Research Article

Xuejia Sang, Linfu Xue, and Xiaoshun Li*

Considering the geological significance in data preprocessing and improving the prediction accuracy of hot springs by deep learning

<https://doi.org/10.1515/geo-2020-0237>

received July 15, 2019; accepted February 23, 2021

Abstract: The geothermal gradient in the eastern area of Liaoning Province is very low, but hot springs resources are variable. The reason is not clear till now but leads to the fact that a few strong influence factors can cause imbalances in the results of many prediction algorithms. It can be found as a black-box algorithm, deep learning will obtain a more unbalanced result with the fault influence factors. To tackle this issue, the role of preprocessing during the process of profound learning was enhanced and four comparative experiments were carried out. The results show that compared with the unprocessed experiment, the accuracy rate of the experiment with fully processed data increased by 11.9 p.p., and the area under the curve increased by 0.086 (0.796–0.882). This inspires us that even though the deep learning method can achieve high accuracy in the prediction of geological resources, we still need to pay attention to the analysis and pretreatment of data with expertise according to local conditions.

Keywords: hot spring, deep learning, pretreatment, prediction

1 Introduction

Geothermal resources are indeed clean so that they have been paid more and more attention worldwide. Besides, to develop and utilize geothermal resources in an effective way, evaluating geothermal resources in various

regions plays a fundamental role [1]. It is a relatively effective method to evaluate the potential of geothermal resources by integrating multidisciplinary data analysis. Among the multi-source data analysis algorithms, deep learning is the fastest-growing algorithm in recent years. Plenty of research studies related to this field have been conducted by scientists worldwide [2,3]. Recently, deep learning has attracted the attention of researchers from all fields [4–6]. After the first design and optimization of LeCun et al. [7] in 1998, the performance of the deep learning model has increased, which is often designed for the processing of complex signals, such as computer vision; it has shown its superiority compared to other technologies [8]. Deep learning is widely used in image classification, speech recognition, traffic sign recognition, medical image analysis, and other applications [9]. Almost all researchers have found that deep learning far exceeds the detection rate or accuracy of traditional algorithms [10–16]. But deep learning is essentially a black-box algorithm, and many researchers still question its theoretical support. Many open-source deep learning frameworks such as Tensorflow and TensorRT take “zero coding” as a development goal, which makes the use of deep learning easy but bewildered. Researchers just need to provide data to the model and adjust the parameters to get better results. The prosperity of this industry has brought many hidden dangers, such as making researchers more inclined to optimize parameters with unknown physical meanings rather than data preprocessing.

When there is a lack of treatment in preprocessing, the prediction error will be significantly increased. Also, the prediction response level of the factor in the high potential orientation will be reduced by the factor in the low potential orientation, resulting in the decrease of predicting accurately [17]. Data of Weights of Evidence (WoE) are necessary to check for problems and assess assumptions [18]. Previous research shows that the pre-processed data can significantly improve the prediction accuracy of the WoE. In the WoE process, if the geological significance of the predicted target is not considered,

* **Corresponding author: Xiaoshun Li**, School of Public Policy & Management, China University of Mining and Technology, Xuzhou, Jiangsu, China, e-mail: lixiaoshun1983@sohu.com

Xuejia Sang: School of Public Policy & Management, China University of Mining and Technology, Xuzhou, Jiangsu, China, e-mail: sangxj@cumt.edu.cn

Linfu Xue: College of Earth Sciences, Jilin University, Changchun, Jilin, China, e-mail: xuelf@jlu.edu.cn

only the distance from the fault is selected as the influencing factor of geothermal (hit rate 60.6%), compared to the method of selectively increasing the weight of the fault in the direction of the local principal stress (hit 72.7%), its predicted hit rate is 12.1 p.p. (percentage points) lower [19].

This question appears not just in the WofE method, but still in lots of multivariable analysis algorithms. To enhance the application of data, preprocessing the data itself can be regarded as one of the primary orientations to develop related to the algorithm in the future [20,21]. This study was developed to verify this theory.

This study is based on multivariable data, such as geological observation, gravity, and remote sensing, and digital elevation model (DEM) data. For this purpose, four deep learning tests are considered as the control experiment. For the first experiment, a traditional approach was used to directly load each layer as input data. In the second experiment, sufficient preprocessing of the data was performed and then deep learning. For the third and fourth experiments, it was only optimized with a fault or DEM data, respectively. Finally, the prediction accuracy of the four experiments was compared.

The research discussion and test on the places in which the distribution of natural resources is under control by the fault are conducted. In this research, geological knowledge and deep learning concepts are introduced into the WofE method, and the high potential geothermal prospects in eastern Liaoning are drawn. The regional hit rate quantifies the difference in accuracy between data-driven algorithms similar to WofE and deep learning algorithms with knowledge-driven optimization. And this difference cannot be ignored even in the deep learning process.

