

Copyright Undertaking

This thesis is protected by copyright, with all rights reserved.

By reading and using the thesis, the reader understands and agrees to the following terms:

1. The reader will abide by the rules and legal ordinances governing copyright regarding the use of the thesis.
2. The reader will use the thesis for the purpose of research or private study only and not for distribution or further reproduction or any other purpose.
3. The reader agrees to indemnify and hold the University harmless from and against any loss, damage, cost, liability or expenses arising from copyright infringement or unauthorized usage.

IMPORTANT

If you have reasons to believe that any materials in this thesis are deemed not suitable to be distributed in this form, or a copyright owner having difficulty with the material being included in our database, please contact lbsys@polyu.edu.hk providing details. The Library will look into your claim and consider taking remedial action upon receipt of the written requests.

MACHINE LEARNING BASED GEOLOGICAL MODELLING AND APPLICATION

QIU YASHI

PhD

The Hong Kong Polytechnic University

This program is jointly offered by The Hong Kong Polytechnic University and Zhejiang
University

2026

The Hong Kong Polytechnic University

Department of Civil and Environmental Engineering

Zhejiang University
College of Agriculture and Biotechnology

**MACHINE LEARNING BASED GEOLOGICAL MODELLING AND
APPLICATION**

QIU YASHI

A thesis submitted in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

December 2025

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

Signature:

Name: QIU YASHI

Abstract of thesis entitled

Machine Learning based Geological Modelling and Application

In recent years, with the rapid development of China's economy and the advancement of modernization construction, the demand for underground space development has been increasing day by day, which has also attracted more and more attention from scholars to the problem of geological profile prediction. Accurate geological profile prediction can reveal the distribution characteristics of soil types and strata in a site, which has a significant impact on the design and construction in geotechnical engineering structures. However, the uncertainty and inherent variability of geological formations make accurate geological profile prediction based on sparse borehole data a highly challenging problem.

This study proposes two machine learning algorithms based on sparse boreholes for accurate prediction of geological profiles, each with its own advantages. These algorithms are applied to the prediction of geological profiles in multiple practical engineering projects, obtaining accurate geological profiles to guide engineering practice. Comparing the predictions of the proposed algorithm with those from other machine learning methods to demonstrate its superior performance, particularly in significantly improving boundary accuracy. Furthermore, by evaluating the differences between simulated and monitored surface settlements outside the foundation pit by comparing a linearly interpolated geological model with a machine learning-derived nonlinear profile, both based on sparse borehole data. The actual monitoring data shows that the machine learning-predicted nonlinear geology leads to more accurate finite element simulations.

The main contributions and innovative achievements of this study are as follows:

(1) The MSRF-XGBoost algorithm was proposed, which includes two parts: classification module and filtering module. The classification module is based on the traditional XGBoost algorithm and adds multi receptive field sampling to learn spatial features of different scales between sparse boreholes, enabling accurate prediction of geological profiles through sparse equidistant boreholes alone; The filtering module automatically selects the optimal prediction result from different results predicted by spatial features at different scales through Gaussian filtering, boundary similarity comparison, and other steps. This algorithm solves the dependence of existing machine learning algorithms on training profiles and improves the boundary accuracy of prediction results by about 40% in nonlinear geological profile prediction.

(2) An improved GCN algorithm has been proposed. By improving and optimizing the feature expression, adjacency matrix design, and model training method of the GCN algorithm, the improved GCN algorithm can predict the soil type of unknown points through their coordinates, and thus predict the complete geological profile, solving the problem of existing machine learning algorithms relying on equidistant drilling for training. This algorithm can predict accurate geological profiles through sparse non equidistant boreholes. Compared with other machine learning methods, the boundary accuracy of nonlinear geological profile prediction has been improved by about 30%.

(3) The two proposed machine learning methods are applied to multiple practical engineering projects. Validation through borehole accuracy assessment of verification boreholes shows prediction accuracy between 80%-99% across three different prediction profiles. Comparative

analysis with other machine learning algorithms confirms the proposed methods' advantages in reducing training data requirements while improving prediction accuracy.

(4) Taking a foundation pit in Hangzhou as an example, linear geological profiles based on sparse boreholes linear interpolation and nonlinear geological profile finite element models based on algorithm prediction are established. During the simulation of excavation process, the non-linear geological profile finite element model showed a difference of 2% between the surface settlement value outside the pit and the actual monitored settlement value, while the linear geological profile finite element model simulation result differed from the actual settlement value by 12%, indicating that accurate non-linear geological profiles can make finite element model simulation more accurate.

LIST OF PUBLICATIONS DURING THE CANDIDATURE

Research Papers

- [1] **Yashi Qiu**, Ning Zhang, Zhenyu Yin, Yu Wang, Changjie Xu, Pin Zhang. Novel multi-spatial receptive field (MSRF) XGBoost method for predicting geological cross-section based on sparse borehole data[J]. **Engineering Geology**, 2024, 338: 107604. (SCI, JCR Q1 ☒)
- [2] **Yashi Qiu**, Jianing Yin, Ning Zhang, Hongchi Liu, Changjie Xu. Novel graph convolutional network for geological profile prediction using non-equidistant borehole data[J]. **Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards**, 2024: 1-15. (SCI, JCR Q1 ☒)

ACKNOWLEDGEMENTS

The time in the Hong Kong Polytechnic University is precious and unforgettable. I would like to express my deepest gratitude to my chief supervisor Prof. Zhen-Yu Yin and co-supervisor Chang-Jie Xu for their supervision and guidance of my research. I am grateful to those who supported and accompanied me throughout this special period. With the support of my family and friends, I was able to get through this period of time. So, I would like to express my sincere thanks to my parents, boyfriend, colleagues and friends.

TABLE OF CONTENTS

CHAPTER 1 INTRODUCTION	1
1.1 Research background.....	1
1.2 Research objectives	5
1.3 Structure of the thesis	6
CHAPTER 2 LITERATURE REVIEW	9
2.1 Introduction	9
2.2 Research on geological profile prediction	9
2.2.1 Interpolation methods.....	9
2.2.2 Statistical methods.....	9
2.2.3 Machine learning methods	11
2.3 Research on numerical simulation of foundation excavation	16
2.4 Summary	17
CHAPTER 3 MULTI-SPATIAL RECEPTIVE FIELD (MSRF) XGBOOST METHOD	19
3.1 Introduction	19
3.2 MSRF-XGBoost methodology.....	20
3.2.1 XGBoost method.....	23
3.2.2 Definition of MSRF	24
3.2.3 Classification module of MSRF-XGBoost method	27
3.2.4 Identification module of MSRF-XGBoost method	29
3.2.5 Accuracy criteria	32
3.3 Linear geological profile prediction results	34
3.4 Nonlinear geological profile prediction results	35
3.4.1 Prediction results	35

3.4.2 Identification process	38
3.5 Analysis of factors affecting the accuracy of MSRF-XGBoost method prediction	42
3.5.1 Effect of borehole space	42
3.5.2 Effect of vertical resolution	43
3.6 Summary	45
CHAPTER 4 GRAPH CONVOLUTIONAL NETWORK FOR NON-EQUIDISTANT BOREHOLE DATA	47
4.1 Introduction	47
4.2 Novel graph convolutional network methodology	49
4.2.1 GCN method	50
4.2.2 Adaptive improvement of GCN algorithm.....	52
4.2.3 Geological Profile Prediction based on novel GCN method.....	54
4.3 Nonlinear geological profile prediction results	56
4.4 Analysis of factors affecting the accuracy of novel GCN method prediction	59
4.4.1 Effect of the number of boreholes	59
4.4.2 Effect of borehole location	61
4.5 Summary	63
CHAPTER 5 ENGINEERING APPLICATION AND COMPARISON WITH DIFFERENT METHODS	65
5.1 Introduction	65
5.2 Engineering application based of MSRF-XGBoost method.....	67
5.2.1 Borehole source and layout.....	67
5.2.2 Geological profile prediction	68

5.3 Engineering application based on novel GCN method	71
5.3.1 Borehole source and layout.....	71
5.3.2 Geological profile IX prediction	72
5.3.3 Geological profile XI prediction	75
5.4 Comparison and analysis of prediction results of different machine learning algorithms	77
5.4.1 Comparative analysis of nonlinear geological profile prediction	77
5.4.2 Comparative analysis of geological profile prediction for Hangzhou foundation pit engineering	79
5.5 Summary	82
CHAPTER 6 FINITE ELEMENT ANALYSIS OF FOUNDATION EXCAVATION BASED ON ACCURATE GEOLOGICAL PROFILE	84
6.1 Introduction	84
6.2 Model development	85
6.2.1 Model dimensions and boundary conditions.....	85
6.2.2 Selection of soil and retaining structure parameters	87
6.2.3 Model validation	90
6.3 Simulation results	91
6.3.1 Ground settlement during the first excavation layer	93
6.3.2 Ground settlement during the second excavation layer	94
6.3.3 Ground settlement during the third excavation layer	96
6.3.4 Total ground settlement.....	97
6.4 Summary	98

CHAPTER 7 CONCLUSIONS AND FUTURE WORK.....	100
7.1 Findings and conclusions	100
7.2 Recommendations for future studies	103
REFERENCES.....	104

LIST OF FIGURES

Figure 1-1(a) Leaning Tower of Pisa (b) Vietnamese Bridge	2
Figure 1-2 Geological profile of a foundation pit project in Hangzhou.....	3
Figure 1-3 Technology roadmap	8
Figure 2.2 IC-XGBoost algorithm	12
Figure 2.3 IC-XGBoost3D algorithm.....	13
Figure 2.4 DNN model application	14
Figure 3.1 Flow chart of the proposed MSRF XGBoost method	22
Figure 3.3 Predictions of nonlinear soil boundary by different SRF models: (a) ground truth; (b) model SRF-3; (c) model SRF-5; (d) model SRF-7; (e) model SRF-9; (f) model SRF-11; (g) model SRF-13; (h) model SRF-15; (i) model SRF-17	27
Fig. 3.4. Illustration on classification module of MSRF XGBoost method.....	29
Fig. 3.5. Illustration on identification module of MSRF XGBoost method	32
3.7. Predictions of case 2 by different MSRF models: (a) ground truth; (b) MRF method; (c) IC-XGBoost method; (d) recommended model SRF-3; (e) model SRF-5; (f) model SRF-7 .	37
Fig. 3.9. Boundary similarity between candidates and predictions: (a) Candidate 1 from model SRF-3 with filter size 3; (b) Candidate 2 from model SRF-5 with filter size 3; (c) predictions from model SRF-7 with filter size 7; (d) predictions from model SRF-9 with filter size 5	41
Fig. 3.11. Predictions of optimal SRF model for different vertical resolutions: (a) model SRF-3 for resolution at 60; (b) model SRF-3 for resolution at 75; (c) model SRF-5 for resolution	

at 100; (d) model SRF-9 for resolution at 150	45
Fig. 4.1. Flow chart for the proposed improved GCN model	50
Fig. 4.2. Illustration on graph convolutional network model.....	52
Fig. 4.4. Comparing the results of the different models in the real case: (a) the actual geological profile; (b) MRF method; (c) IC-XGBoost method; and (d) the proposed GCN method .	58
Fig. 4.5. Predictions of different numbers of boreholes: (a) four boreholes; (b) five boreholes; (c) six boreholes; and (d) seven boreholes	60
61	
Fig. 4.6. Accuracy variations versus the borehole distances	61
Fig. 4.7. Predictions of different positions of boreholes: (a)–(c) different unequal spacings; (d) MSRF-XGBoost method.....	63
Fig. 5.2. Predictions of case 3 by MSRF XGBoost models with filter size 3: (a) measurement; (b) model SRF-3; (c) model SRF-5; (d) model SRF-7.....	69
Fig. 5.3. Layout plan of boreholes in a Hangzhou foundation.....	71
Fig. 5.5. Predictions of Hangzhou Profile 1: (a) nine boreholes (white); (b) six boreholes (white); (c) six boreholes (white); and (d) six boreholes (white)	76
Fig. 5.7 Prediction results of actual engineering geological profiles using different algorithms (a) Improved GCN algorithm prediction (b) SRF-3 model prediction (c) SRF-5 model prediction (optimal recommendation after screening) (d) SRF-7 model prediction	81
Fig 6.3 Deep horizontal displacement of the diaphragm wall.....	91
Fig 6.4 Comparison of different geological profiles	92

(a) linear geological profile, (b) nonlinear geological profile	92
Fig 6.5 Finite element model with linear stratum interface	92
Fig 6.7 Mesh division of the linear profile finite element model.....	93
Fig 6.8 Mesh division of the nonlinear profile finite element model.....	93
Fig 6.9 Ground settlement of the linear profile model during the first excavation layer.....	94
Fig 6.10 Ground settlement of the nonlinear profile model during the first excavation layer...	94
Fig 6.11 Ground settlement of the linear profile model during the second excavation layer	95
Fig 6.14 Ground settlement of the nonlinear profile model during the third excavation layer..	97
Fig 6.15 Relationship between ground settlement and excavation depth at the pit edge.....	98

LIST OF TABLES

Table 5.1. Prediction accuracy of the models for verification boreholes	91
Table 5.2. Average relative difference between the prediction and verification boreholes	91
Table 5.3. Measurement accuracy of the different predictions and validating boreholes	91
Table 5.4. Global accuracy and boundary accuracy between the different predictions and Profile 1 (b1-b10).....	92
Table 5.5. Measurement accuracy of the different predictions and validating boreholes	92
Table 5.6. Global accuracy and boundary accuracy between the different predictions and Profile 2 (b1-b9).....	92
Table 5.7 Prediction accuracy of different models for validation boreholes	93
Table 6.1 Physical parameters of soil layers	108
Table 6.2 Hoek-Brown model parameters for bedrock	108
Table 6.3 Calculation parameters of retaining structures	108
Table 6.4 Excavation conditions	109

CHAPTER 1 INTRODUCTION

1.1 Research background

Geotechnical engineering construction involves modifying existing rock and soil masses to meet human daily needs. The formation of rock and soil masses is a complex process that has evolved over hundreds of millions of years, shaped by prolonged and intricate geological processes within the lithosphere. Under the influence of geological processes across different times and spaces, the spatial distribution of rock and soil masses in various regions exhibits strong spatial variability and uncertainty. Therefore, detailed geotechnical engineering investigations are crucial before commencing geotechnical engineering projects (Luo et al., 2020; Nguyen T S and Likitlersuang S., 2021). Based on the engineering data from detailed geotechnical investigations, geological profiles of the site can be plotted to understand the stratigraphic distribution, which is essential for the design and construction of geotechnical projects.

Inaccurate stratigraphic distribution revealed by geological profiles can lead to numerous engineering issues. For example, the famous Leaning Tower of Pisa tilts due to the failure to account for the unevenness of the foundation and the excessive softness of the soil during its construction (Xu et al., 2014). A Vietnamese bridge, costing 15 million dollars, collapsed before completion because the investigation failed to detect karst caves in the bearing layer. In the United States, an analysis of 300 engineering contracts revealed that insufficient site investigations leading to unclear stratigraphic distribution are the primary cause of geotechnical change orders (Prezzi M et al., 2011).



Fig. 1-1(a) Leaning Tower of Pisa (b) Vietnamese Bridge

In the UK, a survey indicated that 40% of underground engineering problems stem from the inherent variability and uncertainty of strata (Clayton C R I, 2001). In China, many foundation pit collapse incidents are attributed to inadequate preliminary investigations, such as misjudging the spatial distribution of confined aquifers and weak layers, leading to severe engineering accidents. For instance, the collapse of the Xianghu Station foundation pit on the Hangzhou Metro was primarily caused by insufficient understanding of the site's soft soil. Therefore, obtaining accurate spatial distribution of site strata based on geotechnical investigation data is of great significance for geotechnical engineering design and construction.

Accurate geotechnical investigation data are obtained from boreholes drilled at the site. The soil type and layer thickness revealed by the drill cores provide insights into the stratigraphic conditions at the borehole location. However, due to the high cost of drilling and limited investigation budgets, only a limited number of boreholes can be drilled to roughly estimate the spatial distribution of the site's strata. How to obtain a relatively accurate geological profile from limited boreholes is an urgent problem to solve. The linear interpolation method commonly used in engineering connects the same soil layers between boreholes with straight lines. While this method provides a rough estimate of the site's stratigraphic distribution, it does not account for

the uncertainty and spatial variability of soil layers. Moreover, for cross-bedded or intermittently occurring strata in different boreholes, this method fails to connect them properly.

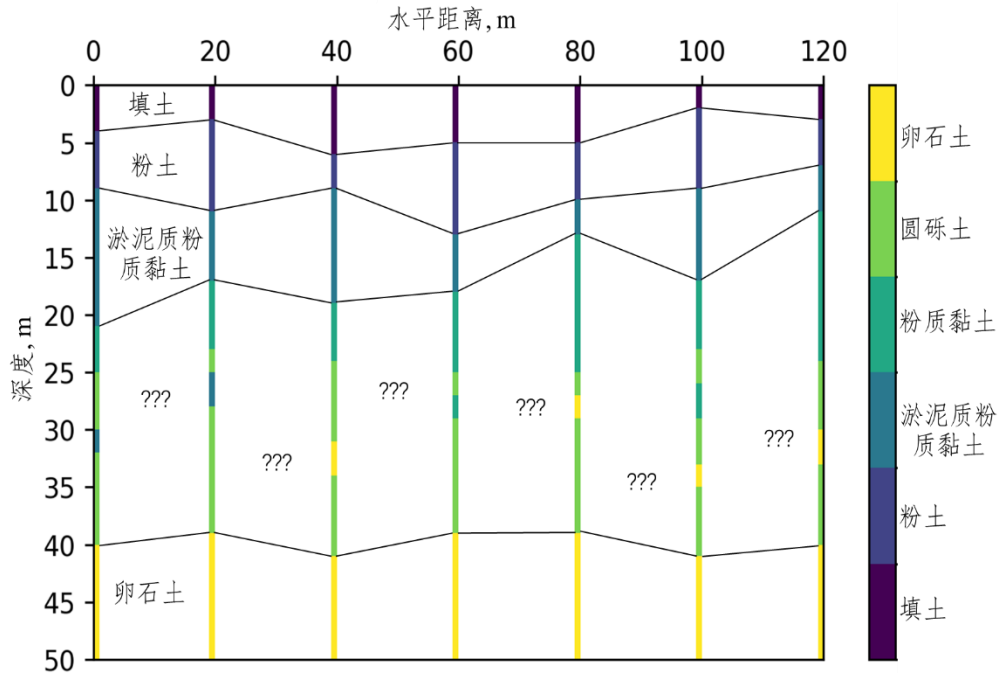


Fig. 1.2 Geological profile of a foundation pit project in Hangzhou

Statistical methods predict geological profiles by constructing random fields but require numerous physical parameters that are difficult to obtain from sparse borehole data. Existing machine learning algorithms, such as IC-XGBoost, assume adjacent geological profiles have similar spatial connectivity and require equally spaced boreholes. These algorithms train models using adjacent geological profiles and predict profiles based on sparse, equally spaced boreholes, resulting in low accuracy for nonlinear boundary predictions.

To address these challenges, this thesis proposes the MSRF-XGBoost algorithm, which consists of a classification module and a screening module. The classification module improves upon the XGBoost algorithm by incorporating multi-scale receptive field sampling (MSRF) code, enabling the model to capture spatial features of different scales between boreholes. This

enhancement allows the model to better learn the spatial features between boreholes, thereby achieving accurate predictions of geological profiles. The screening module is entirely independently developed by the author. It employs Gaussian filtering, boundary similarity comparison, and other steps to automatically select the optimal prediction result from different outcomes generated by various receptive fields. In nonlinear case predictions, the algorithm improves boundary accuracy by approximately 40% compared to other methods. However, this algorithm requires boreholes to be equally spaced. To address this limitation, an improved GCN algorithm is proposed. Leveraging the characteristics of the GCN algorithm, the problem of predicting geological profiles based on boreholes is transformed into predicting the soil type at unknown points based on their coordinates and the soil types of adjacent points. This approach enables the prediction of geological profiles using non-equally spaced sparse boreholes. The feature representation, adjacency matrix design, and training iterations of the GCN algorithm are optimized to achieve better prediction results. In nonlinear case predictions, the improved GCN algorithm enhances boundary accuracy by approximately 30% compared to other methods. The two proposed machine learning methods are applied to predict stratigraphic interfaces in multiple practical engineering projects, providing valuable assistance for engineering practice.

Assuming that accurate geological profiles predicted by machine learning are available, this thesis introduces the multi-point layered summation method, which better accounts for the complex geological conditions beneath foundations, yielding more accurate maximum foundation settlements. Furthermore, the machine learning-predicted geological profiles are applied to finite element simulations. Compared to finite element models based on linearly

interpolated geological profiles, the nonlinear geological profile models provide more accurate simulations of ground surface settlements outside foundation pits, demonstrating the importance of accurate geological profiles for finite element simulations.

In summary, this thesis employs machine learning methods to predict relatively accurate geological profiles based on sparse boreholes and applies these profiles to practical engineering, theoretical calculations, and finite element simulations. The proposed methods are validated across multiple dimensions, providing a basis for geotechnical engineering design and construction. The research holds significant scientific and engineering value for mitigating geotechnical hazards such as foundation pit collapses and tunnel settlements caused by unclear stratigraphic distribution, thereby safeguarding lives and property.

1.2 Research objectives

The main issue in current geological profile prediction research is that existing algorithms cannot accurately predict geological profiles using only sparse boreholes, especially for nonlinear boundaries between soil layers. In line with these focuses, the following research objectives are proposed:

- (1) To develop a ML algorithm to predict accurate geological profile by sparse boreholes.
- (2) To develop a deep learning algorithm to predict accurate geological profile by unequal-distance boreholes.
- (3) To prove the ability of proposed methods by using in real geological engineering and

comparing with other existing methods.

(4) To develop a finite element model by predicted geological profile to prove the application of proposed methods.

1.3 Structure of the thesis

This study investigates geological profile prediction based on limited borehole data. Multiple machine learning algorithms are proposed to predict geological profiles from limited borehole data. These algorithms are applied to practical engineering projects, theoretical settlement calculations, and finite element simulations of foundation pit excavations, demonstrating the importance of accurate geological profile prediction for engineering practice, theoretical calculations, and numerical modeling. The research content is divided into seven parts:

Chapter 1 serves as the introduction, providing the research background, research objectives, research methodologies, and the thesis structure.

Chapter 2 reviews existing research in studies on geological profile predicting methods, which including interpolation methods, statistical methods, and machine learning methods. Furthermore, numerical simulation method applicated in foundation pits introduced.

Chapter 3 proposes the MSRF-XGBoost algorithm, which consisting of a classification module and a screening module. The classification module enhances the traditional XGBoost algorithm with multi-scale receptive field sampling to learn spatial features of different scales between sparse, equally spaced boreholes, enabling accurate geological profile prediction. The screening module uses Gaussian filtering and boundary similarity comparison to automatically select the optimal prediction from results generated by different receptive fields. The algorithm achieves

82% boundary accuracy for nonlinear geological profiles, compared to 36% for existing machine learning algorithms.

Chapter 4 proposes novel GCN algorithm to address the requirement for equally spaced boreholes in other machine learning algorithms. The feature representation, adjacency matrix design, and training iterations of the GCN algorithm are optimized to predict soil types at unknown points based on their coordinates and adjacent points' soil types, enabling accurate geological profile prediction using non-equally spaced boreholes. The improved GCN algorithm improves boundary accuracy by 30% compared to existing machine learning algorithms.

Chapter 5 applied the proposed machine learning algorithms to predict geological profiles for a publicly available borehole dataset in the Netherlands and a foundation pit project in Hangzhou. The Netherlands dataset, with deep boreholes and simple geological structures, achieves over 90% accuracy. The Hangzhou project, with complex stratigraphy, achieves over 80% accuracy. The results demonstrate the algorithms' practical utility for engineering projects.

Chapter 6 applied the machine learning-predicted geological profiles in finite element modeling. Finite element models based on linearly interpolated profiles and machine learning-predicted nonlinear profiles are compared with actual settlement monitoring data. The nonlinear profile models yield closer matches to monitoring data, highlighting the importance of accurate geological profiles for finite element simulations.

Chapter 7 summarizes the findings in this thesis and presents some recommendations for future research in the related area.

CHAPTER 2 LITERATURE REVIEW

2.1 Introduction

The problem of predicting geological profiles from sparse boreholes has been studied by numerous domestic and international scholars, who have proposed various methods to address it. These methods can be broadly categorized into interpolation methods, statistical methods, and machine learning methods. The numerical simulation method widely applied in foundation pits.

2.2 Research on geological profile prediction

2.2.1 Interpolation methods

The application of conventional interpolation methods for establishing stratigraphic profiles between boreholes has been documented in geotechnical engineering practice. Dasaka et al. (2012) utilized linear interpolation to connect identical soil layers between adjacent boreholes with straight-line segments, deriving preliminary stratigraphic distribution models for sites. Building upon this, Aswar et al. (2017) implemented spline interpolation techniques to divide inter-borehole distances into smaller segments, fitting continuous curves for each segment and connecting them to generate more refined representations of subsurface stratigraphy. These deterministic interpolation approaches have provided fundamental methodologies for preliminary site characterization in geotechnical investigations.

2.2.2 Statistical methods

The application of advanced statistical and probabilistic methods for characterizing stratigraphic uncertainty using limited borehole data has formed a significant research trend in geotechnical

engineering. Statistical methods can still be subdivided into multiple different methods, such as multiple-point statistics which discretizes boreholes into points and predicts unknown soil types through distance-weighted statistical analysis of surrounding points (Shi et al., 2021);

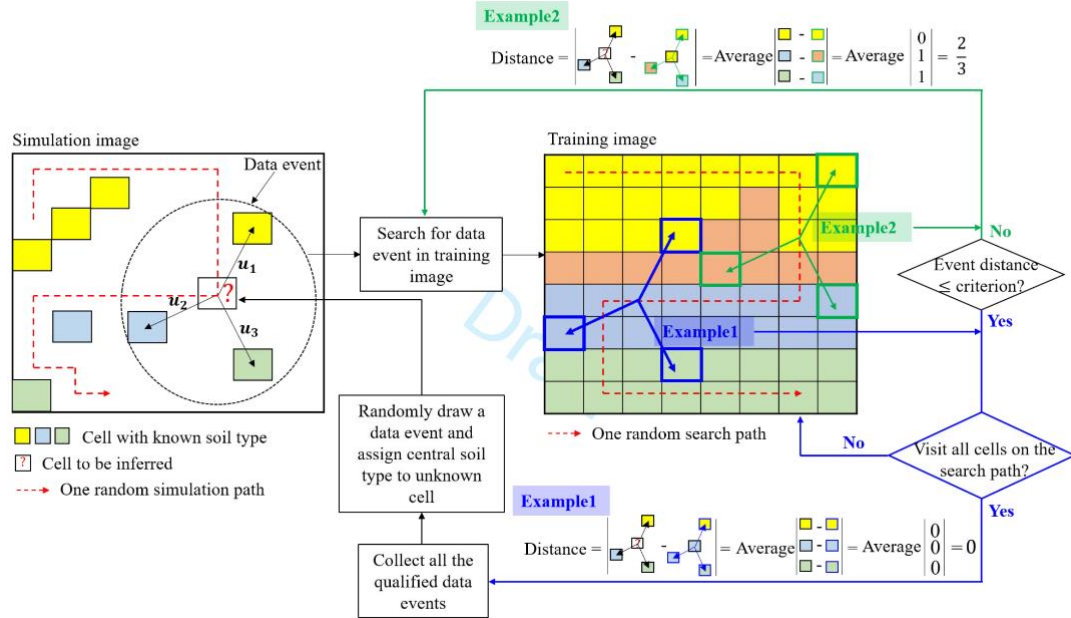


Fig. 2.1 Multiple-point statistics method

Conditional random fields that couple stratigraphic distribution and property uncertainties using autocorrelation functions based on maximum likelihood principles (Gong et al., 2021; Zhao et al., 2021); coupled Markov chain models improved through horizontal transition probability matrices for reducing geotechnical uncertainties (Qi et al., 2022; Li et al., 2023; Deng et al., 2017); spatial variability zoning techniques that distinguish regions with different variability characteristics using in-situ test data (Qi et al., 2019); hidden Markov random fields for representing stratigraphic variability in 3D spatial segmentation (Wang et al., 2020; Wang et al., 2018); integrated Bayesian frameworks combining Markov random fields with classifiers for stochastic stratigraphic simulation (Wei et al., 2022; Jiang et al., 2020); Bayesian compressed

sensing/sampling methods for reconstructing signals from sparse measurements with explicit uncertainty quantification (Wang et al., 2017; Zhao et al., 2018; Hu et al., 2020; Huang et al., 2014; Lyu et al., 2023); cross-correlated random field generators leveraging autocorrelation and cross-correlation structures from sparse boreholes (Guan et al., 2023; Li et al., 2023); and multivariate adaptive regression splines capable of automatically modeling nonlinearities for spatially predicting geological interfaces (Qi et al., 2021; Qi et al., 2022). These advanced techniques have been successfully employed to address critical geotechnical challenges including stratigraphic modeling under data sparsity, slope stability and reliability analysis under uncertainty, spatial interpolation of soil properties, and characterization of complex geological interfaces across various site conditions.

2.2.3 Machine learning methods

In recent years, the geotechnical field has witnessed remarkable progress in developing machine learning methodologies for stratigraphic modeling and uncertainty quantification. These approaches effectively integrate limited site investigation data with advanced computational algorithms to address complex geological challenges. The developed methods can be broadly categorized into several innovative frameworks.

First, supervised learning and ensemble methods Shi et al. (2021a) pioneered the IC-XGBoost algorithm that leverages prior geological experience and limited borehole data, training models with adjacent geological profiles and iteratively predicting soil types at unknown locations. This approach was successfully validated through application to Australian geological profiles.

