A Semisupervised Deep Learning Neural Network Using Pseudolabels for Three-Dimensional Shallow Strata Modelling and Uncertainty Analysis in Urban Areas from Borehole Data

Jiateng Guo1*, Xuechuang Xu1, Luyuan Wang1, Xulei Wang1, Lixin Wu2, Mark Jessell3, Vitaliy Ogarko3, Zhibin Liu1 and Yufei Zheng1

[1]College of Resources and Civil Engineering, Northeastern University, Shenyang, China; 2000985@stu.neu.edu.cn (X. X.); wangluyuan9@163.com (L. W.); wxldbdx@163.com (X. W.); 207158587@qq.com (Z. L.); 15333134708@163.com (Y. Z.)

[2]School of Geosciences and Info-Physics, Central South University, Lushan Nanlu 932, Yuelu District, Changsha 410012, China; awulixin@263.net (L. W.)

[3]Centre for Exploration Targeting / Mineral Exploration Cooperative Research Centre / DARE Centre, School of Earth Sciences, The University of Western Australia, Perth, Australia; mark.jessell@uwa.edu.au (M. J.); vitaliy.ogarko@uwa.edu.au (V. O.)

[*]*Correspondence to*: Jiateng Guo (guojiateng@mail.neu.edu.cn; Tel.: +86-24-8368-7693)

Abstract. Borehole data are essential for conducting precise urban geological surveys and large-scale geological investigations. Traditionally, explicit and implicit modelling have been the primary methods for visualizing borehole data and constructing 3D geological models. However, explicit modelling requires substantial manual labour, while implicit modelling faces challenges related to uncertainty. Recently, machine learning approaches have emerged as effective solutions to address these issues in 3D geological modelling. Nevertheless, the use of machine learning to build 3D geological models is often limited by insufficient training data. In this paper, we propose the semisupervised deep learning using pseudolabels (SDLP) algorithm to overcome the issue of insufficient training data. Specifically, we construct the pseudolabels in the training dataset using the triangular irregular network (TIN) method. A 3D geological model is constructed using borehole data obtained from a geological survey of urban areas in Shenyang, Liaoning Province, NE China. Additionally, we compare the results of the 3D geological model built based on SDLP with those obtained from a support vector machine (SVM) method and an implicit HRBF modelling method. The findings demonstrate that our proposed method effectively resolves issues with insufficient training data. Moreover, compared to the 3D geological models constructed using the HRBF algorithm and SVM algorithm, the 3D geological model built based on the SDLP algorithm better conforms to the sedimentation patterns of the region and supports uncertainty analysis. In conclusion, the semisupervised deep learning method with pseudolabelling proposed in this paper provides a solution for 3D geological modelling in sparsely distributed areas with borehole data.

1. Introduction

Three-dimensional (3D) urban geological models are digital representations of subsurface strata and their associated features (Houlding, 1994). In recent years, the utilization of 3D geological models has expanded across various geological fields, such as mineral exploration (Zhang et al., 2021), geological storage (Thanh et al., 2019), groundwater resource estimation (Thibaut et al., 2021), geological disaster early warning generation (Høyer et al., 2019; Livani et al., 2022), and engineering geological condition evaluation (Chen et al., 2018; Guo et al., 2021; Lyu et al., 2021; Marz´an et al., 2021).

The commonly used data for 3D geological modelling include borehole data, geophysical data, survey and mapping data, and other types of data. Among these, borehole data provide the most accurate reflection of subsurface geological information (Guo et al.,2022). Notably, 3D geological modelling from borehole data can be divided into explicit modelling and implicit modelling (Jessell, 2001; Caumon et al., 2007; Wang et al., 2018). The explicit modelling approach can be used to directly delineate geological formations and interpret tectonics based on borehole data. Explicit 3D geological modelling methods are widely used in the 3D modelling of mines and regional geological structures, and they include the interactive 3D forward modelling method (Yang et al., 2011), generalized tri-prism (GTP) modelling method (Wu et al., 2004; Che et al., 2009) and parametric surface method (Lyu et al., 2021). However, these approaches heavily relies on the expertise of geologists and often proves time-consuming and labour-intensive when dealing with large-scale borehole data.

Implicit modelling methods are used to construct a 3D geological model by establishing the implicit equation of the isosurface representing the geometric shape of a geological body and using a series of implicit function visualization methods (Jessell M. et al., 2022). That is, a complex 3D geological object is represented as a continuous function of geological coordinates (Wang G.W. et al., 2011; Zhong D. Y. et al., 2021). This method does not require extensive human–computer interaction and has the advantages of high modelling accuracy, excellent smoothness and high spatial analysis efficiency (Sun H. et al., 2023). It is widely used in the field of geological modelling (Hillier M. J. et al., 2014; Calcagno P. et al., 2008; Shi T. D. et al., 2021) and provides results to complement the results of most urban geological surveys (de la Varga M. et al., 2019). Common implicit modelling methods include nearest neighbour value interpolation (Olivier R. et al., 2012), inverse distance weighted (IDW) interpolation (Liu H. et al., 2020; Liu Z et al., 2021), discrete smooth interpolation (DSI) (Mallet J. 1997), kriging (Wang G.W. et al., 2011; Thanh H. V. et al., 2019), the MLS method (Manchuk J. G. et al., 2019), and the radial basis function (RBF) method (Caumon G. et al., 2013; Hillier M. et al., 2014; Cuomo S. et al., 2017; Martin R. et al., 2017; Skala et al., 2017; Zhong D. Y. et al., 2019).However, the sparsity of borehole data, the complexity of geological bodies or geological phenomena, and the limitations of human cognition and expression lead to uncertainty in the relationship between the geometric form of a 3D geological model and the corresponding geological system (Caumon et al., 2007; Caers, 2011; Pakyuz-Charrier et al., 2018). The construction of 3D geological models by establishing implicit equations cannot effectively address this uncertain relationship.

Machine learning methods have been widely used in 3D geological modelling, and they are generally applied in unsupervised or supervised 3D geological modelling (Wang et al., 2023). Unsupervised machine learning algorithms (e.g., k-means clustering, self-organizing maps, and Gaussian mixture models) can be used to translate multisource geophysical datasets into 3D lithological models by measuring similarity between properties in feature space (Hellman et al., 2017; Giraud et al., 2020; Whiteley et al., 2021; Zhang et al., 2022). Supervised machine learning algorithms (e.g., random forests and artificial neural networks) can be applied to construct 3D lithological models by training from labelled geophysical and geological datasets (Jia et al., 2021; Lysdahl et al., 2022). Despite obtaining encouraging results with supervised machine learning algorithms, most studies have not addressed the following critical challenges regarding supervised machine learning algorithms for 3D geological modelling:

(1) In the scope of 3D modelling, precise geological information, such as relationship of stratigraphy, and tectonic in study area, revealed by borehole data is much less than not revealed by borehole data. Utilizing the precise information obtained via boreholes as labelled data to predict the correctness of the results in many unknown areas still requires further research.

(2) Labelled geological datasets were mainly composed of borehole data from early exploration phases (Jia et al., 2021; Lysdahl et al., 2022). The number of categories of lithological samples in drilling datasets is commonly imbalanced. A classification dataset with skewed class proportions can influence the performance of machine learning algorithms (Chawla et al., 2002; Batista et al., 2004). However, there is very little published research that addresses the sample imbalance issue in the context of training supervised machine learning algorithms for 3D lithological modelling.

Compared with machine learning methods, deep learning algorithms improve the ability to learn from mining data and are often combined with complex geophysical and geochemical data for modelling. Currently, there is a wealth of research on neural network-based deep learning methods for addressing geological issues such as tectonic recognition (Titos et al., 2018), mineral identification and classification (Xu and Zhou, 2018), seismic data inversion (Huang et al., 2020), and others.

Furthermore, in the realm of constructing 3D geological models, the deep learning approach using neural networks has also gradually garnered significant attention from numerous scholars (Laloy et al., 2017; Zhang et al., 2019; Ran et al., 2020; Zhang et al., 2018; Michael Hillier et al., 2021, 2022; S Avalos and Ortiz, 2020). However, the issue of insufficient training data has yet to be adequately addressed.

In this paper, we propose a semisupervised deep learning using pseudolabels (SDLP) algorithm for constructing 3D geological models and to overcome the lack of accurate labelled data. Upsampling is used to resolve label imbalance issues in the training dataset. The shallow borehole data from Shenyang, Liaoning Province, are used to construct 3D geological models using the proposed algorithm. To demonstrate the applicability of SDLP, the accuracy, precision, recall, and F1 score results of SDLP are compared with those of a classic support vector machine (SVM) algorithm based on a test dataset. To further assess the accuracy of SDLP, the profiles of the 3D geological models constructed by SDLP, SVM, and Hermite radial basis function (HRBF) are compared. Furthermore, the SDLP algorithm addresses the uncertainty limitations in the implicit HRBF modelling method.

2. 3D Modelling Method Based on Deep Learning

2.1. Borehole data preprocessing

Table 1. Table of average thickness, maximum thickness, minimum thickness and frequency of occurrence of different strata

	Frequency	Average (m)	Maximum (m)	Minimum (m)
fill	167	1.14	4.1	0.4
Clay-1	128	2.21	6	0.7
Clay-2	58	3.46	9.8	0.5
Clay-3	107	5.94	12.8	0.5
Clay-4	54	2.86	5.8	0.5
Sand-1	25	3.34	8.1	1.2
Stone-1	71	6.30	14	1.3
Stone-2	104	3.91	10	0.5
Stone-3	72	6.22	12.5	1.2
Residual-1	52	10.98	16.1	4.8
Residual-2	50	4.77	13.8	2
Residual-3	44	5.47	13.9	1

A total of 167 boreholes from engineering projects in Shenyang city were used to build the 3D geological model in this study. These boreholes are distributed in a 305×264 m area, with an average spacing of approximately 23 metres between adjacent boreholes. The average depth of the boreholes is 29.5 metres. The minimum thickness of the formations revealed by the boreholes is 0.4 metres, and the maximum thickness is 16.1 metres. The maximum average thickness is 10.98 metres, and the minimum average thickness is 1.14 metres (Table 1). The original borehole data mainly include borehole coordinates (X, Y), elevation, lithological thickness, lithological bottom depth, borehole number, lithological ID, etc.