The selected area is found in the northeast section of North China Craton (NCC), with an area of 24262.9 km², which mainly includes the fault zones: Hanling-Pianling fault (HP), Hanling-Benxi fault (HB), Haicheng-Caohekou fault (HC), Haicheng Ximucheng-Xiuyan fault (HX), Daheisha-Zhangjiapu fault (DZ), Liujiache-qingduizi fault (LQ), Sipingjie-Fengcheng fault (SF), Zhuanghe-Huanren fault (ZH), Erpengdianzi-Changdian fault (EC), and Yalu River fault (YL) [22–25]. The research on the distribution law of hot springs is much less. The faults developed in the eastern area of Liaoning Province in China, with the wide distribution of Mesozoic intrusions and Archean metamorphic rocks [22,26,27]. Main faults in the region are dominated in NE-trending (Figure 1). The main fractures there contain ZH, SF, EC, YL, and LQ, which divide tectonic regions into some long strip faulted blocks and get a command of modern landform generating and developing. The NE trending faults generated afterward break the previous original structure. The EW-trending

faults generated at the earliest periods [28] have activities in the early Proterozoic [29]. Most of the NE trending faults come into being during Indosinian movements at the end of the middle three Triassic periods [30]. The crustal movement achieved the summit in the early Cretaceous, the eastern parts of China including the study area are influenced by the E–W extrusion effect, resulting from the North–South couple [31]. Therefore, both NE strike-slip faults and NW–NNW trending tensional faults appear [32].

2 Data

In this study, ArcGIS and SPSS software were implemented to collect and evaluate the data. Then, data related to hot springs points, rivers, lithology, Bouguer gravity, and faults of remote sensing interpretation are collected.

There are 33 hot springs in the selected area, with 24 springs being acquired by digitizing historical data, 3 being discovered by field survey, and 6 being gained from the Internet. Besides, there are 28 artesian springs and 5 artificial mining hot springs. In terms of data of geology, they can all be traced back to a geological map of 1:2,50,000 provided by the Research Institute of Geology and a survey of the mineral resources of Liaoning province and Jilin province. Additionally, the revision and supplementation of this map are performed by the project named “Deep geological survey of Linjiang and Benxi” by the China Geological Survey.

Four Landsat8 remote sensing images are applied, and the map numbers are LC81180312016090, LC81180322014276, LC81180312016138, and LC81190322016241. Characterized by less cloud cover and false-color composition, they are appropriate for being applied to have visual interpretation (red-green-blue false-color channel: 752) and make surface features more recognizable. Compared with a geological map, the interpretation results are divided into several parts.

The DEM data in the research belong to ASTER GDEM (global digital elevation model), with the mesh accuracy of 30 m and the elevation accuracy of 7 m. Valley data originated from DEM data after being treated in the hydrological aspect. Based on the fill and analysis of the flow direction with the DEM data, the computation of flow accumulation data, and the extraction of the grid and vectorization, all vector data of the river network are thus acquired. The source of the two data sets above is the Geospatial Data Cloud site of the Computer Network Information Center at the Chinese Academy of Sciences (<http://www.gscloud.cn>) (Figure 2).

Figure 1: The general situation of geology about this research [19].

3 Methods

The use of deep learning lies in predicting the geothermal potential in this study, which is based on the multi-data after pretreatment (Figure 3). The pretreatment process of the fry method is shown in light blue rectangle, which involved calculation of the structural stress of the study area and then the trend orientation related to the points distribution of hot springs and the curve fitting is acquired. First, the calculation of weight of fault depends on the fitting curve, then the calculation of the fault buffer area under the influence of weight is performed with weight value, eventually, it can conduct the calculation of the weighted fault layer. The DEM pretreatment is shown in light green rectangle, which includes the calculation of slope, river network, and Euclidean distance. The deep learning process is shown in light brown rectangle, which includes data enhancement, verification, modeling, and prediction.

The core workflow of pretreatment is as follows:

1. Processing rudimentary data are related to these aspects of geocoding the hot springs points, the extraction of data from DEM, the implementation of STD (standard deviations) in accordance with the gravity anomaly data, the utilization of remote sensing to make the existing fault abundant, and the classification of the geological layer which can influence the distribution of the hot spring.
2. The fry method is used to get the relationship between fault and azimuth and build an anisotropic buffer by fitting the curve. The fault factor layers finally obtain.

3. The DEM data were analyzed to obtain the statistical significance of the location of hot springs in the study area.
4. Combine the layers into multidimensional data and analyze them by deep learning model, then obtain the forecast result.