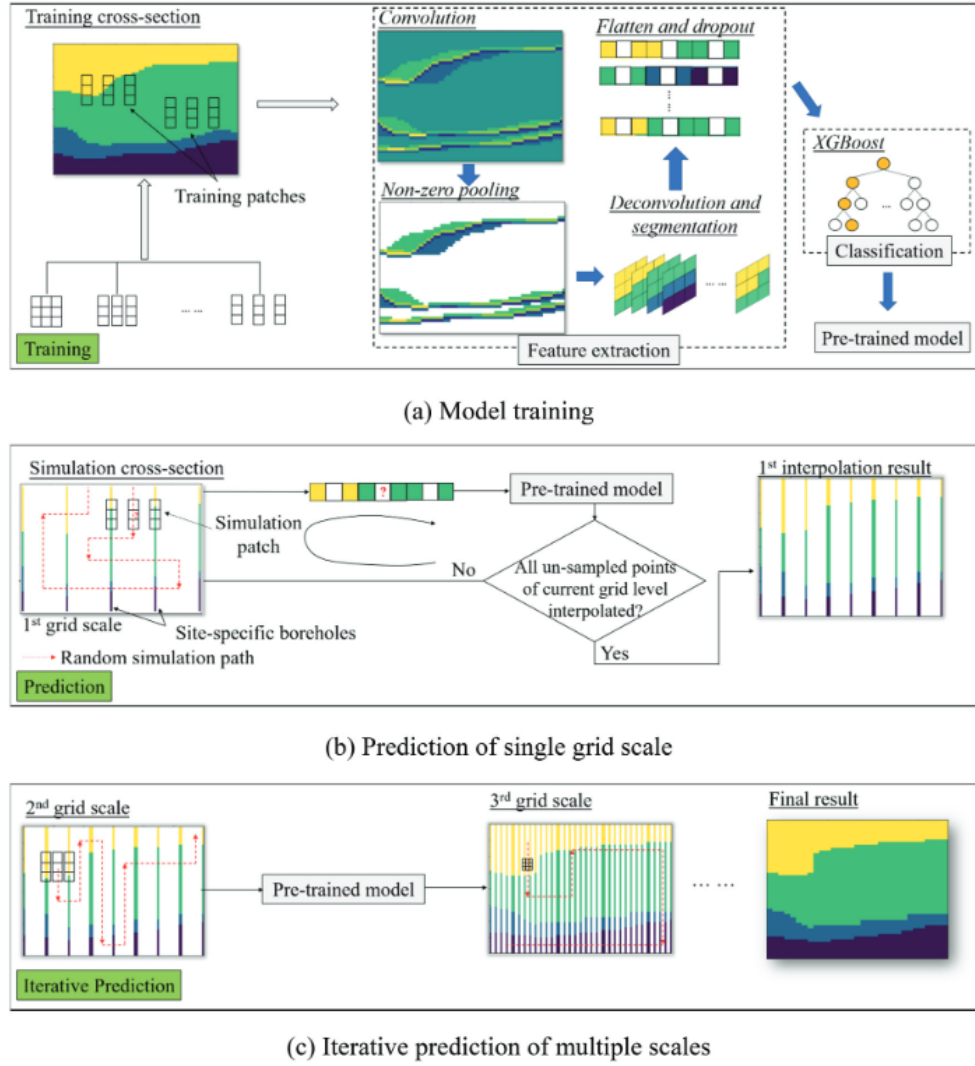


Fig. 2.2 IC-XGBoost algorithm

The methodology was further refined by Shi et al. (2021b) through a data-driven approach for optimal training image selection using edge direction detection, significantly enhancing model training accuracy. The framework evolved into IC-XGBoost3D (Shi et al., 2022a) for constructing three-dimensional geological models from sparse borehole data, demonstrating particular effectiveness in modeling Hong Kong's complex sedimentary strata.

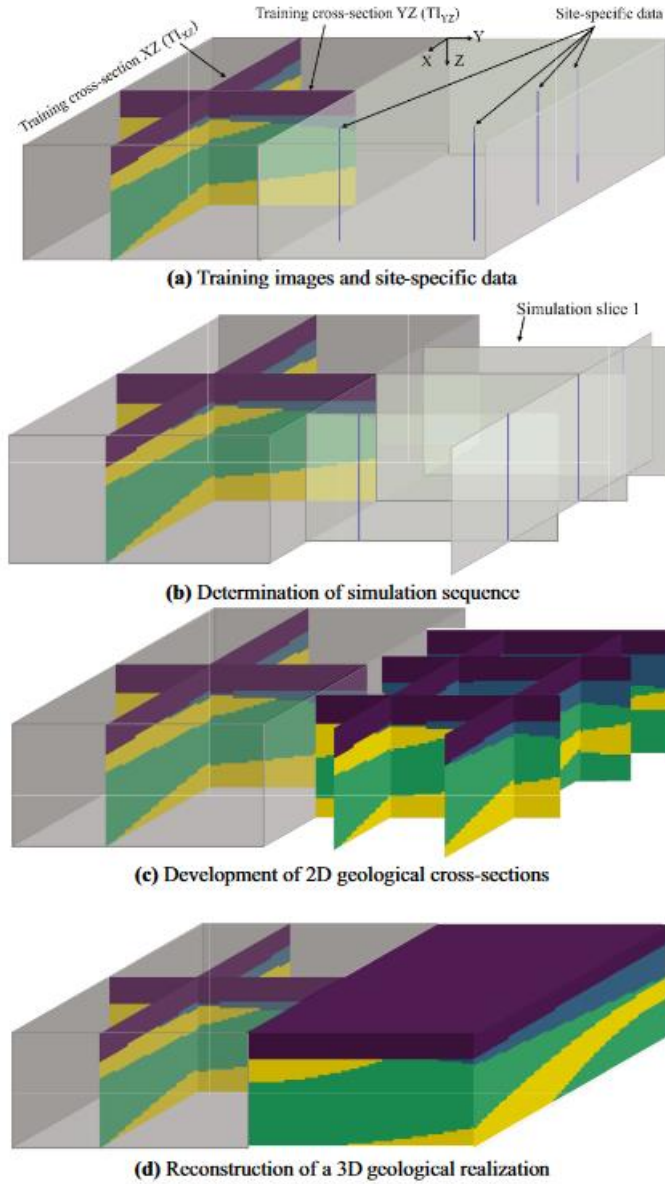


Fig. 2.3 IC-XGBoost3D algorithm

Building upon this foundation, Shi et al. (2022b) established a stochastic framework integrating machine learning and random field simulation for consolidation settlement assessment in reclamation areas.

Second, regression and interpolation techniques Qi and colleagues made significant contributions through Multivariate Adaptive Regression Splines (MARS) implementation. Qi et al. (2021) developed MARS for automatically modeling nonlinearities in predicting geological

interface elevations, achieving remarkable performance with 95% confidence intervals covering 95% of test data in Singapore case studies. Subsequent enhancement through zoning methods (Qi et al., 2022) demonstrated superior capability in revealing spatial trends of geological interfaces compared to conditional random fields and thin-plate spline interpolation.

Integrated Bayesian Frameworks and Deep Learning Approaches. Tian et al. (2023) introduced a groundbreaking data- and physics-driven Bayesian learning framework that automatically constructs stratigraphic models from sparse investigation data and integrates spatiotemporally varying settlement monitoring data.

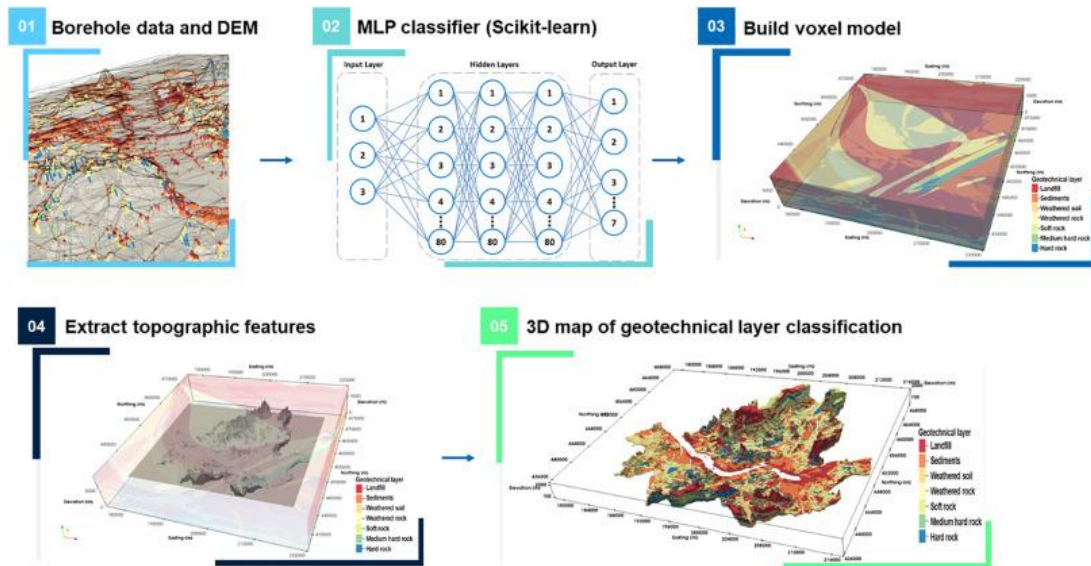


Fig. 2.4 DNN model application

Kim et al. (2022) developed an optimal Deep Neural Network-based classifier for reliable rock and soil layer classification in Seoul's underground space, employing comprehensive data preprocessing and hyperparameter optimization. The research was extended by Liu et al. (2022) through a comprehensive framework quantifying stratigraphic boundary uncertainty and its impact on slope stability analysis.

Last one, hybrid modeling and advanced computational methods. Leu et al. (2011) established an innovative probabilistic general model combining hidden Markov models with neural networks for tunnel stratigraphy determination. Wang et al. (2023) introduced a data-driven analysis framework considering stratigraphic uncertainty for consolidation settlement assessment in ground improvement projects. Xu et al. (2022) proposed multi-source Bayesian compressed sampling for statistically characterizing 2D geological profiles by integrating geotechnical and geophysical data. Yang et al. (2022) developed a 3D geological model auto-reconstruction method based on Deep Convolutional Generative Adversarial Networks (DCGAN), automatically generating 3D geological models from 2D profiles. Xie et al. (2022) presented a general framework for geotechnical geological modeling incorporating spatial autocorrelation using locally mapped machine learning methods.

These advanced methodologies demonstrate significant capabilities in addressing critical geotechnical challenges, including stratigraphic modeling under data sparsity conditions, slope stability analysis under uncertainty, spatial interpolation of soil properties, and characterization of complex geological interfaces across diverse site conditions. The integration of machine learning with traditional geotechnical principles has opened new frontiers in reliable underground characterization, providing robust tools for engineering decision-making while explicitly quantifying associated uncertainties. Future research directions likely involve enhanced multi-source data fusion, improved uncertainty quantification in complex geological settings, and real-time model updating capabilities during construction phases.

2.3 Research on numerical simulation of foundation excavation

The geotechnical field has witnessed significant advancements in the application of numerical simulation for foundation pit analysis. These approaches effectively integrate site investigation data with sophisticated computational mechanics to address the challenge of foundation pit deformation. The developed methodologies can be broadly categorized into several key technical frameworks.

First, fluid-solid coupling and 3D finite element analysis. Zeng et al. (2021) established a fluid-solid coupling numerical model based on actual dewatering tests to reveal the influence of groundwater level decline on maximum ground settlement, with extensive simulations elucidating its evolution under various conditions. Luo et al. (2008) designed an optimized dewatering scheme, simulated using 3D finite element analysis and validating its efficacy in the Shanghai Dongjiadu Tunnel project. Similarly, Luo et al. (2006) conducted a 3D numerical simulation of unstable seepage fields for the Shanghai World Financial Center pit, calibrating model parameters via pumping tests to analyze the optimized dewatering scheme. Cheng et al. (2011) conducted 3D fluid-solid coupling simulations for a tunnel open-cut pit, analyzing seepage field characteristics with cutoff walls and identifying soil deformation modulus as the most sensitive factor affecting settlement through orthogonal experiments.

Second, simulation based on Biot's consolidation theory. Feng et al. (2012) studied the time-dependent effects of dewatering and excavation in deep foundation pits under complex geological conditions using Biot's consolidation theory and the modified Cam-clay model, with finite element calculations validating the fluid-solid coupling model against field measurements.

Ni et al. (2008) simulated seepage coupling during the excavation of a hydropower station foundation pit based on Biot's consolidation theory and the Mohr-Coulomb criterion, highlighting the critical importance of anti-seepage measures.

Third, commercial software application and finite difference methods. Yuan et al. (2021) utilized MIDAS/GTS software to simulate an ultra-deep, water-rich metro station pit, revealing the evolution of seepage paths and their impact on pit deformation during construction. Zhou et al. (2010) simulated foundation pit dewatering using 3D finite difference methods based on field tests and inverse analysis, demonstrating that increasing diaphragm wall depth effectively controlled external groundwater level decline and settlement.

These advanced numerical methodologies demonstrate robust capabilities in addressing critical geotechnical challenges, including predicting ground settlement induced by dewatering, optimizing dewatering and seepage control schemes, analyzing complex fluid-solid interaction mechanisms, and validating design parameters against field monitoring data. The integration of numerical simulation with traditional geotechnical principles has provided powerful tools for engineering design and risk mitigation, offering insights into system behavior that are difficult to obtain through analytical methods alone. Future research directions likely involve enhanced coupling of multi-physical processes, improved model calibration through real-time data assimilation, and the application of stochastic methods to quantify uncertainty in predictions.

2.4 Summary

This section reviews prevailing methodologies for predicting geological profiles from borehole data, highlighting their respective strengths and limitations. Conventional approaches, including

interpolation and statistical methods, alongside emerging machine learning techniques, are examined to identify persistent challenges in accurate stratigraphic modeling.

The primary limitations of existing methods can be summarized as follows: First, traditional interpolation techniques, while simple and widely practiced, rely on deterministic connections between boreholes and fundamentally fail to account for inherent stratum variability and uncertainty. Second, statistical methods, such as those based on random field theory, often require numerous physical parameters that are difficult to constrain with limited site investigation data and may overlook complex, hidden spatial information within the dataset. Third, although machine learning (ML) algorithms show promise in capturing non-linear relationships and reducing subjective assumptions, several issues remain. Certain ML frameworks, like IC-XGBoost, depend on adjacent geological profiles as prior knowledge rather than operating on the target borehole data alone. Other hybrid models that integrate ML with random fields introduce numerous parameters that are challenging to determine robustly, potentially compromising their practical utility. Finally, a significant gap persists between stratigraphic prediction and numerical modeling in engineering practice. Most geotechnical numerical models continue to incorporate strata by simplistically connecting boreholes with straight lines, thereby neglecting spatial uncertainty and variability, which undermines the reliability of subsequent simulations for deformation and settlement analysis.

CHAPTER 3 MULTI-SPATIAL RECEPTIVE FIELD (MSRF) XGBOOST METHOD

3.1 Introduction

Accurately determining the spatial distribution of soil layers from borehole data is a fundamental prerequisite for the design and analysis of geotechnical engineering projects, such as slope stability assessment, foundation design, and excavation support. However, the high cost of drilling typically results in spatially sparse and discontinuous borehole data, making the precise interpretation of subsurface stratigraphy a significant challenge.

Existing methodologies for stratum estimation can be broadly categorized into interpolation, statistical, and machine learning (ML) approaches. While widely used, conventional interpolation and statistical methods often struggle to capture the complex, nonlinear nature of soil layers. They may rely on simplistic assumptions, overlook hidden spatial information in the data, or require numerous parameters that are difficult to identify from limited investigations.

In recent years, ML methods have demonstrated considerable promise in overcoming these limitations due to their superior capacity to learn complex, nonlinear relationships directly from data. Among them, the eXtreme Gradient Boosting (XGBoost) algorithm has shown notable performance in stratigraphic prediction. Nevertheless, a critical gap persists in their practical application. A prevalent but often unstated assumption in many ML-based models, such as the Iterative Convolution XGBoost (IC-XGBoost), is the requirement for regularly spaced training data. Specifically, the IC-XGBoost

method necessitates a complete geological cross-section from an adjacent site as prior knowledge for model training, rather than operating on the target borehole data alone. This reliance on densely sampled, equidistant data severely limits its applicability to real-world projects where boreholes are typically sparse and irregularly spaced. Furthermore, the fixed receptive field of conventional models constrains their ability to represent the complex spatial features of strata.

To address these gaps, this study introduces a novel Multi-Spatial Receptive Field XGBoost (MSRF-XGBoost) method. The primary objective of this method is to achieve accurate prediction of complete geological cross-sections using only the sparse borehole data from the target site, thereby eliminating the dependency on adjacent profiles. The core innovation lies in employing multiple spatial receptive fields (SRFs) to better capture the nonlinear spatial characteristics of soil layers from the regular boreholes. This thesis is structured as follows: the methodology of the MSRF-XGBoost algorithm is first introduced, followed by its validation in both linear and nonlinear stratigraphic scenarios. Finally, the influence of key factors on the model's performance is discussed.

3.2 MSRF-XGBoost methodology

This chapter concisely introduces how the new MSRF XGBoost method predicts the distribution of soil layers. Fig. 3.1 presents the flowchart of the MSRF XGBoost model. At first, the training and testing datasets are constructed by the borehole data after data preprocessing. The training samples consist of data from three neighboring boreholes, with the data from the middle borehole serving as the output and the data from the other

two boreholes serving as the input. On the other hand, the input of the testing samples comprises data from two adjacent boreholes, while the output is the unknown data that needs to be predicted between the boreholes. The training dataset is used to train XGBoost classification models with various spatial receptive fields in the classification module. Subsequently, how the identification module utilizes the trained classification models to predict the unknown soil layers using the testing dataset is described in detail. A series of predictions are obtained from different classification models and then filtered using Gaussian filters to get candidates. The standard of selecting candidates from different filter results is the difference between the predictions and filter results. The boundary similarity algorithm is proposed to ensure all the candidates have similar soil layer boundaries. Finally, the candidate with the smallest error is recommended as the optimal prediction of soil layers.

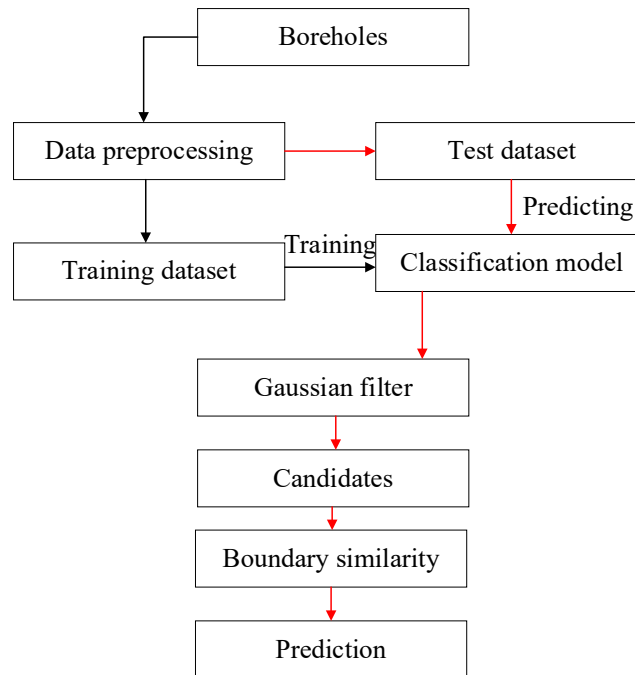


Figure 3.1 Flow chart of the proposed MSRF XGBoost method

The MSRF-XGBoost approach differs from the standard XGBoost approach in terms of training data, training strategy, and identification module. In the standard XGBoost approach, a training image containing the complete geological information from an adjacent area is used as the training data. On the other hand, the MSRF-XGBoost approach utilizes only the borehole data as the training data. Regarding the training strategy, MSRF-XGBoost employs multiple spatial receptive fields (SRFs) for sampling and trains multiple SRF models to achieve optimal geology prediction. In contrast, the standard XGBoost approach uses a fixed 3x3 sampling and training patch. Additionally, MSRF-XGBoost incorporates an identification module that utilizes Gaussian filter and Boundary similarity algorithms to determine the optimal predictions based on the results obtained from multiple SRF models.

3.2.1 XGBoost method

The XGBoost method is a typical decision tree model that combines many base decision trees to improve prediction accuracy. Compared to other boosting tree models (e.g., the gradient boosting decision tree (Ke et al. 2017) and adaptive boosting (Margineantu and Dietterich 1997)), XGBoost employs the second-order gradient approximation and regularization term to accelerate optimization and prevent overfitting, respectively. A boosting tree is a sequential ensemble technique for regression and classification problems. This model can combine new decision trees to enhance modeling capability until the desired accuracy or the maximum number of trees is attained. XGBoost models in Fig. 3.4 illustrates a schematic of a typical boosting tree. In this model, the second decision tree (f_2) is trained to minimize the residual difference between the

ground truth (y) and the prediction of the first decision tree (f_1). The residual error can be reduced by integrating new trees. According to Chen and Guestrin (2016), the final prediction is expressed by Eq. (3.1):

$$\hat{y} = \sum_{k=1}^K f_k(x_i), \quad x_i \in \mathbb{R}^m, \quad f_k \in F \quad (3.1)$$

where $\hat{y}$ is the prediction from the boosting tree model, K indicates the total number of trees, x_i represents the i th training vector with m features, $\mathbb{R}^m$ denotes the m -dimensional real space, and F stands for the space of the decision trees.

Each soil type in multilayer soils can be predicted using Eq. (3.1). Then, the final probability (p_c) is determined by transforming $\hat{y}_c$ using the softmax function (Hastie et al. 2009), as expressed in Eq. (3.2):

$$p_c = \frac{e^{\hat{y}_c}}{\sum_{c=c_1}^{c_N} e^{\hat{y}_c}}, \quad c = c_1, c_2, \dots, c_N \quad (3.2)$$

where $\hat{y}_c$ is the prediction of category c from the boosting tree model, and c_N indicates the total number of categories.

The XGBoost method uses a regularization term in the loss function to penalize the complex tree against overfitting, as provided in Eq. (3.3) and Eq. (3.4):

$$L \equiv \sum_{i=1}^n l(\hat{y}_i, y_i) + \sum_{k=1}^K \Omega(T_k, \omega_k) \quad (3.3)$$

$$\Omega(T_k, \omega_k) = \gamma T_k + \frac{1}{2} \lambda \|\omega_k\|^2 \quad (3.4)$$

where $\hat{y}_i$ is the i th training target, y_i indicates the i th prediction, $l(\hat{y}_i, y_i)$ denotes the total training loss between the prediction and training targets, $\Omega(T_k, \omega_k)$ represents the regularization term to punish the hyperparameters of the k th complex decision tree, T_k

stands for the number of leaves, ω_k indicates the weight of leaves, λ is a coefficient with a default value of 1.0, and γ is the tuning parameter. The cross-entropy loss (Hastie et al. 2009) is applied in the training process.

3.2.2 Definition of MSRF

Fig. 3.2 shows how the MSRF extracts the spatial feature of the input on a local soil boundary. In the representation, the rows correspond to horizontal cells, while the columns represent vertical cells. Soil 1 is depicted by yellow cells, and soil 2 is represented by green cells. The gray masks indicate the extracted cells corresponding to a SRF within the boundary samples. For a given SRF of three, two SRF3 samples may extract the same input but different outputs; in that case, the soil boundary presents a significant stratigraphic change. These types of samples disrupt the one-to-one mapping relationship between the input and output, which can undermine the reliability of the data-driven model. This issue resources from a small SRF that fails to consider sufficient spatial features for a significant stratigraphic change. Increasing the SRF can remarkably improve the quality of samples, as seen in Fig. 3.3 By increasing SRF to five, the two SRF5 samples take different inputs corresponding to different outputs, fulfilling the one-to-one mapping rules. This result indicates that a large SRF can extract more spatial features to predict significant stratigraphic changes. However, stratigraphic changes in the ground can involve various spatial features, which may be small for deposited soils and large for weathered soils. Therefore, the MSRF is required to accurately predict geological cross-sections.

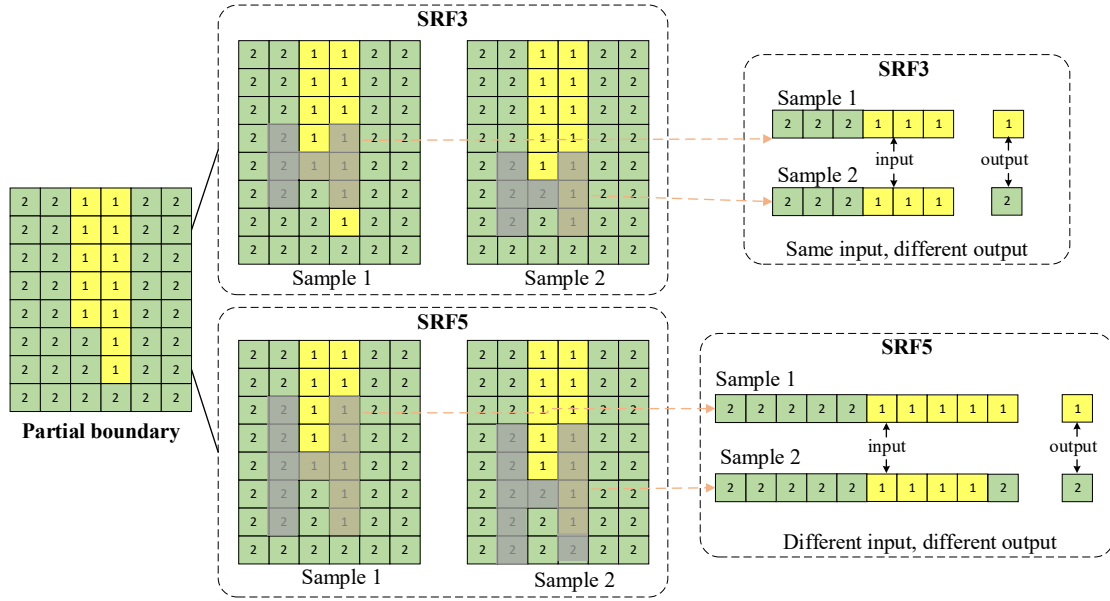


Fig. 3.2 Effect of different spatial receptive fields on sample quality

Fig. 3.3 depicts the results of the MSRF in extracting the nonlinear geological cross-section, commonly corresponding to a significant stratigraphic change. An artificial soil boundary with strong nonlinearity is assigned using a polynomial function to analyze the capability of the MSRF to extract nonlinear boundary features as follows:

$$v = A \times h^6 + B \times h^5 + C \times h^4 + D \times h^3 + E \times h^2 + F \times h + G \quad (3.5)$$

where h and v are the horizontal distance and the vertical depth, respectively, and A, B, C, D, E, F , and G indicate the coefficients of the polynomial function. In this case, $B = 5.225 \times 10^{-9}$, $C = -1.101 \times 10^{-6}$, $D = 5.805 \times 10^{-5}$, $E = 3.767 \times 10^{-3}$, $F = -0.061$, and $G = 13.91$.

Fig. 3.3(a) illustrates the original artificial cross-section, where the input boreholes are represented by four white lines. According to Fig. 3.3(b) – 3.3(i), the boundary accuracy of the different SRF models ranges from 0.47 to 0.70, with a maximum of 0.70 for model SRF-13 and a minimum of 0.47 for model SRF-9. The

difference in the boundary accuracies of the various SRF models demonstrates that the SRFs considerably impact the prediction of the nonlinear soil boundary in the XGBoost models. The boundary accuracy is defined as the different cells between the prediction and the original cross-section divided by all the boundary cells of original cross-section, as described in Section 2.4.

Once the classification module yields multiple predicted results by MSRF XGBoost models (e.g., SRF-3, SRF-5, SRF-7), the optimal predictive result could be recommended by the identification module.

