IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received April 13, 2019, accepted May 14, 2019, date of publication May 20, 2019, date of current version June 4, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2917756

A Data-Driven Framework for Tunnel Geological-Type Prediction Based on TBM Operating Data

JUNHONG ZHAO[1], MAOLIN SHI[2], GANG HU[3], XUEGUAN SONG[2], CHAO ZHANG[1], (Member, IEEE), DACHENG TAO[4], (Fellow, IEEE), AND WEI WU[1]

[1]School of Mathematical Sciences, Dalian University of Technology, Dalian 116024, China
[2]School of Mechanical Engineering, Dalian University of Technology, Dalian 116024, China
[3]School of Civil Engineering, The University of Sydney, Darlington, NSW 2008, Australia
[4]UBTECH Sydney Artificial Intelligence Centre, School of Computer Science, Faculty of Engineering and Information Technologies, University of Sydney, Darlington, NSW 2008, Australia

Corresponding author: Chao Zhang (chao.zhang@dlut.edu.cn)

This work was supported by the National Natural Science Foundation of China under Grant 11401076, Grant 61473328, Grant 51505061, and Grant U1608256.

ABSTRACT One main challenge in tunnel constructions is to predict the tunnel geological conditions without excavation to ensure safety during the construction process. This paper proposes a data-driven framework for real-time interpreting the operating data of tunnel boring machines (TBMs) without interrupting tunneling operations, and eventually automate the tunneling operation. In this framework, we first convert the indexes of the original data from discontinuous operating time to continuous operating displacement. After screening outliers, to more exhaustively explore the inherent characteristics of the TBM operating data, we then augment features by using the first-order and the second-order difference information. There are two main concerns for developing a desired geological-type predictor: 1) since multiple geological types could coexist in one tunnel section, the predictor should have multiple outputs and 2) since the geological types are specified by the values of 7 kinds of physical-mechanical indexes of geological types, this geological characteristic should also be encoded into the predictor's structure. Therefore, we design a feed-forward multiple-output artificial neural network (ANN) with two hidden layers as the predictor, where the second hidden layer has 7 nodes that correspond to 7 kinds of physical-mechanical indexes. The experimental results show that: 1) the feature augmentation (FA) method indeed improves the prediction performance; 2) the ANN predictor has the best performance on the test set when the second hidden layer has 7 nodes; 3) the proposed ANN predictor outperforms many widely-used learning models, *e.g.,* XGboost, random forest (RF), and support vector regression (SVR); and 4) the predictor is capable of accurately predicting the geological types of stratum.

INDEX TERMS Tunnel boring machines, geological-type prediction, operating parameter, neural network, physical-mechanical indexes.

I. INTRODUCTION

In the past decades, a large number of tunnels have been constructed and are currently under construction to create short cuts for transportation and public traffic. To date, the conventional tunneling technique, i.e. drilling and blasting, is still widely used and continuously improved. In general, it is a mature, reliable and well-built tunneling approach. However, its shortcomings are prominent. For example, the tunnel contour is quite rough due to the blasting and therefore additional efforts are needed to polish the tunnel profile; its efficiency is quite low. By contrast, mechanical tunneling normally utilizing tunnel boring machines (TBM) has a significant higher efficiency and is capable of continuously tunneling under various ground environments. Furthermore, tunneling using TBM has been considered to be the safest approach regarding

The associate editor coordinating the review of this manuscript and approving it for publication was Tao Zhang.

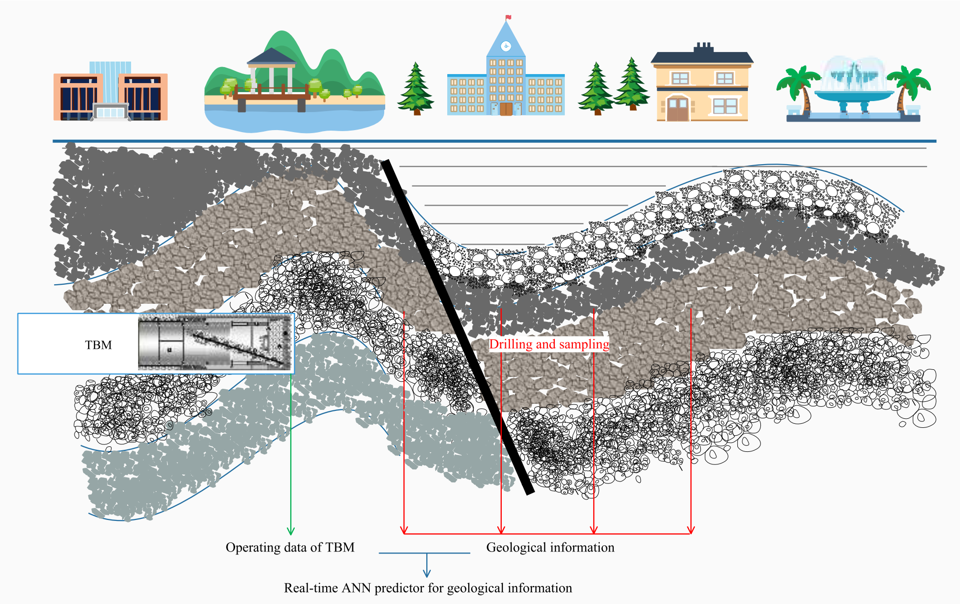


FIGURE 1. Collection of TBM operating data and geological information.

security in the workplace [1]. As a consequence, TBM has been increasingly used in various geological environments, such as urban environments, in close proximity to other tunnels and foundations, and in more complex geologies, *e.g.* mixed face conditions and karstic rock.

However, under these complex geological conditions, the excavation process incurs very high risk because of an unanticipated presence of water leakage, weak regions or voids, break outs of the tunnel wall, collapse of the working front or blocking of the cutting wheel. A significant limitation of the current TBM tunneling technique is the inability to accurately and efficiently foresee these unexpected changes [2]. An encounter with unforeseen ground conditions during tunneling process seriously delays the construction schedule and induces large extra costs, causes instrument damages, introduces additional hazards or even tremendous casualties. For instance, 19 times of water inrush hazards occurred during the construction of Maluqing Tunnel of Yiwan Railway in China, which disastrously causes 15 deaths [3]. Apparently, it is crucial to fully understand the geological condition before tunneling to ensure the safety and efficiency of the tunneling construction.

The geological conditions are always unknown before tunneling. To understand the geological condition, a number of approaches have been developed, including hard methods and soft methods [4]. Hard methods, such as subsurface boring, pilot drilling, and utilizing in-site equipment to obtain geological information in a few specific locations along the tunnel alignment. Soft methods build prediction models using such as the Markov process approach [4], real-time Bayesian approach [5], statistical approach [6], artificial neural network [7], ground penetrating radar [8] for ahead forecasting the geological conditions to complement hard methods. On the other hand, TBM itself has also integrated a number of different kinds of sensors into the cutting wheel to real-time collect data associated with the geological information of the tunnel face. TBM operating data recode the continuous geological information of the entire construction tunnel and contain a quantity of features that are strongly associated with the geological situation of tunnel sections [9]. However, currently the interpretation of these real-time data entirely relies on the experience of the operators. This old-fashion human-machine interaction is not only very inefficient but also not reliable. Different operators may make quite different judgements based on the same dataset. This imperceptibly increases the risk during the excavation process. Therefore, it is promising to conduct the tunnel geological-type prediction by exploring the mathematical relation between the TBM operating data and the geological types. However, to the best of our knowledge, there are limited works on the geological-type prediction based on TBM operating data to automate the interpretation of the data. Inspired by the self-driving concept, this paper aims to build an intelligent model of real-time interpreting the operating data of TBM without interrupting tunneling operations, and eventually automate the tunneling operation, as demonstrated in Fig. 1.

