008
- Wang H, Wellmann JF, Li Z, Wang X, Liang RY (2017) A segmentation approach for stochastic geological modeling using hidden Markov random fields. Math Geosci 49(2):145–177. https://doi. org/10.1007/s11004-016-9663-9

- Wang X, Yang S, Zhao Y, Wang Y (2018) Lithology identification using an optimized KNN clustering method based on entropyweighed cosine distance in Mesozoic strata of Gaoqing field, Jiyang depression. J Petrol Sci Eng 166:157–174. https://doi.org/ 10.1016/j.petrol.2018.03.034
- Wang H, Zhang X, Zhou L, Lu X, Wang C (2020) Intersection detection algorithm based on hybrid bounding box for geological modeling with faults. IEEE Access 8:29538–29546. https://doi. org/10.1109/ACCESS.2020.2972317
- Wang J, Mao X, Peng C, Chen J, Deng H, Liu Z, Wang W, Fu Z, Wang C (2023) Three-dimensional refined modelling of deep structures by using the level set method: Application to the Zhaoping Detachment Fault, Jiaodong Peninsula, China. Math Geosci 55(2):229–262. https://doi.org/10.1007/s11004-022-10031-z
- Wang S, Nguyen Q, Lu Y et al (2022) Evaluation of geological model uncertainty caused by data sufficiency using groundwater flow and land subsidence modeling as example. Bull Eng Geol Environ 81:331. https://doi.org/10.1007/s10064-022-02832-7
- Wellmann JF (2013) Information theory for correlation analysis and estimation of uncertainty reduction in maps and models. Entropy 15(4):1464–1485. https://doi.org/10.3390/e15041464
- Wellmann JF, Regenauer-Lieb K (2012) Uncertainties have a meaning: Information entropy as a quality measure for 3-D geological models. Tectonophysics 526–529:207–216. https://doi.org/10.1016/j. tecto.2011.05.001
- Yang L, Hyde D, Grujic O, Scheidt C, Caers J (2019) Assessing and visualizing uncertainty of 3D geological surfaces using level sets with stochastic motion. Comput Geosci 122:54–67. https://doi. org/10.1016/j.cageo.2018.10.006
- Yang Z, Chen Q, Cui Z, Liu G, Dong S, Tian Y (2022) Automatic reconstruction method of 3D geological models based on deep convolutional generative adversarial networks. Comput Geosci 26(5):1135–1150. https://doi.org/10.1007/ s10596-022-10152-8
- Zhang Z, Wang G, Carranza EJM, Liu C, Li J, Fu C, Liu X, Chen C, Fan J, Dong Y (2023) An integrated machine learning framework with uncertainty quantification for three-dimensional lithological modeling from multi-source geophysical data and drilling data. Eng Geol 324:107255. https://doi.org/10.1016/j.enggeo.2023. 107255
- Zhao L, Zhuo S, Shen B (2023a) An efficient model to estimate the soil profile and stratigraphic uncertainty quantification. Eng Geol 315:107025. https://doi.org/10.1016/j.enggeo.2023.107025
- Zhao Z, Chen Y, Zhang Y, Mei G, Luo J, Yan H, Onibudo OO (2023b) A deep learning model for predicting the production of coalbed methane considering time, space, and geological features. Comput Geosci 173:105312. https://doi.org/10.1016/j. cageo.2023.105312
- Zhong D, Wang L, Wang J (2021) Combination constraints of multiple fields for implicit modeling of ore bodies. Appl Sci 11(3):1321. https://doi.org/10.1016/S1003-6326(19)65145-9
- Zhou C, Ouyang J, Ming W, Zhang G, Du ZI, Liu Z (2019) A stratigraphic prediction method based on machine learning. Appl Sci 9(17). https://doi.org/10.3390/app9173553
- Zhu X, Zhang H, Ren Q, Zhang D, Zeng F, Zhu X, Zhang L (2023) An automatic identification method of imbalanced lithology based on deep forest and K-means SMOTE. Geoenergy Sci Eng 224:211595. https://doi.org/10.1016/j.geoen.2023.211595