3.1 Fry preprocessing of fault data

The fry method finite strain measurement uses deformed particles, such as deformed gravel, granules, and near-isometric minerals as strain markers, which are easier to obtain in geological practice and are the most common means of finite strain measurement. Many researchers achieved the extension of an applied scene to spatial distribution analysis [19,33,34] and acquired well effects [3]. With the use of Fry analysis [35], the mutual operation can be achieved through a sheet of tracing paper which can present a series of parallel reference lines (typically pointing north on a map) that are drawn and then the position of points related to each set of data is collected. Similar to the first sheet, there are a set of marked lines which are also parallel on the second sheet of paper with a center point (or origin). The origin of the second sheet is put on one of the data points in the first sheet, and all the locations of points on the first one are marked on the second sheet. Then, the origin of the second sheet is put on a different data point from the first and the locations are once again signed on the second sheet.

Figure 2: Part of the data utilized in this research. (a) Bouguer gravity data after STD processing, it can reflect the contact boundary of different density geological bodies. (b) Fault data, including faults found in the field, are interpreted from remote sensing images and inferred by earth exploration methods. (c) The acquisition of River network data by DEM. (d) Lithology data, which are roughly divided into six types of lithological combinations based on the difference between sedimentary rocks and intrusive rocks, as well as the different formation ages, including Late Triassic intrusive rocks, Balidianzi intrusive rocks, and Neoproterozoic granite, the three phases of granite invasion events represented, and the three most representative stratigraphic combinations: Qianshan super-units, Dashiqiao group, and Gaixian group (Appendix and ref. 19).

Figure 3: Workflow of geothermal anomaly analysis with pretreatment then deep learning.

This kind of procedure is successive with the same orientation until all the points on the first sheet are utilized as the origin on the second. In terms of n data points, $n^2 - n$ translations are generated. A further analysis of the resulting Fry plot probably can be performed through constructing a rose diagram that records joint frequency versus directional sector. Based on the reconstruction of distributive characteristics without visibility, the fry method can offer a chance to visualize spatial distribution. It is worth noting that although the fry method uses hot spring data in this research, it is not used for spatial location information extraction. The use of the fry method is just one of the methods to quantify the structural stress in the study area.

The fault is not only a hot springs heat source but also a hot water channel for the hot springs. The premise for reasoning the converse inference of the influencing extent of fault can be regarded as the spatial distribution of hot springs. The emergence of Fry analysis [35] can be found in the research studies of lot of researchers such as Noorollahi, Moghaddam, and Wibowo [34,36–38]; however, utilizing it merely is to present the distribution trend of objects or to perform spatial analysis as an auxiliary method. The distribution of hot springs is mainly affected by the stress azimuth, and this effect can be quantified using the fry method. The fault that conforms to this trend should be paid more attention when calculating afterward. Figure 4 introduces the course of enabling Fry analysis to be applied in hot springs data.

The purpose of the fry method is that, before the calculation of the fault layer in deep learning, parameters for the subdivision layer should be split with a strain weighting factor. The prior possibility of the subdivision layer can be acquired through the Fry model statistics and reduced to

$$IP = \sum_{k=1}^n F(a_k)f(L_k) + C_k, \quad (1)$$

$$F(a_k) = \frac{a_k}{n^2 - n}, \quad (2)$$

$$L = \sum_1^n f(L_n), \quad (3)$$

where C_k reflects the crucial extent of the geothermal anomaly at the k section in a layer. $F(a_k)$ is the distribution

frequency of faults in Fry analysis rose diagram at the k section, n means the sum of sample points, and L is the fault layer.

Through the above approach, when the possibility factors and prediction emerge simultaneously, the predictive target under their control will become more essential and precise, and in the meanwhile, the weight value of the consistent weight factor will become greater. Each grid layer is used as the predictive factor to predict the target, and then different levels of expectation maps can be obtained.

To establish the relationship between the fitting curve and the weight, it is necessary for us to set the fitting curve of the standardized y values within the range of 0 to 1. MATLAB is used to let $x = 0:0.01:180$ from 0 to 180, with 0.01 for each procedure for the fitting results.

Then, $f(x)$ being able to fit the fault distribution frequency is acquired by using λ th degree polynomial:

$$f(x) = \beta_0 a_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_\lambda x_i^\lambda + \varepsilon_i \quad (4)$$

$$(i = 1, 2, \dots, n)$$

and then λ should be satisfied:

$$\begin{cases} \lambda \leq n_{\text{Fry}} \\ |\text{normr}_\lambda - \text{normr}_{\lambda+1}| < \sigma \end{cases} \quad (5)$$

where n means the number of Fry points and normr is the matrix norm residuals; the smaller the normr , the better the fit of the curve. ε is an unobserved random error with mean zero conditioned on a scalar variable x . β is the parameter in the regression function. The σ may be a small constant to ensure that there will not be excessive change for the normr value (comparing between normr_λ and $\text{normr}_{\lambda+1}$), which can converge to σ in the local scope.