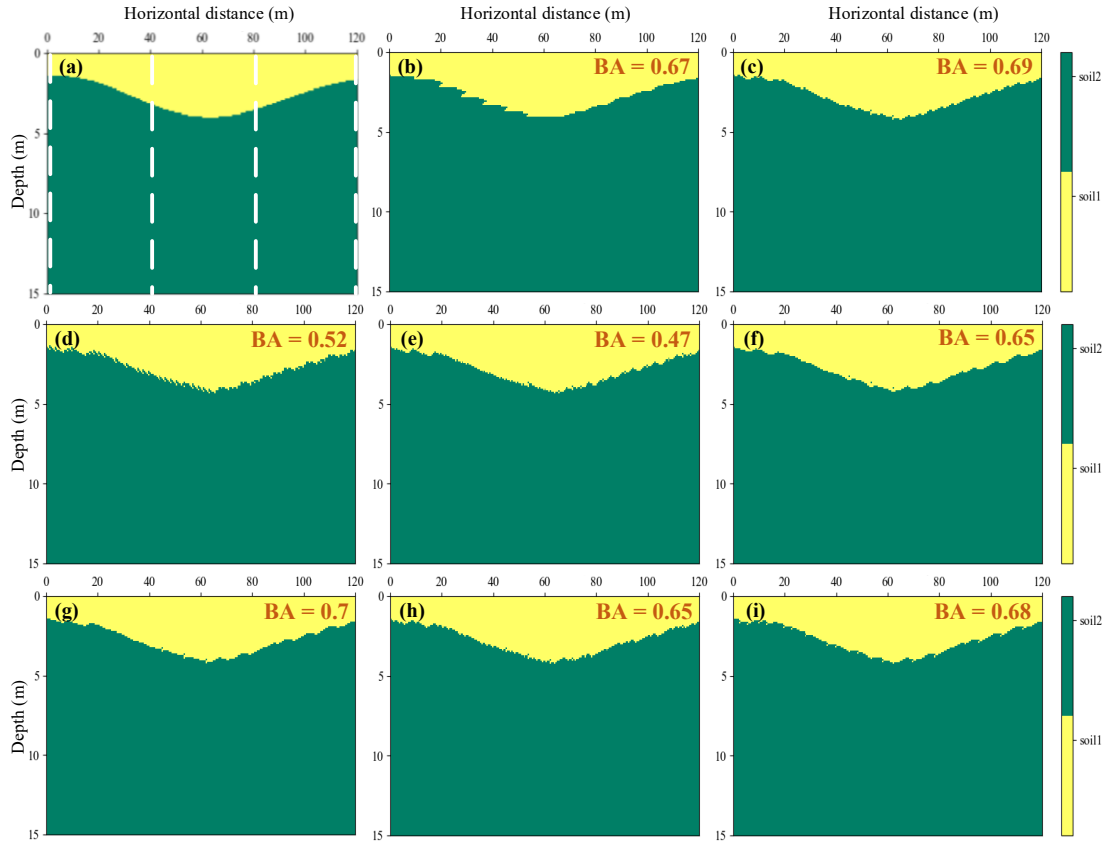


Fig. 3.3 Predictions of nonlinear soil boundary by different SRF models: (a) ground truth; (b) model SRF-3; (c) model SRF-5; (d) model SRF-7; (e) model SRF-9; (f) model SRF-11; (g) model SRF-13; (h) model SRF-15; (i) model SRF-17

3.2.3 Classification module of MSRF-XGBoost method

The classification module commences with collecting borehole data from site investigation and generating a raw training database using the multi-spatial receptive field (MSRF) sampling method, as depicted in Fig. 3.4. In comparison with the fixed 3×3 training patches in existing XGBoost method, multiple sizes of training patches are utilized to extract the nonlinear features of soil layers in terms of multi-spatial receptive field, e.g., 5×3 and 7×3 . The MSRF sampling method consists of two stages: borehole discretion and sliding sampling. In the borehole discretion phase, the borehole data are split into numerous uniform cells with soil types assigned as labels. In sliding sampling, several cells from every three adjacent boreholes are selected along their depth to create raw samples. The number of cells depends on the size of the SRFs. For example, when the size of the SRF equals 3, 5, and 7, the raw samples cover three, five, and seven cells along the depth corresponding to the dimensions 3×3 , 5×3 , and 7×3 , respectively. After sampling all the boreholes, Laplacian operators are utilized to select boundary samples from the raw samples. The standard Laplacian operator is a derivative operator

which is usually applied to find edges in an image, defining as $\begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix}$. The

standard Laplacian operator can be only suitable for samples with size of 3. In this chapter, the modified Laplacian operator has been utilized to extract the soil boundary corresponding to different SRF samples. For example, the extended Laplacian operator

for SRF3 is $\begin{bmatrix} 1 & 0 & 1 \\ 1 & -6 & 1 \\ 1 & 0 & 1 \end{bmatrix}$. For a given raw sample, SRF5, in Fig. 3.4, the Laplacian

operator yields a nonzero value because sample SRF5 contains more than one soil type;

otherwise, the Laplacian operator yields zeros. In data preprocessing, only raw samples with a nonzero value of the Laplacian operator are divided into boundary samples. In the training strategy, only the boundary samples are required to train the model, avoiding the disturbance of a large number of non-boundary samples. These non-boundary samples from one soil layer can be taken as one sample because these samples have the same input and output. If all these non-boundary samples are utilized to train the model, it will cause sample imbalance problems increasing the training cost. The boundary samples are further processed into the training samples, where cells from boreholes on both sides are concatenated and flattened as the input and the middle cell from the middle borehole as the output. The training data set comprises samples from different SRFs, namely SRF3, SRF5, and SRF7. The MSRF XGBoost classification models trained by training samples SRF3, SRF5, and SRF7 can be named models SRF-3, SRF-5, and SRF-7, respectively.

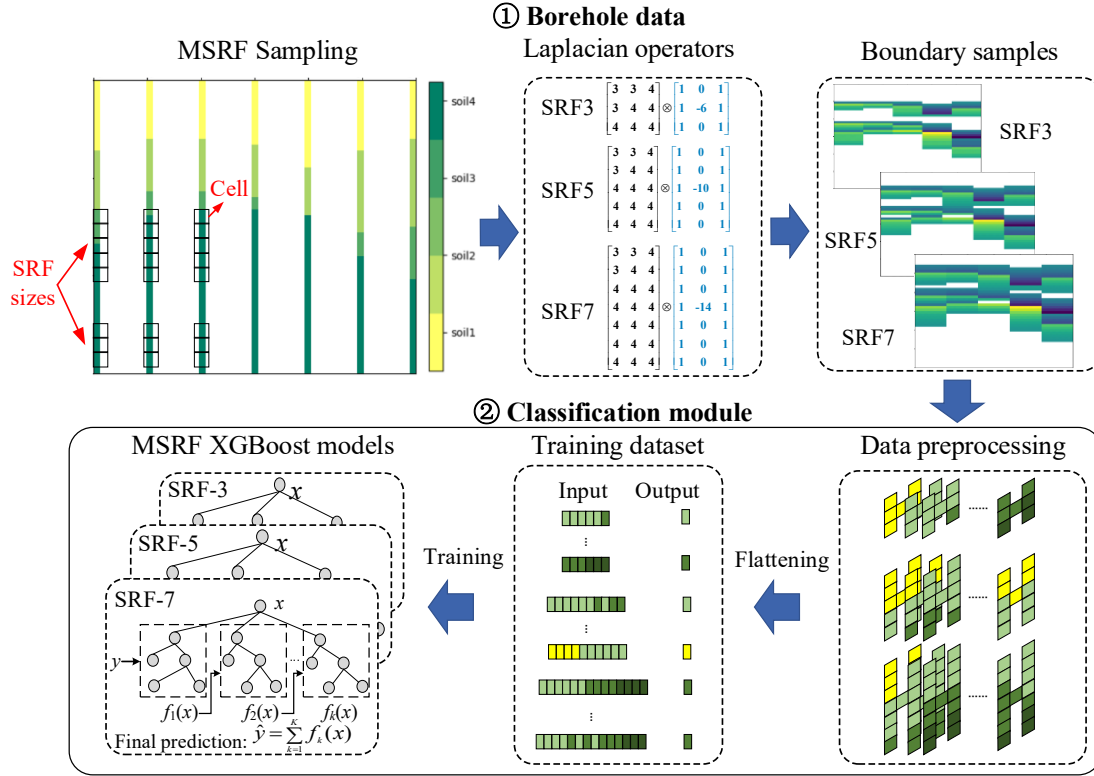


Fig. 3.4. Illustration on classification module of MSRF XGBoost method

3.2.4 Identification module of MSRF-XGBoost method

The identification module is a crucial component of the MSRF XGBoost model. Fig. 3.5 demonstrates how this module recommends the best prediction from the MSRF classification results. Much like the classification module, the adjacent boreholes are initially sampled slidingly from top to bottom. The prediction process is iteratively conducted by the identification module. After obtaining a trained model in the classification module, the identification module concatenates cells from two neighboring boreholes (referred to as father boreholes) as the input. This input is used to predict the soil type of unknown cells in the middle column (referred to as the child borehole) located between the two known boreholes. Subsequently, the father boreholes and the predicted child boreholes are combined as the new input for the next iteration,

aiming to predict the soil types of the unknown cells in the subsequent middle columns.

The iteration continues until all cells have been predicted, ensuring that every cell in the target area is assigned a soil type based on the available information from the neighboring boreholes. Instead of using the Laplacian operator, a judgment criterion is proposed to generate the predicting data set. If the input feature of a predicting sample only contains one soil type, the target soil type defaults to the same. Otherwise, the predicting sample is imported into the trained classification model for prediction. As shown in Fig. 5, the predicting data set includes SRF3, SRF5, and SRF7 predicting samples, which are input into models SRF-3, SRF-5, and SRF-7, respectively, for soil prediction. Next step is to select the best one from different SRF results. The Gaussian filters algorithm and the boundary similarity algorithm are applied to do the selection.

At first, gaussian filters denoise the predictions of the classification models. The effect of denoising is associated with the size of the Gaussian kernel: the larger the kernel, the more the noises are removed. Each prediction usually requires a specific filter depending on the various noise levels. Based on the difference between the prediction results before and after filtering, candidates of the optimal prediction can be selected. Subsequently, all the filtered predictions undergo a boundary similarity comparison with one another, facilitating the removal of any abnormal predictions. The boundary similarity algorithm aims to evaluate the similarity among predictions generated by different SRF models (such as SRF-3, SRF-5, SRF-7). This evaluation is achieved by measuring the perceptual difference based on structural similarity index measure (SSIM) (Wang et al., 2004), which is widely used to measure the similarity between two images.

Through the boundary similarity check, the algorithm identifies the most prevalent predictions while disregarding any abnormal or outlier predictions. This process ensures that the optimal prediction, with a preferred probability level, is selected by retaining predictions that exhibit high similarity along the boundaries. Then, the optimal result is the prediction with the slightest error. The identification module can accurately predict a complex geological cross-section based on the classification module, the Gaussian filtering method, and the boundary similarity comparison.

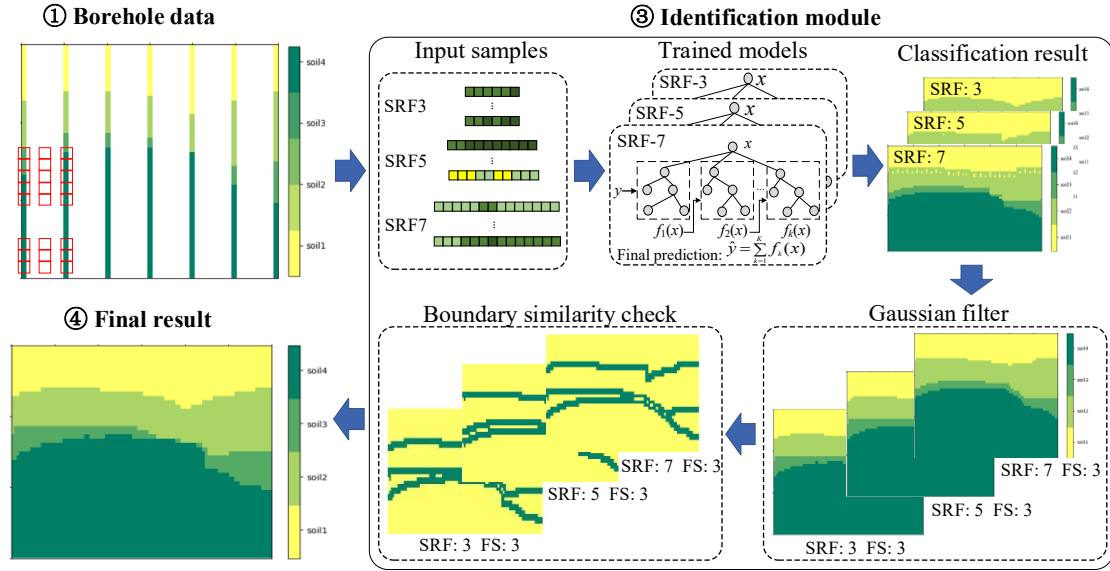


Fig. 3.5. Illustration on identification module of MSRF XGBoost method

3.2.5 Accuracy criteria

A practical accuracy criterion is in demand for assessing the performance of MSRF XGBoost models. In existing methods, a global accuracy (GA) is commonly adopted, as expressed in Eq. (3.6):

$$GA = 1 - \frac{\sum_{i=1}^{N_h \times N_v} I(Z_T(x_i) \neq Z_p(x_i))}{N_h \times N_v} \quad (3.6)$$

where Z_T is the referenced geological cross-section; Z_p indicates the predicted result; I

represent the indicator function and has a value of 1.0 when the actual soil type ($Z_T(x_i)$) differs from the predicted result, $Z_p(x_i)$; and N_h and N_v stand for the total cells of the original geological cross-section in the horizontal and vertical directions, respectively.

The GA method can evaluate the prediction accuracy over the entire area. However, since the soil boundary only takes a small part of the entire area, global accuracy tends to overestimate the prediction accuracy of the soil boundaries. Thus, this study proposes a new metric called boundary accuracy (BA) to address this issue, as follows:

$$BA = 1 - \frac{\sum_{i=1}^{N_h \times N_v} I(Z_T(x_i) \neq Z_p(x_i))}{N_b} \quad (7)$$

where N_b is the number of boundary cells from original geological cross-section.

Boundary accuracy only considers the soil boundary and can assess the predictive accuracy of the soil boundary more precisely. When the prediction closely matches the ground truth, BA will approach a value of 1. Conversely, when there is a deviation between the prediction and the ground truth, BA will deviate from 1.

Both GA and BA are based on the assumption of a known geological cross-section, so they are unsuitable for borehole data. An accuracy index for borehole data is still required in practice. The measurement accuracy (MA) of boreholes should be calculated as follows:

$$MA = 1 - \frac{\sum_{i=1}^{N_m} I(M_T(x_i) \neq M_p(x_i))}{N_m} \quad (8)$$

where M_T is the actual verification data from boreholes which are not divided into the training dataset; M_p denotes the predicted measurement; I is the indicator function and

has a value of 1.0 when the actual soil type ($M_T(x_i)$) differs from the predicted result, $M_p(x_i)$; and N_m represents the total number of cells from actual verification measurement.

3.3 Linear geological profile prediction results

The synthetic case data utilized in this study were derived from the research conducted by Shi et al. (2021b), which focused on soil layers characterized by quasi-linear boundaries, commonly referred to as the linear case. To train the model, four boreholes located at different horizontal positions (0m, 10m, 20m, and 30m) were employed, as illustrated in Fig. 3.6. Fig. 3.6 compares the geological cross-sections predicted by the conventional XGBoost method and the MSRF XGBoost method. Fig. 3.6 (a) depicts the original geological cross-section, and Fig. 3.6 (b) illustrates the prediction of the conventional XGBoost method. Fig. 3.6 (c) – 3.6 (e) shows the predictions of the proposed MSRF XGBoost method, and Fig. 3.6 (d) presents the optimal result the identification module recommends. The conventional and MSRF XGBoost methods can represent the spatial features of the soil layers with a quasi-linear boundary. According to Fig. 3.6 (b) – 3.6 (e), the soil boundaries predicted by the MSRF XGBoost method are much smoother than the discontinuous soil boundaries estimated by the conventional XGBoost method. The GA of the conventional method is 0.97, while its BA is 0.69, nearly 30% lower than its GA. In contrast, the GA and BA of models SRF-3 and SRF-7 are 0.98 and 0.75, respectively, higher than those of the conventional method. Recommended model SRF-5 has the highest GA of 0.99 and BA of 0.82, with an improvement of 2%–13% compared with the existing method. The reason for the

significant improvement in BA of the proposed algorithm is that it obtains accurate spatial feature relationships between adjacent boreholes through SRF of different sizes, achieving one-to-one mapping of samples.

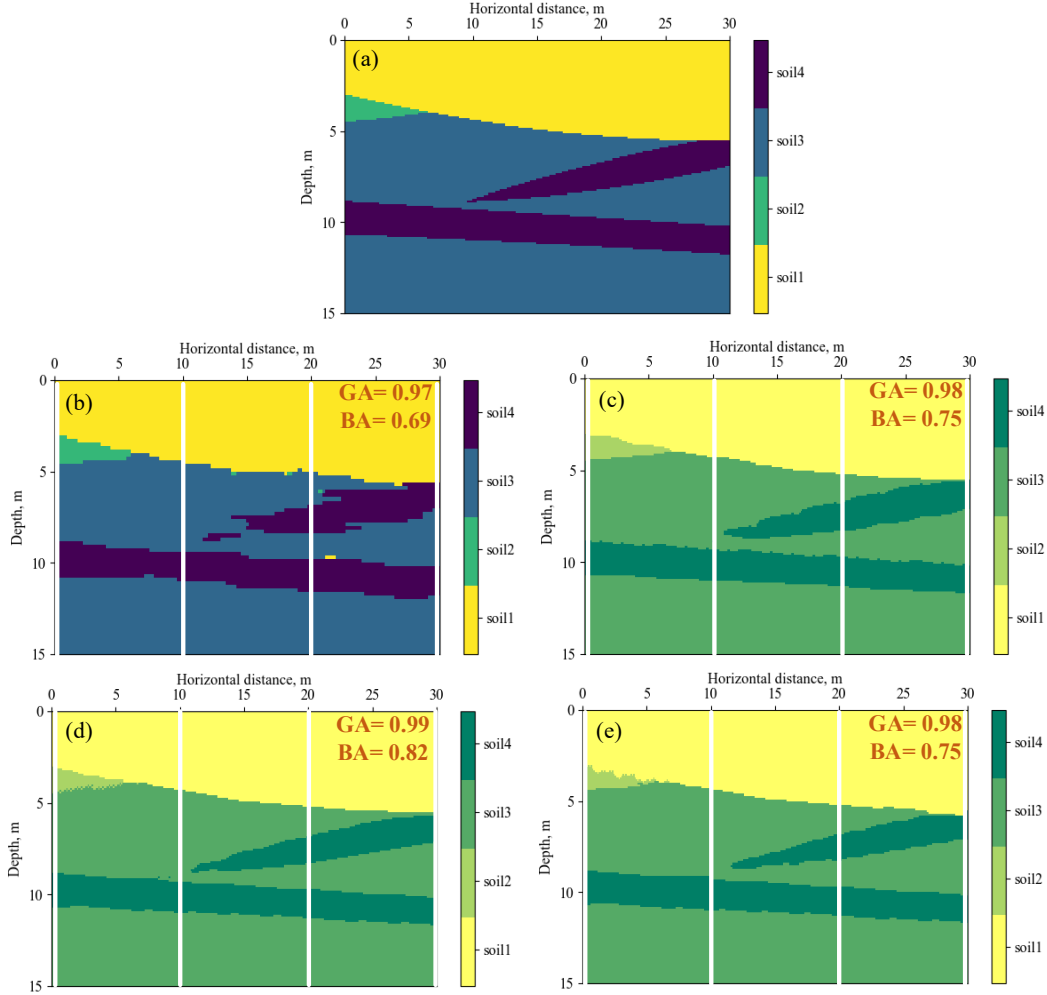


Fig. 3.6. Predictions of case 1 by different MSRF models: (a) Ground truth; (b) IC-XGBoost model; (c) model SRF-3; (d) recommended model SRF-5; (e) model SRF-7

3.4 Nonlinear geological profile prediction results

3.4.1 Prediction results

The data in the nonlinear case are also taken from the work of Shi et al (2021b) and characterized by the complex soil layers of the strongly nonlinear boundaries, called

the nonlinear case. In contrast to the synthetic case with quasi-linear soil boundaries, the nonlinear case can evaluate the capacity of the proposed model in the complex ground. In this case, four boreholes are used to train the model which is represented by white lines in Fig. 3.7. Fig. 3.7 compares the results predicted by the proposed MSRF XGBoost model, conventional XGBoost method, and Markov random field (Wei and Wang 2022). Fig. 3.7 (a) shows the actual geological cross-section, where white lines represent the four boreholes used for training. Fig. 3.7 (b) displays the prediction by the MRF method with GA of 0.948. The GA and BA of the IC-XGBoost method are 0.92 and 0.35, respectively, as shown in Fig. 3.7 (c). The low boundary accuracy indicates that precisely predicting nonlinear soil boundaries is difficult for the conventional XGBoost method. The GAs and BAs of predictions of all the MSRF XGBoost models are higher than those of the IC-XGBoost model. The optimal result, as recommended by the identification module, is predicted by model SRF-3 with the highest GA of 0.98 and BA of 0.82. The Model SRF-3 demonstrates a remarkable enhancement in its BA, surpassing that of the IC-XGBoost method by over 40%. Additionally, its GA stands out as being higher than the other two methods employed. According to Fig. 3.7 (d) and 3.7 (e), the global accuracy of models SRF-5 and SRF-7 is 0.97 and 0.95, respectively, and their boundary accuracy equals 0.75 and 0.61, respectively, which are still much higher than those of the conventional model. The high boundary accuracy confirms that the MSRF XGBoost method outperforms the IC-XGBoost method in the nonlinear geological cross-sections.

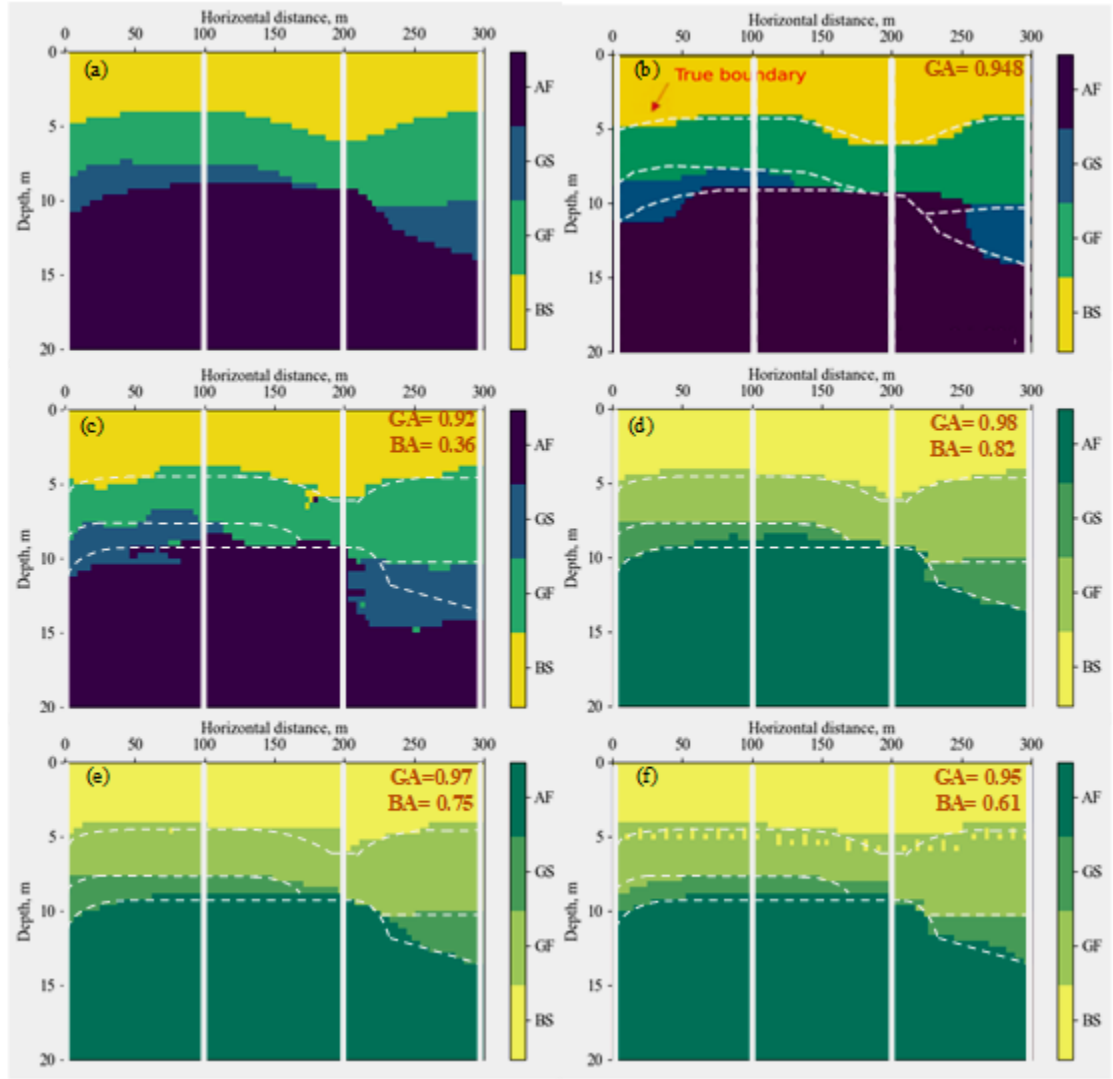


Fig. 3.7. Predictions of case 2 by different MSRF models: (a) ground truth; (b) MRF method; (c) IC-XGBoost method; (d) recommended model SRF-3; (e) model SRF-5; (f) model SRF-7

Additionally, the impact of the various SRFs on determining soil layers can be observed. Model SRF-5 drops some nonlinear features of the soil boundary, while model SRF-7 presents a distortion of the soil boundary characterized by saw teeth. In comparison, model SRF-3 outputs the optimal results with the highest accuracy on the soil boundary. This prediction difference in soil boundaries highlights the importance of an appropriate spatial receptive field.

3.4.2 Identification process

In practice, the ground truth is often unavailable, making the GA and BA of candidates from the classification module unknown. Therefore, the identification module is thus specially designed to choose the optimal result from these predictions. It has been observed that the original predictions generated by XGBoost models usually contain noise at the soil boundaries. To enhance the quality of predictions, applying a Gaussian filter to these original predictions can reduce the boundary noises. This filtering process not only helps in improving the overall prediction quality but also plays a crucial role in selecting the optimal model and prediction. One key objective when selecting the optimal prediction is to ensure smoothness of the predicted soil boundaries. The smallest error between the filtered and unfiltered predictions, the smooth the prediction will be. Then, the prediction that exhibits the smallest errors between the filtered and unfiltered versions can be considered as the candidate of the optimal prediction.

To demonstrate the effectiveness of the identification module, let's consider the non-linear case. Initially, the Gaussian filter intends to select the smooth boundary by reducing local noises of the predicted soil boundaries. The size of the filter plays a significant role in noise reduction. The filtering difference between the original and filtered predictions is employed to determine the filtering quality. To illustrate the impact of different filter sizes, Fig. 3.8 (c), 3.8 (e), and 3.8 (g) showcase the filtered predictions of model SRF-3 using various filter sizes, namely FS3, FS5, and FS7, respectively. Similarly, Fig. 3.8 (d), 3.8 (f), and 3.8 (h) depict the filtered predictions of model SRF-7 using FS3, FS5, and FS7, respectively. Generally, the filtering error

increases with the enlargement of the filter size, which can smooth the predicted soil boundary. In addition, the same filter size can induce various smoothing effects for the different models. For example, the filtering errors of model SRF-7 by FS3 (Fig. 8 (d)) are five times those of model SRF-3 by FS3 (Fig. 3.8 (c)) which means the smooth boundaries in predictions have less errors. Hence, it is challenging to determine a specific filter size for the various prediction models. For model SRF-3 in Fig. 3.8 (c), a small filter size (FS3) is appropriate due to small noise, while a large filter size is over-filtered. In contrast, for model SRF-7 in Fig. 3.8 (h), significant noises on the boundary need a large filter size (FS7) to smoothen the boundary. Thus, various models require different filter sizes. Considering the original smoothness of the boundary, two candidates with the smallest errors are selected from the good filter results.

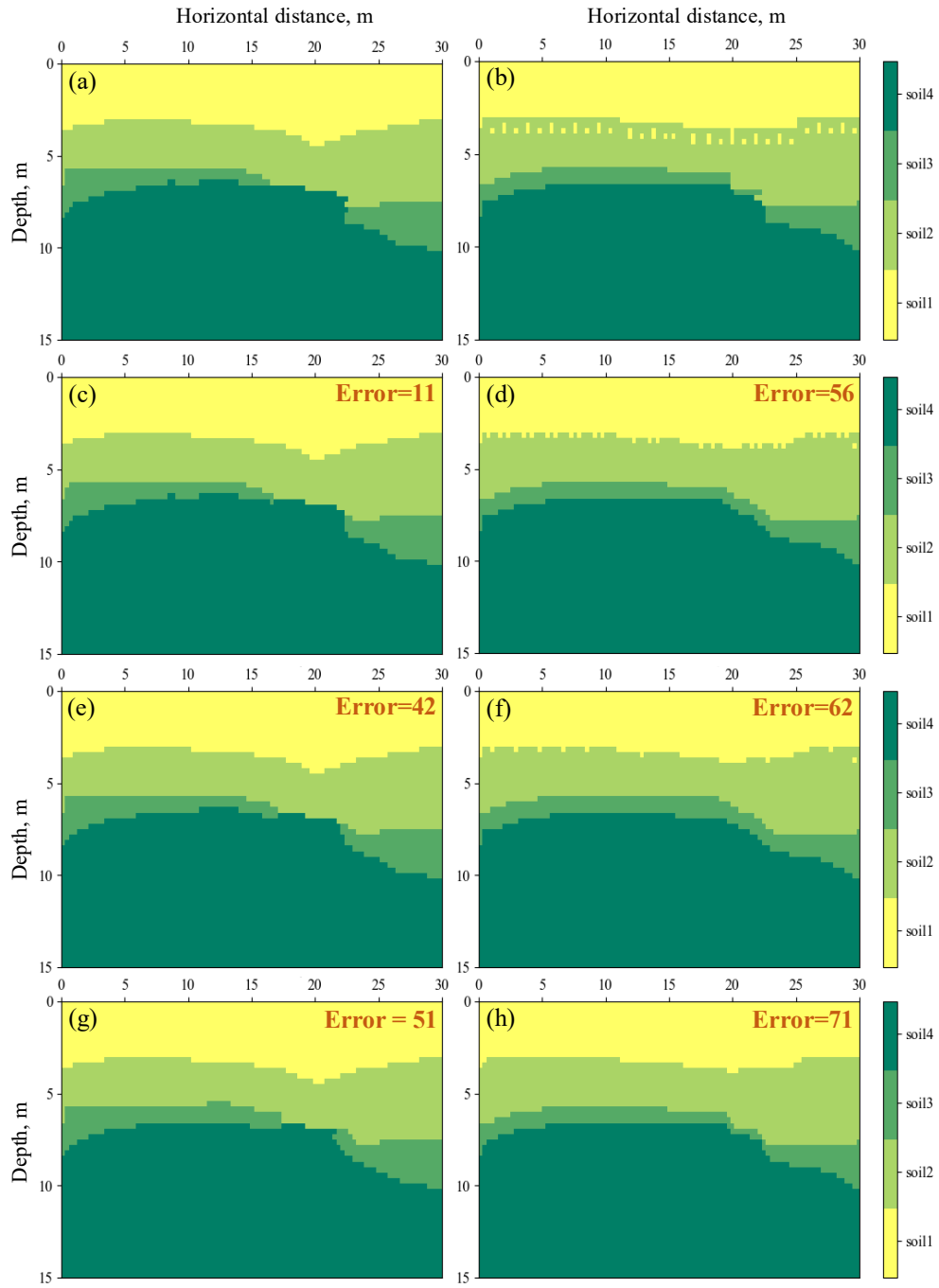


Fig. 3.8. Noise elimination of predictions by MSRF models with different filter sizes: (a) model SRF-3; (b) model SRF-7; (c) model SRF-3 with filter size 3; (d) model SRF-7 with filter size 3; (e) model SRF-3 with filter size 5; (f) model SRF-7 with filter size 5; (g) model SRF-3 with filter size 7; (h) model SRF-7 with filter size 7

Fig. 3.9 presents a comparison of soil boundaries predicted by different MSRF

models using various Gaussian filters. Specifically, Fig. 3.9 (a) and (b) display the soil boundaries of two candidates for the optimal prediction. These two candidates are from SRF-3 and SRF-5 models, respectively. On the other hand, Fig. 3.9 (c) and (d) illustrate the soil boundaries resulting from two specific Gaussian filters. The mean similarity scores of the two candidates with respect to the other filtering results are 0.815 and 0.78, respectively. Notably, the SRF-3 model demonstrates the optimal prediction, exhibiting the highest boundary similarity and the fewest errors, which is consistent with the recommended result with the best GA and BA scores from the identification module.