This paper uses deep learning methods for 3D geological modelling, which can further simplify the modelling problem into a strata classification problem. In this method, the coordinate data and strata depth data obtained from boreholes are used as input vectors, and the lithological attributes of the boreholes are used as output vectors. In this study, the borehole data were simplified into continuous one-dimensional data when creating the dataset. However, there are significant differences in the lengths and frequencies of different formations within the borehole dataset (Table 1). For example, in terms of formation thickness, the maximum thickness is 16.1 m, while the minimum thickness is only 0.4 m. In terms of the formation occurrence frequency, the most frequent label, "fill," occurs 167 times, while the least frequent label, "sand-1," occurs 25 times. This significant difference may lead to the overfitting of the training model and ultimately result in poor training performance. Therefore, preprocessing of the borehole data is needed. An upsampling method is proposed to avoid overfitting in the training model caused by training dataset imbalance in this study.

Based on the above discussion, an unequal interval sampling method is adopted in this paper (Fig. 1). In the figure, H_{11}-H_{35} represents unequal-interval sampling for each stratum in the borehole, while $H_{11}P_1$-$H_{35}P_5$ represents unequal-interval sampling for each stratum on the deterministic section. Compared with equal-interval sampling, unequal-interval sampling involves changes to the sampling interval according to the thickness of different strata, thereby ensuring the balance of the sampled data. Otherwise, thinner strata may be difficult to predict or deemed to be outliers due to insufficient sampling. As shown in Figure 1, different colours in the borehole region represent different strata attributes, and the strata data are displayed in strips that are continuously distributed in the vertical direction. The attributes of a single stratum are continuously unique within the corresponding depth interval, and there are no data gaps between strata.

Due to the high reliability of borehole data, they can be directly or indirectly used for the generation of accurate models. By applying the Delaunay principle to borehole position points, a surface triangular irregular network (TIN) is created. This TIN encompasses the fundamental topological relationships between adjacent boreholes. If the stratum attributes of two neighbouring boreholes within each TIN are similar, they are connected to form a deterministic section. To ensure accurate geological predictions and eliminate the influence of distant and loosely correlated borehole connections, narrow triangles are removed from the TIN. This approach, similar to the generalized tri-prism (GTP) model, preserves the internal connectivity among the three corresponding boreholes and enables the simulation of various complex geological phenomena. Once the deterministic sections are connected, unequal interval sampling is conducted both horizontally and vertically, and the sampling density at the borehole locations is balanced to avoid overly dense sampling that may impact network training. The unequal interval sampling formula for borehole data is expressed as Equation (1), and the unequal interval sampling point coordinate formula for deterministic sections is expressed as Equation (2).

$$\mathrm{Z_{ij}} = \frac{(\mathrm{S_{ij}} - \mathrm{S_{ij-1}})}{\mathrm{n}} \tag{1}$$

$$\begin{cases} P_{ijx} = x_1 + \frac{x_2 - x_1}{n}(2j-1) \\ P_{ijy} = y_1 + \frac{y_2 - y_1}{n}(2j-1) \\ P_{ijz} = \frac{D_1C_2 + A_1C_2P_{ijx} + B_1C_2P_{ijy} - D_2C_1 - A_2C_1P_{ijy} - B_2C_1P_{ijy}}{C_1C_2n}(2i-1) \end{cases} \tag{2}$$

where Sij is the bottom depth of the jth stratum in the ith borehole, n is the number of samples from each stratum, and Zij is the sampling interval of the jth stratum in the ith borehole. Pijx, Pijy, and Pijz represent the x, y, and z coordinates of the sampling point in the ith row and jth column of a section. x1, y1, x2, and y2 are the coordinates of the two connected boreholes in a section. A1, B1, C1, D1, A2, B2, C2, and D2 are the parameters of the straight-line equations representing the top and bottom boundaries of the strata for the connected boreholes.

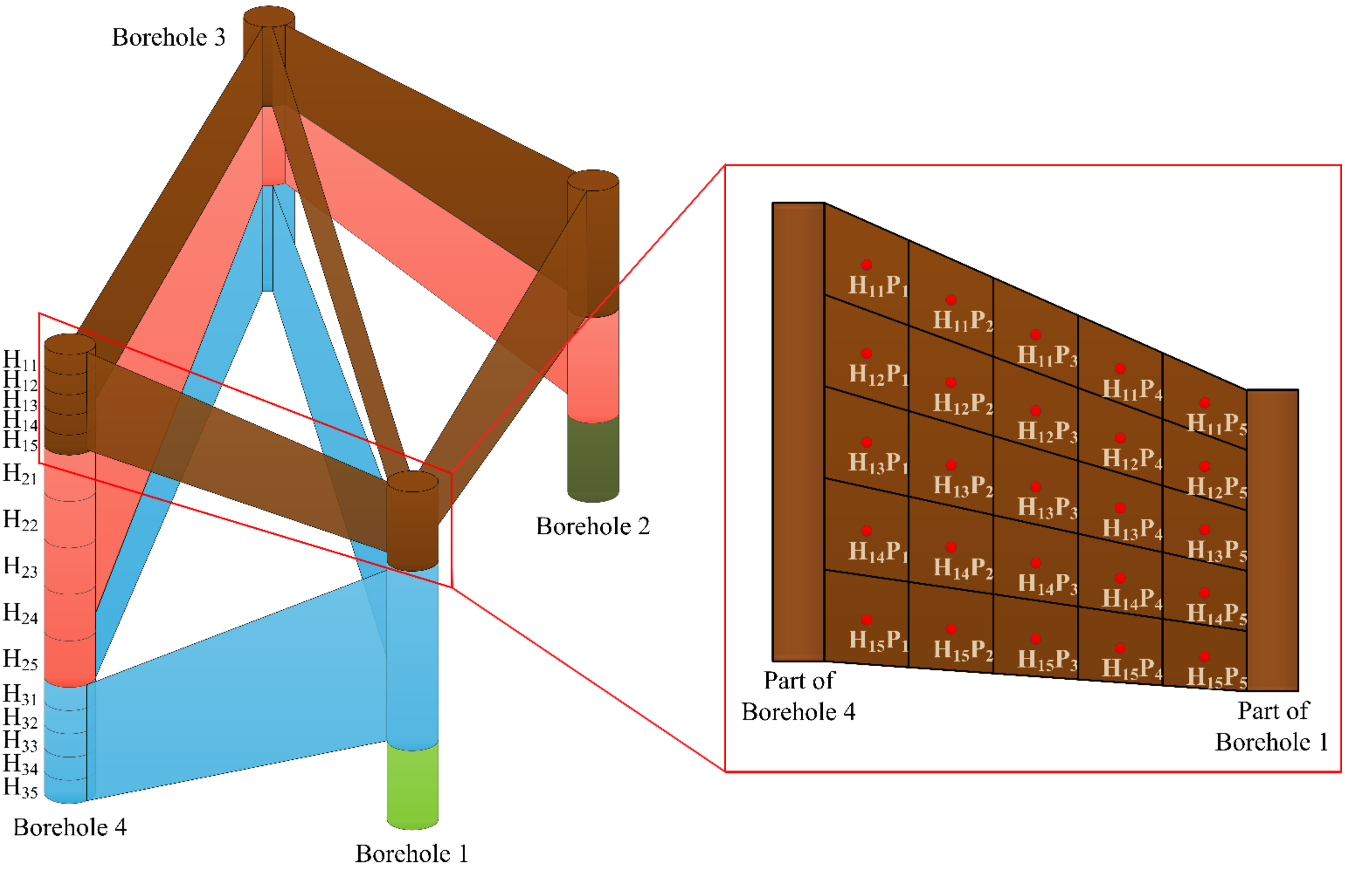


Figure 1. Resampling of borehole data. Upsampling on the boreholes (left); upsampling on the deterministic sections (right).

The difference in the number of digits between coordinate data (typically 7-8 digits with 3 decimal places) and stratum depth (typically 1-2 digits with 1 decimal place) in borehole data can lead to numerical computation issues in computer systems, making it difficult to train the model and adjust parameters, ultimately affecting the training results of the model. After performing data normalization based on raw data, each indicator is scaled to a specific range, allowing for comprehensive comparative evaluation. To eliminate the influence of digit disparity among input features, ensure the equal impact of different features on model training, and achieve convergence, it is necessary to apply min–max normalization to the data and map the resulting values to the range of 0 to 1. For any dataset x, the mapping function is as follows:

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{3}$$

where xmax is the maximum value of the sample data and xmin is the minimum value of the sample data. x' is the normalized result, and x is the input of the model data. Through this normalization method, the convergence speed of the network training model is improved, the training accuracy is improved, and the model training becomes easier.

2.2. Construction of deep neural networks

A multilayer perceptron (MLP) is a feedforward artificial neural network that learns to form certain rules through training based on input and output indicators. Thus, results closest to the expected output are obtained after inputting certain values. An MLP is a multilayer feedforward neural network based on the backpropagation algorithm. Each unit between layers in an MLP has a weight with an initial preset value, and unit training is performed using the backpropagation algorithm to adjust the weights between hidden layers. Input data are output after passing data through multiple hidden layers and compared with the expected labels to obtain the corresponding error, which is then propagated layer by layer backwards to adjust the weight of each layer. After multiple adjustments, suitable weights for the model are obtained. The relationship between layers can be expressed as shown in Equation (4). In the network model, the coordinates of each upsampled spatial point in the prediction area, x, y, and z, are used as inputs, and the geological properties of the spatial points are output. Each input represents a spatial feature dimension, and through four fully connected layers, the input data are processed and transformed. Each hidden layer contains multiple nodes, where each node is connected to all nodes in the previous layer. By multiplying by weights and applying an activation function, the input undergoes nonlinear transformation, resulting in expanded dimensionality. This result encompasses the deep features of the sample, and samples of different categories should have different high-dimensional features. The number of neurons in the hidden layer varies according to the complexity of the model, and the rectified linear unit (RELU) activation function is used between hidden layers. To prevent overfitting, a dropout function is added to the penultimate fully connected layers of the network to randomly reduce the number of neurons. Finally, the output value of each category is normalized using the exponential function through a fully connected layer and a softmax layer, and the sum of the probabilities of all categories is 1. The predicted results of each data point are integrated to form the entire 3D geological model (Figure 2). The network model uses the Adam optimizer, and the loss function adopted is the cross-entropy loss function, which is commonly used in multiclassification tasks.

$$\mathrm{Y_j} = \sum_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{W_{ij} X_i} + \mathrm{b} \qquad (4)$$

where Yj is the input of the next layer, Wij is the connection weight from cell X_i of the previous layer to cell Yj of the next layer, and b denotes the offset value.