The data-driven technique provides an effective way to deal with the geological-type prediction: instead of directly detecting the geological situation of the tunnel sections, one only needs to explore the relationship between the geological types and the observable data. In this manner, the geological-type prediction will be implemented as a classification or regression learning task without being restricted by the aforementioned limitations. In the earlier work [10], Mooney *et al.* recorded the TBM vibration frequency from the sensors set on TBM's bulkhead as a source of information to examine the changes of geological conditions, while their examination is based on the empirical observation on the data graphs rather than the mathematical relation between the vibration frequency and the geological condition. Recently, Shi *et al.* [11] and Zhang *et al.* [12] used TBMs' operating data to build the machine-learning models for classifying the geological types appearing in the TBM construction tunnels. However, their models are unsuitable to (at least cannot be directly used for) predicting the thickness of the geological types.

In this paper, we propose a data-driven framework for predicting the geological-type thickness based on TBM operating data, which are collected from the sensors set on a TBM in an urban subway construction project and the geological-type samples are obtained by drilling method at some discrete locations of the subway construction line. This framework mainly concerns with the following issues: 1) the raw TBM operating data are indexed by the discontinuous operating time because the construction will be intermitted in some cases, *e.g.,* equipment maintenance or rest; 2) there could be some outliers because of the complex construction environment; 3) in contrast with the whole operating data, the size of data labeled by the geological types is relatively small; and 4) because of complex geological conditions, several geological types usually coexist in one tunnel section. Without loss of the generality, the framework contains three stages to deal with these issues:

1) Data acquisition - We make the raw data actionable, where the data indexed by discontinuous operating time are converted to the data indexed by continuous operating displacement;
2) Data preprocessing - We use KNN-based method to screen outliers. To explore the advanced features relevant to geological information from the relatively few labeled data and meanwhile to avoid the redundant information, we augment features by applying their first-order and the second-order difference information and then use principle component analysis (PCA) method to eliminate the redundancy.
3) ANN predictor - We select the feed-forward multiple-output artificial neural network (ANN) as the geological-type predictor. It has two hidden layers, and especially the second hidden layer has 7 nodes, which correspond to 7 physical-mechanical indexes respectively. Moreover, each of its outputs corresponds to the proportion of the specific geological type appearing in the location of interest.

The experimental results show that 1) the proposed ANN predictor outperforms many learning models including XGBoost, CatBoost, random forests (RF), decision tree (DT), support vector regression (SVR), K-nearest neighbors (KNN) and Bayesian linear regression (BLR); and 2) the feature-augmenting (FA) method improves the performance of the geological-type prediction.

The rest of this paper is organized as follows. In Section II, we introduce the data acquisition and the problem background. The stage of data reprocessing is arranged in Section III. Section IV presents the ANN predictor for geological types. In Section V, we show the experimental results to support the validity of our framework and the last section concludes the paper.

II. DATA ACQUISITION

In an urban subway construction project, the operating data are collected by the sensors set on the earth pressure balance shield TBM, which consists of cutter-head, chamber, screw conveyor, tail skin and auxiliaries. The tunnel is about 2000 meter long with diameter of 6.3 meter, and its longitudinal geological profile is shown in Fig. 1. The engineering route, located in alluvial and coastal plain, is divided into 1364 ring sections, each of which is 1.5 meter long. The range of ground surface elevation is 0.2 ∼ 5.8 meter and the depth of the tunnel floor from the ground surface is within 11.8 ∼ 25.4 meter. The stratum can be divided into five layers in terms of the geological types, and each layer can further be divided into 2 ∼ 7 sub-layers according to physical-mechanical indexes. In this tunnel construction, there is likely to be 20 kinds of geological types, each of which is specified by the values of 7 physical-mechanical indexes (see Tab. 1). The 7 indexes include natural severity (Y), internal friction angle (quick direct shear test) (FI), deformation modulus (EM), Poisson's ratio (P), coefficient of lateral pressure (SITA), permeability coefficient (K), and cohesive strength between rock mass and anchors (FRB).[1]

A total of about 4.6 millions of operating data were recorded, and each datum has 72 features including torque, thrust, tunneling speed and fuel tank temperature. It is noteworthy that the raw TBM operating data are indexed by the discontinuous operation time because the construction will be intermitted in some cases, *e.g.,* equipment maintenance or rest. To develop a continuous index system for the operating data, we apply the velocity integral to locate the position of each datum, and then the resulting positions form a continuous index system for the operating data. For the i-th ring section, denote $s^{(i)}(j)$ as TBM's position at time $t_j^{(i)}$:

$$s^{(i)}(j) = \int_{t_0^{(i)}}^{t_j^{(i)}} v^{(i)}(t)dt, \quad i = 1, \cdots, 88, \tag{1}$$

where $v^{(i)}(t)$ is the instantaneous velocity at time t in the i-th ring section and $t_0^{(i)}$ is the initial time of the i-th ring section.

The geological-type samples are drawn from 88 of the 1364 ring sections by the drilling method. Each drilling sample is 30 meter deep and runs through the tunnel section

[1] Besides the aforementioned 7 indexes, there are other rock-soil physical-mechanical indexes, *e.g.,* natural moisture content, pore ratio, cohesive force (quick direct shear test), internal friction angle (consolidation quick direct test), compression modulus, coefficient of subgrade reaction (vertical), coefficient of subgrade reaction (horizontal), uniaxial compressive rock strength (saturation) and uniaxial compressive rock strength (natural). However, since these indexes are irrelevant to the TBM tunneling process, this paper does not consider the relationship between these indexes and TBM operating data.

TABLE 1. Rock-soil physical-mechanical indexes of different geological types.