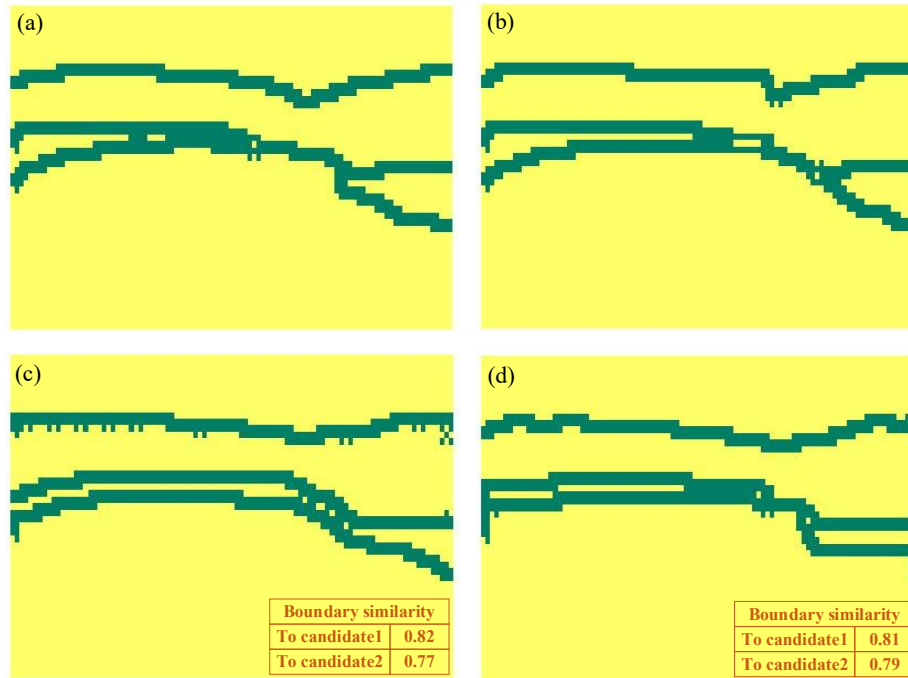


Fig. 3.9. Boundary similarity between candidates and predictions: (a) Candidate 1 from model SRF-3 with filter size 3; (b) Candidate 2 from model SRF-5 with filter size 3; (c) predictions from model SRF-7 with filter size 7; (d) predictions from model SRF-9 with filter size 5

3.5 Analysis of factors affecting the accuracy of MSRF-XGBoost method prediction

In this section, the effects of borehole space and vertical resolution, and potential limitations of the proposed method will be discussed.

3.5.1 Effect of borehole space

A regular space between the boreholes is required in the proposed method. This method is applicable to extract the nonlinear spatial relationship between one borehole and its regular-space adjacent boreholes. When the spacing between boreholes is irregular, the performance of the proposed method is likely to degrade significantly due to the confusing nonlinear spatial relationship that arises.

Taking nonlinear case as an example, the positions of four boreholes are 0 m, 8 m, 21 m, and 30 m, with the unregular space of 8 m, 13 m, and 9 m, respectively. Fig. 3.10 (a), (b), (c), and (d) represent the predictive geology cross-sections from unregular boreholes by SRF3, SRF5, SRF7, and SRF9, respectively. The predictive boundaries shown in Fig. 3.10 (a), (b), and (c), have a lot of noise. Only Fig. 3.10 (d) has relatively little boundary noises, benefiting from the extracted large-scale space features by a large SRF.

Compared with the regular spacing, the global accuracies of these predictions decline by 10% to 0.89, 0.87, 0.88, and 0.88, respectively. Furthermore, all of the boundary accuracies in these cases fall below 0.2, significantly lower than the results achieved with regular-spacing cases. In addition, the global geological features in unregular-space cases exhibit substantial disparities from the ground truth. In general, this

example highlights the limitations of the proposed method in effectively learning geological features from boreholes with irregular spacing.

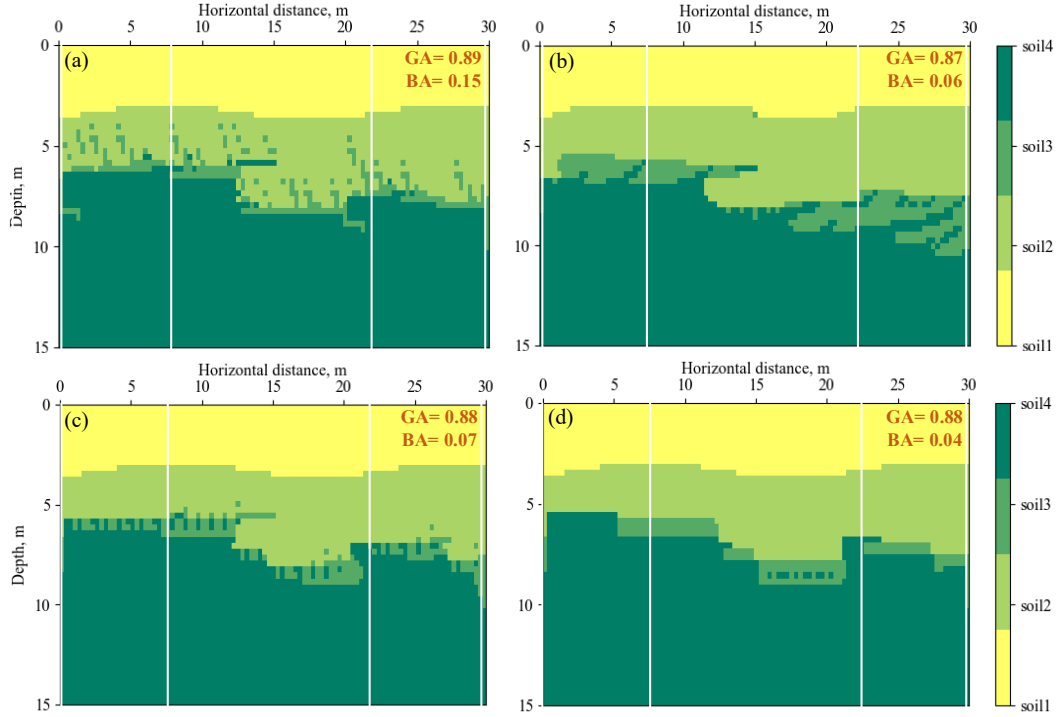


Fig. 3.10 Predictions based on unregular-space boreholes by: (a) model SRF-3; (b) model SRF-5; (c) model SRF-7; (d) model SRF-9

3.5.2 Effect of vertical resolution

The vertical resolution also has a significant impact on the predictive performance and the selection of the optimal SRF size of the MSRF-XGBoost model. The vertical resolution represents the number of cells created by splitting the borehole. With a fixed depth borehole, a higher resolution results in more cells of smaller size. A smaller cell size implies that a geological feature will span across more cells, often necessitating a larger SRF size to accurately represent the feature.

To determine the effect of vertical resolution, different cell sizes ranging from 0.25 m, 0.2 m, 0.15 m, to 0.1 m are set up based on the geological condition in case 2. The

corresponding vertical resolutions were set at 60, 75, 100, and 150, while the original vertical resolution in case 2 was 50 with a cell size of 0.3 m. Fig. 3.11 provides a comparison of the predictive results of the optimal SRF models for different vertical resolution. Notably, the optimal SRF sizes differ for each vertical resolution. Specifically, the optimal SRF size for vertical resolutions of 60 and 75 is 3, which increases to 5 and 9 for vertical resolutions of 100 and 150, respectively.

Generally, the optimal SRF size in the proposed method gradually increases with the growth of the vertical resolution, supporting the previous analysis. The global accuracies of these predictions from different resolution are 0.98, 0.97, 0.97 and 0.97, respectively, similar to that of the original resolution. However, a decrease in boundary accuracies is observed from 0.8, 0.63, 0.62 to 0.40 as the vertical resolutions increase. As shown in Fig. 3.11 (c) and (d), the decreasing accuracy results from the noisy predicted boundaries generated by the MSRF model with a large SRF size. To summarize, as the vertical resolutions increase, the optimal SRF size also increases, while the accuracy of the predicted boundaries decreases. Nonetheless, the overall global accuracies remain consistent with those obtained at the original resolution.

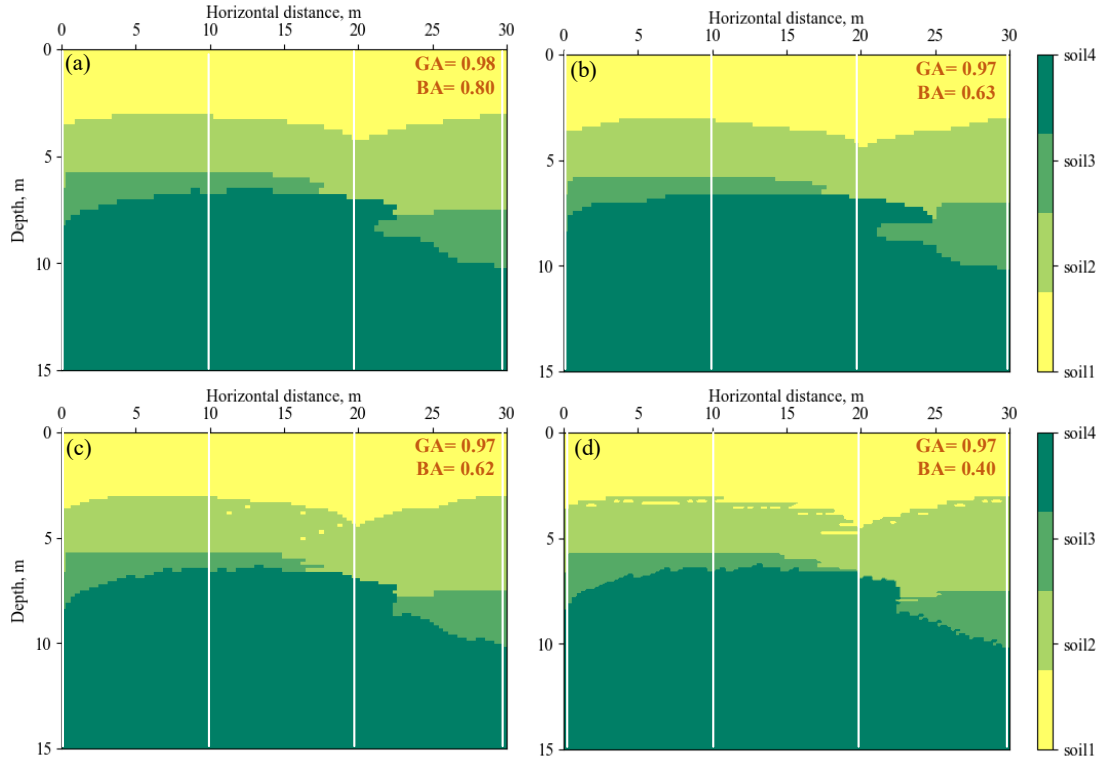


Fig. 3.11. Predictions of optimal SRF model for different vertical resolutions: (a) model SRF-3 for resolution at 60; (b) model SRF-3 for resolution at 75; (c) model SRF-5 for resolution at 100; (d) model SRF-9 for resolution at 150

3.6 Summary

The MSRF-XGBoost algorithm was proposed, which relies solely on limited borehole data to predict geological profiles with high accuracy. The results outperform the existing IC-XGBoost method in terms of global accuracy, boundary accuracy, and reliability. Specific conclusions include:

- (1) Unlike the IC-XGBoost method, which uses geological images as training data, the MSRF-XGBoost method extracts spatial features of strata directly from sparse borehole data.

- (2) A screening module, based on Gaussian filtering and boundary similarity algorithms, automatically recommends the optimal prediction result from multiple outputs generated by different SRF models.
- (3) Compared to IC-XGBoost, MSRF-XGBoost improves boundary prediction accuracy by over 0.4 for non-linear geological profiles and by 0.13 for linear profiles.

CHAPTER 4 GRAPH CONVOLUTIONAL NETWORK FOR NON-EQUIDISTANT BOREHOLE DATA

4.1 Introduction

Predicting the distribution of stratigraphic interfaces using borehole data is crucial for geotechnical engineering applications, such as foundation excavation (Lin et al., 2021; Sindete et al., 2024; Usui et al., 2024), tunnel construction (Zhang et al., 2022; Meschke et al., 2023; Bui et al., 2023; Kong and Zhang, 2024), and slope management (Kennedy et al., 2021; Wei et al., 2024). However, borehole data are typically sparse and spatially discontinuous due to the high cost and practical constraints of drilling (Qiu et al., 2024), leading to a common scenario of non-equidistant borehole spacing. This non-equidistant nature of data poses a significant, yet often unaddressed, challenge for accurate stratigraphic modeling.

Extensive research has produced various estimation methods, which can be broadly categorized as interpolation, statistical, and machine learning approaches. While interpolation methods like Kriging (Dasaka and Zhang, 2012) and spline interpolation (Aswar and Ullagaddi, 2017) are widely used, they can produce significant errors in complex geological settings. Statistical methods, including multipoint statistics (Shi and Wang, 2019), conditional random fields (Qi and Liu, 2019; Gong et al., 2021; Zhao et al., 2021; Juang et al., 2019; Jiang et al., 2020), coupled Markov chains (Qi et al., 2016; Li et al., 2016; Deng et al., 2017; Cao et al., 2021), Markov random fields (Mariethoz and Caers, 2014; Wang et al., 2017; Wang et al., 2018), and Bayesian

compressive sampling (Wang and Zhao, 2017; Zhao et al., 2018; Hu et al., 2020; Huang et al., 2014; Wei and Wang, 2022; Tian and Wang, 2023), often rely on specific mathematical assumptions and can struggle with parameter identification from limited data.

The advent of machine learning has introduced methods like artificial neural networks (ANNs) (McCulloch and Pitts, 1942; Rumelhart and Williams, 1986; Zhou et al., 2018; Li et al., 2022, 2023, 2024), clustering algorithms (Ikotun et al., 2023), stacking classification (Yan et al., 2023; Yan et al., 2023), IC-XGBoost (Shi and Wang, 2021), and Bayesian machine learning (Wei and Wang, 2022). A critical limitation, however, is that methods such as XGBoost and Bayesian machine learning have been developed and validated primarily using equidistant borehole data (Shi and Wang, 2022). This assumption is frequently violated in practice, creating a salient gap between methodological development and real-world application.

This chapter focuses on addressing this gap by developing an approach suitable for the prevalent case of non-equidistant borehole data. We propose a Graph Convolutional Network (GCN)-based framework, which is naturally suited to learning from irregularly structured data. An improved GCN algorithm is presented to predict 2D geological profiles from sparse, non-equidistant boreholes. The method is validated against a practical case study, showing superior boundary accuracy, and the impact of borehole number and spacing on predictions is systematically discussed.

4.2 Novel graph convolutional network methodology

GCN is widely used in various classification tasks within the computer science domain, such as document and word classification. Recognizing the GCN algorithm's superior performance in classification tasks, we enhance the GCN model by incorporating three linear layers and two nonlinear layers to improve its feature extraction capabilities. This enhancement enables the model to more effectively tackle the challenge of geological profile prediction, which is a small-sample classification problem. Fig. 4.1 presents the predictive flowchart of the improved GCN proposed in this study. Since computer programs are unable to directly process entire boreholes for training, it is essential to discretize the borehole along its depth into individual points for input into the GCN model. For example, if a borehole has a depth of 20 meters and is divided into segments with a thickness of 0.4 meters, it will result in 50 individual points. These discretized boreholes are then placed into a Cartesian coordinate system to obtain the coordinates of each point. Given that the soil type within the borehole is known, the soil type represented by each discretized point is also established. The geological profile to be predicted is similarly positioned within the Cartesian coordinate system to obtain the coordinates of each point along the entire profile. For the geological profile under consideration, the soil type of unknown points is highly correlated with the soil type of adjacent points. Therefore, the features of the nine adjacent points to each center point are used to establish an adjacency matrix. The coordinates of the points are provided as inputs to the GCN model, while the adjacency matrix represents the relationships

among different points. The output of the GCN model corresponds to the soil type associated with each point. By training the GCN, the relationship between the coordinates and soil types of the points is established. By inputting the coordinates of unknown points into the trained GCN model, the soil types of unknown points can be retrieved, enabling the prediction of the entire geological profile.

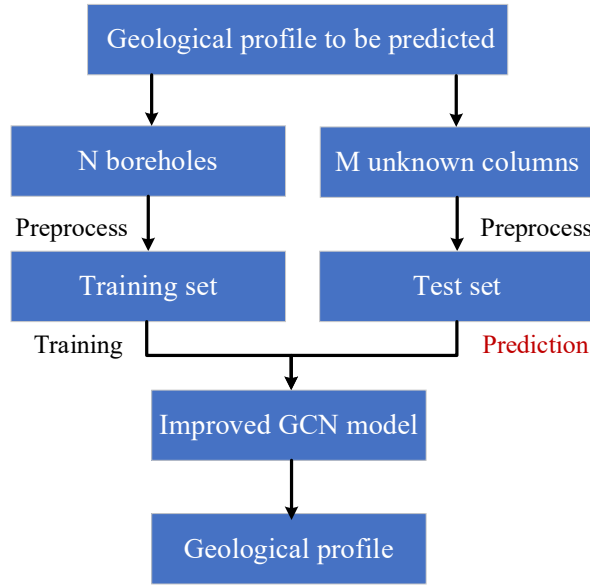


Fig. 4.1. Flow chart for the proposed improved GCN model

4.2.1 GCN method

The GCN was proposed by Kipf (Kipf and Welling, 2016), marking the first instance of utilizing a graph structure within a deep learning method. To define an undirected graph G , as shown in Equation (4.1):

$$G = (V, E) \quad (4.1)$$

where V represents N nodes in the graph $v_i \in V$, and E denotes the edges connecting every pair of nodes $(v_i, v_j \in E)$.

Based on graph G , we define a graph neural network model $f(X, A)$. Let X be a

matrix of node feature vectors $X \in \mathbb{R}^{N \times C}$, where C is the dimension of the feature vector.

A is an adjacency matrix representing the edges among all nodes $A \in \mathbb{R}^{N \times N}$.

The following layer-wise propagation rule of a multi-layer GCN is displayed in Equation (4.2):

$$H^{l+1} = \sigma(\tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}} H^l W^l) \quad (4.2)$$

where $\tilde{A} = A + I_N$ is the adjacent matrix of the undirected graph G with added self-connections, I_N is the identity matrix. $\tilde{D}_{ii} = \sum_j \tilde{A}_{ij}$ is a degree matrix, and W^l is a layer-specific trainable weight matrix. σ denotes an activation function, such as the rectified linear unit (ReLU) or softmax. $H^l \in \mathbb{R}^{N \times D}$ is the matrix of activations in the l^{th} layer, $H^0 = X$.

Research indicates that an excessive number of layers in a GCN is detrimental to prediction performance. In most tasks, a two-layer GCN yields the optimal results. The simplified form of this model is presented in Equation (4.3):

$$Z = f(X, A) = \text{softmax}(\tilde{A} \text{ReLU}(\tilde{A} X W^{(0)}) W^{(1)}) \quad (4.3)$$

where $W^{(0)} \in \mathbb{R}^{C \times H}$ is an input-to-hidden weight matrix for a hidden layer with H feature maps, while $W^{(1)} \in \mathbb{R}^{H \times F}$ is a hidden-to-output weight matrix.

The loss function of GCN method is cross entropy function which displayed in Equation (4.4),

$$H(p, q) = - \sum_{i=1}^n p(x_i) \log q(x_i) \quad (4.4)$$

Where $p(x_i)$ is the real distribution, $q(x_i)$ is the predictive distribution.

Fig. 4.2 illustrates the process of a GCN. This represents a semi-supervised

learning task within the GCN framework (Kipf and Welling, 2016). The inputs consist of C channels, and the output layer contains F feature maps. The edges between nodes are depicted as black lines. The graph structure is consistent across layers, and the labels are denoted by Y_i .

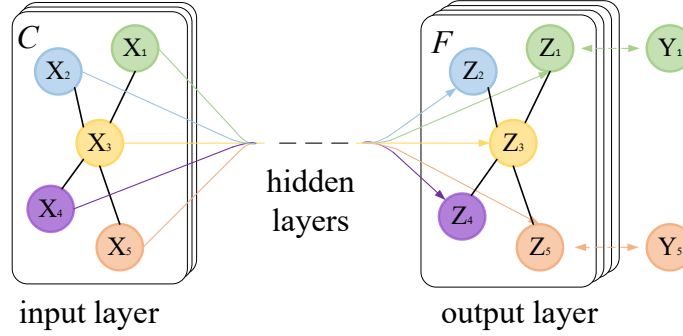


Fig. 4.2. Illustration on graph convolutional network model

4.2.2 Adaptive improvement of GCN algorithm

Applying the GCN algorithm to geological profile prediction problems requires adaptive improvements to existing GCN algorithms. The basic idea of geological profile prediction is to predict the soil type corresponding to each point by its coordinates, that is, the feature of the input point X is the x and y coordinates of each point. For the GCN algorithm, X has fewer input features, which is not conducive to subsequent feature learning and extraction. Therefore, by using linear and nonlinear layers before the input graph convolutional layer, the feature abstraction ability, model depth, and adaptability to complex graph structures can be improved.

The linear layer, also known as the fully connected layer, is used to achieve linear transformation of features through matrix multiplication and bias addition, such as

adjusting feature dimensions or linearly combining features. The formula is shown in equation (4-5):

$$Y = WX + b \quad (4-5)$$

Where X is the input data, W is the weight matrix, and b is the bias vector.

The nonlinear layer is an activation function layer used to introduce nonlinear transformations. Its function is to break the limitations of linear superposition and enable neural networks to approximate any complex function. The formula is shown in equation (4-6):

$$Y = ReLU(X) \quad (4-6)$$

Where X is the input data, $ReLU$ is the nonlinear activation function, and the formula is $f(x) = \max(0, x)$.

For predicting geological profile in this thesis, the X represents the coordinate of every node which is relatively simple features. Before inputting the graph convolutional network model, multiple linear and nonlinear layers are added to transform features and improve the feature extraction ability of graph convolutional networks. By combining different layers, choose to stack three linear layers and two non-linear layers to improve the feature extraction ability of the graph convolutional network. The feature is expressed in Equation (4.6),

$$X = W^3(ReLU(W^2(ReLU(W^1x)))) \quad (4.6)$$

The training hyperparameters of the improved GCN method involve learning rate, weight decay, and bias. After iterative experimentation, the learning rate was determined to be 0.01, weight decay was specified as 1×10^{-5} , and the bias parameter

was configured accordingly. The model will be trained by using the Adam optimizer with 500 epochs in the following sections.

4.2.3 Geological Profile Prediction based on novel GCN method

In accordance with the GCN model's definition, training the GCN model necessitates the construction of an undirected graph with labeled connections between points. Each training point in the training set requires a corresponding label. In the context of geological profile prediction, constructing an undirected graph involves discretizing borehole data into points along depth and placing them into a Cartesian coordinate system. As illustrated in Fig. 4.3, parts of four boreholes are used as an example, with each borehole discretized into five points along its depth. To meet the precision requirements of the geological profile, additional columns are inserted between boreholes for prediction. The geological profile to be predicted is also positioned within the Cartesian coordinate system to obtain the corresponding coordinates for each point. Soil types at borehole points are represented by 1, 2, and 3, with the label for each point indicating the corresponding soil type at that depth. A point sample, which includes a central point and eight neighboring points, slides across the geological profile from point (1,1) to point (5,13) to construct the input dataset. For the central point, eight neighboring points constitute its edges, and the inverse distance between each neighboring point and the central point serves as the edge weight. Although the point sample includes eight neighboring points, this number is not fixed. The actual number of neighboring points depends on the position of the central point; for instance, point

(1,1) in the geological profile only has three neighboring points, resulting in three edges, as shown in Fig. 4.3. After the point sample slides across the entire geological profile, the input graph is constructed, composed of all input point coordinates, edges, and edge weights. The input training points consist of data from the four boreholes, while the output label is the soil type of each input training point. The input graph, input training points, and output labels are fed into the GCN model to predict the soil type of each point in the geological profile. One epoch refers to training the GCN model with the input training points once, while five hundred epochs mean training the GCN model with the input training points five hundred times. As evident from Fig. 4.3, a higher number of epochs correlates with improved prediction accuracy.

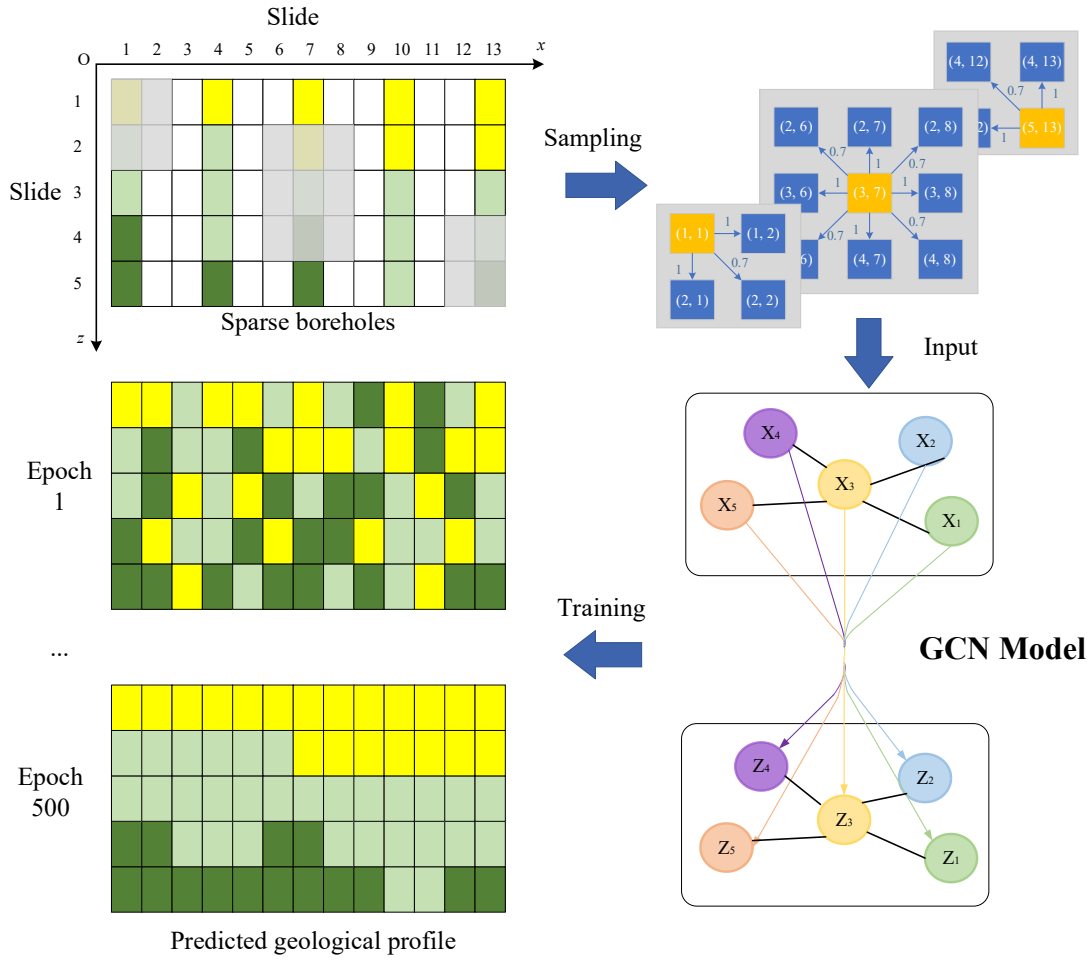


Fig. 4.3. Scheme of sampling and training GCN model for the geological profile

4.3 Nonlinear geological profile prediction results

The data for Case 1 is the test profile, which includes four soil types represented by BS, GF, GS, and AS, respectively. To predict the geological profile of the site and compare it with other results from the literature, four equally spaced boreholes are utilized. The locations of the boreholes are indicated by the white lines in Fig. 4.4 Each borehole is depicted as a column and is discretized into 50 points along its depth. The entire geological section comprises 100 columns, resulting in a total of 5,000 points in the geological profile. For the improved GCN model, the soil types corresponding to the

four boreholes (200 points) are utilized for training to predict the soil types of the entire geological profile. Fig. 4 (a) shows the actual geological profile, (b) displays the geological profile predicted by the MRF method, (c) presents the prediction made by the IC-XGBoost method, and (d) indicates the prediction of the GCN method. The top right corner of the prediction results displays the values of GA and BA, while the white dashed lines mark the boundaries of the actual geological profile. From Fig. 4, it can be observed that the GA of the improved GCN method is slightly higher than that of the IC-XGBoost method and comparable to the MRF method, with GAs exceeding 90% for all methods. Therefore, GA does not distinguish between these three methods. In terms of BA, the improved GCN method significantly outperforms the IC-XGBoost method, with an enhancement of 24%. This demonstrates that the improved GCN method is more effective at extracting spatial features from boreholes and predicting stratum boundaries accurately.

