Machine Learning-Based Evaluation of Susceptibility to Geological Hazards in Yunyang District, Shiyan City, China

Regional geohazard susceptibility evaluation and early warning are effective means of disaster prevention and mitigation. The traditional regional geohazard evaluation has problems such as limited model accuracy and insufficient refinement. With the rapid development of big data and artificial intelligence technology, machine learning algorithms are gradually widely used in geologic hazard evaluation and have achieved better results. The paper uses BP neural network model and support vector machine model in machine learning algorithms to predict regional geologic disaster susceptibility. The paper selects Utopia District of Shiyan City, Hubei Province as the study area, constructs the evaluation database, selects the sample set, and trains the evaluation model with tuning parameter optimization. The results show that the support vector machine model has the highest AUC value and the distribution of geologic hazards in the evaluation results is more accurate. The susceptibility of geologic hazards in Utopia is divided into four categories: low susceptibility, medium susceptibility, medium-high susceptibility and high susceptibility, in which the low susceptibility area accounts for 17.11% of the total area, the medium susceptibility area accounts for 33.57% of the total area, the medium-high susceptibility area accounts for 42.94% of the total area and the high susceptibility area accounts for 36.55% of the total area. The results of the thesis research are of guiding significance for the disaster prevention and mitigation work in Shiyan City Utopia.

Keywords: *geohazard; susceptibility assessment; support vector machine; BP neural network; informativeness modeling*

Introduction

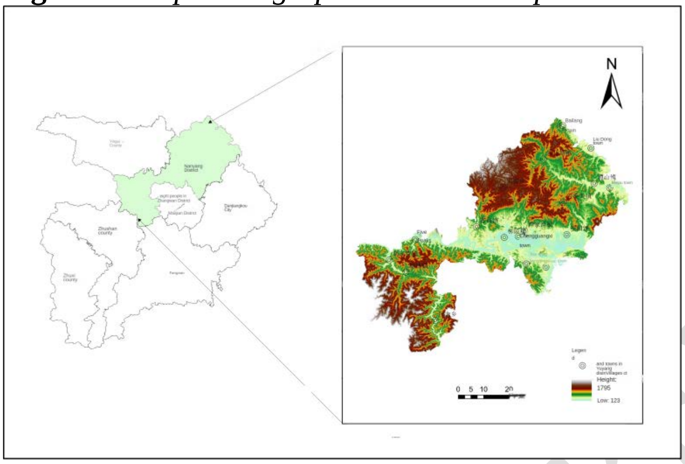
Geohazard risk evaluation can be defined as a systematic process of studying the extent to which a particular impact factor poses a hazard to human society in a given area and time. The main purpose of geohazard risk evaluation is to determine the scope of the risk and to rank the risk in order to provide a scientific and systematic method to reduce the risk. In the evaluation research, researchers have directed their research goals to the improvement of evaluation accuracy in the evaluation process, and with the continuous improvement of machine algorithms, the research on the use of machine learning algorithms in geohazard analysis has been a hot topic nowadays. In 2014, Renneng Bi et al. used an artificial neural network evaluation method to establish an evaluation index system based on the analysis of the distribution and causes of landslides to evaluate landslide susceptibility in the western basin of Hunan[1]; in 2014, Paraskevas Tsangaratos et al. used an artificial neural network in order to better simulate the nonlinear relationship between landslides and geomorphological parameters to evaluate the susceptibility of geologic hazards in the study area in two phases using an artificial neural network model to evaluate the geohazard susceptibility of the study area in two phases[2]; in 2015, Christos Polykretis et al. studied the various factors leading to the genesis of landslides based on 3S technology, established an evaluation index system, and evaluated landslide susceptibility using an artificial network[3]; and in 2019, Joaquín Andrés Valencia Ortiz et al. used a neural network model to evaluate the conditions of the degree of landslide susceptibility in Capitanijo, Colombia, and the results of the evaluation were predictive for the landslide[4]; in 2019, Andang Suryana Soma et al. utilized a combination of logistic regression and artificial neural network evaluation methods to evaluate landslide susceptibility, and the prediction accuracy reached more than 90%[5]; and in 2020. Moayedi Hossein et al. applied the artificial neural network optimized by particle swarm optimization algorithm to the problem of landslide susceptibility map prediction, and the study showed that the artificial neural network optimized by particle swarm optimization algorithm had a good prediction performance[6]; in 2020, L. Bragagnolo et al. selected seven factors such as geomorphology, stratigraphic lithology, etc., and used an artificial neural network model to evaluate the susceptibility of landslide susceptibility map of Brazilian Porto Alegre and Rio de Janeiro regions for landslide susceptibility evaluation[7]; in 2020, Dong Vandao et al. investigated the development and validation of a deep learning neural network model for predicting landslide susceptibility, and the insights provided by this study will be valuable for the further development of landslide prediction models and the spatial evaluation of landslide susceptible areas worldwide[8]; In 2022, Rui Liu et al. selected Zhangzha Town, Sichuan Province and Lantau Island, Hong Kong as the study areas to introduce a convolutional neural network (CNN)-based model for landslide susceptibility assessment, and systematically compared its overall performance with that of traditional random forest, logistic regression, and support vector models, using the ROC curve accuracy test and several statistical metrics to evaluate the model's performance. The results show that both CNN and traditional machine learning based models have satisfactory performance, and the CNN based model has excellent predictive ability and achieves the highest performance 2021[9].In this thesis, two machine algorithms, BP neural network and support vector machine, are used to carry out the evaluation research of geohazard susceptibility in Utopia District of Shiyan City, combining with the evaluation results of the information quantity model, to discuss in depth the performance and differences of the two machine learning algorithms in the evaluation process.