Index \ Type	②$_{3}$	②$_{4}$	⑨$_{1}$	⑨$_{2-1}$	⑨$_{2-2}$	⑨$_{3}$	
Y (kN/m^3)	17.000	18.5	19.50	20.500	22.50	24.50	
FI (°)	4.500	18.0	25.00	27.500	45.00	55.00	
EM (MPa)	4.000	5.5	90.000	10000	10000	40.00	
P	0.650	0.5	0.250	0.250	0.25	0.25	
SITA	0.650	0.5	0.00	0.00	0.00	0.00	
K (m/d)	0.003	3.5	0.80	2.5	15.00	1.50	
FRB (kPa)	10.000	20.0	45.000	60.000	125.00	380.00	
Index \ Type	**④$_{2}$**	**④$_{4}$**	**④$_{5}$**	**④$_{8}$**	**④$_{9}$**	**④$_{10}$**	**④$_{11}$**
Y (kN/m^3)	19.000	18.000	19.00	19.50	20.00	20.50	21.00
FI (°)	15.000	8.000	20.00	26.00	28.00	32.00	35.00
EM (MPa)	15.000	4.500	6.50	20.00	22.00	25.00	18.00
P	0.320	0.420	0.32	0.28	0.25	0.22	0.25
SITA	0.200	0.700	0.48	0.45	0.40	0.35	0.35
K (m/d)	0.0050	0.0050	4.50	6.50	12.00	20.0	30.0
FRB (kPa)	25.00	18.00	22.00	35.00	50.00	55.00	65.00
Index \ Type	**⑦$_{2-1}$**	**⑦$_{2-2}$**	**⑫$_{1}$**	**⑫$_{2-1}$**	**⑫$_{2-2}$**	**⑫$_{3}$**	**⑫$_{4}$**
Y (kN/m^3)	18.50	18.50	19.50	20.50	22.50	24.50	26.50
FI (°)	20.50	22.50	27.00	30.00	45.00	55.00	70.00
EM (MPa)	20.00	40.00	90.00	10000.00	10000.00	10000.00	10000.00
P	0.30	0.28	0.25	0.25	0.25	0.22	0.18
SITA	0.45	0.55	0.00	0.00	0.00	0.00	0.00
K (m/d)	0.50	0.50	1.00	2.50	15.00	1.50	1.20
FRB (kPa)	22.00	28.00	45.00	60.00	125.00	380.00	650.00

whose diameter is 6.3 meter (see Fig. 1). To exploit the geological information exhaustively, we extract operating data within 0.3 meter around the 88 drilling points. Namely, in each of these ring sections, the data labeled by geological information lie in the interval of 0.6 meter length centered at the drilling point. In this manner, we finally get 66226 operating data labeled by geological-type information. Since multiple geological types may coexist in the tunnel sections, we record the geological types and their thicknesses. As shown in Tab. 2(a), some geological types do not appear or have a low number of existence in the geological-type samples. Thus, we emerge the geological types with similar physical-mechanical indexes to generate 6 kinds of geological labels (see Tab. 2(b)), and denote them as

$$\mathcal{P} = \left\{\textcircled{2}_*, \textcircled{4}_*, \textcircled{7}_{2-1}, \textcircled{7}_{2-2}, \textcircled{9}_*, \textcircled{12}_*\right\}.$$

We would like to design a predictor for the thickness of each geological type in the tunnel section. Specifically, denote $a_{i,\alpha}$ as the thickness of the geological type α ($\alpha \in \mathcal{P}$) in the i-th drilling sample ($i = 1, 2, \cdots, 88$). Then, the relevant thickness of the geological type T_α in the tunnel part of the i-th drilling sample can be formulated as

$$y_{i,\alpha} = \frac{a_{i,\alpha}}{\sum_{\alpha\in\mathcal{P}} a_{i,\alpha}} = \frac{a_{i,\alpha}}{6.3},$$

which is also the output form of the resulting predictor.

III. DATA PREPROCESSING

There are two main issues considered in the stage of data preprocessing. First, due to complicated construction environments, the operating data could contain some outliers. Second, since only 88 (out of 1364) ring sections have the corresponding geological information, there is an imbalance between the operating data labeled by geological types and the unlabeled operating data. Thus, to achieve a good predictor, it is crucial to explore the advanced features relevant to geological information from the relatively few labeled data and meanwhile to avoid the redundant information.

In this section, we use the K-nearest neighbors (KNN)-based algorithm, proposed in [13], to detect outliers

TABLE 2. Geological types appearing in drilling samples. (a) The number of times that each geological type appears in 88 drilling samples. (b) The number of times that each geological type appears after being merged.

(a)

Type	②$_{3}$	②$_{4}$	④$_{2}$	④$_{4}$	④$_{5}$
Number	7	0	4	4	0
Type	**④$_{8}$**	**④$_{9}$**	**④$_{10}$**	**④$_{11}$**	**⑦$_{2-1}$**
Number	0	0	14	0	13
Type	**⑦$_{2-2}$**	**⑨$_{1}$**	**⑨$_{2-1}$**	**⑨$_{2-2}$**	**⑨$_{3}$**
Number	40	11	7	0	0
Type	**⑫$_{1}$**	**⑫$_{2-1}$**	**⑫$_{2-2}$**	**⑫$_{3}$**	**⑫$_{4}$**
Number	8	4	0	1	0

(b)

Type	②$_{*}$	④$_{*}$	⑦$_{2-1}$	⑦$_{2-2}$	⑨$_{*}$	⑫$_{*}$
Number	7	19	13	40	16	10

for TBM operating data. Then, we introduce a difference-based method to obtain new features that are beneficial to improving generalization performance of the resulting predictor.

A. OUTLIER DETECTION

Because of complicated construction environments, the TBM operating data usually contain a certain amount of outliers, which seriously influence the data quality. Therefore, the outlier detection plays an essential role in preprocessing TBM operating data. In general, the outlier-detection methods can be divided into two categories: the statistics-based methods (*e.g.* distribution [14] and density [15]) and the distance-based methods (*e.g.* one-class support vector machine (OCSVM) [16], support vector data description (SVDD) [17] and K-nearest neighbor (KNN) [13], [18], [19]). Since it is difficult to explore the distribution of TBM operating data, we instead consider the distance-based methods for outlier detection.

The support-vector methods (*e.g.* OCSVM and SVDD) aim to find the support vectors from the data to form a decision boundary for selecting outliers that are the data points lie outside the decision boundary. Since there are a total of about 4.6 millions of TBM operating data, the efficiency of these methods will be affected by the large amount of data as well as the selection of hyper-parameters. In contrast, the KNN-based outlier detection is processed based on the density of the neighborhoods of the data points, and thus its efficiency could not be significantly affected by the amount of data. Peng and Huang [20] have showed that the KNN-based outlier detection method can is suitable to detecting the outliers from a large amount of data. Therefore, we finally adopt the KNN-based method to detect the outliers from the massive TBM operating data (see Algorithm 1).

Algorithm 1 KNN-Based Outlier Detection [18]

```
Input: sample set $\{x_n\}_{n=1}^N \subset \mathbb{R}^M$, neighbor number $k = 5$
    and percentage $p\% = 5\%$;
Output: list of outliers;
for all samples $x_1, \cdots, x_N$; do
    find the $k$-nearest neighbor set $N_\rho(x_n, k)$ of the sample
        $x_n$, that is, the set of $k$ points belong to $\{x_n\}_{n=1}^N$ and are
        nearest to $x_n$ w.r.t. the metric $\rho$;
    compute the outlier score $\tau$ of $x_n$ for the neighbor
        number $k$:
        $$\tau(x_n, k) = \frac{\sum_{y \in N_\rho(x_n,k)} \rho(x_n, y)}{k};$$
end for
sort the samples $\{x_n\}_{n=1}^N$ in increasing order of
    $\{\tau(x_n, k)\}_{n=1}^N$, and then select the last $p\%$ samples
    as outliers.
return
```

It is noteworthy that the KNN-based algorithm is sensitive to the choice of k: 1) if k is too small, one cannot obtain enough neighbor information of sample points and the time consuming is high; and 2) if k is too large, the outlier score $\tau(x_n, k)$ cannot exactly describe the outlier behavior of the point x_n because $\tau(x_n, k)$ is the averaged distance between x_n and the points lying in its k-nearest neighbor. Therefore, we empirically set the parameters $k = 5$.