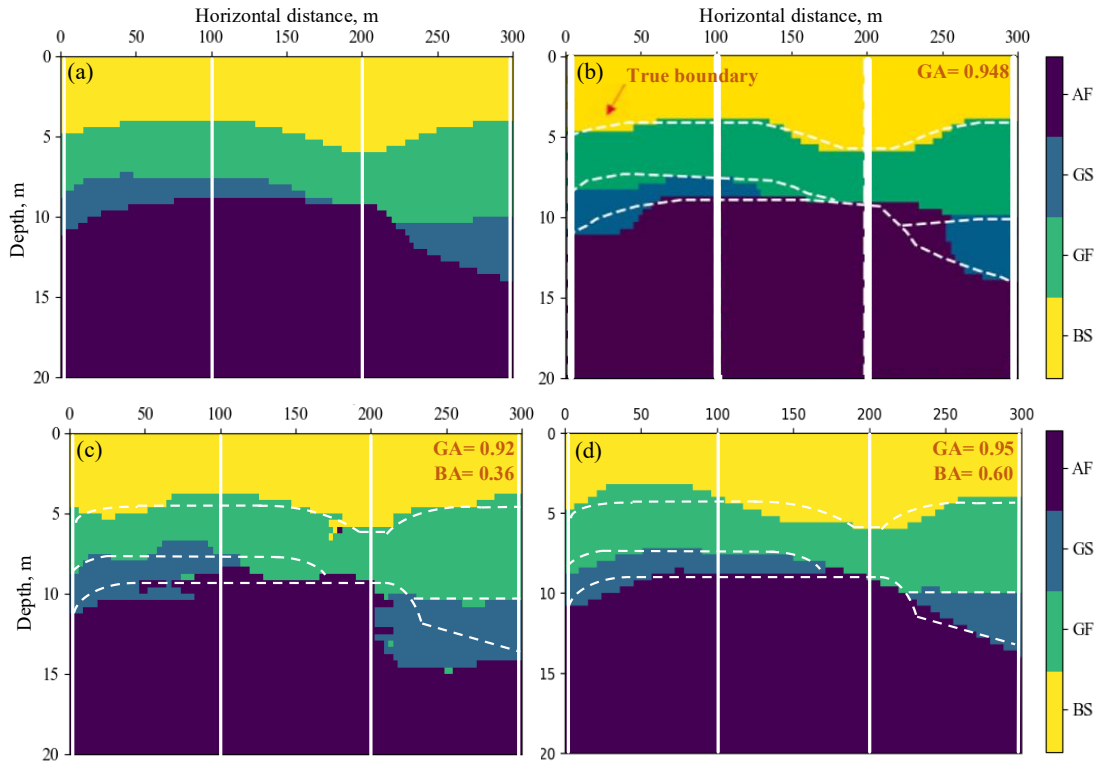


Fig. 4.4. Comparing the results of the different models in the real case: (a) the actual geological profile; (b) MRF method; (c) IC-XGBoost method; and (d) the proposed GCN method

The IC-XGBoost method has notable limitations, primarily because its predictions depend on training with adjacent geological profiles and require equally spaced boreholes. However, in real-world engineering, adjacent geological profiles may not always be accessible, and borehole spacing is often unequal. The enhanced GCN method overcomes these constraints in two ways. First, it directly extracts spatial features from the boreholes themselves to predict stratum distribution, eliminating the IC-XGBoost method's reliance on adjacent profiles for training. Second, the improved GCN approach uses coordinates as point features, meaning that variations in borehole

spacing only alter the coordinates without hindering the model's ability to learn spatial features, thus addressing the IC-XGBoost method's need for equally spaced boreholes.

4.4 Analysis of factors affecting the accuracy of novel GCN method prediction

The improved GCN model can directly predict the entire geological section based on boreholes. Therefore, the number and spacing of boreholes significantly impact the prediction results. To examine how varying the number of boreholes affects prediction results, the improved GCN model is utilized with diverse input configurations: four, five, six, and seven equally spaced boreholes. The changes in prediction accuracy are then compared. To analyze how borehole spacing influences prediction results, the study keeps the number of boreholes constant at four while varying their positions to assess changes in prediction accuracy.

4.4.1 Effect of the number of boreholes

Fig. 4.5 illustrates the prediction results for various numbers of equally spaced boreholes. The positions of the boreholes are marked by white lines, while the white dashed lines denote the boundaries of the actual soil layers. The GA and BA of the predictions are highlighted in red in the top right corner. In Fig. 4.5 (a), with four boreholes, the BA is 0.6 and the GA is 0.95. Fig. 4.5 (b) shows results for five boreholes, where the GA remains the same as in Fig. 4.5 (a) at 0.95, but the BA improves slightly to 0.63. This indicates that while the number of boreholes affects their placement and the spatial information available, it also impacts the prediction of the geological profile. Although the GA remains constant, BA varies with the number of boreholes. Fig. 4.5

(c) depicts results with six boreholes, where both GA and BA increase marginally by 0.01 and 0.03, respectively, compared to the five-borehole scenario. Fig. 4.5 (d) presents the results with seven boreholes, showing a GA improvement of 0.01 and a significant BA increase to 0.76 compared to six boreholes. This demonstrates that the effect of adding boreholes on BA is not linear. Generally, prediction accuracy improves with more boreholes, but the rate of improvement varies. When additional boreholes reveal crucial spatial details about soil layer distribution, BA can increase sharply, as observed between six and seven boreholes. Conversely, if additional boreholes do not provide new spatial insights, as seen from five to six boreholes, BA may remain relatively stable.

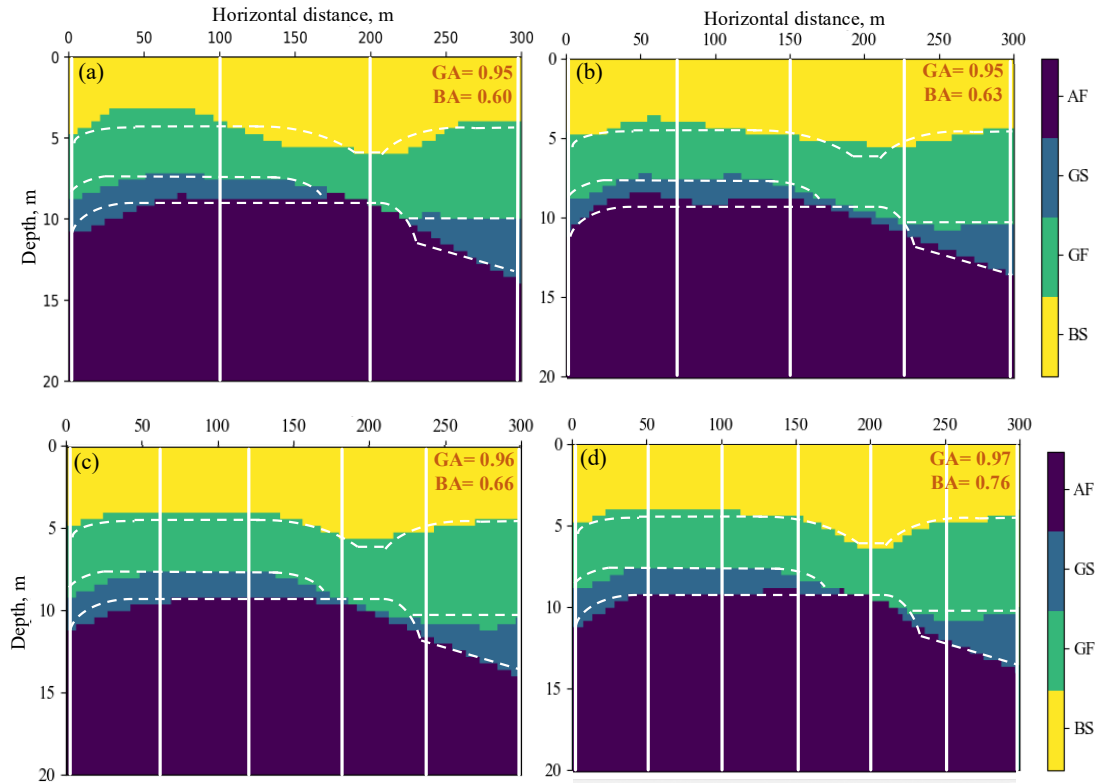


Fig. 4.5. Predictions of different numbers of boreholes: (a) four boreholes; (b) five boreholes;

(c) six boreholes; and (d) seven boreholes

Fig. 4.6 shows the variations of GA and BA with different borehole distances. Overall, the distances between boreholes exhibit a minor influence on the global accuracy. However, the boundary accuracy demonstrates a nonlinear growth as the borehole distance decreases. As the distance diminishes, there is a notable rise in boundary accuracy.

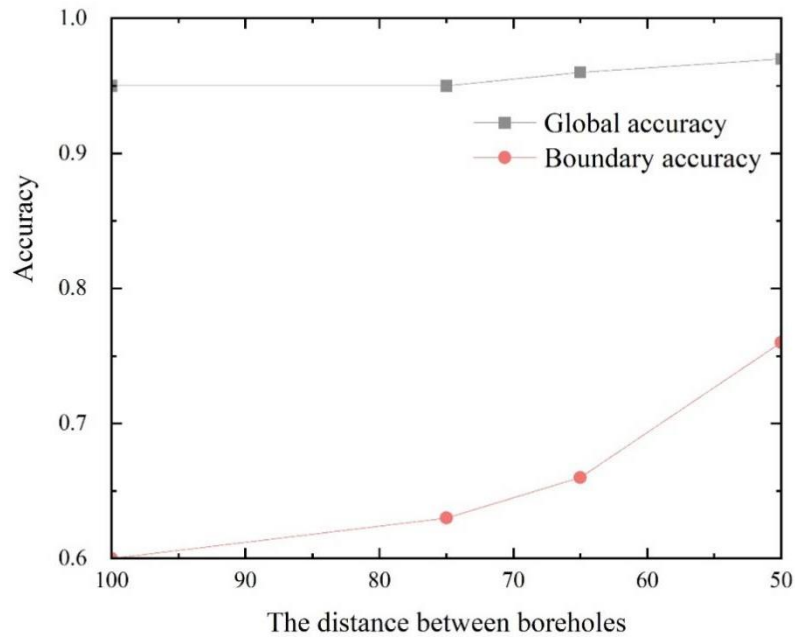
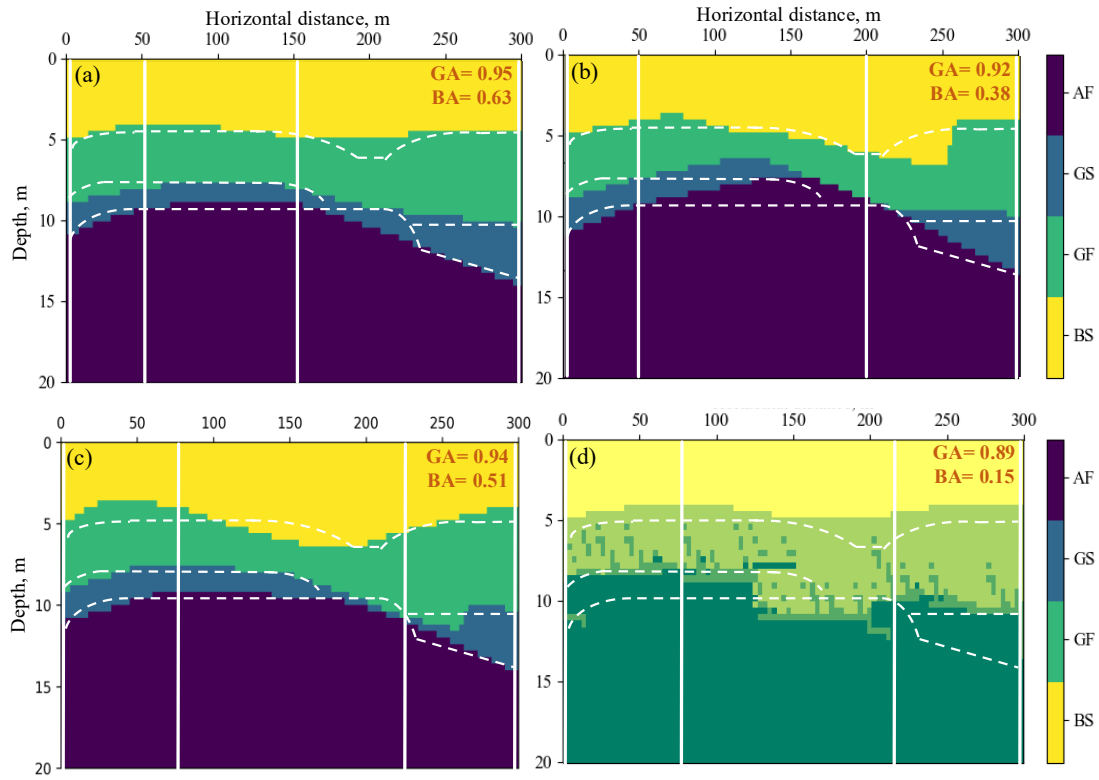


Fig. 4.6. Accuracy variations versus the borehole distances

4.4.2 Effect of borehole location

Fig. 4.7 presents the prediction results for four boreholes placed at various locations. The borehole positions are marked with white lines, while the white dashed lines outline the actual geological stratum boundaries. The GA and BA of the predictions are indicated in red in the top right corner. In Fig. 4.7 (a), the boreholes are unequally spaced with the first three placed at 0 m, 50 m, and 150 m, followed by a borehole at

300 m. This close spacing offers better spatial information, allowing the improved GCN method to capture essential spatial features effectively, resulting in a GA value of 0.95 and a BA value of 0.63. Fig. 4.7 (b) shows boreholes positioned at 0 m, 50 m, 200 m, and 300 m with unequal spacing. This arrangement results in a loss of crucial boundary information, causing a 30% reduction in BA compared to Fig. 4.7 (a). Fig. 4.7 (c) features boreholes at 0 m, 75 m, 225 m, and 300 m. Here, the GA is 0.94 and the BA is 0.51, which is better than Fig. 4.7 (b) but not as good as Fig. 4.7 (a). Fig. 4.7 (d) displays the same borehole positions as Fig. 4.7 (c), but with predictions made using the MSRF-XGBoost method. MSRF-XGBoost performs well with equally spaced boreholes but struggles with unequal distributions. The GA for MSRF-XGBoost is 0.89, and the BA is 0.19, which is significantly lower than the GCN method's BA. As shown in Fig. 9, even with the same number of boreholes, their specific locations and spacing can cause substantial differences in predicted geological sections due to variations in spatial information. Accurate geological predictions require borehole data at critical locations. Thus, the impact of unequal spacing on prediction results is highly dependent on borehole positions. To improve prediction accuracy, it is crucial to gather detailed information from boreholes at strategically important sites. The comparison between Fig. 4.7 (c) and Fig. 4.7 (d) highlights the superior performance of the improved GCN method with unequal borehole distributions.



**Fig. 4.7. Predictions of different positions of boreholes: (a)–(c) different unequal spacings;
(d) MSRF-XGBoost method**

4.5 Summary

An improved Graph Convolutional Neural Network (GCN) algorithm was proposed, capable of predicting accurate geological profiles using non-uniformly spaced boreholes. Key conclusions include:

- (1) The GCN eliminates the dependency on adjacent geological profiles and the requirement for uniformly spaced boreholes, enabling accurate predictions with sparse data.
- (2) Compared to existing machine learning methods and literature cases, the GCN model improves boundary prediction accuracy by 24%.

- (3) Prediction accuracy increases with the number of boreholes, but plateaus if additional boreholes provide no new spatial information.
- (4) The location of boreholes significantly impacts accuracy. For the same number of boreholes, strategically placed ones yield higher accuracy.

CHAPTER 5 ENGINEERING APPLICATION AND COMPARISON WITH DIFFERENT METHODS

5.1 Introduction

With the advent of the big data era, the application of machine learning algorithms in the engineering field has gradually transitioned from theoretical research to practical engineering validation. This shift not only underscores the importance of algorithmic engineering but also highlights the unique value of machine learning technologies in solving complex engineering problems. Engineering application validation, as a critical bridge between theoretical research and industrial practice, focuses on data-driven verification in real-world scenarios to assess the predictive effectiveness and result stability of algorithms in practical engineering applications. In the previous two chapters, the theories and prediction processes of the two proposed machine learning algorithms were introduced in detail. In this chapter, the predictive performance of these two machine learning algorithms for geological profiles will be validated through specific engineering applications.

Among the many machine learning algorithms, the XGBoost algorithm and the Graph Convolutional Network (GCN) algorithm have demonstrated significant potential in engineering classification problems due to their unique advantages. The XGBoost algorithm, as an optimized implementation of gradient-boosted decision trees, exhibits exceptional performance in classification tasks for small- to medium-scale structured data through parallel computing and regularization strategies. Its engineering

value lies in feature importance analysis, mechanisms for handling missing values, and high compatibility with engineering parameters (Yin et al, 2020). For example, in the design of high-strength concrete mixes, XGBoost integrates multi-source parameters such as material ratios, production costs, compressive strength, and shear strength to achieve optimal mix design (Tipu et al, 2023). In the field of bridge structural damage identification, the XGBoost algorithm has been applied to bridge management systems to evaluate active damage (Lim et al, 2019). In landslide studies, XGBoost employs multi-parameter exponential functions—including slope, aspect ratio, curvature, and rainfall as inputs for hierarchical selection of triggering factors, revealing the variable importance of each factor (Badola et al, 2023).

The GCN algorithm overcomes the limitations of traditional machine learning methods that rely on Euclidean space data, making it particularly suitable for modeling non-Euclidean topological data commonly found in engineering systems. Through node feature propagation and aggregation mechanisms in graph structures, GCN effectively captures the spatial correlation characteristics of engineering systems (Kipf et al, 2016). In civil engineering, GCN has been successfully applied to anomaly detection in bridge monitoring networks, reducing false alarm rates by 40% by modeling the spatial correlations among sensors (Li et al, 2023). In geological hazard risk assessment, GCN integrates geospatial topology and multi-source remote sensing data to construct a landslide classification model with spatial propagation properties (Murshed et al, 2024).

A key challenge in engineering application validation lies in establishing a

verification framework that aligns with engineering characteristics. This includes acquiring real-world engineering data, training models with small sample sizes, and defining evaluation criteria for prediction results. In this chapter, the practical engineering applications of the two proposed machine learning algorithms are evaluated using case studies from excavation projects in the Netherlands and Hangzhou. Subsequently, the prediction results of different machine learning algorithms are compared and analyzed through nonlinear geological section prediction and a geological section prediction project in Hangzhou. This comparative analysis elucidates the strengths, weaknesses, and suitable application scenarios of the two proposed machine learning methods.

5.2 Engineering application based of MSRF-XGBoost method

5.2.1 Borehole source and layout

The boreholes used for geological profile prediction come from a publicly available website in the Netherlands, which records information on over 100 boreholes for a certain project. The depth of these boreholes is about 800 meters, with a horizontal span of 2800 meters. Based on their Earth coordinates, we selected eleven equally spaced boreholes on a straight line for geological profile prediction. Extract five boreholes at intervals as validation boreholes, and the other six boreholes as input boreholes. Predict geological profiles through six input boreholes, and compare the predicted results with five validation boreholes to verify the accuracy of the predicted results. The distribution of boreholes is shown in Figure 5.1. The drilling selection is shown by the red line in

the figure.

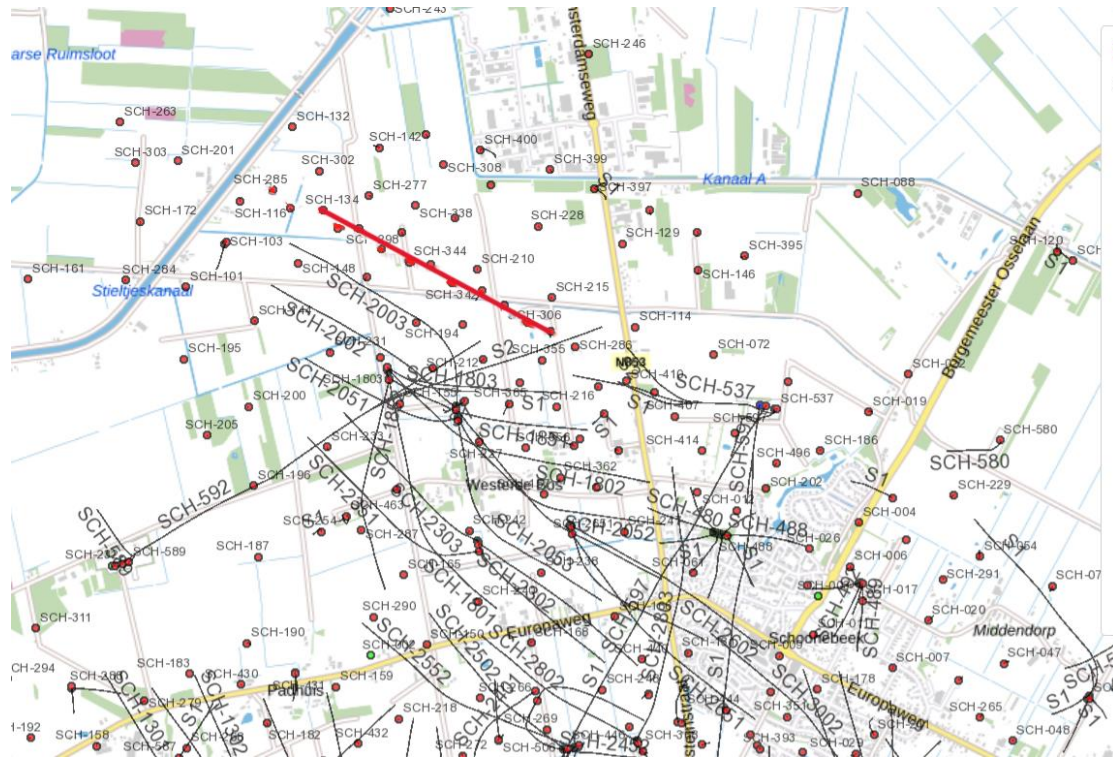


Fig. 5.1. The layout of borehole

5.2.2 Geological profile prediction

In the case of the Netherlands, eleven boreholes are selected from the available dataset of open boreholes. These particular boreholes have a substantial depth of approximately 800 meters and cover a considerable horizontal span of 2800 meters. Each cell within the dataset has a vertical size of 3 meters and a horizontal size of 14.5 meters, resulting in vertical and horizontal resolutions of 267 and 194, respectively. Six known in situ boreholes are selected as the input for the training data set, represented by colorful lines in Fig. 5.2 (a) labeled b1–b6 and solid white lines in Fig. 5.2 (b) – 5.2(d). Five boreholes represented by dash lines labeled v1–v5 are selected for verification. Fig. 5.2 compares the different predictions of the geological cross-section by the MSRF models, where

model SRF-7 with FS3 is considered as the optimal by the identification module.

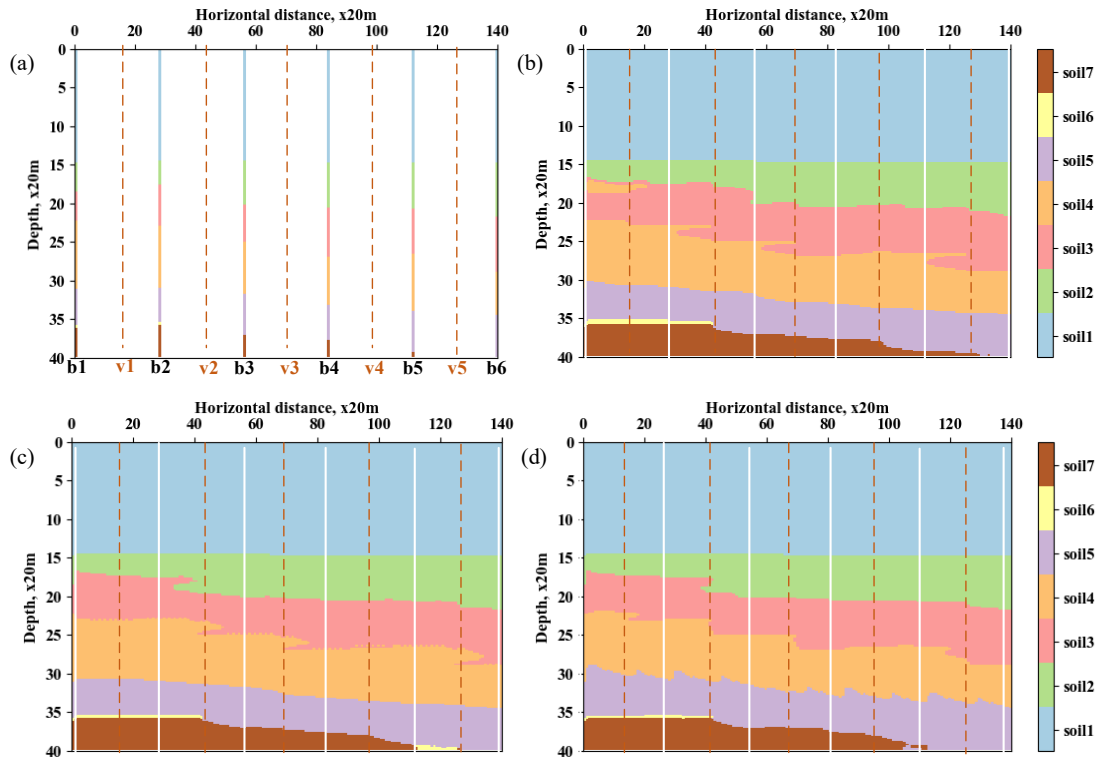


Fig. 5.2. Predictions of case 3 by MSRF XGBoost models with filter size 3: (a) measurement; (b) model SRF-3; (c) model SRF-5; (d) model SRF-7

Table 5.1 tabulates the measurement accuracy of the MSRF models for the verification boreholes. MMA 1 indicates the mean measurement accuracy of all the verification boreholes for each model, and MMA 2 represents the mean measurement accuracy of the various models for each verification borehole. Recommended model SRF-7 achieves the highest MMA 1. All the MMA 2 values are over 0.9, with the lowest value of 0.901 for borehole v1 and the highest value of 0.97 for borehole v4. High MMA 2 values illustrate that the MSRF XGBoost method can accurately predict the geological cross-section using only boreholes from practical geotechnical engineering data.

Table 5.1. Prediction accuracy of the models for verification boreholes

Model	Verification boreholes					MMA 1
	v1	v2	v3	v4	v5	
SRF-3	0.895	0.925	0.925	0.975	0.945	0.93
SRF-5	0.9	0.905	0.92	0.965	0.91	0.92
SRF-7	0.91	0.94	0.935	0.96	0.93	0.94
MMA 2	0.901	0.92	0.93	0.97	0.93	0.93

Table 2 lists the average relative difference between the measured thickness of the soil and that predicted by the MSRF models in the verification boreholes. The relative error is the ratio of the difference between the predicted and measured thicknesses of the soil to the measured thickness.

$$R = (P - M)/M \quad (5.1)$$

Where R is the relative error; P is the predicted thicknesses of the soil; M is the measured thickness of the soil.

A negative value indicates that the predicted thickness of the soil is smaller than the measurement. The relative error of soil 1 is the lowest (0.4%) in all the models, while that of soil 3 is the highest due to the irregular boundaries. Most average relative errors of all the soil layers predicted by the MSRF model are less than 20%.

Table 5.2. Average relative difference between the prediction and verification boreholes

Model	Average relative error (%)						
	Soil 1	Soil 2	Soil 3	Soil 4	Soil 5	Soil 6	Soil 7
SRF3	0.4	-23.8	23.4	05.6	-13.6	10.0	-19.2
SRF5	0.4	-18.6	27.4	10.0	-15.4	0	-25.2
SRF7	0.4	-15.4	21.6	-2.4	14.4	-10.0	-27.8

5.3 Engineering application based on novel GCN method

5.3.1 Borehole source and layout

This project is an underground complex located in the Fuyang district of Hangzhou, China. The length of the foundation is 260m, the width is 206m, and the excavation depth is 12.4m. The foundation consists of six different soil types, with the primary

soils being silt and silty clay. The total number of boreholes is one hundred and twenty. The borehole layout follows a rectangular grid, with twelve boreholes along the long side and ten boreholes along the short side. For analysis, two profiles from the short side, labeled IX and XI, are selected as Profile 1 and Profile 2, as shown in Fig. 5.3. The distance between profile 1 and profile 2 is similar, so the number of soil layers is also quite similar. But there are three interlaced soil layers in the last borehole of profile 2. Comparing the two cross-sections can better reflect the predictive ability of the improved GCN model.