Under the premise of not influencing the precision, smoothing the fault data is indispensable to decrease discontinuous parts of fault data which resulted from extremely little information at the reflection points, and the corners become denser after smoothing. There has been much convenience for dividing the fault data into numerous small faults and then acquiring the position of each minor fault. Based on the combination of the fitting curve gained before, it can achieve the buffer distance of the fault under the orientation weighting correction. λ determines the fitting effect of the curve, and at the

Figure 4: Fry method analysis course for hot springs data.

same time, it also affects the area of the buffer in the fault layer, thereby affecting the weight of the fault layer in the WofE algorithm. In this study, taking the selected study area as an example, the distribution of λ and the weight value w_+ under the same variables are shown in Figure 5. It can be seen in the figure that the value of λ is 15 when the value of w_+ is the largest. Therefore, in the subsequent calculation process, $\lambda = 15$.

Then, the relationship between the azimuth and the structural stress is established through fitting the curve, while the structural stress is positively connected with the spatial distribution of hot springs. Therefore, it creates a chance for building the relationship between the fault with structural stress and the geothermal distribution tendency (Figure 6).

3.2 Statistics preprocessing of DEM

In addition to geothermal sources and transportation channels, sufficient water sources and pressure are also necessary conditions for the emergence of hot springs. Therefore, the high-drop terrain that can provide water pressure is often the best prospecting area of hot springs. The topography of some hot springs in the selected area was analyzed and it was found that the hot springs in this area often appear at the foot of the mountain where the terrain is strongly cut, close to the valley (Figure 7a). To further quantify the impact of terrain, we performed a statistical analysis of the topography of the hot springs in the selected area, including slope and water source distance. We found that most hot springs appear in a triangle within the gradient and water source coordinate system (Figure 7b), which includes 32 hot springs (97% of all). This

Figure 5: The distribution of λ and weight value w_+ calculated with WofE.

further confirms that the hot springs in this area are more likely to appear next to the river and under the cliff.

This may be due to the high water pressure provided by the cliff, and the river can provide a sufficient water source. If there is a heat source underground and a suitable water supply channel, hot springs are very highly likely to appear. Therefore, according to the topographical features, screening the high-potential areas of hot springs from the DEM data can not only reduce the calculation scale but also improve the predicted hit rate (Figure 7c).

By solving the equation, the linear equation of the boundary of the range is $y = -\frac{22}{47}x + 2.2$. According to this equation, the high potential areas were defined where hot springs appear based on the slope and the distance from the river (Figure 7b). In this way, the target area of the terrain layer was reduced to 35.06% of the unprocessed area.

3.3 Deep learning

Deep learning, also known as the deep neural network, has attracted the attention of researchers from all fields [39–41] in the last few years. Numerous approaches related to deep learning have been put forward. A typical deep neural network architecture includes a deep belief network (DBN) [42], Convolutional Neural Networks (CNNs), autoencoder [43], and so on. In 2015, CNN's classification accuracy has surpassed human on the 1,000 class of ImageNet dataset [44], including 12,00,000 training images, 50,000 verification images, and 10,000 test images. The CNNs have shown their superiority compared to other technologies [45]. This effective technique is also well applied to the classification of high-resolution and medium-resolution remote sensing images [46–50]. Deep learning can automatically learn features from large data, rather than manual selection. The deep learning technique is also very suitable for the automatic identification of land types in the field since feature selection is not required.

A deep learning network was modeled based on AlexNet. The main parameters are as follows (Figure 8).

The network includes six weighted layers; the first four layers belong to convolutional layers, so sometimes it is also called CNNs, and the rest belong to completely connected layers. The output of the last full-connection layer is sent to a two-way SoftMax layer, causing a judgment that if there have hot springs. Our network achieves the maximization of the multi-class logistic regression goal, similar to the maximization of the logarithmic

Figure 6: Fault buffer with traditional methods compared with the fry method.

possibility average of the accurate labels in the training sample based on the predictive distribution. The core of the latter convolutional layer is relevant to all the core maps in the previous convolutional layer. The neurons in the completely linked layer are in connection with all neurons in the previous layer. The normalized layer responds in accordance with the first and second convolutional layers. The pooling layer follows the normalized layer and the fourth convolutional layer.