B. FEATURE AUGMENTATION

Since only 88 (out of 1364) ring sections have the corresponding geological information, there is an imbalance between the operating data labeled by geological types and the unlabeled operating data. Thus, to improve the performance of the predictor, it is crucial to explore the advanced features relevant to geological information from the labeled data and meanwhile to avoid the redundant information. It is noteworthy that many features of the operating data have the specific physical meanings. Thus, with the help of the obtained continuous index system $S \subset \mathbb{R}^+$, we introduce the feature augmentation (FA) method to generate new features from the original operating data.

Let $d \in (0, 0.2]$ be the difference gap. The reason why the upper bound of d is set to be 0.2 is because the labeled data in each section ring lie in the 0.6-length interval and the second-order difference is achieved by using three points with the same gap. Given three data with the same gap ($\forall\, s \in \mathcal{S} - 2d$)

$$\begin{aligned} x_s &= \left(x_s^{(1)}, \cdots x_s^{(M)}\right); \\ x_{s+d} &= \left(x_{s+d}^{(1)}, \cdots, x_{s+d}^{(M)}\right); \\ x_{s+2d} &= \left(x_{s+2d}^{(1)}, \cdots, x_{s+2d}^{(M)}\right), \end{aligned}$$

we then compute the first-order difference:

$$\nabla^1(x_s) = \left(x_{s+d}^{(1)} - x_s^{(1)}, \cdots, x_{s+d}^{(M)} - x_s^{(M)}\right),$$

and the second-order difference:

$$\nabla^2(x_s) = \left(x_{s+2d}^{(1)} - 2\, x_{s+d}^{(1)} + x_s^{(1)}, \cdots, x_{s+2d}^{(M)} - 2\, x_{s+d}^{(M)} + x_s^{(M)}\right).$$

Next, we augment the resulted features into the original datum x_s to generate a new datum:

$$\widehat{x_s} := \left(x_s, \nabla^1(x_s), \nabla^2(x_s)\right) \in \mathbb{R}^{3M}.$$

Since the size of feature-augmented datum $\widehat{x_s}$ is twice larger than that of x_s, there could be some redundancy information in $\widehat{x_s}$. We applied the principle component analysis (PCA) method to reduce the dimension and meanwhile to eliminate the correlation among the features of $\widehat{x_s}$. In this paper, we retain 95% variance information of the data set $\{\widehat{x_s}\}$.

IV. ANN PREDICTORS FOR GEOLOGICAL TYPES

Artificial neural networks (ANNs) have been successfully applied in many geological engineering problems [21]–[23]. Compared with other learning models, one main advantage of ANN is the high flexibility of its structure, which can be designed to meet specific requirements of real-world problems. The recent success of deep networks confirm the importance of designing hidden layers - the structure of hidden

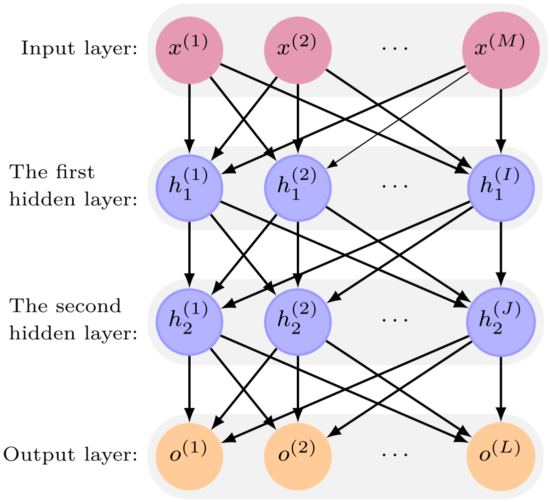


FIGURE 2. **The structure of ANN predictor.**

layers should be interpretable and encodes the inherent characteristics of the problems [24].

To develop a desired predictor for geological types based on TBM operating data, one needs to consider the following issues: 1) although TBM operating status is strongly related with the geological types appearing in the tunnel section, the relationship between them is generally very complex and cannot be directly formulated by using some common functions; 2) since several geological types usually coexist in one tunnel section, the resulting predictor should have multiple outputs; and 3) as shown in Tab. 1, the geological types are specified by 7 physical-mechanical indexes, and thus the geological characteristics should be encoded into the predictor's structure.

Compared with other learning models (*e.g.* support vector regression and random forest), the following advantages of ANN make it more suitable to the geological-type prediction task: 1) ANN has the powerful ability of non-linear mapping to implement complex regression or classification tasks; 2) it has a high flexibility to set multiple outputs without significantly increasing the training difficulty; and 3) the structure of ANN can be designed to meet the specific requirements of practical applications. Therefore, we apply a feed-forward ANN to achieve the geological-type predictor. The network has one input layer, two hidden layers and one output layer, and every two adjacent layers are fully-connected (see Fig. 2). Specifically, the first hidden layer is used to extract the higher features from the inputs and the second hidden layer provides a bridge between the TBM operating information and the geological information.

Let $\sigma : \mathbb{R} \to \mathbb{R}$ be the active function and we set it to be the rectified linear unit (ReLU) function that has been widely used in the current ANN learning models (see [25], [26]):

$$\sigma(x) = \max\{0, x\}, \quad x \in \mathbb{R}.$$

The outputs of the first hidden layer, the second hidden layer and the output layer are respectively computed as follows:

$$h_1^{(i)} = \sigma\Big(\sum_{m=1}^{M} u^{(im)} x^{(m)}\Big), \quad i = 1, 2, \cdots, I; \tag{2}$$

$$h_2^{(j)} = \sigma\Big(\sum_{i=1}^{I} v^{(ji)} h_1^{(i)}\Big), \quad j = 1, 2, \cdots, J; \tag{3}$$

$$o^{(l)} = \sigma\Big(\sum_{j=1}^{J} w^{(lj)} h_2^{(j)}\Big), \quad l = 1, 2, \cdots, L, \tag{4}$$

where $u^{(im)}$, $v^{(ji)}$ and $w^{(lj)}$ are the weights between the two adjacent layers. Denote $U = \big(u^{(im)}\big)_{I\times M}$, $V = \big(v^{(ji)}\big)_{J\times I}$ and $W = \big(w^{(lj)}\big)_{L\times J}$ as the weight matrices.