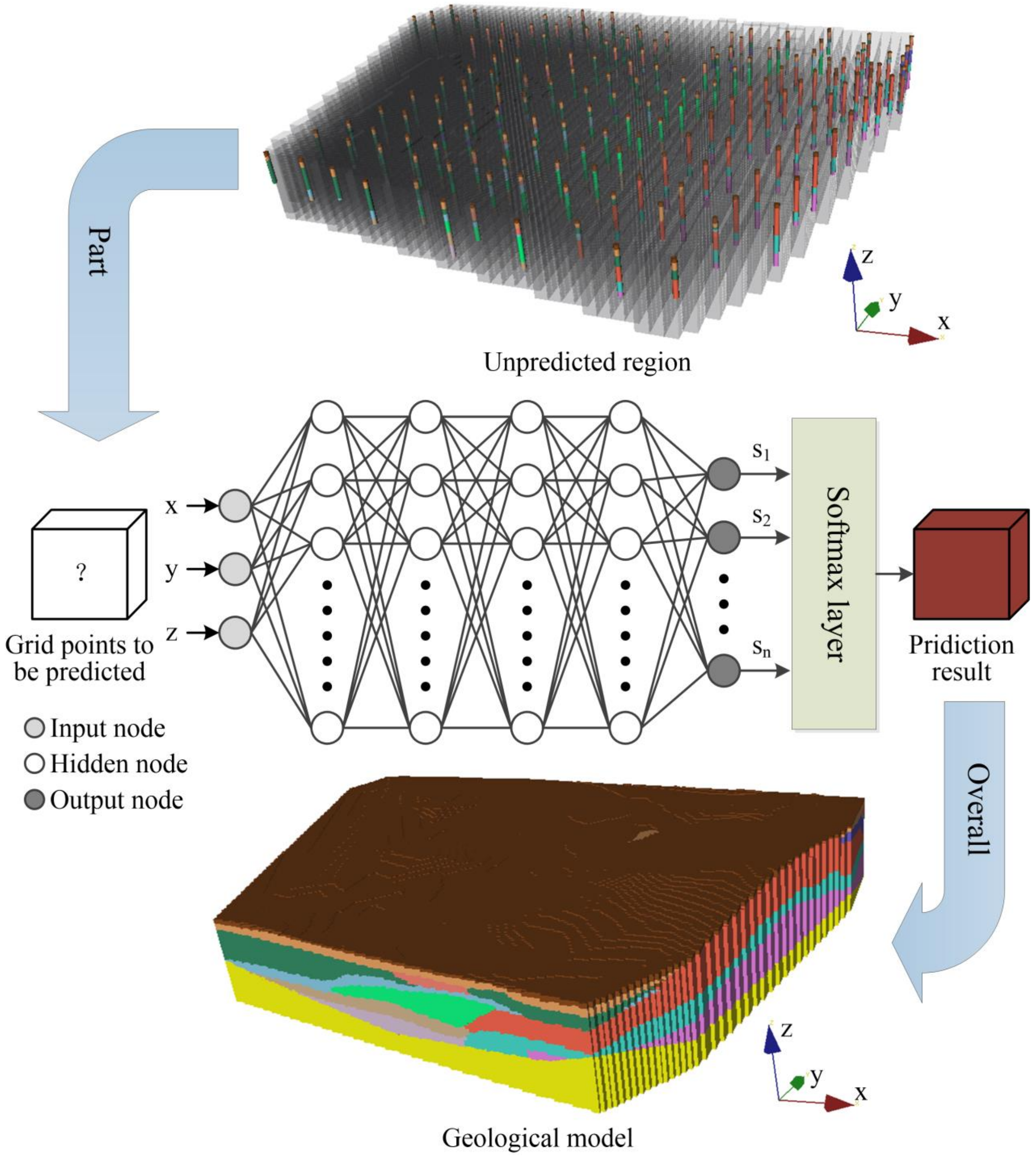


Figure 2. Architecture of a deep neural network. Light grey nodes are input features, dark grey nodes are target outputs, and white nodes are internal network nodes.

2.3 Semisupervised deep learning algorithm using pseudolabels

Compared with that of data from images, point cloud data, etc., borehole data displays a clustered characteristic with local concentrations but overall dispersion. Due to the large amount of missing point data between boreholes, it is difficult to accurately express the changing features of stratigraphic boundaries and inclination angles. Supervised learning depends on a large quantity of labelled data to enhance model performance. The labelled data used for training 3D geological models are obtained by upsampling limited borehole points and deterministic borehole profiles. Labelled data associated with spatial grid points in urban areas, which require high modelling precision, are scarce and contain very few features. To effectively solve the labelling problem, semisupervised learning is combined with deep learning , and a model is constructed using a small amount of labelled data and a large amount of unlabelled data with pseudolabels for prediction. This is beneficial for expanding the training data.

The attributes of strata are difficult to determine based on a single mathematical formula. Based on the topological relationships established with the TINs of three boreholes, three prisms are constructed using a method similar to the GTP approach by connecting the boreholes based on their stratigraphic properties, and the stratigraphic properties of the interior grid points of the prisms are obtained. For the predicted grid points within the prisms, it is assumed that their stratigraphic properties are similar to the properties of the prism, and when adding pseudolabels, it is assumed that the confidence level for each predicted stratigraphic property is high. Based on this approach, a semisupervised learning method based on pseudolabels is used to generate pseudolabels for the unlabelled data and improve learning performance. First, the model is trained using labelled data. When the model achieves a relatively high accuracy after training for a certain number of rounds, the trained model is used to predict the unlabelled data, and the results predicted with high confidence are selected as the pseudolabels. The pseudolabelled data and labelled data are combined and used in training for a certain number of rounds. The above process is repeated until the proportion of newly added pseudolabelled data in each round is lower than a certain threshold. At this point, high-confidence labels are considered to be obtained, and the model has been sufficiently trained on all data.

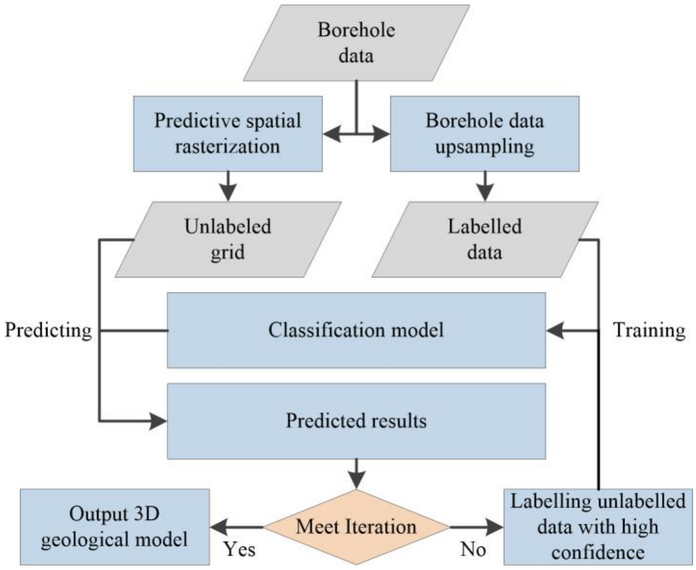


Figure 3. Algorithm flow chart.

2.4. Analysis of model uncertainty

The last layer of the neural network classifier normalizes the probability of the output through the softmax layer, and the softmax normalized result can be approximated as the probability corresponding to each stratum at a given data point. Therefore, when analysing the uncertainty of each data point in the raster model, the normalized information entropy can be introduced to quantitatively evaluate the uncertainty of the geological model. The normalized information entropy formula is as follows:

$$\mathrm{H(X)} = -\frac{\sum_{x\in S} \mathrm{p(x)ln(p(x))}}{\mathrm{S_{max}}} \qquad (5)$$

where S is the number of possible geological attributes for each data point, Smax is equal to ln(n), and n is the number of possible geological attributes. The information entropy of each data point is obtained by calculating the probability p(x) of each data point over all geological attributes. The magnitude of information entropy reflects the degree of complexity at a certain location in the geological model. The closer the information entropy is to 0, the higher the certainty of a data point for a certain stratum attribute, and the closer the information entropy is to 1, the higher the uncertainty of a data point for multiple geological attributes.

In addition, the data can be analysed based on an estimated confusion index (Burrough et al., 1997), and the ambiguity of classification can be evaluated by selecting the results of the two prediction categories with the highest probability for each data point. The confusion index formula is as follows:

$$\mathrm{CI} = [1 - (\mu_{\mathrm{max}} - \mu_{\mathrm{max-1}})] \qquad (6)$$

where μmax is the probability of the class with the highest predicted probability and μmax-1 is the probability of the class with the second highest predicted probability. CI values range from 0-1 to indicate the degree of confusion predicted for a certain data point, with 0 indicating that a classification result with a low confusion index is not ambiguous and 1 indicating that a classification result with a high confusion index is highly ambiguous.

3. Experimental method and verification

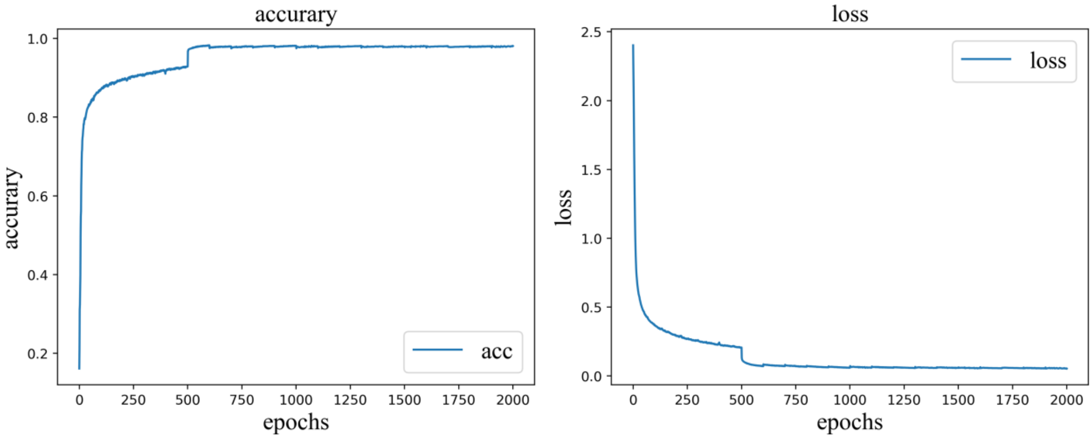
The Shenyang city 3D geological models were built using the SDLP, SVM, and HRBF algorithms. All test experiments in this chapter were performed on the same device: an Intel(R) Core (TM) i7-10750H CPU @2.60 GHz with an NVIDIA GeForce RTX 2060, 16.0 GB RAM, and Windows 10 (64-bit).