We randomly divide the 33 hot springs into two groups, which are the training data of 22 hot springs and the verification data of 11 hot springs, and extract 128×128 multi-dimensional raster data centered on these points, including gravity, river, fault, geological layer, and through window sliding (64 pixels each step, $\times 9$), flip ($\times 2$), and rotation (0, 90, 180, 360 degrees, $\times 4$) for data enhancement, and the data are expanded to 72 times of original data. This ensures that under the limit of the number of hot springs, the model can get enough training to optimize internal parameters.

4 Results

With the use of the fry method, the hot springs are repeatedly drawn in the selected area, and afterward, the results of spatial distribution can be revealed in the following parts. The Fry points orientated from 120° to 160°

and 0° (Figure 9), which reflects an obvious tendency of NW and NS for the distribution of hot springs points. The idea conforms to the research studies of predecessors who claimed that the primary distributions of hot springs focus on both sides of faults in the NW and nearly NNE directions [51]. It can be seen that the geographical data are both inconsecutive and successive [52]. And it is inconsecutive for the frequency distribution of faults in all directions. We adopted a curve-fitting approach to fit the Fry point's distribution data to achieve better fluency of various directions of faults and decrease the "cliff intermittent" of the predicted consequences.

Therefore, the function of the hot springs point aims at performing the verification of reasonability of predicting geothermal regions. This paper analyzed and normalized the weighted layers and then acquired the geothermal potential evaluation layer. According to the past experience, 10, 20, and 80% of the total area are chosen as the anticipated division threshold. The top 10% of the area potential value is chosen as the high potential area of the forecast model, and 10–20% is regarded as the middle potential area (Figure 10).

5 Verification and prediction

All data points were utilized to validate, and the results of statistics related to the four approaches were compared to

Figure 7: (a) Three typical hot springs terrain sections, (b) the relationship between the slope of the location of the hot springs and the distance from the river, and (c) hot springs target area after topographical analysis.

Figure 8: Parametric schematic diagram of deep learning model.

make a comparison between the predictive capacity and the hit condition of the four approaches. Then, receiver operating characteristic (ROC) curve and area under the

curve (AUC) are used as a comparison and verification scheme. ROC is very suitable to be used in the evaluation of binary classifiers. Its precision and recall are popular

Figure 9: Fry points and fitting curve with azimuth from 0° to 180°, the image indicates the azimuth of the structural stress, which affects the stretching or extrusion parameters of faults.

metrics to evaluate the quality of a classification system. ROC curves can be used to evaluate the tradeoff between true- and false-positive rates of classification algorithms.

AUC evaluation is to evaluate the performance of the classifier by measuring the area under the ROC curve (Figure 11).

Although deep learning has achieved a high prediction accuracy, the prediction results after preprocessing for DEM or fault data are still greatly improved. In particular, after fully pre-processing the data, the AUC reached 0.882 from 0.796, when accuracy reached 0.954 from 0.835.

Figure 12 shows the details of the differences in the prediction results of the four experiments. After the prediction results of the fault and DEM optimization, the target area drawn is the smallest under the premise of ensuring hits in areas 1 and 3, where hot springs are exposed. At the same time, in area 2, it is known that no hot springs outcrop was found during the survey, and the smallest target area was drawn. It is worth noting that in the *1c* and *2d* subgraph, there are large blue patterns (marked as medium potential area). After comparison, this is probably because there are wide river valleys and river terraces in the area. Subgraph *2b* reflects the strong interference of an inactive fault with non-principal

Figure 10: Contrast of prediction consequences of geothermal potential among four experiments: (a) the results with both fault and DEM optimized, (b) the results just with fault optimized, (c) the results just with DEM optimized, and (d) the results without fault or DEM optimized.

Figure 11: The prediction accuracy (a) and ROC curve (b) of the four methods.

stress on the prediction potential area. The drawing of these patterns in 1c, 2b, and 2d is more likely to be unnecessary for field surveys.

The high-potential and median-potential areas of the forecasted consequences based on the river valley data are superimposed, and the forecasted areas of hot springs

Figure 12: Detailed comparison of the prediction results of the four methods. The meanings of a–d in the figure are (a) the results with both fault and DEM optimized, (b) the results just with fault optimized, (c) the results just with DEM optimized, and (d) the results without fault or DEM optimized.

in the eastern part of Liaoning Province are acquired. Figure 13 shows key targets marked as green.

After comprehensively considering the prediction results of this experiment and the local government planning and geo-exploration status, we marked 13 target regions; however, for the newly explored geothermal prospecting areas, including III to VIII, IX, X, XII, and XIII, hot springs and geothermal resources have not been detected yet. With the help of sufficient data pre-processing and deep learning, the target area of geothermal and hot springs is reduced to one-third when compared with the previous work [19].