Given a sample set $\{(x_n, y_n)\}_{n=1}^{M} \subset \mathbb{R}^M \times \mathbb{R}^L$, we adopt the mean squared error (MSE) to measure the difference between the predictor outputs $o_n^{(l)}$ and the real outputs $y_n^{(l)}$:

$$E(U, V, W) = \frac{1}{2}\sum_{n=1}^{N}\sum_{l=1}^{L}(y_n^{(l)} - o_n^{(l)})^2 + \lambda R(U, V, W), \tag{5}$$

where the coefficient $\lambda > 0$ and

$$R(U, V, W) = \|U\|_F^2 + \|V\|_F^2 + \|W\|_F^2.$$

is the regularization term for preventing the overfitting. By minimizing the objective function (5), we then obtain the appropriate weights U, V and W to achieve the desired predictor. Following the back-propagation algorithm, the weights are renewed in the $(k+1)^{th}$ iteration as follows: for any $k = 1, 2, 3, \cdots$,

$$\begin{aligned} u_{k+1}^{(im)} &= u_k^{(im)} + \Delta u_k^{(im)}; \\ v_{k+1}^{(ji)} &= v_k^{(ji)} + \Delta v_k^{(ji)}; \\ w_{k+1}^{(lj)} &= w_k^{(lj)} + \Delta w_k^{(lj)}, \end{aligned}$$

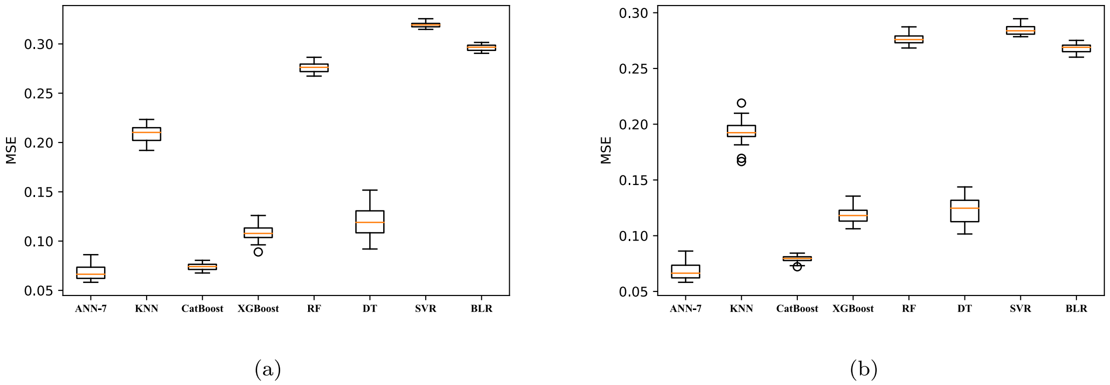
where

$$\begin{aligned} \Delta u_k^{(im)} &= -\eta\Big(\frac{\partial E}{\partial u_k^{(im)}} + 2\lambda u_k^{(im)}\Big); \\ \Delta v_k^{(ji)} &= -\eta\Big(\frac{\partial E}{\partial v_k^{(ji)}} + 2\lambda v_k^{(ji)}\Big); \\ \Delta w_k^{(lj)} &= -\eta\Big(\frac{\partial E}{\partial w_k^{(lj)}} + 2\lambda w_k^{(lj)}\Big). \end{aligned}$$

V. NUMERICAL EXPERIMENTS

In this section, we present the experimental results to support the validity of the proposed framework. The experiments are conducted to verify the following issues: (i)

1) The performance of the geological-type predictor, *i.e.*, whether the ANN model outperforms other learning models for the geological-type prediction task.
2) The node number of the second hidden layer, *i.e.*, whether the nodes should correspond to the 7 kinds of physical-mechanical indexes specifying geological types;



(a)

(b)

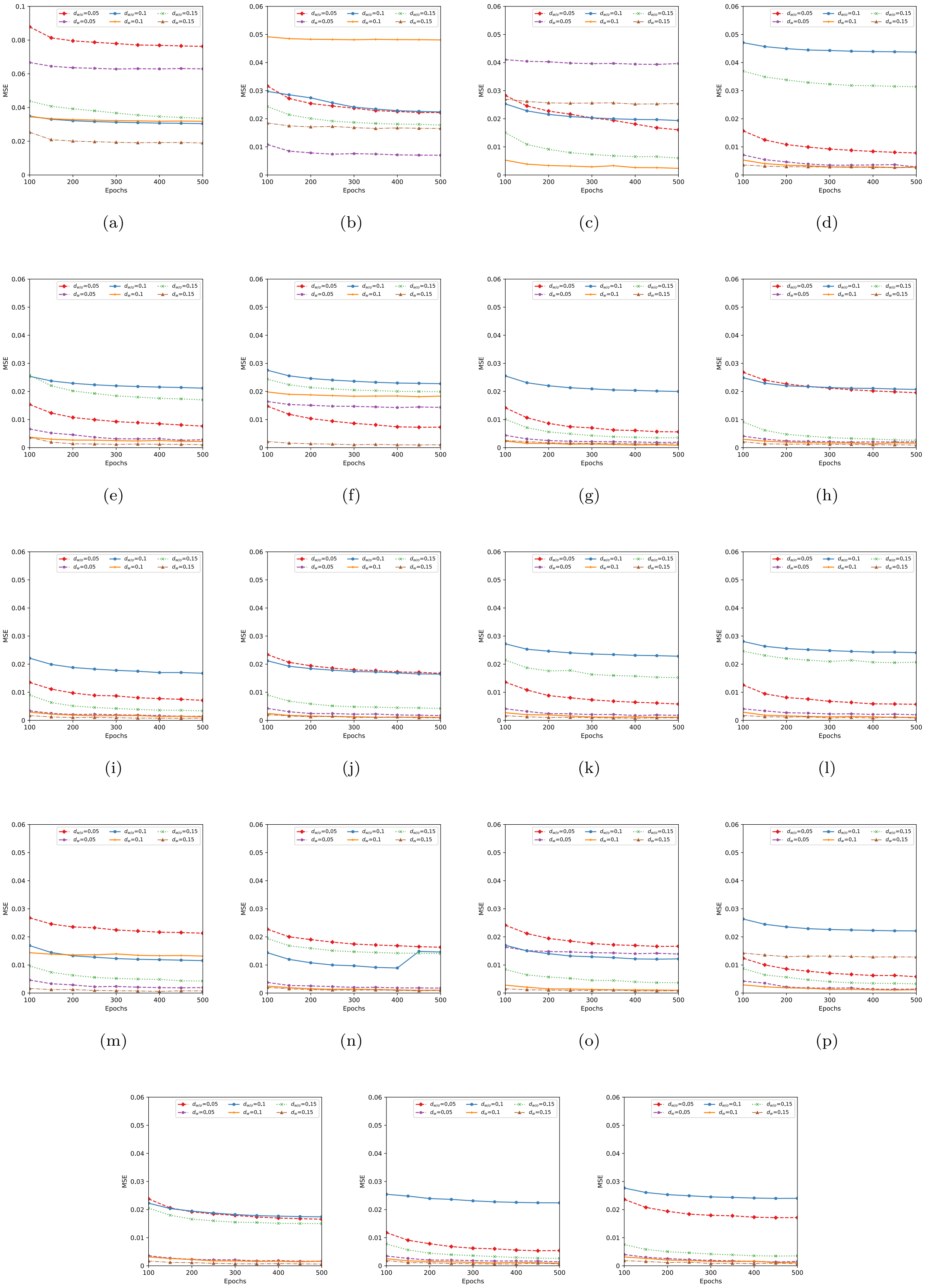
FIGURE 3. **Boxplots of different models on FA ($d_w = 0.15$) and Non-FA ($d_{w/o} = 0.15$) samples. (a) Boxplot on FA samples ($d_w = 0.15$). (b) Boxplot on Non-FA samples ($d_{w/o} = 0.15$).**