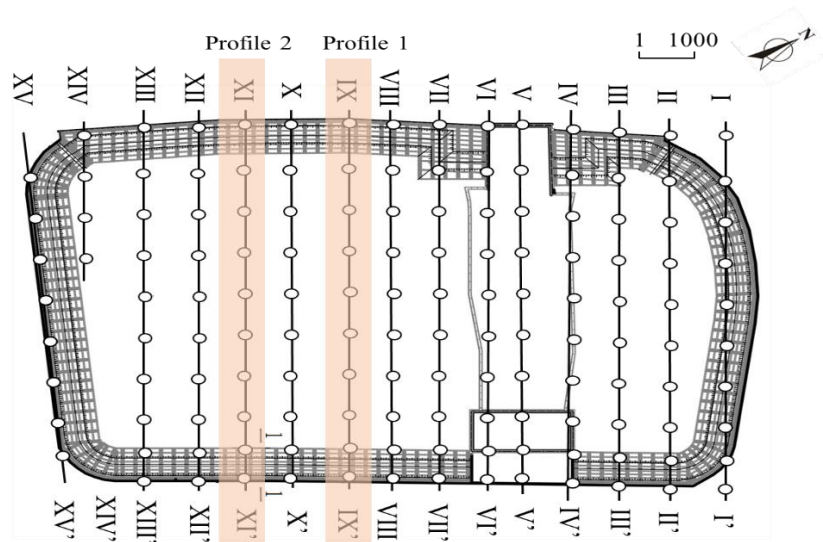


Fig. 5.3. Layout plan of boreholes in a Hangzhou foundation

5.3.2 Geological profile IX prediction

Three boreholes are selected from a total of ten to validate the proposed method. Given the various combinations available for selecting validating boreholes, three typical combinations are chosen. Fig. 5.4 illustrates predictions derived from different numbers of boreholes and varying positions. The white lines represent training boreholes labeled

b1 to b10, while the red dashed lines indicate validating boreholes, which differ in Figs. 5.4 (b), (c), and (d). Fig. 5.4 (a) presents the prediction based on ten boreholes. Fig. 5.4 (b) shows the predicted result using seven boreholes, with validating boreholes (v2, v4, and v6) all on the left side, positioned at 28 m, 86 m, and 144 m along the horizontal coordinate axis. Consequently, the right side of Fig. 5.4 (b) exhibits more spatial features, closely resembling the right side of Fig. 5.4 (a). Fig. 5.4 (c) presents the predicted result using seven boreholes, with validating boreholes (v3, v6, and v9) equally distributed along the profile, positioned at 57 m, 144 m, and 231 m on the horizontal coordinate axis, resulting in loss features also equally distributed along the profile. Fig. 5.4 (c) is generally similar to Fig. 5.4 (a). Fig. 5.4 (d) is the prediction based on seven boreholes, with validating boreholes (v5, v7, and v9) all on the right side, positioned at 155 m, 173 m, and 231 m on the horizontal coordinate axis. This results in loss of spatial features primarily on the right side, with the left side of Fig. 5.4 (d) closely resembling that of Fig. 5.4 (a). It is evident that a greater number of boreholes in the profile allows for more spatial information to be extracted, leading to a more accurate prediction of the profile.

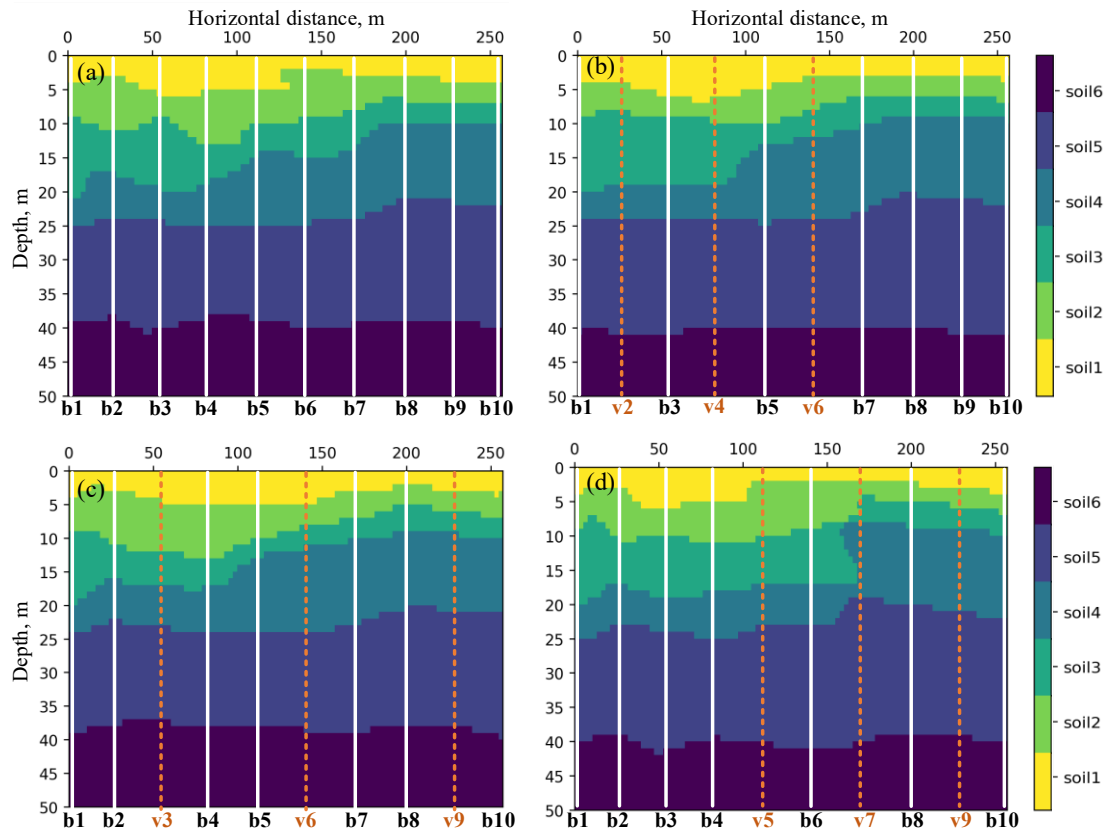


Fig. 5.4. Predictions of Hangzhou Profile 1: (a) ten boreholes (white); (b) seven boreholes (white); (c) seven boreholes (white); and (d) seven boreholes (white)

Table 5.3 summarizes the MA of different predictions for the validating boreholes. Mean MA refers to the mean measurement accuracy of all validating boreholes for each predicted profile. Based on Table 5.3, Fig. 5.4 (c) has the highest mean MA among all predictions, indicating that a uniform distribution of missing spatial features is preferable to a concentrated one. All MAs exceed 0.8, suggesting that the proposed method effectively predicts geological profiles based on unequally spaced boreholes.

Table 5.3. Measurement accuracy of the different predictions and validating boreholes

Validating borehole position	Measurement accuracy of the validating boreholes		
	Profile1(v2, v4, v6)	Profile1(v3, v6, v9)	Profile1(v5, v7, v9)
28 m	0.82	\	\

57 m	\	0.94	\
86 m	0.86	\	\
115 m	\	\	1
144 m	0.82	0.96	\
173 m	\	\	0.8
231 m	\	0.88	0.9
Mean MA	0.83	0.93	0.9

Table 5.4 presents the GA and BA between Fig. 5.4 (a), predicted by ten boreholes, and Figs. 5.4 (b), (c), and (d), predicted by seven boreholes with different positions. The GA and BA of the different predictions are very similar, with the GA near 0.89 and the BA near 0.51. However, Fig. 5.4 (c) is slightly better than the others, consistent with the conclusion from Table 1 that Fig. 5.4 (c) has the highest mean MA. In conclusion, using more training boreholes enhances the prediction of geological profiles when employing the proposed method, and the distribution of training boreholes significantly affects the accuracy of the prediction results. Finally, MA serves as a reliable criterion for evaluating predictions, based on the same optimal results recommended by Tables 5.3 and 5.4.

Table 5.4. Global accuracy and boundary accuracy between the different predictions and

Profile 1 (b1-b10)			
Accuracy criteria	Accuracy of Profile 1 (b1-b10)		
	Profile 1 (left loss three)	Profile 1 (middle loss three)	Profile 1 (right loss three)
Global accuracy	0.88	0.89	0.89
Boundary accuracy	0.512	0.527	0.526

5.3.3 Geological profile XI prediction

This case involves the XI profile of the Hangzhou foundation, which contains nine boreholes due to construction-related issues. A unique aspect of this profile is that the last borehole contains two parts of soil 3. Thus, the shape of soil 3 in the last borehole is predicted by spatial features derived from other boreholes. Two boreholes are selected as validating from the nine to test the efficacy of the proposed method, with three typical combinations chosen, as in geological profile IX. Fig. 5.5 illustrates predictions from different numbers of boreholes and varying positions. The white lines represent training boreholes labeled b1 to b9, while the red dashed lines indicate validating boreholes. Fig. 5.5 (a) presents the prediction based on nine boreholes. Fig. 5.5 (b) shows the predicted result using six boreholes, with validating boreholes (v2 and v4) all on the left side, positioned at 28 m and 86 m along the horizontal coordinate axis. Due to the loss of spatial features, Fig. 5.5 (b) presents a different boundary shape compared to Fig. 5.5 (a), particularly the shape of soil 3 in the last borehole. Fig. 5.5 (c) presents the predicted result using six boreholes, with validating boreholes (v4 and v6) equally distributed along the profile at 86 m and 144 m on the horizontal coordinate axis, resulting in loss features also equally distributed along the profile. Fig. 5.5 (c) is generally similar to Fig. 5.5 (a). Fig. 5.5 (d) presents the prediction based on six boreholes, with validating boreholes (v6 and v8) all on the right side, positioned at 144 m and 208 m on the horizontal coordinate axis, which also presents a different boundary shape compared to Fig. 5.5 (a).

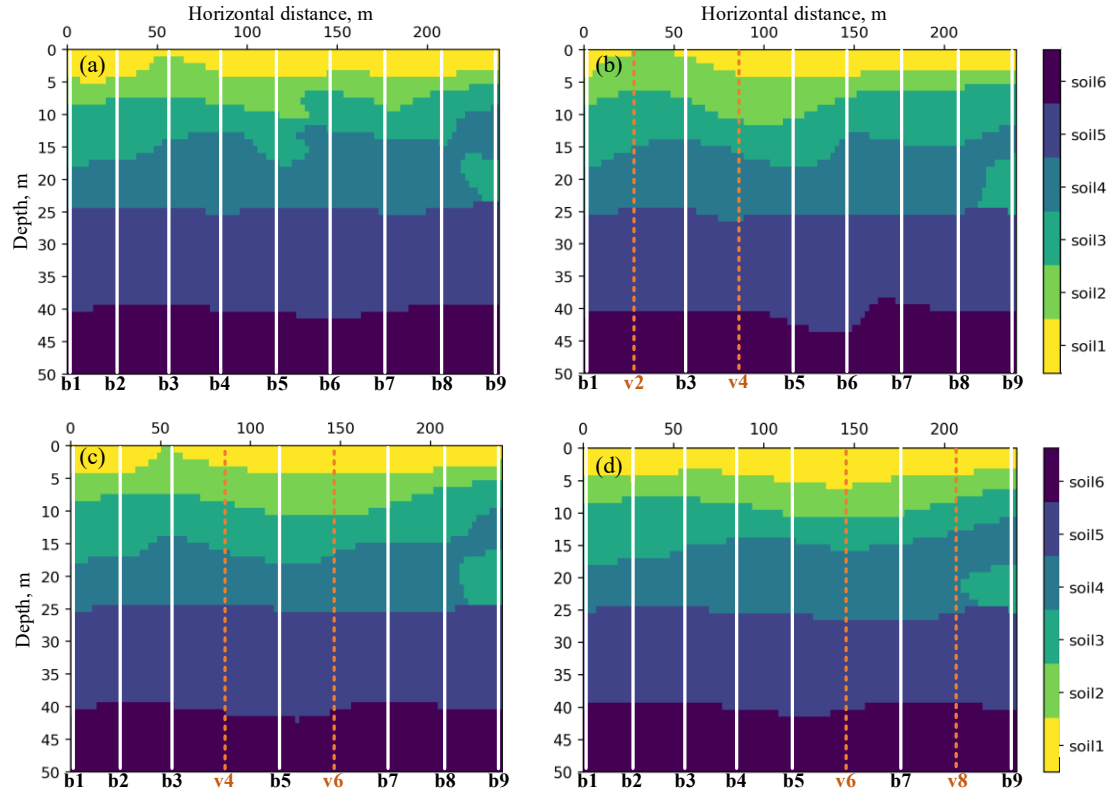


Fig. 5.5. Predictions of Hangzhou Profile 1: (a) nine boreholes (white); (b) six boreholes (white); (c) six boreholes (white); and (d) six boreholes (white)

Table 5.5 summarizes the MA of different predictions for the validating boreholes. Mean MA refers to the mean measurement accuracy of all validating boreholes for each predicted profile. Based on Table 5.5, Fig. 5.5 (c) exhibits higher MA compared to Fig. 5.5 (a) at v4 and Fig. 5.5 (d) at v6, indicating that a uniform distribution of missing spatial features is preferable to a concentrated one. Table 5.6 presents the GA and BA between Fig. 5.5 (a), predicted by nine boreholes, and Figs. 5.5 (b), (c), and (d), predicted by six boreholes with different positions. The GAs of the different predictions are very similar, approximately 0.89, but the BAs differ significantly, being 0.53, 0.65, and 0.58 for Figs. 5.5 (b), (c), and (d), respectively. In conclusion, the shape of the

stratigraphic interface depends on the spatial features of all boreholes, especially for soil layers that appear in only one borehole. MA is an effective method for evaluating predictions; however, too many validating boreholes can lead to a loss of spatial information. Therefore, careful consideration is necessary when selecting the number of validating boreholes.

Table 5.5. Measurement accuracy of the different predictions and validating boreholes

Validating borehole position	Measurement accuracy of the validating boreholes		
	Profile 2(v2, v4)	Profile 2(v4, v6)	Profile 2(v6, v8)
28 m	0.85	\	\
86 m	0.83	0.87	\
144 m	\	0.77	0.72
208 m	\	\	0.83
Mean MA	0.84	0.82	0.78

5.4 Comparison and analysis of prediction results of different machine learning algorithms

5.4.1 Comparative analysis of nonlinear geological profile prediction

The non-linear geological profile prediction case was used to validate the accuracy of both the MSRF-XGBoost algorithm and the improved GCN algorithm. By comparing the results of these algorithms, their respective strengths and weaknesses can be evaluated.

Figure 5.6 shows the prediction results of different algorithms for the non-linear geological profile: (a) the true geological profile, (b) the profile predicted by the MRF

method, (c) the profile predicted by the IC-XGBoost algorithm, (d) the profile predicted by the MSRF-XGBoost algorithm, and (e) the profile predicted by the improved GCN algorithm. In terms of global accuracy, boundary accuracy, and boundary fitting, the MSRF-XGBoost algorithm outperformed the other methods, showing significant improvements. However, the MSRF-XGBoost algorithm relies on uniformly spaced boreholes to learn spatial features and predict the complete geological profile, and its performance declines sharply for non-uniformly spaced boreholes. Aside from MSRF-XGBoost, the improved GCN algorithm also demonstrated improvements in global accuracy, boundary accuracy, and boundary fitting compared to the other two methods. The key advantage of the improved GCN algorithm lies in its ability to handle non-uniformly spaced boreholes effectively, making it particularly suitable for specific site conditions, such as missing or unevenly distributed boreholes.

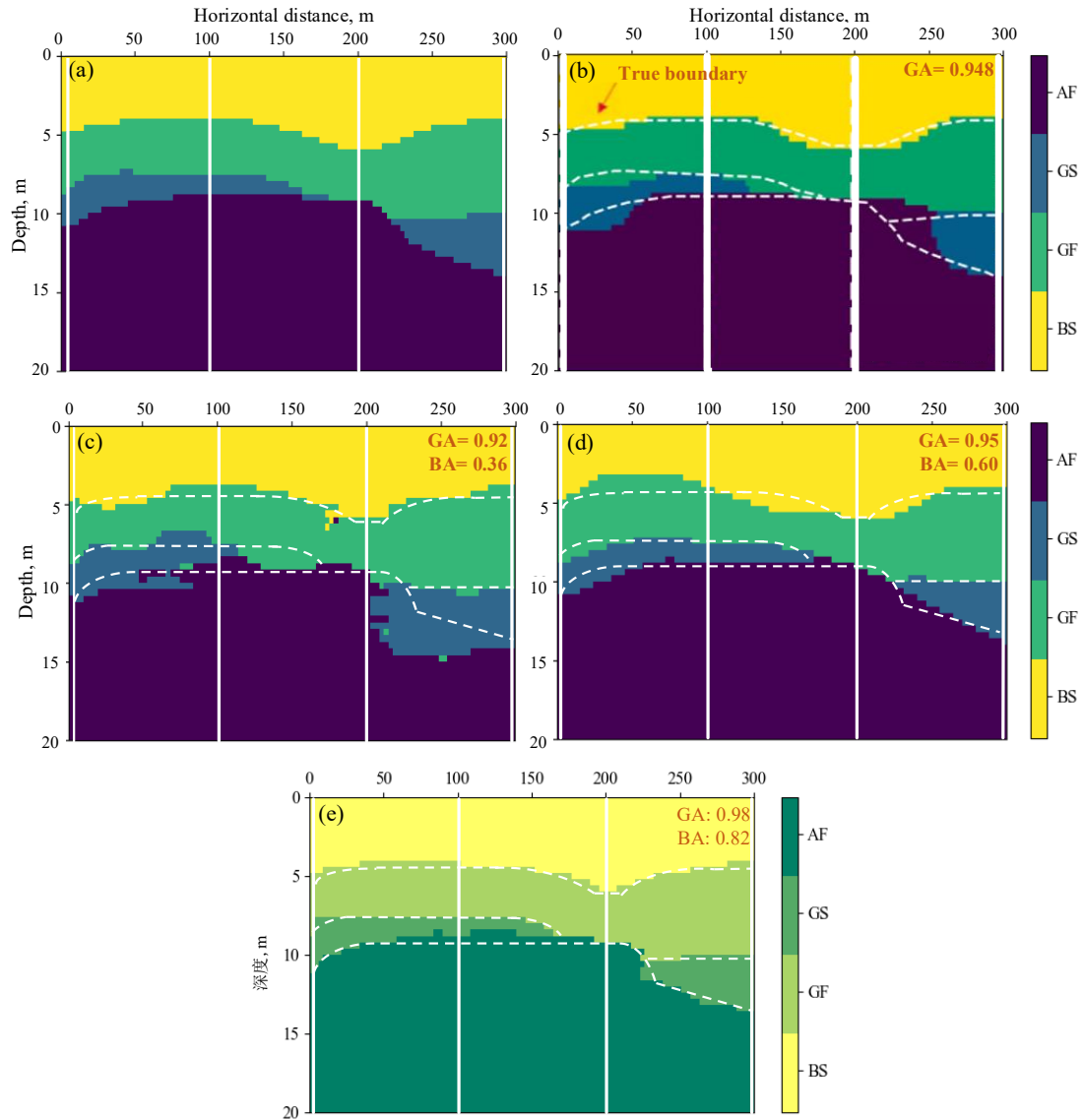


Figure 5.6 Nonlinear Geological Profile Prediction Results of Different Prediction

Algorithms (a) true geological profile (b) MRF method prediction (c) IC-XGBoost method prediction (d) novel GCN method prediction (e) MSRF-XGBoost method prediction

5.4.2 Comparative analysis of geological profile prediction for Hangzhou foundation pit engineering

A case study of a Hangzhou excavation project was conducted using seven boreholes to compare the prediction results of the MSRF-XGBoost algorithm and the improved

GCN algorithm. For a more accurate comparison, the seven boreholes were divided into two groups: four boreholes were used as training data to predict the geological profile, and the remaining three were used as validation boreholes to assess prediction accuracy. During the discretization process, the horizontal unit size was set to 3.5 meters to ensure the continuity of stratum boundaries in the horizontal direction, while the vertical unit size was fixed at 0.5 meters to accurately display stratum boundaries in each borehole. In Fig. 5.7 (a), (b), (c), and (d), the input boreholes are marked with solid white lines, and the validation boreholes are marked with red dashed lines. Since the MSRF-XGBoost algorithm performs better with uniformly spaced boreholes, the validation boreholes were selected at intervals to ensure uniform spacing for the training boreholes.

Fig. 5.7 (a) shows the geological profile predicted by the improved GCN model, while Fig. 5.7 (b), (c), and (d) display the profiles predicted by the SRF-3, SRF-5, and SRF-7 models, respectively. A comparison reveals that the MSRF-XGBoost algorithm produces smoother stratum boundaries than the improved GCN algorithm. Except for Stratum 3, the boundary shapes predicted by both algorithms are similar. If the prediction results are divided into left and right sections based on the position of validation borehole V2, the left sections of all predictions are highly consistent.

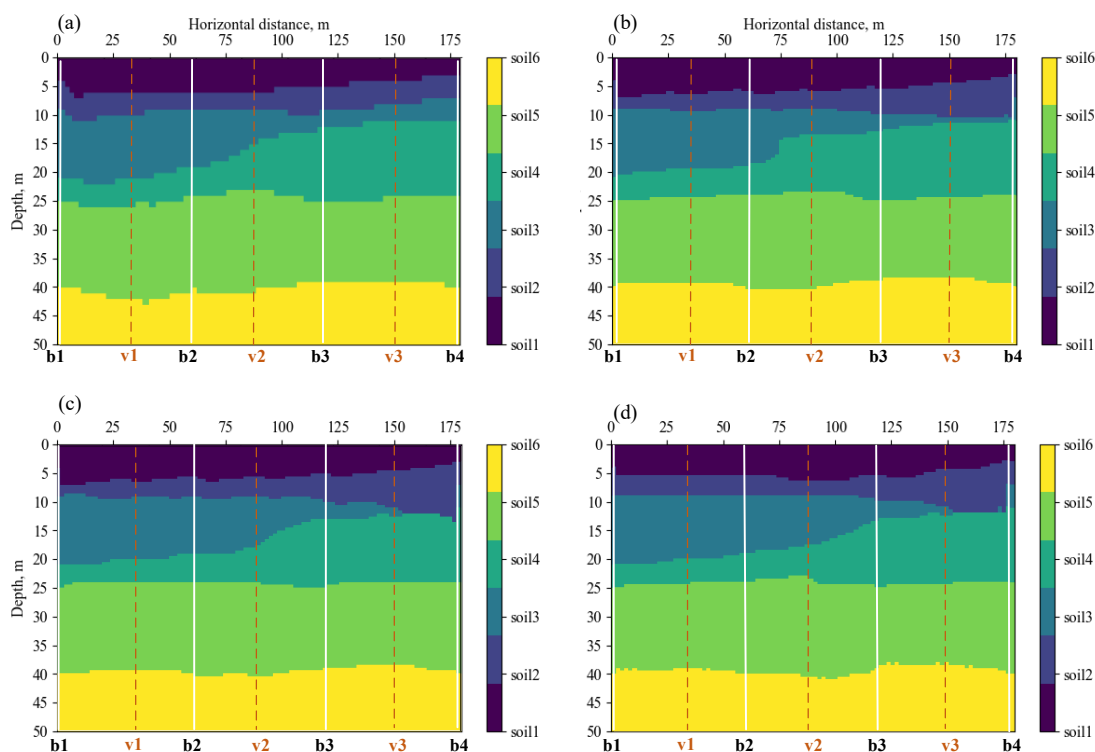


Fig. 5.7 Prediction results of actual engineering geological profiles using different algorithms

(a) Improved GCN algorithm prediction (b) SRF-3 model prediction (c) SRF-5 model prediction (optimal recommendation after screening) (d) SRF-7 model prediction

Table 5.7 presents the prediction accuracy of different SRF models and the improved GCN model for the three validation boreholes. The screening module identified SRF-5 as the optimal model, with the highest Mean Model Accuracy (MMA1) score of 0.82. In contrast, the improved GCN model had lower accuracy than all SRF models, indicating that SRF models perform better for uniformly spaced boreholes. Additionally, for the three SRF models, the Mean Model Accuracy (MMA2) scores for validation boreholes V1 and V2 were 0.83 and 0.84, respectively, while the score for V3 was only 0.74. This suggests that the stratigraphic features of V3 differ significantly from those of the training boreholes (B1–B4). Although the GCN algorithm underperformed the

SRF models for V1 and V2, it achieved slightly higher accuracy for V3, demonstrating its advantage in localized predictions.

Table 5.7 Prediction accuracy of different models for validation boreholes

Model	v1	v2	v3	MMA 1
SRF-3	0.82	0.78	0.73	0.78
SRF-5	0.83	0.88	0.75	0.82
SRF-7	0.83	0.85	0.74	0.81
MMA 2	0.83	0.84	0.74	0.81
GCN	0.72	0.76	0.76	0.75

In summary, for uniformly spaced geological profiles, the MSRF-XGBoost algorithm outperforms the improved GCN algorithm in terms of boundary smoothness and prediction accuracy. However, the MSRF-XGBoost algorithm lacks flexibility for non-uniformly spaced boreholes, whereas the improved GCN algorithm shows greater potential for broader applications.

5.5 Summary

The two algorithms were applied to multiple real-world projects, and their performance in predicting non-linear geological profiles was compared. The results demonstrate their practical value: (1) MSRF-XGBoost achieved an average accuracy of over 0.9 for five validation boreholes using six known boreholes.

(2) The improved GCN algorithm accurately predicted geological profiles for non-uniformly spaced boreholes in Hangzhou excavation sites, with validation accuracies ranging from 0.72 to 0.96.

(3) For uniformly spaced boreholes, MSRF-XGBoost outperformed other algorithms

in accuracy and boundary fitting, while the improved GCN showed moderate improvements.

(4) In comparative studies, MSRF-XGBoost achieved higher overall accuracy for uniformly spaced boreholes, but the improved GCN exhibited advantages in specific scenarios, such as uneven or missing borehole distributions.

CHAPTER 6 FINITE ELEMENT ANALYSIS OF FOUNDATION EXCAVATION BASED ON ACCURATE GEOLOGICAL PROFILE

6.1 Introduction

Foundation pit engineering is a core component of urban underground space development, and its deformation control directly affects the safety of surrounding buildings. The finite element analysis method integrates mathematical calculation techniques, mechanical analysis theories, and computer analysis technologies, playing a significant role in the field of geotechnical engineering. Many scholars have established constitutive models for geotechnical materials and designed calculation units for foundation pit support structures to simulate and analyze practical engineering issues such as deformation and stress of support structures, pit bottom heave, ground settlement outside the pit, and deformation of adjacent buildings ^[160-163]. Finite element analysis is implemented through finite element software, among which PLAXIS 2D, with its efficient plane strain analysis capability, has become an important tool for simulating the interaction between foundation excavation and support systems.

This chapter utilizes the accurate geological profile predicted by the machine learning method proposed earlier to establish a numerical model using PLAXIS 2D software, studying the influence of different shapes of stratum interfaces on the ground settlement changes at the edge of the foundation pit. The study is based on a foundation pit project in Hangzhou, analyzing the ground settlement changes during the excavation process.

The conclusions drawn can demonstrate the accuracy of the geological profile predicted by the machine learning method and provide a reference for similar foundation pit excavation cases in Hangzhou.

6.2 Model development

The numerical simulation software used in this chapter is PLAXIS 2D finite element software. Compared to other numerical software, PLAXIS 2D offers advantages such as an intuitive interface, simplified modeling processes, and clear data results. PLAXIS supports various classical or advanced soil constitutive models, enabling the simulation of geotechnical structures and construction processes under complex conditions, including steady (transient) seepage and fluid-solid coupling. It can also simulate interactions between soil and structures, making it widely used by researchers and geotechnical engineers.

6.2.1 Model dimensions and boundary conditions

As shown in Figures 6.1, the model is based on a foundation pit project in Hangzhou, establishing a 1:1 numerical model according to the project dimensions. The foundation pit is 260 m long, 206 m wide, and has an excavation depth of approximately 12.4 m.

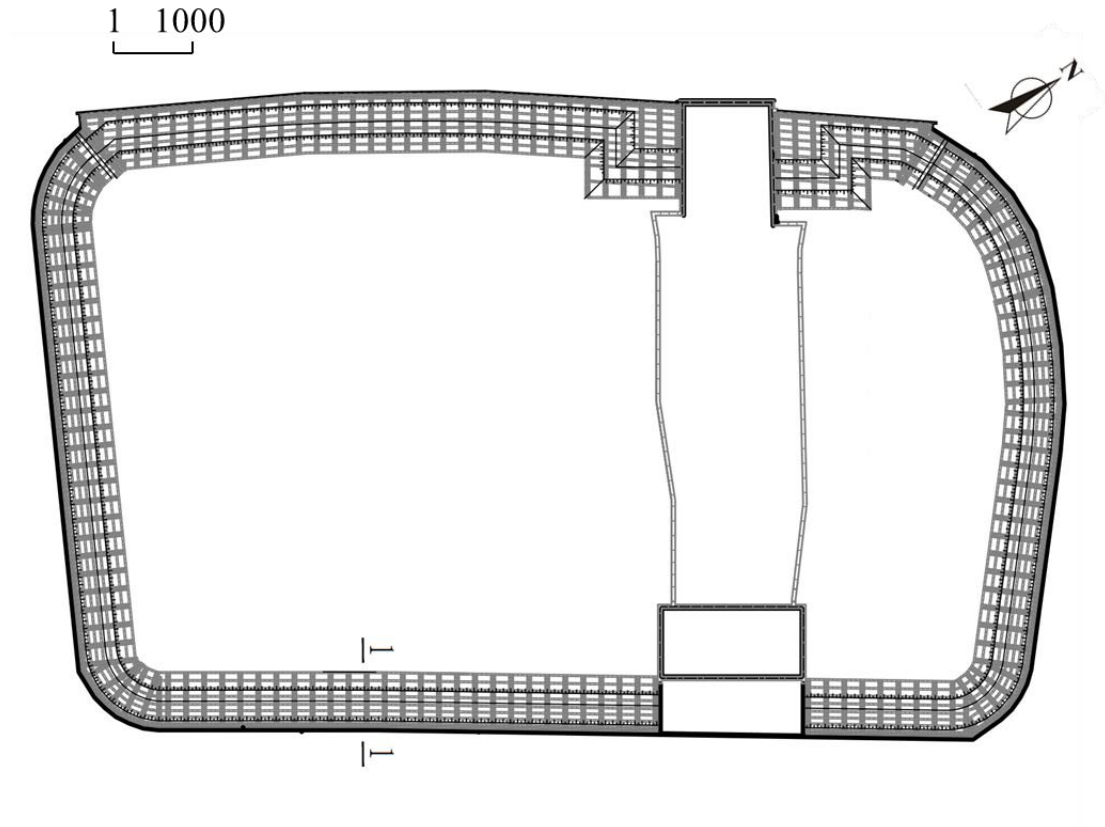


Fig 6.1 Plan view of the foundation pit

The pit is adjacent to the Fuchun River, with a developed water system, resulting in a significant influence of groundwater on the excavation. The site is predominantly located in the alluvial-marine plain area on both sides of the Fuchun River, influenced by marine transgression, forming a river island and piedmont plain composed of Quaternary mid-late alluvial-marine deposits. The riverbanks are mostly artificial flood control embankments. The site includes six soil types, listed from top to bottom as: (1)1 miscellaneous fill, (2)1 silty clay, (2)2 silt, (3)2 muddy silty clay, (4)1 silt, (4)2 silty clay, (4)3 fine silty sand, (5)1 gravelly soil, (5)2 pebble soil; (12)1 completely weathered quartz diorite, (12)2 strongly weathered quartz diorite, and (12)3 moderately weathered quartz diorite. The foundation pit base is located in the (5)2 pebble soil layer.