6 Discussion

The eastern part of Liaoning is rich in geothermal resources, but hot springs outcrops mostly appear along hidden faults and secondary faults [19,22,23]. Artificial geological reconnaissance is very difficult to find hot springs in large mountainous areas. Many researchers try to use infrared remote sensing images to find geothermal fields, but due to their spatial resolution, the hot springs exposed through granite cracks on riversides in the mountains like eastern Liaoning can hardly be identified [53,54]. This makes us realize the deficiencies of technology and data-driven methods. In this case, it is very necessary to rely on geological knowledge to optimize the existing technical solutions.

The upper limit about the accuracy of prediction is determined by data, then the lower limit by the method. The nature of deep learning determines that it only discovers patterns from the data provided [44]. For data-driven algorithms, if the data are not preprocessed with geological significance, the predictions will not have much geological significance but only data significance [55]. The fry method is only used to compute the weight of azimuth, and DEM is only used to find the target area. But many data are characterized by multidimension or even high dimension. For instance, fault has fault property, tendency, fault throw, forming times, and other characteristics. Since a suitable preprocessing method can improve the prediction accuracy, there is still a lot of work required to be accomplished for seeking the appropriate preprocessing approaches for mastering various kinds of characteristics of diverse data [55]. For the time being, in the course of processing several aspects of surface characteristics, pattern characteristics, or three-dimensional geologic bodies, if we introduce proper approaches to deal with data, there will be a quite positive role in the consequences of many methods. It is particularly real once subjects are controlled by a small number of elements. It has been proven that even if deep learning models are more inclined to pure data-driven black-box algorithms, researchers can still intervene at a knowledge-driven level by providing pre-processed data, to obtain higher accuracy and more robust results.

The economic level of Eastern Liaoning is not good even for China. For local residents, geothermal or hot

Figure 13: Geothermal key target area in Liaodong based on adequate data preprocessing and deep learning.

springs are not so much an energy source as it is a tourism product with great attraction. There are already many tourists who have been attracted by the hot springs, which has greatly promoted local economic development and poverty alleviation. Therefore, it is believed that this research will promote the further exploration and utilization of local geothermal resources and promote the improvement of the quality of life of local residents.

7 Conclusion

1. The introduction of Geological Significance in Data Preprocessing of deep learning is very significant for Improving the Prediction Accuracy of Hot Springs. In the case of this study, compared with the unprocessed experiment, the accuracy rate of the experiment with fully processed data increased by 11.9 p.p. (83.5 to 95.4%), and the AUC increased by 0.086 (0.796 to 0.882).
2. Geothermal key target regions were forecasted by adequate data preprocessing and deep learning in the eastern part of Liaoning, providing related data to refer to for the sake of further exploring and developing geothermal resources in this area. This is of great significance to the development of local tourism, local poverty alleviation, and improvement of people's quality of life.

Acknowledgments: This study was supported by the National Natural Science Foundation of China (71704177), the Fundamental Research Funds for the Central Universities (2020QN16, 2019CXNL08, and 2019ZDPY-RH02), and the Key Teaching Research Project of China University of Mining and Technology (2019ZD04). The main content of this article was written by Sang Xuejia. Xue Linfu is mainly responsible for the design of the experimental program. Li Xiaoshun helped correct the language and description problems.