TABLE 3. **Averaged MSE of different models for geological-type prediction.**

	Without feature augmentation						With feature augmentation					
MSE \ Gap	$d_{w/o} = 0.05$		$d_{w/o} = 0.1$		$d_{w/o} = 0.15$		$d_w = 0.05$		$d_w = 0.1$		$d_w = 0.15$	
Model	train	test	train	test	train	test	train	test	train	test	train	test
XGBoost	0.0667	0.0936	0.0667	0.0854	0.0672	0.1246	0.0665	0.0922	0.0663	0.0815	0.0674	0.1105
CatBoost	0.0652	0.0778	0.0649	0.0772	0.0463	0.0792	0.0591	0.0735	0.0567	0.0740	0.0446	0.0739
RF	0.0423	0.0688	0.0447	0.0711	0.0533	0.0946	0.0401	0.0689	0.0346	0.0649	0.0376	0.0581
DT	0.0354	0.0757	0.0454	0.0704	0.0569	0.1116	0.0397	0.0735	0.0257	0.0674	0.0476	0.0951
SVR	0.3215	0.3264	0.2401	0.2397	0.3133	0.3212	0.3026	0.3143	0.2301	0.2265	0.2692	0.2875
KNN	0.2632	0.3093	0.2019	0.2349	0.1628	0.1936	0.2245	0.2380	0.2132	0.2394	0.1982	0.2090
BLR	0.2819	0.3002	0.2452	0.2877	0.2198	0.2685	0.2657	0.3058	0.2671	0.3047	0.2322	0.2962
ANN-2	0.0833	0.0973	0.0672	0.0582	0.0763	0.0751	0.0483	0.0618	0.0429	0.0520	0.0426	0.0383
ANN-3	0.0636	0.0757	0.0622	0.0739	0.0762	0.0791	0.0398	0.0424	0.0471	0.0533	0.0358	0.0533
ANN-4	0.0614	0.0672	0.0549	0.0624	0.0573	0.0532	0.0329	0.0383	0.0352	0.0372	0.0372	0.0354
ANN-5	0.0528	0.0539	0.0540	0.0594	0.0499	0.0522	0.0368	0.0381	0.0339	0.0373	0.0323	0.0353
ANN-6	0.0518	0.0525	0.0516	0.0533	0.0483	0.0502	0.0332	0.0346	0.0293	0.0327	0.0236	0.0294
ANN-7	0.0413	**0.0430**	0.0417	**0.0436**	0.0354	**0.0393**	0.0266	**0.0260**	0.0222	**0.0212**	0.0202	**0.0236**
ANN-8	0.0414	0.0455	0.0438	0.0449	0.0414	0.0443	0.0305	0.0372	0.0319	0.0315	0.0227	0.0281
ANN-9	0.0424	0.0462	0.0418	0.0453	0.0475	0.0502	0.0275	0.0346	0.0348	0.0362	0.0244	0.0323
ANN-10	0.0408	0.0479	0.0426	0.0504	0.0465	0.0522	0.0321	0.0341	0.0284	0.0307	0.0361	0.0367
ANN-11	0.0390	0.0505	0.0417	0.0524	0.0467	0.0529	0.0313	0.0328	0.0236	0.0248	0.0305	0.0329
ANN-12	0.0406	0.0488	0.0405	0.0491	0.0464	0.0480	0.0419	0.0369	0.0259	0.0274	0.0266	0.0271
ANN-13	0.0383	0.0457	0.0412	0.0485	0.0424	0.0564	0.0294	0.0346	0.0212	0.0273	0.0285	0.0291
ANN-14	0.0397	0.0503	0.0407	0.0516	0.0419	0.0520	0.0273	0.0381	0.0326	0.0334	0.0282	0.0302
ANN-15	0.0381	0.0485	0.0385	0.0504	0.0348	0.0484	0.0322	0.0361	0.0318	0.0378	0.0336	0.0348
ANN-16	0.0378	0.0510	0.0386	0.0524	0.0353	0.0501	0.0342	0.0372	0.0346	0.0357	0.0334	0.0374
ANN-17	0.0368	0.0540	0.0372	0.0514	0.0342	0.0494	0.0284	0.0369	0.0410	0.0510	0.0446	0.0327
ANN-18	0.0354	0.0480	0.0381	0.0517	0.0352	0.0483	0.0282	0.0316	0.0291	0.0461	0.0291	0.0370
ANN-19	0.0321	0.0564	0.0334	0.0502	0.0371	0.0516	0.0272	0.0361	0.0325	0.0348	0.0252	0.0283
ANN-20	0.0315	0.0531	0.0345	0.0497	0.0337	0.0572	0.0307	0.0358	0.0297	0.0351	0.0323	0.0401

3) The effectiveness of the feature augmentation (FA), *i.e.,* whether the FA method can improve the predictor performance;

We use Keras (ver. 2.1.4) to process all experiments in a computer with Intel®i7-6700K CPU at 4.0GHz×8, 64GB RAM and two Nvidia®GTX-1080 graphic cards. We split the samples into two parts: 70% of them are treated as the training set and the remaining 30% are used as the test set. For the ANN model of interest (*cf.* Fig. 2), the node number of the first hidden layer is set to be 20 according to the empirical observation of the experiments, and the node number of the second hidden layer is selected from 2 to 20, respectively. As a comparison, we also consider other learning models for the prediction task including XGBoost [27], CatBoost [28], random forest (RF), decision tree (DT), support vector regression (SVR), K-Nearest Neighbor (KNN) [29], [30] and Bayesian linear regression (BLR) [31]. To examine the validity of the FA method, we implement these models by using the FA samples and the non-FA samples respectively. Specifically, the FA samples are taken from TBM operating data w.r.t. the difference gap $d_w \in \{0.05, 0.1, 0.15\}$. To make the experimental results comparable, the non-FA samples are also taken w.r.t. the same the difference gap $d_{w/o} \in \{0.05, 0.1, 0.15\}$. In the



(a)

(b)

(c)

(d)

(e)

(f)

(g)

(h)

(i)

(j)

(k)

(l)

(m)

(n)

(o)

(p)

(q)

(r)

(s)

FIGURE 4. MSE curves of ANN predictors with different numbers of nodes in the second hidden layer. (a) Node = 2. (b) Node = 3. (c) Node = 4. (d) Node = 5. (e) Node = 6. (f) Node = 7. (g) Node = 8. (h) Node = 9. (i) Node = 10. (j) Node = 11. (k) Node = 12. (l) Node = 13. (m) Node = 14. (n) Node = 15. (o) Node = 16. (p) Node = 17. (q) Node = 18. (r) Node = 19. (s) Node = 20.

back-propagation algorithm, the learning rate η is 0.01; the batch size is 10; and the maximum number of iterations is 500. The parameters of XGBoost and CatBoost are obtained by the grid-search method and the parameters of the other models are selected based on the empirical observation of the experiments.

In Fig. 4, we first show the mean square error (MSE) curves of ANN predictors with different numbers of nodes in the second hidden layer on FA samples and non-FA samples, respectively. We observe that when the node number of the second hidden layer is suitably chosen, the geological-type predictor performs well on the test set. Moreover, the FA method improves the performance of the predictor, and the performance of difference gap $d = 0.15$ is superior to other values of d in most cases. Interestingly, the ANN predictor has the best performance when its second hidden layer has 7 nodes. Recall that the geological types are specified by the values of 7 kinds of physical-mechanical indexes, which suggests that the structure of the second hidden layer should correspond to these physical-mechanical indexes, *i.e.,* the node number should be 7 accordingly.