The ReLU function was used as the activation function in the SDLP algorithm, the initial learning rate was set to 0.001, and the batch size for training was set to 512. When the model training accuracy reached 90% or after 500 epochs, the unlabelled grids were labelled with pseudolabels. When the newly added pseudolabels accounted for less than 10% of the number of grids lacking labels in a given epoch, the model was trained for a total of 2000 epochs more before stopping. The training accuracy and loss values are shown in Fig. 4. The accuracy, precision, recall, and F1 score of the SVM, SDLP, and DL (the neural networks is same with SDLP but without pseudolabel) algorithms for the test dataset are shown in Table 2.

In the training process, when the labelled data and pseudolabel data are fused, the boundaries of stratigraphic categories are finely delineated, the final model training accuracy is above 95%, the loss function is close to zero, and the precision of the model for the test set is 98.16%. A confusion matrix is obtained from the test set (Fig. 5), which reflects the reliability of the evaluation results of the model. The classification accuracy of the model is high for all layers. Some strata are more likely to be confused due to being thin, displaying similar boundaries as other strata, or the influence of geological phenomena, such as depositional termination. The receiver operating characteristic curve (ROC) is another performance indicator that summarizes the performance of the binary classification model in the positive class and thus can be used to evaluate the diagnostic ability of the classifier according to the threshold change (Fawcett, 2006). The area under the ROC curve (AUC) (Fig. 6) represents a comprehensive measure of all possible classification thresholds. AUC values greater than 90%, from 75-90%, from 50-75% and less than 50% are considered to represent excellent, good, poor and unacceptable performance, respectively (Ray et al., 2010). The AUC values of the model are all above 90%, indicating that the classification performance of the model is excellent.

Table 2. The accuracy, precision, recall, and F1 score values for SVM and SDLP algorithms based on the test dataset

	Accuracy	Precision	Recall	F1 score
SAM	0.955	0.948	0.940	0.944
SDLP	0.982	0.983	0.980	0.982
DL	0.973	0.967	0.968	0.968



(a) (b)

Figure 4. Model training accuracy and loss variation curves

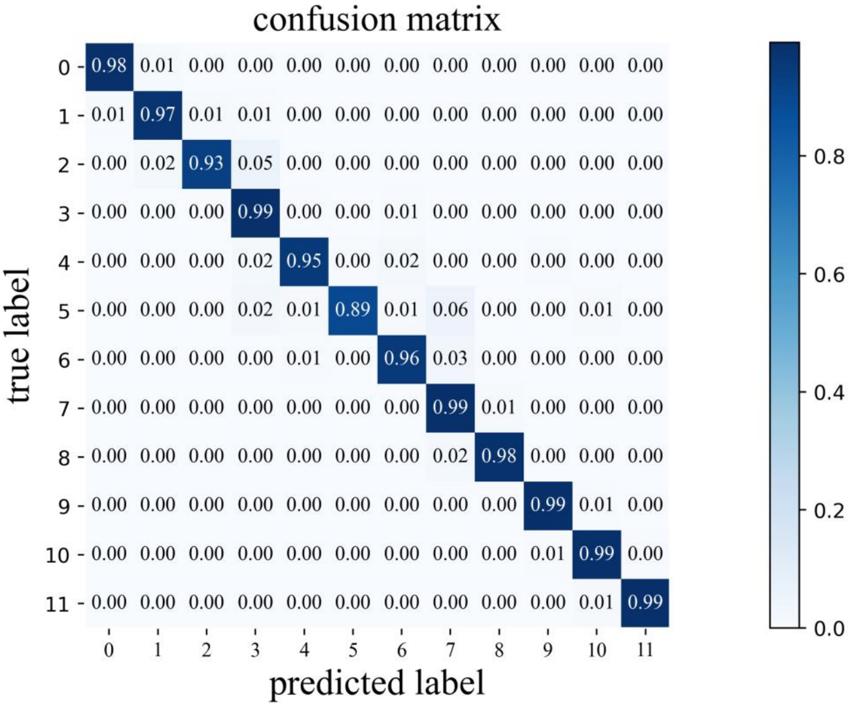


Figure 5. Confusion matrix of the classification results when the model is applied to the test dataset

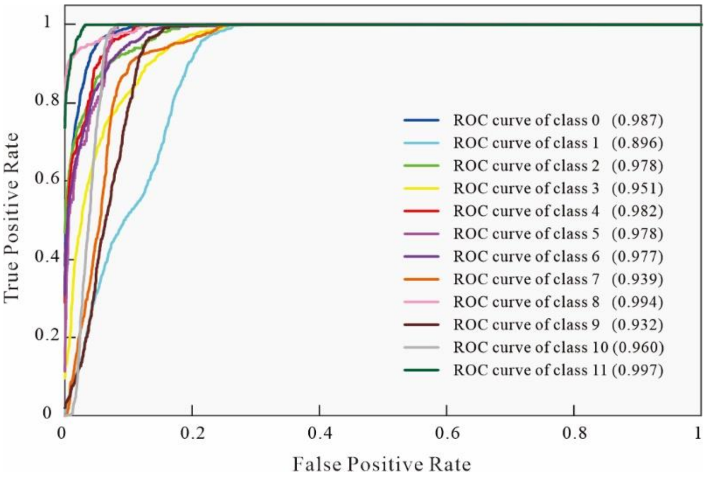


Figure 6. ROC curve for classification

The grid used in modelling is 1.5 m×1.3 m×0.3 m. The model uses the Tin mesh constructed from the top of boreholes to restrict the surface. The modelling range is determined according to a convex hull built by the borehole data, and the base of the model is determined according to a convex hull built by the bottoms of borehole data. Fig. 7 shows the modelling results for the study area. The model reveals the coverage relationships among the strata and reproduces the contact relationship between the depositional termination and unconformity of the strata.

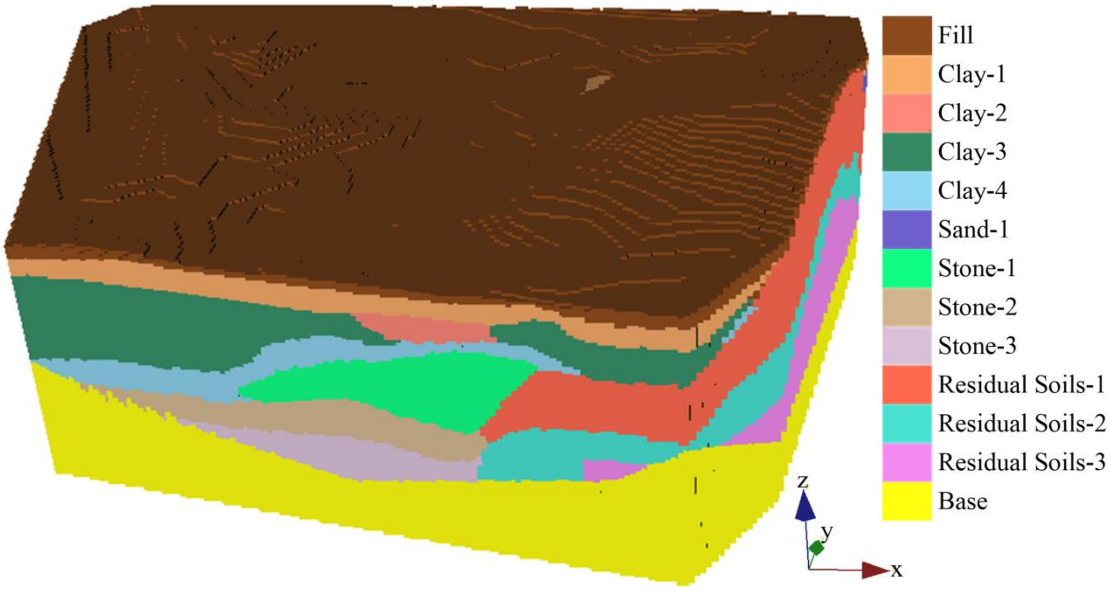


Figure 7. Model built using deep neural networks and the model legend.

To test the estimation accuracy at nonborehole locations using the proposed method, the borehole data were divided into a training set and a test set through k-fold cross validation, learning was performed with the training set of borehole data, and the test set accuracy was compared and analysed, where K was set to 10.