References

- [1] Bellotti F, Capra L, Sarocchi D, D'Antonio M. Geostatistics and multivariate analysis as a tool to characterize volcanoclastic deposits: application to Nevado de Toluca volcano, Mexico. *J Volcanol Geotherm Res.* 2010;191:117–28.
- [2] Gudmundsdottir H, Horne R. Prediction modeling for geothermal reservoirs using deep learning. 45th workshop on geothermal reservoir engineering. Stanford, California: Stanford University; February 10–12, 2020.
- [3] Assouline D, Mohajeri N, Gudmundsson A, Scartezzini J-L. A machine learning approach for mapping the very shallow theoretical geothermal potential. *Geotherm Energy.* 2019;7:19.
- [4] Litjens G, Kooi T, Bejnordi BE, Setio A, Ciampi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. *Med Image Anal.* 2017;40:60.
- [5] Hinton GE, Osindero S, Teh YW. A fast learning algorithm for deep belief nets. *Neural Comput.* 2014;18:1527–54.
- [6] He Z, Liu H, Wang Y, Hu J. Generative adversarial networks-based semi-supervised learning for hyperspectral image classification. *Remote Sens.* 2017;9:1042.
- [7] LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. *Proc IEEE.* 1998;86(11):2278–324.
- [8] Yang M, Liu Y, You Z. The Euclidean embedding learning based on convolutional neural network for stereo matching. *Neurocomputing.* 2017;6:195–200.
- [9] Santos JA, Faria F, Calumby R, Torres R, Lamparelli R. A genetic programming approach for coffeecrop recognition. *Geosci Remote Sens Symp.* 2010;1:3418–21.
- [10] Pradhan B, Oh HJ, Buchroithner M. Weights-of-evidence model applied to landslide susceptibility mapping in a tropical hilly area. *Geomat Nat Hazards Risk.* 2010;1:199–223.
- [11] Vearncombe JR, Vearncombe S. The spatial distribution of mineralization: applications of fry analysis. *Econ Geol.* 1999;94:475–86.
- [12] Jianming L. Prediction of geothermal resources in China based on evidence weighting method. Changchun, China: Jilin University; 2012.
- [13] Coolbaugh MF, Zehner RE, Raines GL, Oppliger GL, Kreemer C. Regional prediction of geothermal systems in the Great Basin, USA using weights of evidence and logistic regression in a geographic information system (GIS). *GIS and spatial analysis - 2005 annual conference of the international association for mathematical geology, IAMG 2005;* 2005. p. 505–10.
- [14] Dang W. Evaluation method and application of mineral resources potential based on SVM model and spatial reasoning. Chengdu, China: University of Electronic Science and technology of China; 2014 (in Chinese with English summary).
- [15] Jing Z. Study on comprehensive evaluation of mineral resources (gold ore) in typical areas of Zhejiang section. Hangzhou, China: Zhejiang University; 2015 (in Chinese with English summary).
- [16] He B, Wang D, Chen C. A novel method for mineral prospectivity mapping integrating spatial-scene similarity and weights-of-evidence. *Earth Sci Inform.* 2015;8:393–409.
- [17] Yousefi H, Noorollahi Y, Ehara S, Ehara S, Itoi R, Yousefi A, et al. Developing the geothermal resources map of Iran. *Geothermics.* 2010;39:140–51.
- [18] Smith EP, Lipkovich I, Ye K. Weight-of-evidence (WOE): Quantitative estimation of probability of impairment for individual and multiple lines of evidence. *Hum Ecol Risk Assess.* 2002;8:1585–96.
- [19] Sang X, Xue L, Liu J, Zhan L. A novel workflow for geothermal prospectively mapping weights-of-evidence in Liaoning Province, Northeast China. *Energies.* 2017;10:1069.
- [20] Zdravetski E, Lameski P, Kulakov A, eds., Weight of evidence as a tool for attribute transformation in the preprocessing