The retaining structure of the foundation pit consists of a diaphragm wall and an internal bracing system. The diaphragm wall is a reinforced concrete structure with a thickness of 1500 mm. The internal bracing system includes one reinforced concrete support and one steel support. The first support is an 800 mm × 800 mm reinforced concrete support, and the second support uses a steel pipe with a diameter of 609 mm and a thickness of 16 mm.

Since the foundation pit is approximately square, the central cross-section can be selected, combined with the two-dimensional geological profile predicted in Chapter 4, to establish a two-dimensional model for simulating the excavation process. Due to the symmetry of the pit, half of the central cross-section is modeled to reduce computational costs. The numerical model extends four times the excavation depth in the width direction and three times the excavation depth in the depth direction to avoid boundary effects. The dimensions of the two-dimensional finite element model are as follows: 160 m in the X (horizontal) direction and 44 m in the Y (vertical) direction.

Boundary conditions are assumed as follows:

1. The bottom of the model is fixed in the X, Y, and Z directions.
2. The two sides of the model are constrained only in the normal direction.
3. The top surface of the model is free, with no constraints.

6.2.2 Selection of soil and retaining structure parameters

The two-dimensional finite element soil model adopts the Hardening Soil model with small-strain stiffness (HSS). Geotechnical engineering practice shows that soil

deformation before failure is typically in the small-strain range, and the HSS model can distinguish between loading and unloading stiffness. Therefore, from a theoretical perspective, the HSS model is more suitable for simulating foundation pit excavation. The HSS model parameters are determined based on the geological survey report. The soil parameters used in the calculations are listed in Table 6.1.

Table 6.1 Physical parameters of soil layers

	Miscella neous Fil	Silt	Muddy Silty Clay	Pebble Soil	Quartz Diorite
Elevation/ <i>m</i>	-3.0	-11.2	-17.3	-22.0	-37.9
$\gamma/(kN/m^2)$	19.2	19.1	18.1	19.7	20.3
$E_{50}^{ref}/(kN/m^2)$	6000	6540	4272	8364	2400
$E_{oed}^{ref}/(kN/m^2)$	7200	5450	3560	6970	2000
$E_{ur}^{ref}/(kN/m^2)$	26400	23980	33100	30670	88000
c'/kPa	9.8	20.5	13.02	30.14	0
$\varphi'/(^\circ)$	18	14.5	11.6	18.6	30
ν	0.2	0.2	0.2	0.2	0.2
$G_0^{ref}/(kN/m^2)$	132000	119900	165500	153300	440000
$K/(m/d)$	1	0.01	0.0003	0.4	21.6

The behavior of bedrock differs significantly from that of soil, so the Hoek-Brown model is used to simulate the bedrock. The Hoek-Brown rock model parameters are determined based on the geological survey report. The specific parameters are listed in Table 6.2.

Table 6.2 Hoek-Brown model parameters for bedrock

CHAPTER 6 FINITE ELEMENT ANALYSIS OF FOUNDATION EXCAVATION BASED ON
ACCURATE GEOLOGICAL PROFILE

Elevation/m	$E/(kN/m^2)$	σ_{ci}/Kpa	GSI	m_i	μ
-49.8	1000000	37500	5	50	0.495

In the finite element model for dewatering and excavation of the foundation pit, the diaphragm wall is simulated using plate elements, and interface elements are created to achieve the effect of groundwater blockage. The reinforced concrete support and steel support are simulated using anchor elements with point-to-point connections. The dimensions of each element are consistent with the foundation pit cross-section. The specific calculation parameters are listed in Table 6.3.

Table 6.3 Calculation parameters of retaining structures

Retaining Structure	Element Type	$E/(kN/m^2)$	$EA/(kN/m^2)$	Installation Position/m
Diaphragm Wall	Plate Element	31×10^6	-	-
Reinforced Concrete Support	Anchor Element	-	30×10^6	-3
Steel Support	Anchor Element	-	20×10^6	-7.2

To replicate the actual construction sequence, the two-dimensional refined model adopts a layered excavation method, simulating dewatering and excavation while synchronously installing supports. The specific excavation steps are listed in Table 6.4.

Table 6.4 Excavation conditions

Condition	Construction Steps
1	Initial geostatic stress balance
2	Construction of the diaphragm wall and capping beam
3	Pre-dewatering to the pit bottom at -13.0 m
4	Excavation to -4.0 m, installation of the first concrete support
5	Excavation to -7.5 m, installation of the second steel support
6	Excavation to -12.4 m

6.2.3 Model validation

The horizontal displacement data of the diaphragm wall at the same location from both field monitoring and finite element model calculations are compared, and the results are plotted in Figure 6.2. As shown in Figure 6.2, the calculated and monitored horizontal displacements of the diaphragm wall exhibit a concave trend. The displacement at the top of the wall is small, increasing with depth and reaching its maximum near the pit bottom. Beyond this point, the displacement gradually decreases and eventually approaches zero. The trends of the calculated and monitored values are consistent, indicating that the numerical model effectively simulates the actual engineering scenario.

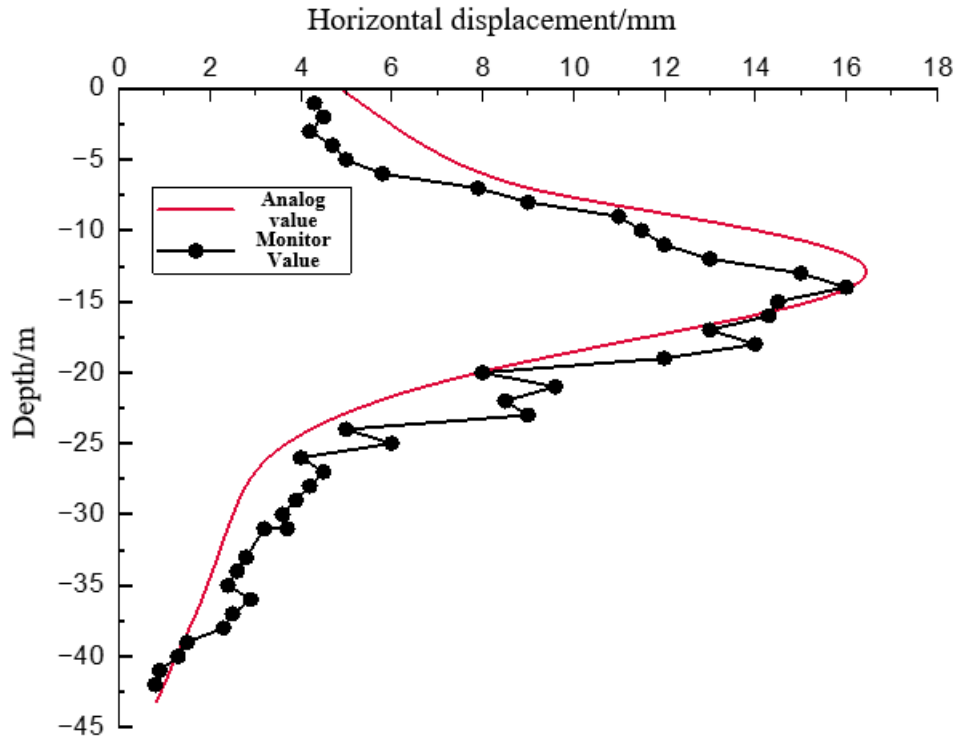


Fig 6.2 Deep horizontal displacement of the diaphragm wall

6.3 Simulation results

To better compare the influence of accurate stratum interfaces on the numerical model, two models are established for the foundation pit project based on different stratum interfaces. Figure 6.3(a) shows the linear stratum interface obtained by connecting five boreholes, while Figure 6.3(b) shows the nonlinear geological profile predicted by the GCN algorithm using the same five boreholes.

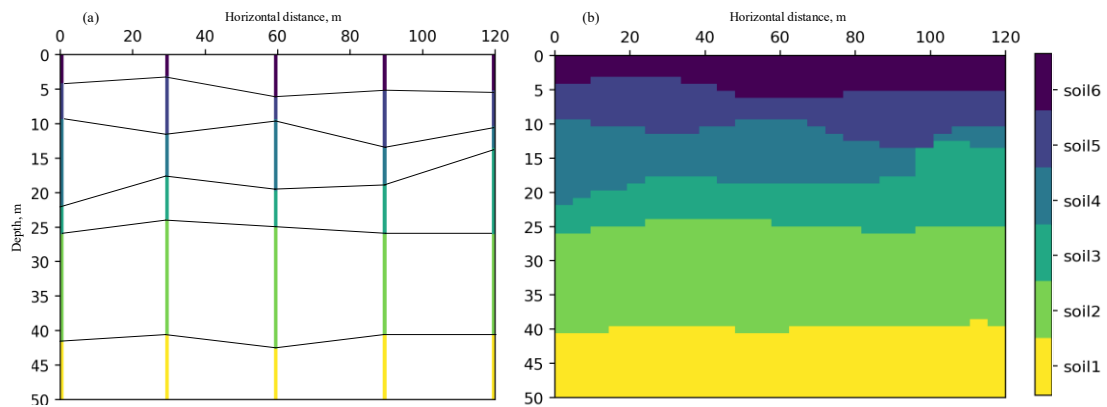


Fig 6.3 Comparison of different geological profiles

(a) linear geological profile, (b) nonlinear geological profile

A linear profile finite element model is established based on the linear geological profile formed by the five boreholes. A nonlinear profile finite element model is established based on the nonlinear geological profile predicted by the machine learning algorithm. The specific modeling profiles are shown in Figures 6.4 and 6.5. The figures demonstrate that the finite element models accurately reflect the differences between the linear and nonlinear geological profiles.

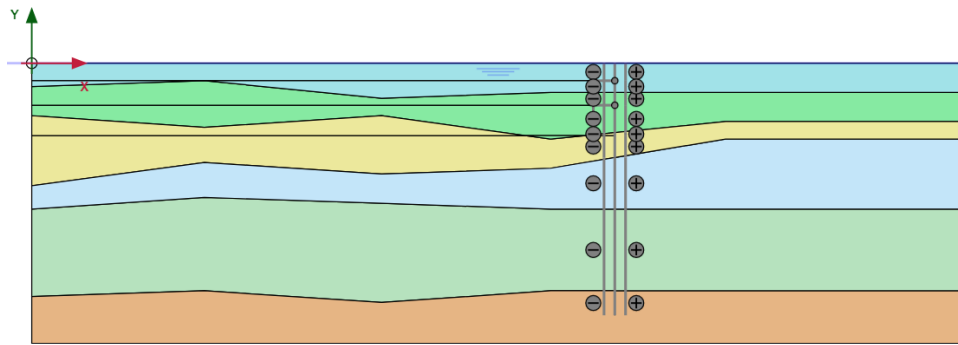


Fig 6.4 Finite element model with linear stratum interface

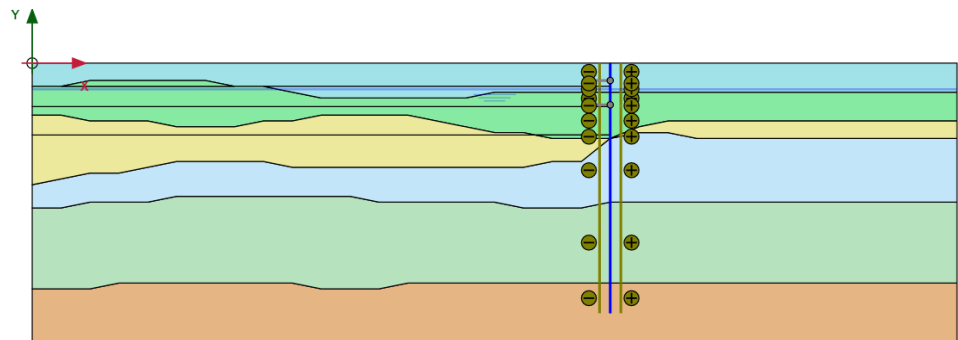


Fig 6.5 Finite element model with nonlinear stratum interface

The two models differ only in the number of boreholes, while other modeling structures

and physical parameters remain identical. The linear and nonlinear profile finite element models are meshed, as shown in Figures 6.6 and 6.7. The figures reveal that the more accurate stratum interface results in denser mesh divisions under the same coarseness, generating more nodes to adapt to the changes in the stratum interface.

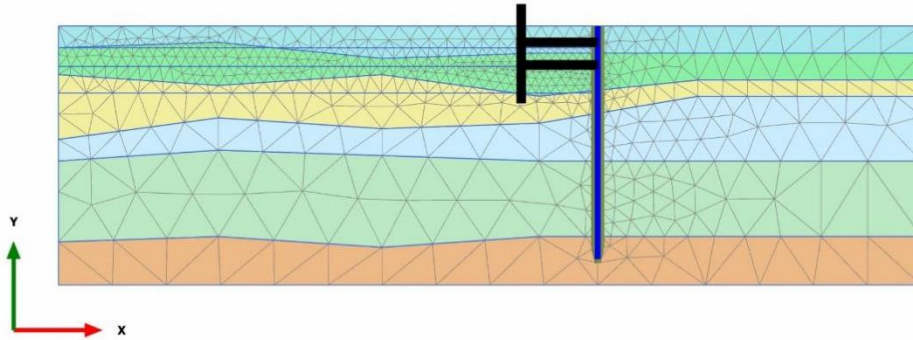


Fig 6.6 Mesh division of the linear profile finite element model

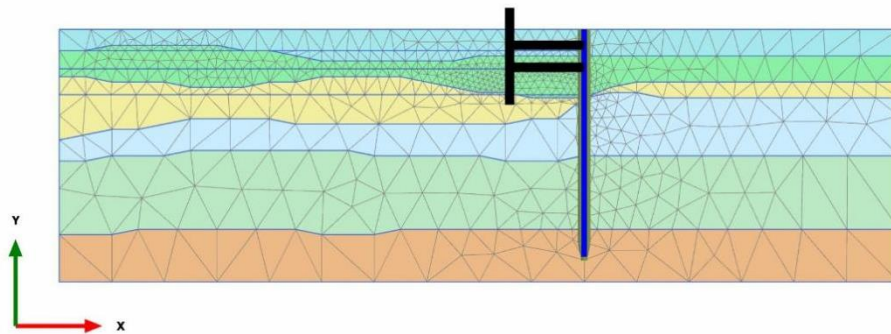


Fig 6.7 Mesh division of the nonlinear profile finite element model

6.3.1 Ground settlement during the first excavation layer

The first excavation layer corresponds to Condition 4, where the pit is excavated to -4 m, followed by the installation of the first concrete support. Figures 6.8 and 6.9 show the ground settlement before support installation. The figures indicate that the

settlement trends outside the pit are similar for both models, with the settlement first increasing and then decreasing. The maximum settlement values differ by only 0.2 mm. Since the excavated soil in this layer is primarily miscellaneous fill and does not involve significant changes in the stratum interface, the settlement differences between the two models are minimal.

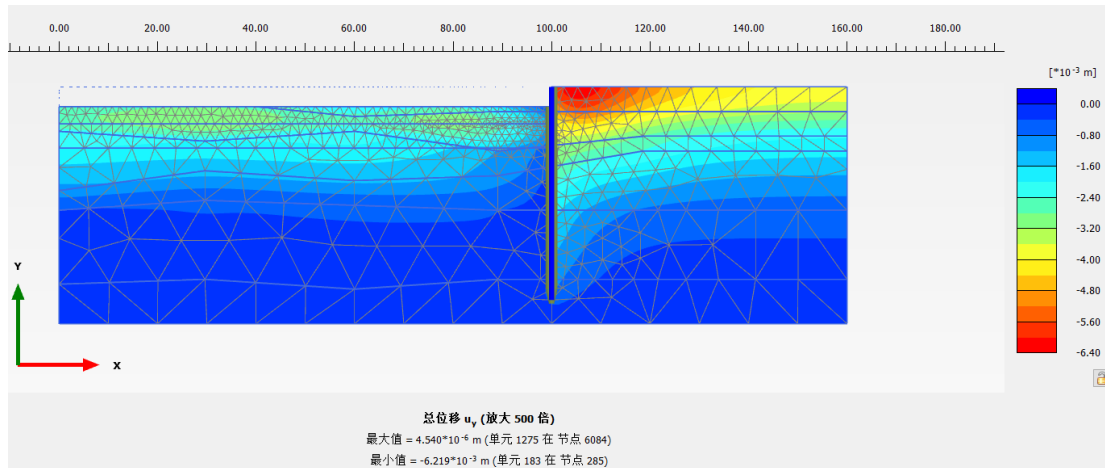


Fig 6.8 Ground settlement of the linear profile model during the first excavation layer

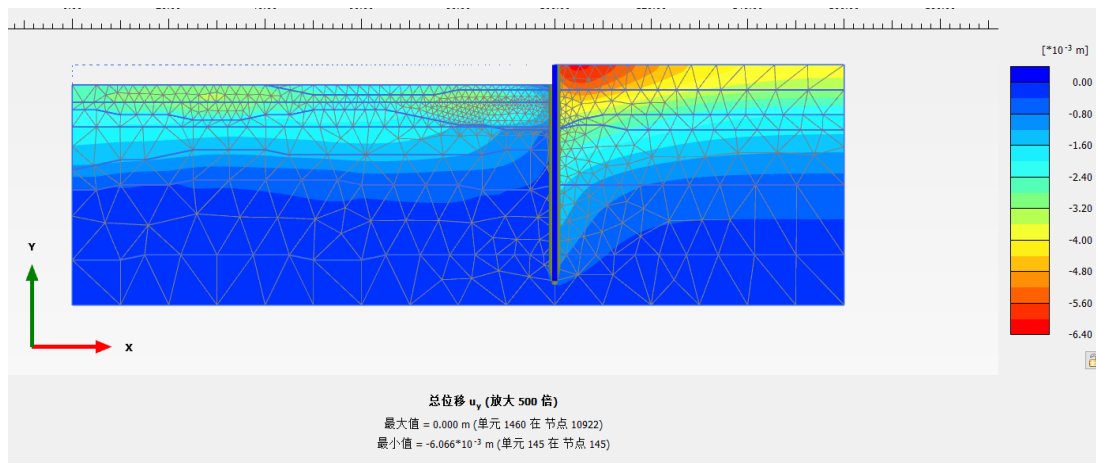


Fig 6.9 Ground settlement of the nonlinear profile model during the first excavation layer

6.3.2 Ground settlement during the second excavation layer

The second excavation layer corresponds to Condition 5, where the pit is excavated to -7.5 m, followed by the installation of the second steel support. Figures 6.10 and 6.11

show that the settlement trends outside the pit remain similar for both models, with a maximum settlement difference of 0.4 mm. The second excavation layer involves the second soil layer, and the settlement differences between the two models are only related to the interface between the first and second soil layers. Since this interface is similar in both models, the maximum settlement values are nearly identical.

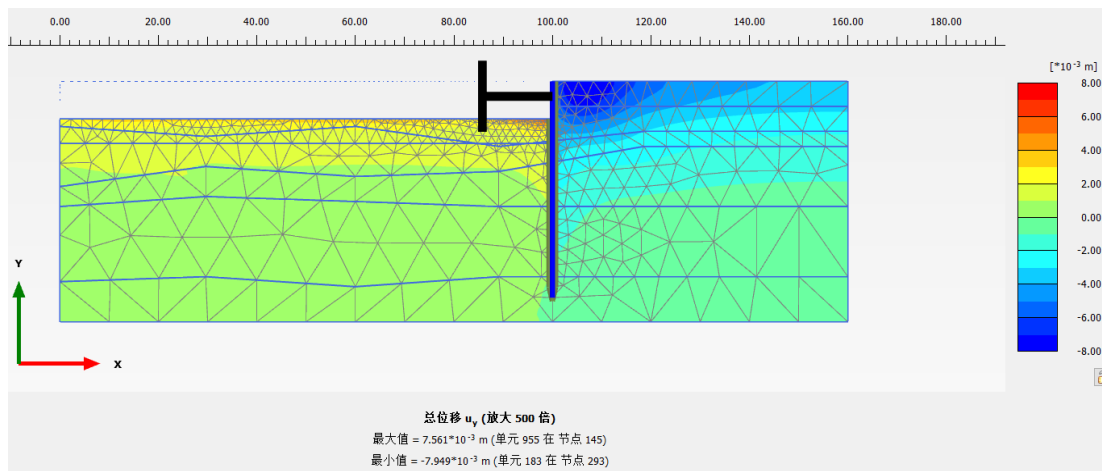


Fig 6.10 Ground settlement of the linear profile model during the second excavation layer

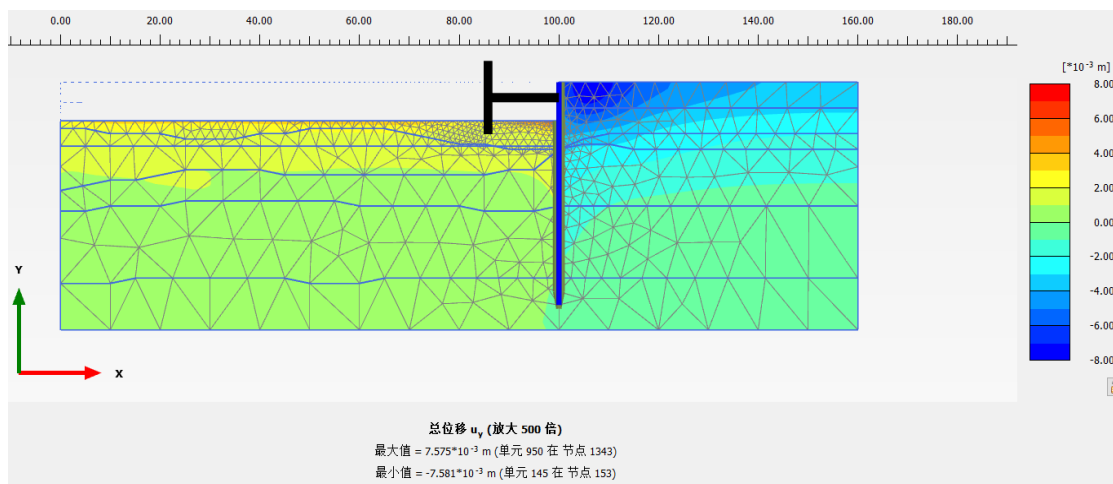


Fig 6.11 Ground settlement of the nonlinear profile model during the second excavation
layer

6.3.3 Ground settlement during the third excavation layer

The third excavation layer corresponds to Condition 6, where the pit is excavated to -12.4 m. Figures 6.12 and 6.13 show that the ground settlement outside the pit first increases and then decreases for both models. However, the linear profile model exhibits larger settlement values with a smaller influence range, while the nonlinear profile model shows smaller settlement values with a larger influence range. The maximum settlement value of the linear profile model is 2 mm larger than that of the nonlinear profile model. The mesh diagrams of the two models reveal significant differences in the interface between the second and third soil layers. Since the third excavation layer fully encompasses this interface, the settlement trends and maximum values differ notably between the two models.

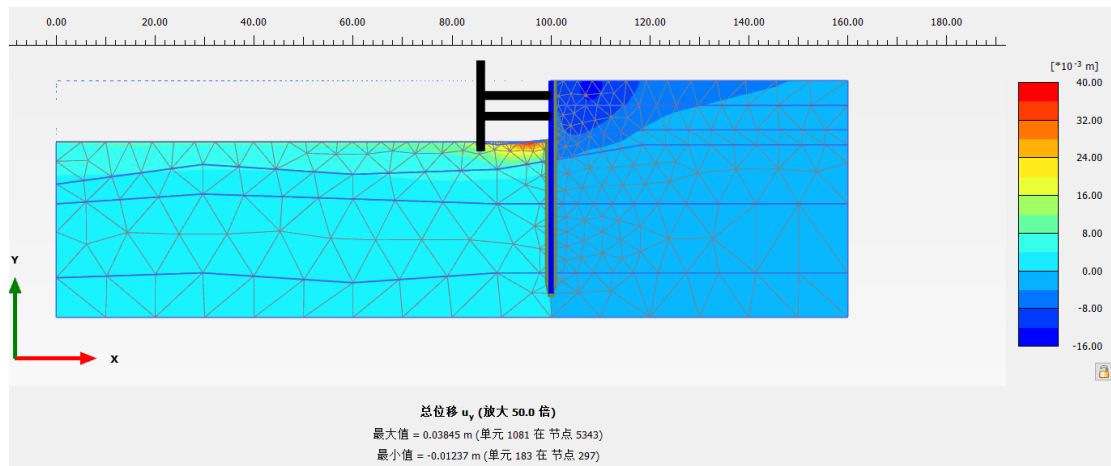


Fig 6.12 Ground settlement of the linear profile model during the third excavation layer

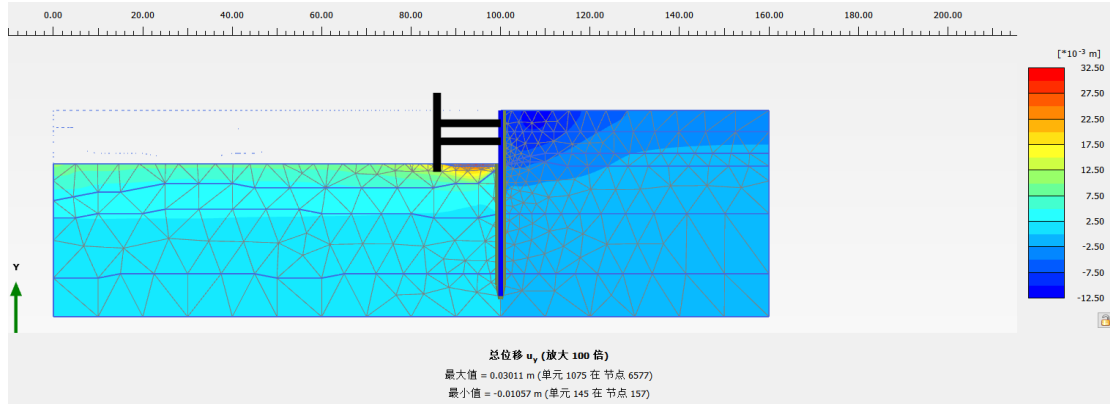


Fig 6.13 Ground settlement of the nonlinear profile model during the third excavation layer

6.3.4 Total ground settlement

Settlement monitoring points are evenly distributed around the foundation pit. The settlement values at the same location on the central cross-section are compared between the monitoring data and the two models. The results are shown in Figure 6.14. The figure indicates that the overall settlement at the pit edge is small, with the monitored settlement at the pit bottom being 8.6 mm, meeting the specification requirements. The settlement value of the linear profile model at the pit bottom is 11.6% larger than the monitored value, while the nonlinear profile model's value is only 2% smaller. Compared to the linear profile model, the nonlinear profile model's final settlement value is closer to the monitored value. Additionally, the settlement curve of the nonlinear profile model aligns better with the monitored data throughout the excavation process. This demonstrates that the nonlinear geological profile predicted by the machine learning algorithm improves the accuracy of the numerical model in simulating actual engineering scenarios.

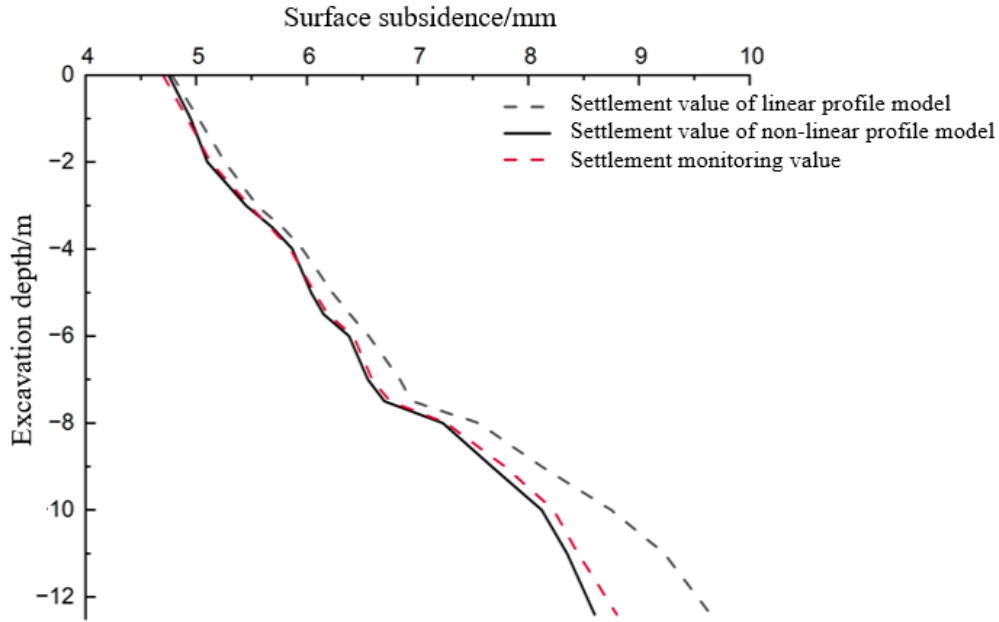


Fig 6.14 Relationship between ground settlement and excavation depth at the pit edge

6.4 Summary

Numerical models based on linear and GCN-predicted non-linear geological profiles were compared to assess their impact on excavation-induced surface settlements. Key findings include:

- (1) Both models exhibited similar trends in surface settlement changes during excavation. Horizontally, settlements increased and then decreased at the same excavation depth; vertically, maximum settlements increased with excavation depth.
- (2) Differences in maximum settlement values between the two models grew as excavation steps involved more stratigraphic interfaces. For example, differences were 0.2 mm and 0.4 mm in the first two steps but reached 2 mm in the final step.
- (3) Compared to field monitoring data, the non-linear profile model's final settlement values deviated by only 2%, while the linear profile model's deviation was 11.6%.

underscoring the importance of accurate geological profiles for realistic simulations.