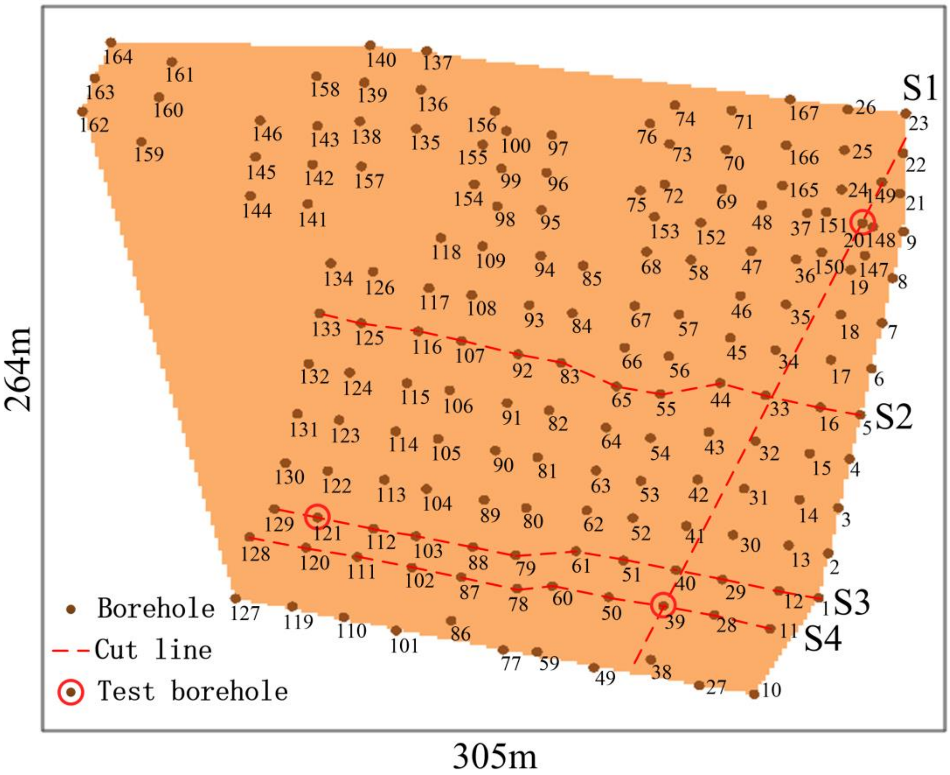


Figure 8. Borehole distribution and experimental analysis based on different profiles. The red dotted line are the profile, and the borehole points circled in red correspond to the boreholes tested using K1

The boreholes in the test set were sampled at equal intervals to determine the data point attributes at the boreholes, and the average accuracy of k-fold cross validation was calculated to be 71.65%. Due to the varying amount of geological information contained in individual borehole data, the importance of different boreholes in constructing the 3D geological model also differs. For instance, the test borehole data contains valuable lens body stratigraphic information and stratigraphic extinction information (Fig. 9). Removing the test borehole data would significantly decrease the accuracy of the prediction results. Therefore, we utilize the surface irregular triangulation method generated by the Delaunay rule to determine the topological relationships between the boreholes. Based on this approach, we ensure that boreholes containing a significant amount of geological information are not excluded during K-fold validation. These operations have improved the accuracy of K-fold validation from 71.65 to 85.9.

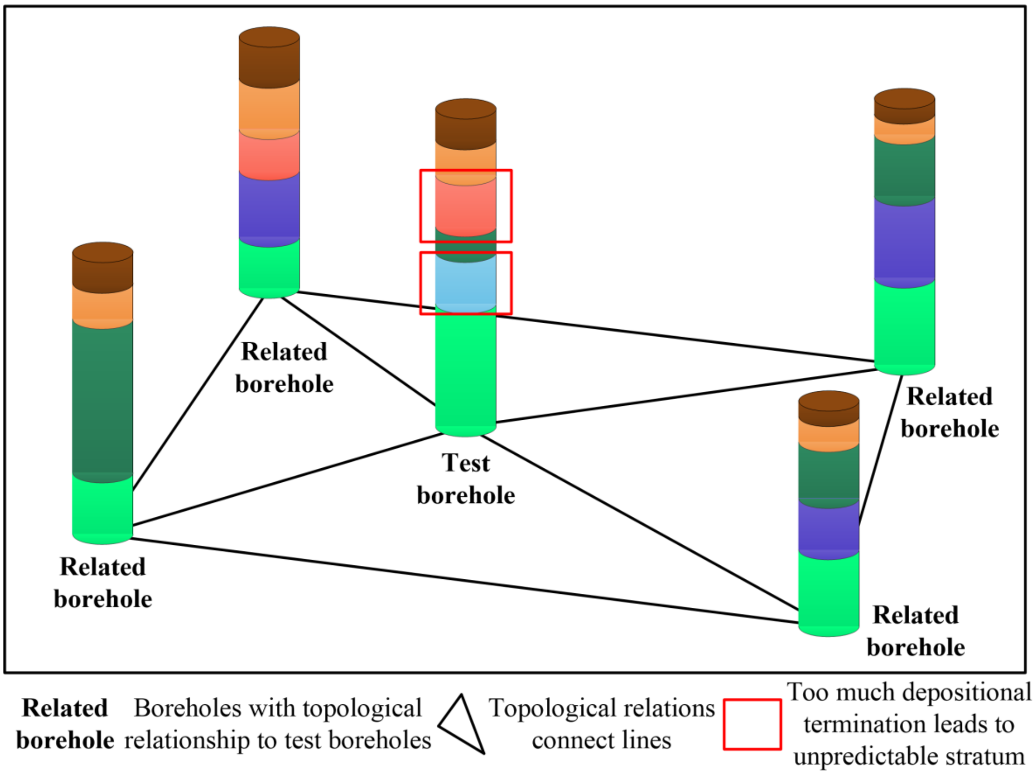

Related borehole
Test borehole
Related borehole
Related borehole
Related borehole
Related borehole
Boreholes with topological relationship to test boreholes
Topological relations connect lines
Too much depositional termination leads to unpredictable stratum

Figure 9. A situation in which too much depositional termination affects the prediction. A related borehole is a borehole that has a topological relationship with the predicted borehole. The red solid frame is the stratum, which is difficult to predict due to the excessive occurrence of depositional termination.

To further analyse the influence of accuracy on the model, a model with complete borehole data and a model with excluded sample K1 test borehole data were established, and the sections of the models through a test borehole were compared (Fig. 10). The Figure 10 show the results for a straight line thought the S1 and S3 profiles. Most areas of the sections at the boreholes in the test set are consistent with the sections built by a complete borehole. Since some test set boreholes are near depositional terminations, there is a certain difference between the model and the data from test boreholes, but the results are close and reasonable. In summary, the SDLP method display a good prediction ability for the neighbouring boreholes and can reveal the distribution characteristics of the strata.

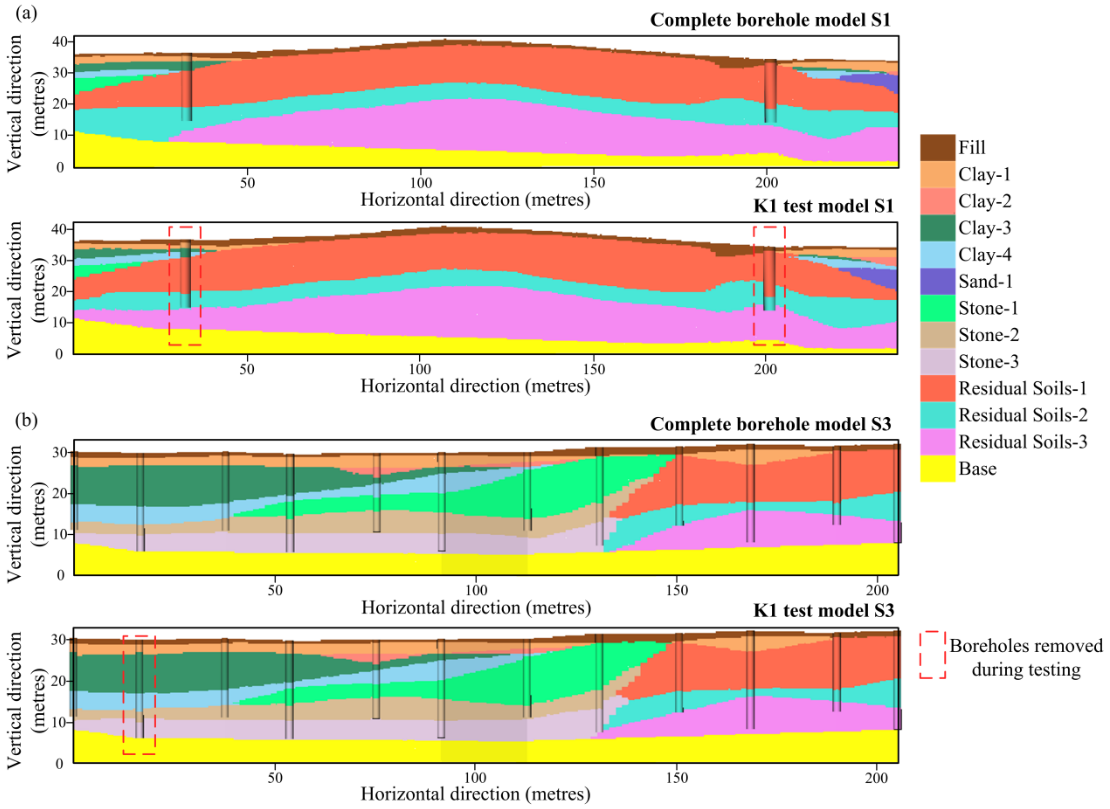


Figure 10. Comparison of the modelling results of sample K1 with the complete drilling results. The dotted box shows the boreholes considered during the test.

4. Discussion

4.1 Verification of the Accuracy of the HRBF Method

Three-dimensional geological modelling based on the Hermite radial basis function (HRBF) is one kind of implicit function modelling, and implicit modelling methods based on HRBF have been widely used in the modelling of ore bodies, regional geological surveys (Guo et al., 2016), urban geological surveys (Guo et al., 2021), tunnelling projects (Xiong et al., 2018), and volcanic formations (Guo et al., 2020). Therefore, in this paper, the HRBF method is used to build a 3D geological model of Shenyang city, and this model is used to compare the accuracy of the SDLP and SVM algorithms. Before evaluating the accuracy of the two algorithms mentioned earlier, it is essential to conduct an accurate analysis of the 3D geological model constructed using the HRBF method. To demonstrate the accuracy of this approach, we first use the HRBF method to build a 3D geological model of Shenyang city. S1, S2, S3, and S4 are profiles within the 3D geological model of Shenyang city, which contain many geological strata and complex geological relationships. The accuracy of these profiles can effectively reflect the accuracy of the HRBF modelling method. In the S1 geological profile, the stratigraphic boundaries contained in the borehole dataset nearly perfectly correspond to the boundaries of the three-dimensional geological model built based on the HRBF method (Figure. 11). This matching effect is also demonstrated for the S2, S3, and S4 geological profiles. The accurate correspondence between the borehole data and the cross-sections of the 3D geological model indicates the precision of the HRBF modelling method in constructing the three-dimensional geological model. Furthermore, 3D geological models of Shenyang city built using the HRBF method have been verified as effective in engineering applications (Guo et al.,2021). In conclusion, the 3D geological model built using the HRBF method can serve as a standard for evaluating the quality of 3D geological models constructed with the SDLP and SVM algorithms.