- stage of supervised learning algorithms. The 2011 international joint conference on neural networks, 31 July–5 Aug. 2011.
- [21] Jeansoulin R. Review of forty years of technological changes in geomatics toward the big data paradigm. *ISPRS Int J Geo-Inf.* 2016;5:155.
 - [22] Zhang G, Cui Y, Yang S, Zuo G. Distribution characteristics of underground hot water in Liaoning province. *Invest Sci Technol.* 2004;2:40–3.
 - [23] Guangzuo L. Journal of Dandong Wulong behind the thermal field of geological structure and reservoir condition analysis. *Dandong Norm Coll.* 1995;3:38–43.
 - [24] Xikui W, Shanwen Q, Changchun S, Aleksey K, Stepan T, Evgeny M. Cenozoic volcanism and geothermal resources in Northeast China. Symposium on volcanism and resources and environment. Chinese Geological Society; 1999.
 - [25] Geng Y, Sarkis J, Wang X, Zhao H, Zhong Y. Regional application of ground source heat pump in China: a case of Shenyang. *Renew Sustain Energy Rev.* 2013;18:95–102.
 - [26] Ruheng F. Basic features of geological structure in Liaoning. *Liaoning Geol.* 1985;3:189–200.
 - [27] Energy Research Institute of Liaoning. Introduction of geothermal resources utilization in Liaoning province. *Gas Heat.* 1984;3:61–4.
 - [28] East F, Wei Q. The potential of geothermal resources in Liaoning province and Countermeasures for development and utilization of land resources. *Land Resources*; 2008;S1:98–9. (in Chinese).
 - [29] Liu J, Ji M, Ni J, Shen L, Zheng Y, Chen X, et al. The ancient proterozoic extensional tectonic model – taking Jiaodong, Liaodong and Jilin southern regions as examples. *J Changchun Inst Geol.* 1997;2:141–6.
 - [30] Hong Z, Xiaofeng W. The mesozoic tectonic evolution of the southeastern Liaoning Province and its relation to the formation of gold deposits. *Geol Precious Met.* 1995;1:41–9.
 - [31] Guanghui Z, Yubo G, Jianjun Z. Tectonic features of liaoning plate and division of tectonic units. *Geol Resour.* 2011;2:101–6.
 - [32] Yadong Z, Davis GA, Cong W, Jinjiang Z, Changhou Z, Gehrels GE. The main tectonic events and the tectonic background of plate tectonics in the mesozoic belt of Yanshan. *Geol J.* 2000;4:289–302.
 - [33] Wibowo H, Carranza EJM. Data-driven evidential belief predictive modelling of regional - scale geothermal prospectivity in West Java, Indonesia. Proceedings of the 5th European congress on regional geoscientific cartography and information systems: earth and water Barcelona. Institut Cartografic de Catalonia; 13–16 June 2006. p. 243–5.
 - [34] Younker LW, Kasameyer PW, Tewhey JD. Geological, geophysical, and thermal characteristics of the Salton Sea geothermal field, California. *J Volcanol Geotherm Res.* 1982;12:221–58.
 - [35] Fry N. Random point distributions and strain measurement in rocks. *Tectonophysics.* 1979;60:89–105.
 - [36] Noorollahi Y, Itoi R, Fujii H, Tanaka T. GIS integration model for geothermal exploration and well siting. *Geothermics.* 2008;37:107–31.
 - [37] Moghaddam MK, Samadzadegan F, Noorollahi Y, Sharifi MA, Itoi R. Spatial analysis and multi-criteria decision making for regional-scale geothermal favorability map. *Geothermics.* 2014;50:189–201.
 - [38] Wibowo H, Carranza EJM, Barritt SD. Spatial data analysis and integration in geothermal prospectivity mapping: a case study in West Java. Indonesia: International Symposium on Mineral Exploration; 2006.
 - [39] Bouvrie J. Notes on convolutional neural networks. *Neural Nets*; 2006. http://people.csail.mit.edu/jvb/pubs/papers/cnn_tutorial.pdf.
 - [40] Ji S, Zhang C, Xu A, Shi Y, Duan Y. 3D convolutional neural networks for crop classification with multi-temporal remote sensing images. *Remote Sens.* 2018;10:75.
 - [41] Lecun Y, Bengio Y, Hinton G. Deep learning. *Nature.* 2015;521:436.
 - [42] Abadi M. Tensorflow: learning functions at scale. *ACM Sigplan Not.* 2016;51:1–1.
 - [43] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In Proceedings of the 25th international conference on neural information processing systems - Volume 1 (NIPS'12). Red Hook, NY, USA: Curran Associates Inc.; 2012. p. 1097–105.
 - [44] Noda K, Yamaguchi Y, Nakadai K, Okuno HG, Ogata T. Audio-visual speech recognition using deep learning. *Appl Intell.* 2015;42:722–37.
 - [45] Yang M, Liu Y, You Z. The Euclidean embedding learning based on convolutional neural network for stereo matching. *Neurocomputing.* 2017;267:195–200.
 - [46] Lv Y, Duan Y, Kang W, Li Z, Wang FY. Traffic flow prediction with big data: a deep learning approach. *IEEE Trans Intell Transp Syst.* 2015;16:865–73.
 - [47] Guo Y, Liu Y, Oerlemans A, Lao S, Wu S, Lew MS. Deep learning for visual understanding: a review. *Neurocomputing.* 2016;187:27–48.
 - [48] Mukherjee DP, Potapovich Y, Levner I, Zhang H. Ore image segmentation by learning image and shape features. *Pattern Recognit Lett.* 2009;30:615–22.
 - [49] Litjens G, Kooi T, Bejnordi BE, Aaa S, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. *Med Image Anal.* 2017;42:60.
 - [50] He Z, Liu H, Wang Y, Hu J. Generative adversarial networks-based semi-supervised learning for hyperspectral image classification. *Remote Sens.* 2017;9:1042.
 - [51] Jing L, Zhao M, Li P, Xu X. A convolutional neural network-based feature learning and fault diagnosis method for the condition monitoring of gearbox. *Measurement.* 2017;111:1–10.
 - [52] Zhou S, Xu H, Chen X, Yuan JF. Structural analysis of Liaoning Tanggangzi geothermal system research: concept model based. *Ground Water.* 2010;32:24–27.
 - [53] van der Meer F, Hecker C, van Ruitenbeek F, van der Werff H, de Wijkerslooth C, Wechsler C. Geologic remote sensing for geothermal exploration: a review. *Int J Appl Earth Observ Geoinf.* 2014;33:255–69.
 - [54] Gaffar EZ. Remote sensing application on geothermal exploration. *AIP Conf Proc.* 2013;1554:261–4.
 - [55] Huerta, Schade, Granell, eds., Connecting a digital Europe through location and place. Proceedings of the AGILE'2014 international conference on geographic information science, Castellón, June, 3–6; 2014.

Appendix

Figure A1: Bouguer gravity data after utilizing STD processing in this research.

Figure A2: Fault data utilized in this research.

Figure A3: Acquisition of River network data by DEM utilized in this research.

Figure A4: Lithology data utilized in this research.