To further verify this finding, we use several state-of-art regression models to implement the geological-type prediction task including XGBoost, CatBoost, random forests (RF), decision tree (DT), support vector regression (SVR), K-nearest neighbor (KNN) and Bayesian linear regression (BLR). For each kind of experimental setting, we repeat 30 simulations on FA samples and non-FA samples respectively, and the experimental results are recorded in Tab. 3 as well. Compared with the other models, the ANN with 7 nodes in the second hidden layer (denoted as ANN-7) has the smallest averaged MSE regardless of the FA samples or the non-FA samples.

Based on the 30 repeated experimental results of these models, we then draw the boxplots to examine the stability of these models in the geological-type prediction task (*cf.* Fig. 3). We find that the introduction of FA method improves the stability of these models on the geological-type prediction task and ANN-7 has the smallest averaged MSE among these models. Note that CatBoost has a high stability in spite of a higher MSE than that of ANN-7.

To sum up, the experimental results support the validity of the framework in the following aspects: 1) the FA method improves the performance of the predictors in most cases; and 2) when the second hidden layer has 7 nodes, the ANN predictor has the best performance on the test set; and 3) the ANN model outperforms the other learning models for the geological-type prediction task.

VI. CONCLUSION

In this paper, we propose a framework to build a predictor for geological types based on TBM operating data. The framework contains three stages: data acquisition, data preprocessing and learning models. In particular, we first convert the indexes of the original data from discontinuous operating time to continuous operating displacement. After screening outliers, to more exhaustively explore the inherent characteristics of TBM operating data, we then augment features by using the first-order and the second-order difference information. To select a suitable predictor for geological types, there are two main concerns:

1) since multiple geological types could coexist in one tunnel section, the predictor should have multiple outputs;
2) since the geological types are specified by the values of 7 kinds of physical-mechanical indicators of geological types, the structure of the predictor should encode these geological characteristics.

Therefore, we adopt a feed-forward multiple-output ANN with two hidden layers to build the predictor, where the second hidden layer has 7 nodes to correspond 7 kinds of physical-mechanical indicators. The experimental results support the validity of the framework. In addition, we also verify that: 1) the FA method indeed improves the performance of the predictors in most cases; 2) when the second hidden layer has 7 nodes, the predictor has the best performance on the test set; and 3) the proposed ANN predictor outperforms other well-known learning models (*e.g.* XGBoost, CatBoost, random forest and SVR) for the geological-type prediction task.

In the future works, we will improve the geological-type data quality by using other data-acquisition methods instead of the drilling method. Then, some soft methods (*e.g.,* the Markov process approach [4] and the real-time Bayesian approach [5]) will be considered to handle the geological-type prediction task. In addition, we will also consider the specific outlier detection methods for the TBM operating data.

REFERENCES

[1] S. Jetschny, "Seismic prediction and imaging of geological structures ahead of a tunnel using surface waves," Ph.D. dissertation, Karlsruhe Inst. Technol., Karlsruhe, Germany, 2010.

[2] K. Schaeffer and M. A. Mooney, "Examining the influence of TBM-ground interaction on electrical resistivity imaging ahead of the TBM," *Tunnelling Underground Space Technol.*, vol. 58, pp. 82–98, Sep. 2016.

[3] S. Li *et al.*, "An overview of ahead geological prospecting in tunneling," *Tunnelling Underground Space Technol.*, vol. 63, pp. 69–94, Mar. 2017.

[4] Z. Guan, T. Deng, S. Du, B. Li, and Y. Jiang, "Markovian geology prediction approach and its application in mountain tunnels," *Tunnelling Underground Space Technol.*, vol. 31, pp. 61–67, Sep. 2012.

[5] S. Leu, T. Joko, and A. Sutanto, "Applied real-time Bayesian analysis in forecasting tunnel geological conditions," in *Proc. IEEE Int. Conf. Ind. Eng. Eng. Manage.*, Macao, China, Dec. 2010, pp. 1505–1508.

[6] G. Papacharalampous, H. Tyralis, and D. Koutsoyiannis, "One-step ahead forecasting of geophysical processes within a purely statistical framework," *Geosci. Lett.*, vol. 5, no. 1, pp. 1–12, 2018.

[7] A. Alimoradi, A. Moradzadeh, R. Naderi, M. Z. Salehi, and A. Etemadi, "Prediction of geological hazardous zones in front of a tunnel face using TSP-203 and artificial neural networks," *Tunnelling Underground Space Technol.*, vol. 23, no. 6, pp. 711–717, 2008.

[8] L. Wei, D. R. Magee, and A. G. Cohn, "An anomalous event detection and tracking method for a tunnel look-ahead ground prediction system," *Autom. Construct.*, vol. 91, pp. 216–225, Jul. 2018.

[9] S. Suwansawat and H. H. Einstein, "Artificial neural networks for predicting the maximum surface settlement caused by EPB shield tunneling," *Tunnelling Underground Space Technol.*, vol. 21, no. 2, pp. 133–150, 2006.

[10] M. Mooney, B. Walter, J. Steele, and D. Cano, "Influence of geological conditions on measured TBM vibration frequency," in *Proc. North Amer. Tunneling*, G. Davidson, A. Howard, L. Jacobs, R. Pintabona, and B. Zernich, Eds. Englewood, CO, USA: Society for Mining, Metallurgy & Exploration, 2014, pp. 32–41.

[11] M. Shi, X. Song, and W. Sun. (2018). "Geology prediction based on operation data of TBM: Comparison between deep neural network and statistical learning methods." [Online]. Available: https://arxiv.org/abs/1809.06688

[12] Q. Zhang, Z. Liu, and J. Tan, "Prediction of geological conditions for a tunnel boring machine using big operational data," *Autom. Construct.*, vol. 100, pp. 73–83, Apr. 2019.

[13] S. Ramaswamy, R. Rastogi, and K. Shim, "Efficient algorithms for mining outliers from large data sets," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, New York, NY, USA, May 2000, pp. 427–438.

[14] A. Koufakou and M. Georgiopoulos, "A fast outlier detection strategy for distributed high-dimensional data sets with mixed attributes," *Data Mining Knowl. Discovery*, vol. 20, no. 2, pp. 259–289, 2010.

[15] M. M. Breunig, H. P. Kriegel, R. T. Ng, and J. Sander, "LOF: Identifying density-based local outliers," in *Proc. ACM Int. Conf. Manage. Data (SIGMOD)*, Dallas, TX, USA, May 2000, pp. 93–104.

[16] H. J. Shin, D.-H. Eom, and S.-S. Kim, "One-class support vector machines—An application in machine fault detection and classification," *Comput. Ind. Eng.*, vol. 48, no. 2, pp. 395–408, 2005.

[17] Y. Xiao *et al.*, "Multi-sphere support vector data description for outliers detection on multi-distribution data," in *Proc. IEEE Int. Conf. Data Mining Workshops*, Miami, FL, USA, Dec. 2009, pp. 82–87.