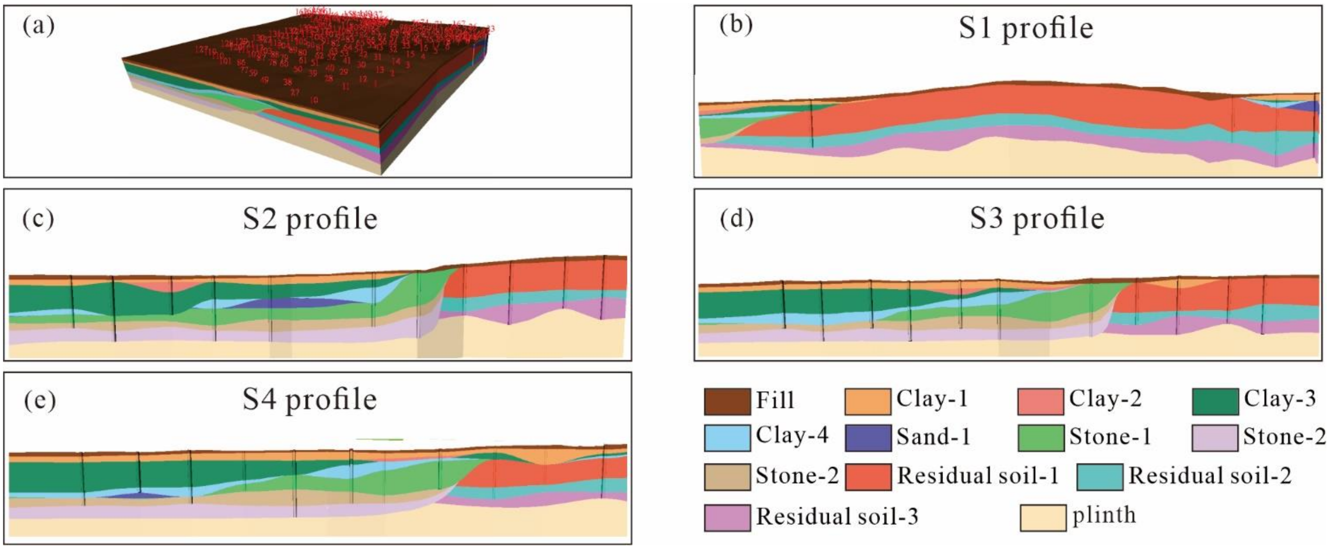


Figure.11 (a) the 3D geological model constructed by HRBF algorithm (b) the S1 profile built by HRBF algorithm; (c) the S2 profile built by HRBF algorithm; (d) the S3 profile built by HRBF algorithm; (e) the S4 profile built by HRBF algorithm

4.2 Comparison of Different Algorithms

Before building the three-dimensional geological model using the SDLP and SVM algorithms, it is necessary to observe the performance of these two algorithms based on the test dataset. According to the prediction results for the test dataset, the accuracy, precision, recall, and F1 score of the SDLP algorithm are 0.982, 0.983, 0.980, and 0.982, respectively, all of which are higher than those of the SVM algorithm (Fig. 12). The reason for these overall results may be that the SDLP algorithm uses more training data, enabling the model to learn patterns with higher generalization ability.

Furthermore, the accuracy, precision, recall, and F1 score of the SDLP algorithm are also higher than those of the DL algorithm (Fig.11). This phenomenon may be attributed to the increased quantity of the training dataset resulting from the use of pseudolabels constructed with the TIN method. The expanded training dataset enables the neural network model to achieve better generalization.

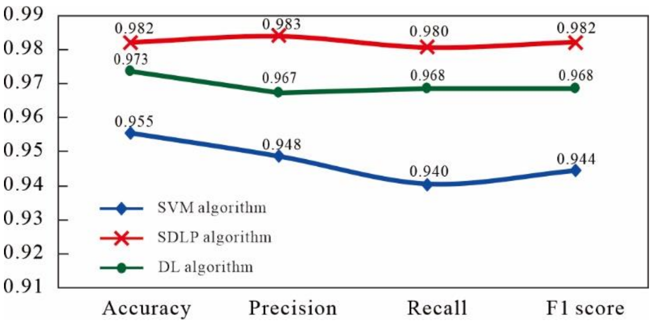


Figure 12. Accuracy, precision, recall, and F1 score of the SDLP and SVM algorithms.

4.3 Comparative Analysis of Models

The profiles of the 3D geological model of Shenyang city are compared to further validate the generalization ability of the SDLP algorithm and the SVM algorithm. The implicit HRBF modelling method exhibits excellent consistency with the borehole data in the profiles, and thus, the profiles constructed with the HRBF method are used as a benchmark for comparison with the profiles generated by machine learning algorithms. In Figure 13, the horizontal axis represents the modelling results of different algorithms for the same geological profile, and the vertical axis represents the geological profiling modelling results of the same algorithm for different geological profiles.

In the S2 geological profile, the 3D geological models built with the HRBF algorithm and the SDLP algorithm demonstrate a high level of consistency with the borehole data. However, the 3D geological model built with the SVM algorithm shows relatively poor correspondence with the borehole data. Furthermore, the morphology of formations in the 3D geological models created with different algorithms is not entirely consistent within the S2 profile. In sedimentary formations without fault structures, the formation boundaries typically undergo gradual changes rather than abrupt changes. The 3D geological models generated using the SDLP algorithm or the HRBF algorithm generally adhere to these geological laws. For instance, the intersection points of the stone-1, stone-2, and stone-3 strata and the residual-1, residual-2, and residual-3 strata in the 3D geological models developed using the SDLP and HRBF algorithms display smooth transitions, aligning well with the sedimentation patterns of sedimentary formations. Conversely, the contact relationships among the strata at these intersections in the 3D geological model built using the SVM algorithm do not conform to the actual sedimentation patterns. Additionally, at the apex of the lens-shaped sand-1 formation, the 3D geological model created with the SVM algorithm is less realistic than the models produced by the HRBF and SDLP algorithms. Guo et al. (2021) demonstrated through 3D geological modelling methods that there are no fault structures in the Shenyang area. This finding implies that the 3D geological model of the S2 profile built with the SVM method is not reasonable. Moreover, the HRBF method produces modelling results that are deemed unreasonable for the lower two layers, stone-3 and residual-3, due to constraints imposed by the implicit model. These constraints involve the stratum interface being defined based on the control points of each borehole and the implicit equation. In conclusion, for the S2 profile, the SDLP algorithm exhibits the most favourable modelling performance.

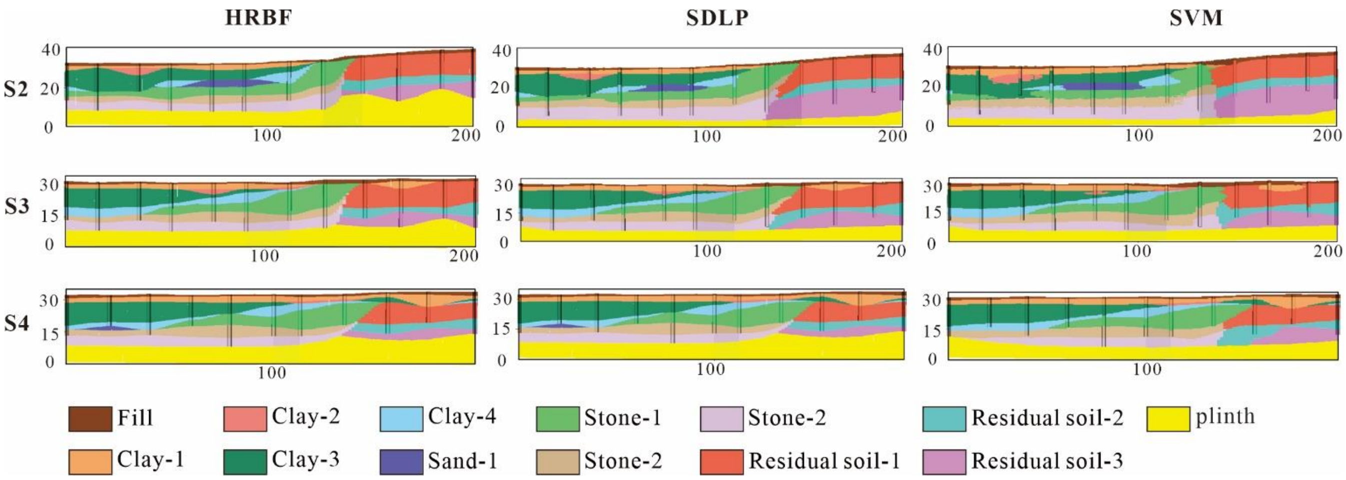


Figure 13. Geological profiles S2, S3, and S4 for Shenyang city built based on the SDLP, SVM, and HRBF algorithms.