CHAPTER 7 CONCLUSIONS AND FUTURE WORK

7.1 Findings and conclusions

In engineering practice, accurately predicting geological profiles and constructing reliable geological cross-sections to guide geotechnical design and construction using limited borehole data remains a critical technical challenge. This thesis proposes two machine learning algorithms, each with distinct advantages, to predict geological profiles with high accuracy using only sparse borehole data. These algorithms were applied to predict geological profiles for an actual excavation site, yielding favorable results and providing valuable support for engineering practice. Compared to existing machine learning methods, the proposed algorithms reduce the demand for training data while improving prediction accuracy. Finally, the machine learning-predicted geological profiles were incorporated into numerical models. A comparison between finite element models based on linearly interpolated geological profiles (using limited boreholes) and those based on non-linear profiles (predicted by machine learning) revealed significant differences in surface settlement predictions outside the excavation area. This highlights the importance of accurate geological profiles for finite element simulations.

The main contributions and conclusions of this thesis are summarized as follows:

1. The MSRF-XGBoost algorithm was proposed, which relies solely on limited borehole data to predict geological profiles with high accuracy. The results outperform

the existing IC-XGBoost method in terms of global accuracy, boundary accuracy, and reliability. Specific conclusions include:

(1) Unlike the IC-XGBoost method, which uses geological images as training data, the MSRF-XGBoost method extracts spatial features of strata directly from sparse borehole data.

(2) A screening module, based on Gaussian filtering and boundary similarity algorithms, automatically recommends the optimal prediction result from multiple outputs generated by different SRF models.

(3) Compared to IC-XGBoost, MSRF-XGBoost improves boundary prediction accuracy by over 0.4 for non-linear geological profiles and by 0.13 for linear profiles.

2. An improved Graph Convolutional Neural Network (GCN) algorithm was proposed, capable of predicting accurate geological profiles using non-uniformly spaced boreholes. Key conclusions include:

(1) The GCN eliminates the dependency on adjacent geological profiles and the requirement for uniformly spaced boreholes, enabling accurate predictions with sparse data.

(2) Compared to existing machine learning methods and literature cases, the GCN model improves boundary prediction accuracy by 24%.

(3) Prediction accuracy increases with the number of boreholes, but plateaus if additional boreholes provide no new spatial information.

(4) The location of boreholes significantly impacts accuracy. For the same number of

boreholes, strategically placed ones yield higher accuracy.

3. The two algorithms were applied to multiple real-world projects, and their performance in predicting non-linear geological profiles was compared. The results demonstrate their practical value: (1) MSRF-XGBoost achieved an average accuracy of over 0.9 for five validation boreholes using six known boreholes.

(2) The improved GCN algorithm accurately predicted geological profiles for non-uniformly spaced boreholes in Hangzhou excavation sites, with validation accuracies ranging from 0.72 to 0.96.

(3) For uniformly spaced boreholes, MSRF-XGBoost outperformed other algorithms in accuracy and boundary fitting, while the improved GCN showed moderate improvements.

(4) In comparative studies, MSRF-XGBoost achieved higher overall accuracy for uniformly spaced boreholes, but the improved GCN exhibited advantages in specific scenarios, such as uneven or missing borehole distributions.

4. Numerical models based on linear and GCN-predicted non-linear geological profiles were compared to assess their impact on excavation-induced surface settlements. Key findings include:

(1) Both models exhibited similar trends in surface settlement changes during excavation. Horizontally, settlements increased and then decreased at the same excavation depth; vertically, maximum settlements increased with excavation depth.

(2) Differences in maximum settlement values between the two models grew as

excavation steps involved more stratigraphic interfaces. For example, differences were 0.2 mm and 0.4 mm in the first two steps but reached 2 mm in the final step.

(3) Compared to field monitoring data, the non-linear profile model's final settlement values deviated by only 2%, while the linear profile model's deviation was 11.6%, underscoring the importance of accurate geological profiles for realistic simulations.

7.2 Recommendations for future studies

Due to limitations in the author's expertise and resources, this study leaves room for improvement in predicting accurate geological profiles from limited borehole data. Future research could explore the following directions:

(1) While the proposed algorithms rely solely on borehole data, other field investigation data (e.g., geophysical surveys, CPT data) could be integrated to enhance prediction accuracy through multi-source data fusion.

(2) The current algorithms focus on geological profile prediction. Future work could extend them to predict spatial variations in soil mechanical parameters (e.g., cohesion, friction angle) by incorporating in-situ and laboratory test data.

(3) The algorithms are limited to 2D geological profiles. Expanding them to 3D geological modeling using borehole data and integrating these models with 3D finite element simulations would enable more comprehensive engineering analyses.

REFERENCES

- [1] Xu C , Wang L , Tien Y M ,et al. Robust design of rock slopes with multiple failure modes: modeling uncertainty of estimated parameter statistics with fuzzy number. *Environmental Earth Sciences*, 2014, 72(8):2957-2969.
- [2] Luo Z, Di H, Kamalzare M, et al. Effects of soil spatial variability on structural reliability assessment in excavations. *Underground Space*, 2020, 5(1): 71-83.
- [3] Nguyen T S, Likitlersuang S. Influence of the spatial variability of soil shear strength on deep excavation: A case study of a Bangkok underground MRT station. *International Journal of Geomechanics*, 2021, 21(2): 04020248.
- [4] Prezzi M, McCullouch B, Mohan V K D. Analysis of change orders in geotechnical engineering work at INDOT. 2011.
- [5] Clayton C R I. Managing geotechnical risk: improving productivity in UK building and construction. Thomas Telford, 2001.
- [6] Dasaka S M, Zhang L M. Spatial variability of in situ weathered soil. *Géotechnique*, 2012, 62(5): 375-384.
- [7] Aswar D S, Ullagaddi P B. An overview of 3-D geological modelling part II. Summary of major 3-d geological modelling methodologies. *International Journal of Latest Engineering and Management Research*, 2017, 2(11): 15-27.
- [8] Shi C, Wang Y. Nonparametric and data-driven interpolation of subsurface soil stratigraphy from limited data using multiple point statistics. *Canadian Geotechnical Journal*, 2021, 58(2): 261-280.
- [9] Gong W , Zhao C , Juang C H ,et al. Coupled characterization of stratigraphic and geo-properties uncertainties – A conditional random field approach. *Engineering Geology*, 2021(8):106348.
- [10] Ching J, Phoon K K. Constructing a site-specific multivariate probability distribution using sparse, incomplete, and spatially variable (MUSIC-X) data. *Journal of Engineering Mechanics*, 2020, 146(7): 04020061.
- [11] Zhao C, Gong W, Li T, et al. Probabilistic characterization of subsurface stratigraphic configuration with modified random field approach. *Engineering Geology*, 2021, 288: 106138.
- [12] Qi X H, Li D Q, Phoon K K, et al. Simulation of geologic uncertainty using coupled Markov chain. *Engineering geology*, 2016, 207: 129-140.

- [13] Qi X, Wang H, Chu J, et al. Two-dimensional prediction of the interface of geological formations: A comparative study. *Tunnelling and Underground Space Technology*, 2022, 121: 104329.
- [14] Qi X, Wang H, Chu J, et al. Effect of autocorrelation function model on spatial prediction of geological interfaces. *Canadian Geotechnical Journal*, 2022, 59(4): 583-600.
- [15] Li D Q, Qi X H, Cao Z J, et al. Evaluating slope stability uncertainty using coupled Markov chain. *Computers and Geotechnics*, 2016, 73: 72-82.
- [16] Li X, Wu L, Wang Y, et al. A data driven real-time perception method of rock condition in TBM construction. *Canadian Geotechnical Journal*, 2023, 61(5): 1018-1034.
- [17] Hicks M A, Nuttall J D, Chen J. Influence of heterogeneity on 3D slope reliability and failure consequence. *Computers and Geotechnics*, 2014, 61: 198-208.
- [18] Deng Z P, Li D Q, Qi X H, et al. Reliability evaluation of slope considering geological uncertainty and inherent variability of soil parameters. *Computers and Geotechnics*, 2017, 92: 121-131.
- [19] Li D Q, Qi X H, Cao Z J, et al. Evaluating slope stability uncertainty using coupled Markov chain. *Computers and Geotechnics*, 2016, 73: 72-82.
- [20] Wang Y, Cao Z. Probabilistic characterization of Young's modulus of soil using equivalent samples. *Engineering Geology*, 2013, 159: 106-118.
- [21] Qi X H, Liu H X. Estimation of autocorrelation distances for in-situ geotechnical properties using limited data. *Structural Safety*, 2019, 79: 26-38.
- [22] Liu Y, Lee F H, Quek S T, et al. Effect of spatial variation of strength and modulus on the lateral compression response of cement-admixed clay slab. *Géotechnique*, 2015, 65(10): 851-865.
- [23] Ching J, Wu T J, Stuedlein A W, et al. Estimating horizontal scale of fluctuation with limited CPT soundings. *Geosci. Front.* 9 (6), 1597–1608.(2018)
- [24] Zhao C, Gong W, Li T, et al. Probabilistic characterization of subsurface stratigraphic configuration with modified random field approach. *Engineering Geology*, 2021, 288: 106138.
- [25] Gong W, Tang H, Juang C H, et al. Optimization design of stabilizing piles in slopes considering spatial variability. *Acta Geotechnica*, 2020, 15: 3243-3259.
- [26] Wang X. Uncertainty quantification and reduction in the characterization of subsurface stratigraphy using limited geotechnical investigation data. *Underground Space*, 2020, 5(2): 125-143.

- [27] Wang H, Wellmann J F, Li Z, et al. A segmentation approach for stochastic geological modeling using hidden Markov random fields. *Mathematical Geosciences*, 2017, 49: 145-177.
- [28] Zhu H, Zhang L M. Characterizing geotechnical anisotropic spatial variations using random field theory. *Canadian Geotechnical Journal*, 2013, 50(7): 723-734.
- [29] Wellmann J F, Thiele S T, Lindsay M D, et al. pynoddy 1.0: an experimental platform for automated 3-D kinematic and potential field modelling. *Geoscientific Model Development*, 2016, 9(3): 1019-1035.
- [30] Wang X, Wang H, Liang R Y, et al. A hidden Markov random field model based approach for probabilistic site characterization using multiple cone penetration test data. *Structural Safety*, 2018, 70: 128-138.
- [31] Wang X, Li Z, Wang H, et al. Probabilistic analysis of shield-driven tunnel in multiple strata considering stratigraphic uncertainty. *Structural safety*, 2016, 62: 88-100.
- [32] Liao T, Mayne P W. Stratigraphic delineation by three-dimensional clustering of piezocone data. *Georisk*, 1 (2), 102-119.(2007)
- [33] Wei X, Wang H. Stochastic stratigraphic modeling using Bayesian machine learning. *Engineering Geology*, 2022, 307: 106789.
- [34] Zhang J Z, Huang H W, Zhang D M, et al. Quantitative evaluation of geological uncertainty and its influence on tunnel structural performance using improved coupled Markov chain. *Acta Geotechnica*, 2021, 16: 3709-3724.
- [35] Jiang S H, Papaioannou I, Straub D. Bayesian updating of slope reliability in spatially variable soils with in-situ measurements. *Engineering Geology*, 2018, 239: 310-320.
- [36] Jiang S H, Huang J, Qi X H, et al. Efficient probabilistic back analysis of spatially varying soil parameters for slope reliability assessment. *Engineering Geology*, 2020, 271: 105597.
- [37] Gao X, Yan E C, Yeh T C J, et al. Sequential back analysis of spatial distribution of geomechanical properties around an unlined rock cavern. *Computers and Geotechnics*, 2018, 99: 177-190.
- [38] Liu K, Vardon P J, Hicks M A. Sequential reduction of slope stability uncertainty based on temporal hydraulic measurements via the ensemble Kalman filter. *Computers and Geotechnics*, 2018, 95: 147-161.
- [39] Wang, Y., & Zhao, T. Statistical interpretation of soil property cross-sections from sparse data using Bayesian compressive sampling. *Géotechnique*, 2017), 67(6), 523-536.

- [40] Rahman M H, Abu-Farsakh M Y, Jafari N. Generation and evaluation of synthetic cone penetration test (CPT) data using various spatial interpolation techniques. *Canadian Geotechnical Journal*, 2021, 58(2): 224-237.
- [41] Yin K S, Zhang L M, Wang H J, et al. Marine soil behaviour classification using piezocone penetration test (CPTu) and borehole records. *Canadian Geotechnical Journal*, 2021, 58(2): 190-199.
- [42] Zhao T, Hu Y, Wang Y. Statistical interpretation of spatially varying 2D geo-data from sparse measurements using Bayesian compressive sampling. *Engineering Geology*, 2018, 246: 162-175.
- [43] Tang X S, Li D Q, Cao Z J, et al. Impact of sample size on geotechnical probabilistic model identification. *Computers and Geotechnics*, 2017, 87: 229-240.
- [44] Wang Y, Zhao T. Interpretation of soil property profile from limited measurement data: a compressive sampling perspective. *Canadian Geotechnical Journal*, 2016, 53(9): 1547-1559.
- [45] Hu Y, Wang Y, Zhao T, et al. Bayesian supervised learning of site-specific geotechnical spatial variability from sparse measurements. *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering*, 2020, 6(2): 04020019.
- [46] Phoon K K. Role of reliability calculations in geotechnical design. *Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards*, 2017, 11(1): 4-21.
- [47] Wang Y, Zhao T, Hu Y, et al. Simulation of random fields with trend from sparse measurements without detrending. *Journal of Engineering Mechanics*, 2019, 145(2): 04018130.
- [48] Guan Z, Wang Y. Data-driven simulation of two-dimensional cross-correlated random fields from limited measurements using joint sparse representation. *Reliability Engineering & System Safety*, 2023, 238: 109408.
- [49] Chen D, Xu D, Ren G, et al. Simulation of cross-correlated non-Gaussian random fields for layered rock mass mechanical parameters. *Computers and Geotechnics*, 2019, 112: 104-119.
- [50] Alibeikloo M, Khabbaz H, Fatahi B, et al. Reliability assessment for time-dependent behaviour of soft soils considering cross correlation between viscoplastic model parameters. *Reliability Engineering & System Safety*, 2021, 213: 107680.

- [51] Huang Y, Beck J L, Wu S, et al. Robust Bayesian compressive sensing for signals in structural health monitoring. *Computer-Aided Civil and Infrastructure Engineering*, 2014, 29(3): 160-179.
- [52] Gangone M V, Whelan M J, Janoyan K D. Wireless monitoring of a multispan bridge superstructure for diagnostic load testing and system identification. *Computer-Aided Civil and Infrastructure Engineering*, 2011, 26(7): 560-579.
- [53] Ji S, Xue Y, Carin L. Bayesian compressive sensing. *IEEE Transactions on signal processing*, 2008, 56(6): 2346-2356.
- [54] Lyu B, Hu Y, Wang Y. Data-driven development of three-dimensional subsurface models from sparse measurements using Bayesian compressive sampling: A benchmarking study. *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering*, 2023, 9(2): 04023010.
- [55] Shuku T, Phoon K K. Three-dimensional subsurface modeling using Geotechnical Lasso. *Computers and Geotechnics*, 2021, 133: 104068.
- [56] Bhagoji A N, Cullina D, Sitawarin C, et al. Enhancing robustness of machine learning systems via data transformations, 2018 52nd Annual Conference on Information Sciences and Systems (CISS). IEEE, 2018: 1-5.
- [57] Li P, Wang Y, Guan Z. Non-parametric generation of multivariate cross-correlated random fields directly from sparse measurements using Bayesian compressive sensing and Markov chain Monte Carlo simulation. *Stochastic Environmental Research and Risk Assessment*, 2023, 37(12): 4607-4628.
- [58] Montoya-Noguera S, Zhao T, Hu Y, et al. Simulation of non-stationary non-Gaussian random fields from sparse measurements using Bayesian compressive sampling and Karhunen-Loève expansion. *Structural Safety*, 2019, 79: 66-79.
- [59] Ribeiro A M T, Ribeiro Junior P J, Bonat W H. A Kronecker-based covariance specification for spatially continuous multivariate data. *Stochastic Environmental Research and Risk Assessment*, 2022, 36(12): 4087-4102.
- [60] Shi C, Wang Y. Development of subsurface geological cross-section from limited site-specific boreholes and prior geological knowledge using iterative convolution XGBoost. *Journal of Geotechnical and Geoenvironmental Engineering*, 2021, 147(9): 04021082.
- [61] Shi C, Wang Y. Training image selection for development of subsurface geological cross-section by conditional simulations. *Engineering Geology*, 2021, 295: 106415.

- [62] Abdollahifard M J, Baharvand M, Mariéthoz G. Efficient training image selection for multiple-point geostatistics via analysis of contours. *Computers & Geosciences*, 2019, 128: 41-50.
- [63] Shi C, Wang Y. Machine learning of three-dimensional subsurface geological model for a reclamation site in Hong Kong. *Bulletin of Engineering Geology and the Environment*, 2022, 81(12): 504.
- [64] Han L, Wang L, Zhang W, et al. Rockhead profile simulation using an improved generation method of conditional random field. *Journal of Rock Mechanics and Geotechnical Engineering*, 2022, 14(3): 896-908.
- [65] Yeung A T. Geotechnical works of the Hong Kong-Zhuhai-Macao bridge project. *Japanese Geotechnical Society Special Publication*, 2016, 2(2): 109-121.
- [66] Shi C, Wang Y. Assessment of reclamation-induced consolidation settlement considering stratigraphic uncertainty and spatial variability of soil properties. *Canadian Geotechnical Journal*, 2022, 59(7): 1215-1230.
- [67] Houmadi Y, Benmoussa M Y C, Cherifi W N E H, et al. Probabilistic analysis of consolidation problems using subset simulation. *Computers and Geotechnics*, 2020, 124: 103612.
- [68] Qin Y, Zhu F, Xu D. Effect of the spatial variability of soil parameters on the deformation behavior of excavated slopes. *Computers and Geotechnics*, 2021, 136: 104246.
- [69] Qi X, Wang H, Pan X, et al. Prediction of interfaces of geological formations using the multivariate adaptive regression spline method. *Underground Space*, 2021, 6(3): 252-266.
- [70] Qi X, Wang H, Chu J, et al. Two-dimensional prediction of the interface of geological formations: A comparative study. *Tunnelling and Underground Space Technology*, 2022, 121: 104329.
- [71] Tian H, Wang Y. Data-driven and physics-informed Bayesian learning of spatiotemporally varying consolidation settlement from sparse site investigation and settlement monitoring data. *Computers and Geotechnics*, 2023, 157: 105328.
- [72] Grieves M, Vickers J. Digital twin: Mitigating unpredictable, undesirable emergent behavior in complex systems. *Transdisciplinary perspectives on complex systems: New findings and approaches*, 2017: 85-113.
- [73] Depina I, Jain S, Mar Valsson S, et al. Application of physics-informed neural networks to inverse problems in unsaturated groundwater flow. *Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards*, 2022, 16(1): 21-36.

- [74] Kim H S, Ji Y. Three-dimensional geotechnical-layer mapping in Seoul using borehole database and deep neural network-based model. *Engineering Geology*, 2022, 297: 106489.
- [75] Marzán I, Martí D, Lobo A, et al. Joint interpretation of geophysical data: Applying machine learning to the modeling of an evaporitic sequence in Villar de Cañas (Spain). *Engineering Geology*, 2021, 288: 106126.
- [76] Bressan T S, de Souza M K, Girelli T J, et al. Evaluation of machine learning methods for lithology classification using geophysical data. *Computers & Geosciences*, 2020, 139: 104475.
- [77] Liu L L, Wang Y. Quantification of stratigraphic boundary uncertainty from limited boreholes and its effect on slope stability analysis. *Engineering Geology*, 2022, 306: 106770.
- [78] Molina-Gómez F, da Fonseca A V, Ferreira C, et al. Defining the soil stratigraphy from seismic piezocone data: A clustering approach. *Engineering Geology*, 2021, 287: 106111.
- [79] Cao Z J, Zheng S, Li D Q, et al. Bayesian identification of soil stratigraphy based on soil behaviour type index. *Canadian Geotechnical Journal*, 2019, 56(4): 570-586.
- [80] Leu S S, Adi T J W. Probabilistic prediction of tunnel geology using a Hybrid Neural-HMM. *Engineering Applications of Artificial Intelligence*, 2011, 24(4): 658-665.
- [81] Yu L, Sun X, Ge Y. The application of hidden Markov model in classifying novice and experienced drivers by driving behavioral features, *International Conference on Transportation Engineering 2009*. 2009: 3160-3165.
- [82] Liu D H, Zhou Y Q, Wang S, et al. Stochastic simulation and risk analysis of water tunnel TBM construction scheduling based on geologic prediction using Markov process. *Journal of System Simulation*, 2009, 21(2): 558-562.
- [83] Wang Y, Shi C. Data-driven analysis of soil consolidation with prefabricated vertical drains considering stratigraphic variation. *Computers and Geotechnics*, 2023, 161: 105569.
- [84] Chai J, Fu H, Wang J, et al. Behaviour of a PVD unit cell under vacuum pressure and a new method for consolidation analysis. *Computers and Geotechnics*, 2020, 120: 103415.
- [85] Xu J, Wang Y, Zhang L. Fusion of geotechnical and geophysical data for 2D subsurface site characterization using multi-source Bayesian compressive sampling. *Canadian Geotechnical Journal*, 2022, 59(10): 1756-1773.

- [86] Yang Z, Chen Q, Cui Z, et al. Automatic reconstruction method of 3D geological models based on deep convolutional generative adversarial networks. *Computational Geosciences*, 2022, 26(5): 1135-1150.
- [87] Chen Q, Liu G, Ma X, et al. 3D stochastic modeling framework for Quaternary sediments using multiple-point statistics: A case study in Minjiang Estuary area, southeast China. *Computers & geosciences*, 2020, 136: 104404.
- [88] Chan S, Elsheikh A H. Parametrization of stochastic inputs using generative adversarial networks with application in geology. *Frontiers in Water*, 2020, 2: 5.
- [89] Xie J, Huang J, Zeng C, et al. A generic framework for geotechnical subsurface modeling with machine learning. *Journal of Rock Mechanics and Geotechnical Engineering*, 2022, 14(5): 1366-1379.
- [90] Pieczyńska-Kozłowska J, Vessia G. Spatially variable soils affecting geotechnical strip foundation design. *Journal of Rock Mechanics and Geotechnical Engineering*, 2022, 14(3): 886-895.
- [91] Zhang W, Phoon K K. Editorial for Advances and applications of deep learning and soft computing in geotechnical underground engineering. *Journal of Rock Mechanics and Geotechnical Engineering*, 2022, 14(3): 671-673.
- [92] Zeng, C., Wang, S., Xue, X., Zheng, G., Mei, G. Evolution of deep ground settlement subject to groundwater drawdown during dewatering in a multi-layered aquifer-aquitard system: Insights from numerical modelling. *Journal of hydrology (Amsterdam)*. 2021, 603: 127078.
- [93] Zhang Y Q, Wang J H, Chen J J, et al. Numerical study on the responses of groundwater and strata to pumping and recharge in a deep confined aquifer. *Journal of Hydrology*, 2017, 548: 342-352.
- [94] Zeng C F, Song W W, Xue X L, et al. Construction dewatering in a metro station incorporating buttress retaining wall to limit ground settlement: Insights from experimental modelling. *Tunnelling and Underground Space Technology*, 2021, 116: 104124.
- [95] Yuan C, Hu Z, Zhu Z, et al. Numerical Simulation of Seepage and Deformation in Excavation of a Deep Foundation Pit under Water-Rich Fractured Intrusive Rock. *Geofluids*, 2021, 2021(1): 6628882.
- [96] Yuan C F, Yuan Z J, Wang Y T, et al. Analysis of the diffusion process of mining overburden separation strata based on the digital speckle correlation coefficient field. *International Journal of Rock Mechanics and Mining Sciences*, 2019, 119: 13-21.

- [97] Li Z, Liu S, Ren W, et al. Multiscale Laboratory Study and Numerical Analysis of Water-Weakening Effect on Shale. *Advances in Materials Science and Engineering*, 2020, 2020(1): 5263431.
- [98] Luo Z, Zhang Y, Wu Y. Finite element numerical simulation of three-dimensional seepage control for deep foundation pit dewatering. *Journal of Hydrodynamics*, 2008, 20(5): 596-602.
- [99] Luo Z, Wu Y. Three-dimension transient seepage numerical simulation of deep foundation pit dewatering in quaternary loose sediment layer. *Journal of Shenyang Jianzhu University (Natural Science Edition)*, 2006, 22(2): 181-185.
- [100] Bing C, Liang C. Numerical investigation of three-dimensional flow and bed shear stress distribution around the span shoulder of pipeline. *JOURNAL OF HYDRODYNAMICS*, 2004, 16(6): 687-694.
- [101] Zhou N, Vermeer P A, Lou R, et al. Numerical simulation of deep foundation pit dewatering and optimization of controlling land subsidence. *Engineering geology*, 2010, 114(3-4): 251-260.
- [102] Wang J, Hu L, Wu L, et al. Hydraulic barrier function of the underground continuous concrete wall in the pit of subway station and its optimization. *Environmental Geology*, 2009, 57: 447-453.
- [103] Luo Z, Liu J, Li L. Three-dimensional full coupling numerical simulation of groundwater dewatering and land-subsidence in quaternary loose sediments. *chinese journal of geotechnical engineering-chinese edition-*, 2008, 30(2): 193.
- [104] Li D Q, Qi X H, Phoon K K, et al. Effect of spatially variable shear strength parameters with linearly increasing mean trend on reliability of infinite slopes. *Structural safety*, 2014, 49: 45-55.
- [105] Popescu R, Deodatis G, Nobahar A. Effects of random heterogeneity of soil properties on bearing capacity. *Probabilistic engineering mechanics*, 2005, 20(4): 324-341.
- [106] Kasama K, Whittle A J, Zen K. Effect of spatial variability on the bearing capacity of cement-treated ground. *Soils and Foundations*, 2012, 52(4): 600-619.
- [107] Yin Z Y, Wang P, Zhang F. Effect of particle shape on the progressive failure of shield tunnel face in granular soils by coupled FDM-DEM method. *Tunnelling and Underground Space Technology*, 2020, 100: 103394.
- [108] Chen T, Guestrin C. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016.
- [109] Zhou W H, Zhao L S, Chen G M, et al. 3D Geologic Modeling with Borehole Data by General Regression Neural Network. 2018.

- [110] Shangguan W, Hengl T, Mendes de Jesus J, et al. Map** the global depth to bedrock for land surface modeling. *Journal of Advances in Modeling Earth Systems*, 2017, 9(1): 65-88.
- [111] Zhang P, Yin Z Y. A novel deep learning-based modelling strategy from image of particles to mechanical properties for granular materials with CNN and Bi-LSTM. *Computer Methods in Applied Mechanics and Engineering*, 2021, 382: 113858.
- [112] Zhang N, Shen S L, Zhou A, et al. Application of LSTM approach for modelling stress-strain behavior of soil. *Applied Soft Computing*, 2021, 100: 106959.
- [113] Li K Q, Yin Z Y, Liu Y. Influences of spatial variability of hydrothermal properties on the freezing process in artificial ground freezing technique. *Computers and Geotechnics*, 2023, 159: 105448.
- [114] Ke G, Meng Q, Finley T, et al. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 2017, 30.
- [115] Margineantu D D, Dietterich T G. Pruning adaptive boosting, *ICML*. 1997, 97: 211-218.
- [116] Juang C H, Zhang J, Shen M, et al. Probabilistic methods for unified treatment of geotechnical and geological uncertainties in a geotechnical analysis. *Engineering geology*, 2019, 249: 148-161.
- [117] Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks. *arxiv preprint arxiv:1609.02907*, 2016.
- [118] Tipu R K, Panchal V R, Pandya K S. Multi-objective optimized high-strength concrete mix design using a hybrid machine learning and metaheuristic algorithm. *Asian Journal of Civil Engineering*, 2023, 24(3): 849-867.
- [119] Lim S, Chi S. Xgboost application on bridge management systems for proactive damage estimation. *Advanced Engineering Informatics*, 2019, 41: 100922.
- [120] Badola S, Mishra V N, Parkash S. Landslide susceptibility map** using XGBoost machine learning method, 2023 international conference on machine intelligence for geoanalytics and remote sensing (MIGARS). *IEEE*, 2023, 1: 1-4.
- [121] Li X, Xu L, Guo H, et al. Application of graph convolutional neural networks combined with single-model decision-making fusion neural networks in structural damage recognition. *Sensors*, 2023, 23(23): 9327.

- [122] Murshed R U, Noshin K, Zakaria M A, et al. Real-time seismic intensity prediction using self-supervised contrastive GNN for earthquake early warning. IEEE Transactions on Geoscience and Remote Sensing, 2024.