[18] V. Hautamaki, I. Karkkainen, and P. Franti, "Outlier detection using k-nearest neighbour graph," in *Proc. 17th Int. Conf. Pattern Recognit.*, Cambridge, U.K., Aug. 2004, pp. 430–433.

[19] Y. Chen, D. Miao, and H. Zhang, "Neighborhood outlier detection," *Expert Syst. Appl.*, vol. 37, no. 12, pp. 8745–8749, Dec. 2010.

[20] Y. Peng and B. Huang, "KNN based outlier detection algorithm in large dataset," in *Proc. Int. Workshop Educ. Technol. Training*, Dec. 2009, pp. 611–613.

[21] P. Q. Memon, S.-P. Yong, W. Pao, and P. J. Seanl, "Prediction of bottom-hole flowing pressure using general regression neural network," in *Proc. Int. Conf. Comput. Inf. Sci. (ICCOINS)*, Kuala Lumpur, Malaysia, Jun. 2014, pp. 1–5.

[22] M. Zare, H. R. Pourghasemi, M. Vafakhah, and B. Pradhan, "Landslide susceptibility mapping at VAZ watershed (Iran) using an artificial neural network model: A comparison between multilayer perceptron (MLP) and radial basic function (RBF) algorithms," *Arabian J. Geosci.*, vol. 6, no. 8, pp. 2873–2888, 2013.

[23] J. Dou *et al.*, "An integrated artificial neural network model for the landslide susceptibility assessment of Osado Island, Japan," *Natural Hazards*, vol. 78, no. 3, pp. 1749–1776, 2015.

[24] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, p. 436, 2015.

[25] S. S. Talathi and A. Vartak, "Improving performance of recurrent neural network with relu nonlinearity," *CoRR*, vol. abs/1511.03771, pp. 1–12, Nov. 2015.

[26] G. E. Dahl, T. N. Sainath, and G. E. Hinton, "Improving deep neural networks for LVCSR using rectified linear units and dropout," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Vancouver, BC, Canada, May 2013, pp. 8609–8613.

[27] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, B. Krishnapuram, M. Shah, A. J. Smola, C. C. Aggarwal, D. Shen, and R. Rastogi, Eds. New York, NY, USA: ACM, Aug. 2016, pp. 785–794.

[28] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: Unbiased boosting with categorical features," in *Proc. Adv. Neural Inf. Process. Syst.*, Montréal, QC, Canada, Dec. 2018, pp. 6637–6647.

[29] W. Liu, D. Xu, I. Tsang, and W. Zhang, "Metric learning for multi-output tasks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 408–422, Feb. 2019.

[30] H. Borchani, G. Varando, C. Bielza, and P. Larrañaga, "A survey on multi-output regression," *Data Mining Knowl. Discovery*, vol. 5, no. 5, pp. 216–233, 2015.

[31] X. Yan and X. Su, "Bayesian linear regression," in *Linear Regression Analysis: Theory and Computing*. Singapore: World Scientific, 2009.

JUNHONG ZHAO is currently pursuing the Ph.D. degree with the School of Mathematical Sciences, Dalian University of Technology. Her current research focuses on data mining, deep learning, and learning methods of spiking neural networks.

MAOLIN SHI received the B.S. degree in mechatronic engineering from Chongqing University, Chongqing, China, in 2015, and the M.S. degree in mechanical manufacturing and automation from Huaqiao University, Xiamen, China. He is currently pursuing the Ph.D. degree with the School of Mechanical Engineering, Dalian University of Technology. His current research interests include fuzzy clustering, data mining, and mechanical design.

GANG HU received the bachelor's degree from Central South University, in 2009, the M.Phil. degree from the Harbin Institute of Technology, in 2011, and the Ph.D. degree from the Hong Kong University of Science and Technology, in 2015. In 2015, he joined the CLP Power Wind/Wave Tunnel Facility, HKUST, as a Research Associate, where he served as a Postdoctoral Fellow with the Department of Civil and Environmental Engineering, from 2016 to 2017. Since 2017, he has been a Postdoctoral Research Associate with the School of Civil Engineering, The University of Sydney. He is currently a Postdoctoral Research Associate with the School of Civil Engineering, The University of Sydney. He is an Executive Committee Member of the Hong Kong Wind Engineering Society. He has published more than 20 papers as a senior author in international reputable journals. His research interests include structural wind engineering, wind energy by using wind tunnel test, CFD, and AI techniques.

XUEGUAN SONG received the B.S. degree in mechanical engineering from the Dalian University of Technology, Dalian, China, in 2004, and the M.S. and Ph.D. degrees in mechanical engineering from DongA University, Busan, South Korea, in 2007 and 2010, respectively. He is currently a Professor with the School of Mechanical Engineering, Dalian University of Technology, China. He has published more than 80 peer-reviewed papers, one book, and one book chapter on *engineering optimization, computational fluid dynamics analysis, and thermal management of power electronics*. His research interests include multidisciplinary design optimization, numerical modeling, data analysis, and data-driven design. He was selected in the Young Thousand Talents Program, in 2016. In addition, he received the Best Paper Award from the International Symposium on Linear Drives for Industry Applications (LDIA 2013), the Best Poster Awards from the International Joint Conference on Computational Sciences and Optimization (CSO 2011), and the Global Conference on Power Control and Optimization (PCO 2010), and the Honorable Mention Award from the American Society of Mechanical Engineers Pressure Vessels and Piping Conference (ASME-PVP 2010).

CHAO ZHANG (M'16) received the B.S. degree in applied mathematics and the Ph.D. degrees in computational mathematics from the Dalian University of Technology, Dalian, China, in 2004 and 2009, respectively, where he is currently an Associate Professor with the School of Mathematical Sciences. His research interests include deep learning, machine learning, data analysis, learning theory, and random matrices.

DACHENG TAO (F'15) is currently a Professor of computer science and an ARC Laureate Fellow of the Faculty of Engineering and Information Technologies and the School of Computer Science, and the Inaugural Director of the UBTECH Sydney Artificial Intelligence Centre, The University of Sydney. He mainly applies statistics and mathematics to artificial intelligence and data sciences. His research results have expounded in one monograph and 200+ publications at prestigious journals and prominent conferences, such as the IEEE TPAMI, TIP, TNNLS, TCYB, IJCV, JMLR, NIPS, ICML, CVPR, ICCV, ECCV, ICDM, and ACM SIGKDD, with several best paper awards, such as the Best Theory/Algorithm Paper Runner Up Award from the IEEE ICDM 2007, the Best Student Paper Award from the IEEE ICDM 2013, the 2014 ICDM 10-Year Highest-Impact Paper Award, the 2017 IEEE Signal Processing Society Best Paper Award, and the Distinguished Paper Award from the 2018 IJCAI. He has received the 2015 Australian Scopus Eureka Prize and the 2018 IEEE ICDM Research Contributions Award. He is a Fellow of the Australian Academy of Sciences, AAAS, IAPR, OSA, and SPIE.

WEI WU received the bachelor's and master's degrees from Jilin University, Changchun, China, in 1977 and 1981, respectively, and the Ph.D. degree from Oxford University, Oxford, U.K., in 1987. He is currently with the School of Mathematical Sciences, Dalian University of Technology, Dalian, China. He has published four books and 90 research papers. His current research interest includes learning methods of neural networks.