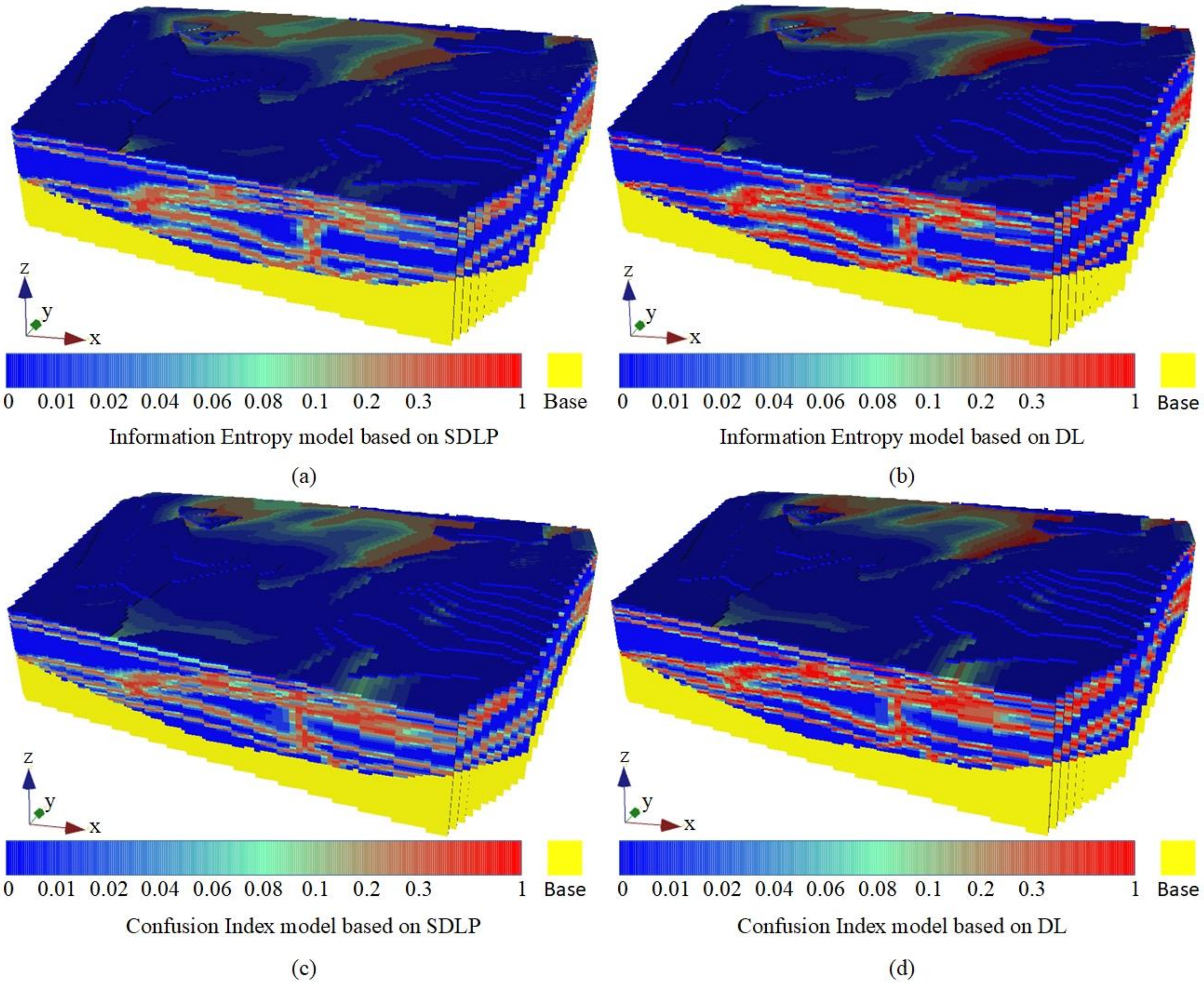
The situation for the S3 and S4 geological profiles is generally similar to that of the S2 profile. The 3D geological models built using the HRBF algorithm and the SDLP algorithm demonstrate a high level of consistency with the borehole data, and the correspondence between the 3D geological model built with the SVM algorithm and the borehole data is comparatively poor. The boundaries of sedimentary formations in the 3D geological models built using the HRBF algorithm or the SDLP algorithm adhere more closely to the actual sedimentation patterns compared to the boundaries of the 3D geological models built using the SVM algorithm. At the lowermost layer boundary, the 3D geological model built using the SDLP algorithm is more reasonable than the one built using the HRBF algorithm.

Based on a comparison of the results of the S2, S3, and S4 profiles, the SDLP algorithm demonstrates better ability to reflect the borehole data when building the 3D geological model. Additionally, the 3D geological model created using the SDLP algorithm better aligns with the sedimentation patterns in terms of the morphology of the formations.

4.4 Analysis of Model Uncertainty

For a 3D geological model, only the strata boundary information reflected by borehole data is accurate, and the strata boundaries in areas outside the borehole data region are either artificially inferred or based on constructed basis functions. Therefore, it is necessary to analyse the strata boundaries established based on borehole data in certain areas in the three-dimensional geological model. The implicit HRBF modelling algorithm can be used to effectively visualize borehole data. However, because it is based on implicit basis functions for visualization, it may not effectively process the undisclosed geological information associated with borehole data. In this study, information entropy and a confusion index are introduced to address the HRBF algorithm's inability to consider uncertainty in areas without borehole data. The information entropy is calculated based on the probability distribution of all data points in the normalized model. A visualized information entropy model can reflect the uncertainty at different locations within the model.

In addition, the results of the information entropy and confusion index models of SDLP and DL algorithms are compared. These results are used to demonstrate the impact of pseudolabelling on the stability of building 3D geological models using neural network methods.



z
y
x
0
0.01
0.02
0.04
0.06
0.08
0.1
0.2
0.3
1
Base
Information Entropy model based on SDLP
Information Entropy model based on DL
Confusion Index model based on SDLP
Confusion Index model based on DL

(a) (b) (c) (d)

Figure 14. Models of uncertainty: (a) information entropy model based on SDLP; (b) information entropy model based on DL; (c) confusion index model based on SDLP; and (d) confusion index model based on DL

The information entropy and confusion index models reflect the uncertainty of the semisupervised learning method using pseudolabels and the supervised learning method used to build the models (Fig. 14). The blue part of the information entropy model (Fig. 14a, c), where the information entropy is close to 0, indicates that the uncertainty of the stratum attribute values in the region is low, and the entropy value is small, mainly between the model stratum boundaries. The red part, where the information entropy is close to 1, indicates that the region has a high probability of being influenced by stratum attribute values, and the entropy value is large, mainly distributed near the stratum boundary obtained through training. In the confusion index model (Fig. 14b, d), the blue part indicates a low confusion index, and the red part indicates a high confusion index.

In the confusion index model, the three-dimensional geological models built by SDLP algorithm and DL algorithm both exhibit a confusion index close to 0 within strata but increases in the confusion index at the boundaries of the strata. The difference lies in the fact that at the boundaries of strata, the confusion index of the three-dimensional geological model built with the deep learning algorithm without pseudolabelling is closer to 1, indicating lower accuracy than that of the 3D geological model built with the deep learning algorithm with pseudolabelling. Additionally, the information entropy model exhibits similar characteristics to the confusion index model. To visually illustrate the differences between the 3D geological models constructed by the SDLP algorithm and the DL algorithm in terms of information entropy and confusion index, the number of stable grids (with information entropy ranging from 0 to 0.01 and confusion index ranging from 0 to 0.01, Fig.15a, b) and unstable grids (with information entropy ranging from 0.3 to 1 and confusion index ranging from 0.3 to 1, Fig.15a, b) are recorded and compared. The results show that compared to the DL algorithm, the 3D geological model constructed by the SDLP algorithm has a higher proportion of stable grids and a lower proportion of unstable grids. The findings demonstrate that utilizing the TIN algorithm to construct pseudolabels can enhance the stability of the model.

The information entropy and confusion index models can be used to overcome the HRBF algorithm's inability to consider uncertainty, and the results demonstrate that the SDLP algorithm is superior to the deep learning algorithm without pseudolabelling for constructing 3D geological models from the perspectives of information entropy and the confusion index.

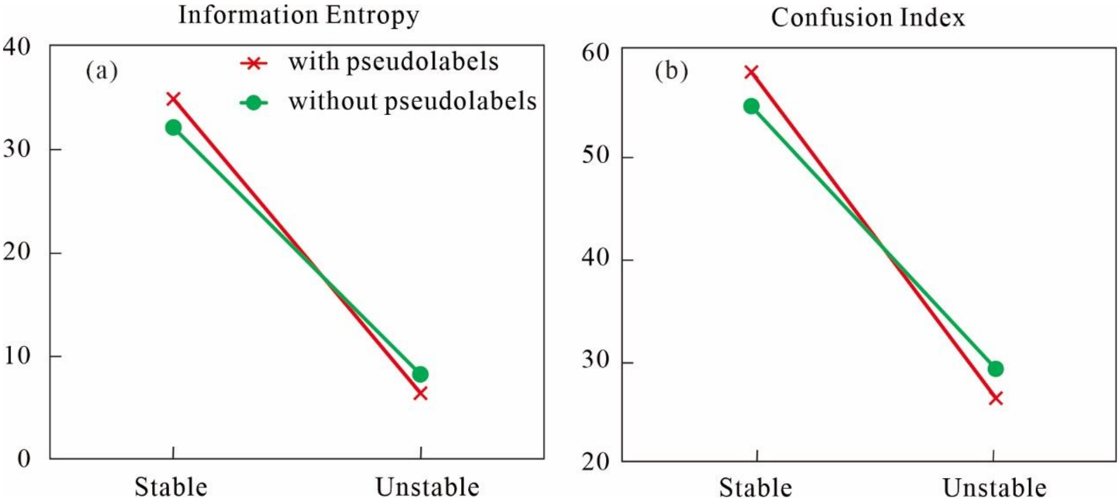


Figure 15. Line plot of information entropy(a) and confusion index (b).

5. Conclusion

In this study, we propose semisupervised deep learning using a pseudolabelling algorithm to build a 3D geological model based on borehole data. By labelling the grid data with high accuracy using the explicit TIN modelling method, we address the lack of labelled training data for building deep learning models. The original data for this study are from engineering borehole dataset from Shenyang city, and 3D geological models of Shenyang city were constructed using the SDLP, SVM, and HRBF algorithms. The SDLP algorithm achieved an accuracy of 98.16% for the test dataset, outperforming a classic

SVM machine learning algorithm. Moreover, the 3D geological model constructed using the SDLP algorithm accurately reflects the boundaries of the formations in the borehole data and aligns well with the real sedimentation patterns. the 3D geological models built with the SDLP algorithm resolve the inability of the implicit HRBF modelling algorithm to consider uncertainty. In conclusion, the proposed SDLP algorithm provides a solution for the lack of training data in deep learning and fills the gap of the HRBF method regarding uncertainty.

Code availability. The PDNN was written in the Python programming language. The program reads borehole data and preprocesses the borehole data with upsampling and normalization. By using the DNN to train the model and predict the attributes of data points, Pseudolabels with high confidence scores were added to the unlabelled grid points. The code is available for download from the following public repository: https://zenodo.org/deposit/7833570.

Data availability. The model data and terrain data used in the case study in this paper are available at: https://doi.org/10.5281/zenodo.7535214.

Video supplement. We have provided web links to download the video recordings of our case studies. The case study of a real area verifies the feasibility of the proposed approach. The video supplement can be viewed at: https://drive.google.com/file/d/13VERDXM6YJmP7xMabQy3IjhCExuQSWzk/view?usp=sharing.

Author contributions. Xuechuang Xu and Jiateng Guo conceived the manuscript; Jiateng Guo provided funding support and ideas; Xuechuang Xu was responsible for the research method and program development; Jiateng Guo provided the data used in this research; Xuechuang Xu, Jiateng Guo, Xulei Wang, Lixin Wu, Mark Jessell, Zhibin Liu and Yufei Zheng helped improve the manuscript; Luyuan Wang helped modified the manuscript. All authors have read and agreed to the published version of the manuscript.

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

Acknowledgements. This work was financially supported by the National Natural Science Foundation of China [grant number: 42172327] and the Fundamental Research Funds for the Central Universities [grant number: N2201022].

References

Avalos, S., & Ortiz, J. M.: Recursive Convolutional Neural Networks in a Multiple-Point Statistics Framework, Computers & Geosciences, 141, 104522, http://doi.org/https://doi.org/10.1016/j.cageo.2020.104522, 2020.

Batista, G. E. A. P., Prati, R. C., & Monard, M. C.: A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data. Sigkdd Explor. Newsl., 6(1), 20-29. http://doi.org/10.1145/1007730.1007735, 2004

Burrough, P.A., vanGaans, P.F.M. and Hootsmans, R.: Continuous classification in soil survey: Spatial correlation, confusion and boundaries, GEODERMA, 77(2-4): 115-135, https://doi.org/10.1016/S0016-7061(97)00018-9, 1997.

Caers, J. Modeling Uncertainty in the Earth Sciences. Wiley., 2011.

Calcagno, P., Chiles, J.P., Courrioux, G. and Guillen, A.: Geological modelling from field data and geological knowledge Pa rt I. Modelling method coupling 3D potential-field interpolation and geological rules, Physics of the Earth and Planetary Interiors, 171(1-4): 147-157, https://doi.org/10.1016/j.pepi.2008.06.013, 2008.

Caumon, G., L. Tertois, A., & Zhang, L.: Elements for Stochastic Structural Perturbation of Stratigraphic Models, European Association of Geoscientists & Engineers, http://doi.org/https://doi.org/10.3997/2214-4609.201403041, 2007.

Carle, S.F. and Fogg, G.E.: Modeling spatial variability with one and multidimensional continuous-lag Markov chains, MATHEMATICAL GEOLOGY, 29(7): 891-918, 1997.

Caumon, G., Gray, G., Antoine, C. and Titeux, M.O.: Three-Dimensional Implicit Stratigraphic Model Building From Remote Sensing Data on Tetrahedral Meshes: Theory and Appli cation to a Regional Model of La Popa Basin, NE Mexico, IEEE Transactions on Geoscience and Remote Sensing, 51(3): 1613-1621, https://doi.org/10.1109/TGRS.2012.2207727, 2013.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P.: Smote: Synthetic Minority Over-Sampling Technique. J. Artif. Int. Res., 16(1), 321-357., 2002.

Che, D. F., Wu, L. X., & Yin, Z. R.: 3D Spatial Modeling for Urban Surface and Subsurface Seamless Integration, 2009 Ieee International Geoscience and Remote Sensing Symposium, Vols 1-5, 1694, http://doi.org/10.1109/IGARSS.2009.5417787, 2009.

Chen, G., Zhu, J., Qiang, M., & Gong, W.: Three-Dimensional Site Characterization with Borehole Data – a Case Study of Suzhou Area, Engineering Geology, 234, 65-82, http://doi.org/https://doi.org/10.1016/j.enggeo.2017.12.019, 2018.

Cuomo, S., Galletti, A., Giunta, G., & Marcellino, L.: Reconstruction of Implicit Curves and Surfaces Via Rbf Interpolation, Applied Numerical Mathematics, 116, 157-171, http://doi.org/https://doi.org/10.1016/j.apnum.2016.10.016, 2017.

de la Varga, M., Schaaf, A., & Wellmann, F.: Gempy 1.0: Open-Source Stochastic Geological Modeling and Inversion, Geoscientific Model Development, 12(1), 1-32, http://doi.org/10.5194/gmd-12-1-2019, 2019.

Fawcett, T.: An introduction to ROC analysis, Pattern Recognition Letters, 27(8): 861-874, 2006.

Guo, J.T., Liu, Y.H., Han, Y.F. and Wang, X.L.: Implicit 3D Geological Modeling Method for Borehole Data Based on Mechine Learning, Journal of Northeastern University (Natural Science), 40(9): 1337-1342, https://doi.org/10.12068/j.issn.1005-3026.2019.09.021, 2019.

Guo, J.T., Wang, X.L., Wang, J.M., Dai, X.W., Wu, L.X., Li, C.L., Li, F.D., Liu, S.J. and Jessell, M.W.: Three-dimensional geological modeling and spatial analysis from geotechnical borehole data using an implicit surface and marching tetrahedra algorithm, Engineering Geology, 284, https://doi.org/10.1016/j.enggeo.2021.106047, 2021.

Guo, J., Wang, Z., Li, C., Li, F., Jessell, M.W., Wu, L. and Wang, J.: Multiple-Point Geostatistics-Based Three-Dimensional Automatic Geological Modeling and Uncertainty Analysis for Borehole Data, Natural Resources Research, 31(5): 2347-2367, https://doi.org/10.1007/s11053-022-10071-6, 2022.

Hillier, M.J., Schetselaar, E.M., de Kemp, E.A. and Perron, G.: Three-Dimensional Modelling of Geological Surfaces Using Generalized Interpolation with Radial Basis Functions, Mathematical Geosciences, 46(8): 931-953, https://doi.org/10.1007/s11004-014-9540-3, 2014.

Huang, X.R., Dai, Y., Xu, Y.G. and Tang, J.: Seismic Inversion Experiments Based on Deep Learning Algorithm Using Different Datasets, Journal of Soutwest Petroleum University (Science & Technology Edition), 42(6): 16-25, 2020.

Laloy, E., Herault, R., Lee, J., Jacques, D. and Linde, N.: Inversion using a new low-dimensional representation of complex binary geological media based on a deep neural network, Advances in Water Resources, 110: 387-405, https://doi.org/10.1016/j.advwatres.2017.09.029, 2017.

Lancaster, S.T. and Bras, R.L.: A simple model of river meandering and its comparison to natural channels, HYDROLOGICAL PROCESSES, 16(1): 1-26, https://doi.org/10.1002/hyp.273, 2002.

Lantuéjoul, C.: Geostatistical simulation: models and algorithms (No. 1139), Springer Science & Business Media, https://doi.org/10.1007/978-3-662-04808-5, 2001.

Liu, H., Chen, S.Z., Hou, M.Q. and He, L.: Improved inverse distance weighting method application considering spatial autocorrelation in 3D geological modeling, Earth Science Informatics, 13(3): 619-632, https://doi.org/10.1007/s12145-019-00436-6, 2020.

Ran, X.J. and Xue, L.F.: The research of method and system of regional three-dimensional geological modeling, Doctor Thesis, Jilin University, 2020.

Ray, P., Manach, Y.L., Riou, B. and Houle, T.T.: Statistical evaluation of a biomarker, Anesthesiology, 112(4): 1023-1040, https://doi.org/10.1097/ALN.0b013e3181d47604, 2010.

Titos, M., Bueno, A., Garcia, L. and Benitez, C.: A Deep Neural Networks Approach to Automatic Recognition Systems for Volcano-Seismic Events, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 11(5): 1533-1544, https://doi.org/10.1109/JSTARS.2018.2803198, 2018.

Wang, J.M., Zhao, H., Bi, L. and Wang, L.G.: Implicit 3D Modeling of Ore Body from Geological Boreholes Data Using Hermite Radial Basis Functions, Minerals, 8(10), https://doi.org/10.3390/min8100443, 2018.

Wu, L.X.: Topological relations embodied in a generalized tri-prism (GTP) model for a 3D geoscience modeling system, Computers & Geosciences, 30(4): 405-418, https://doi.org/10.1016/j.cageo.2003.06.005, 2004.

Xu, S.T. and Zhou, Y.Z.: Artificial intelligence identification of ore minerals under microscope based on deep learning algorithm, Acta Petrologica Sinica, 34(11): 3244-3252, 2018.

Yang, Y.S., Li, Y.Y., Liu, T.Y., Zhan, Y.L. and Feng, J.: Interactive 3D forward modeling of total field surface and three-component borehole magnetic data for the Daye iron-ore deposit (Central China), Journal of Applied Geophysics, 75(2): 254-263, https://doi.org/10.1016/j.jappgeo.2011.07.010, 2011.

Zhang, T.F., Tilke, P., Dupont, E., Zhu, L.C., Liang, L. and Bailey, W.: Generating geologically realistic 3D reservoir facies models using deep learning of sedimentary architecture with generative adversarial networks, Petroleum Science, 16(3): 541-549, https://doi.org/10.1007/s12182-019-0328-4, 2019.

Zhang, X.Y., Ye, P., Wang, S. and Du, M.: Geological entity recognition method based on Deep Belief Networks, Acta Petrologica Sinica, 34(2): 343-351, 2018.

Zhang, Y. and Chu, B.Z.: Automatic Borehole Comparison Technology Based on Machine Learning, Master Thesis, China University of Geosciences (Beijing), 2021.

Zhang, Z., Wang, G., Carranza, E. J. M., Yang, S., Zhao, K., Yang, W., & Sha, D.: Three-Dimensional Pseudo-Lithologic Modeling Via Adaptive Feature Weighted K-Means Algorithm From Multi-Source Geophysical Datasets, Qingchengzi Pb–Zn–Ag–Au District, China. Natural Resources Research, 31(4), 2163-2179, http://doi.org/10.1007/s11053-021-09927-0, 2022.

Zhang, Z., Wang, G., Carranza, E. J. M., Liu, C., Li, J., Fu, C., Liu, X., Chen, C., Fan, J., & Dong, Y.: An Integrated Machine Learning Framework with Uncertainty Quantification for Three-Dimensional Lithological Modeling From Multi-Source Geophysical Data and Drilling Data. Engineering Geology, 324, 107255. http://doi.org/10.1016/j.enggeo.2023.107255, 2023.