Study Area and Data

Regional situation

Utopia was renamed Shiyan Utopia in 2014 from Utopia County, Utopia is located in northwestern Shiyan City, Hubei Province, upstream of the Hanjiang River, known as "the barrier of E, the gateway to Yu, the throat of Shaanxi, outside the Bureau of Shu". Northeast and Henan Province Xichuan County, southwest and Zhushan County adjacent to the west and Shaanxi Province Baihe County junction, northwest and Uyutsi County intersection, north and Shaanxi Province Shangnan County (Figure 1). It is 92km wide in the north and south, 108km long in the east and west, wide at both ends, narrow in the middle, and only 6km at the narrowest point, resembling the shape of a goldfish, with a land area of 3863km^2.

Figure 1. *Utopia Geographic Location Map*



Grid Division

Considering the area of Utopia and the distribution of evaluation indexes, the grid division unit size of the study area is selected to be 500m*500m, and the Utopia is divided into 16,046 evaluation units according to the grid size of 500m*500m in ArcGIS software, and the attribute data of the evaluation indexes are assigned to each grid unit by the tool of multi-value extraction to the point in ArcGIS software, and the attribute database of the evaluation indexes is established to facilitate the subsequent evaluation study. The attribute data of evaluation indexes are assigned to each grid cell through the multi-value extraction to point tool in ArcGIS software, and the attribute database of evaluation indexes in the study area is established, which is convenient for the subsequent evaluation research.

Selection of Evaluation Factors

According to the principle of evaluation index selection, in order to select representative evaluation indexes and eliminate highly correlated evaluation indexes, therefore the correlation analysis of evaluation indexes is carried out. The thesis uses ArcGIS software to extract the attribute data of 8 evaluation indexes, and conducts Kendall correlation analysis on the 8 evaluation indexes initially selected by SPSS software respectively.The range of the Kendall correlation coefficient τ value is $\left[-1,1\right]$. When $\tau>0$, the evaluation indicators are positively correlated with each other, when $\tau < 0$, the evaluation indicators are negatively correlated with each other, when τ=0, it means there is no correlation, and when τ is close to 1, it means the correlation is highly correlated . The results of Kendall correlation analysis of evaluation indicators are shown in Table 1.

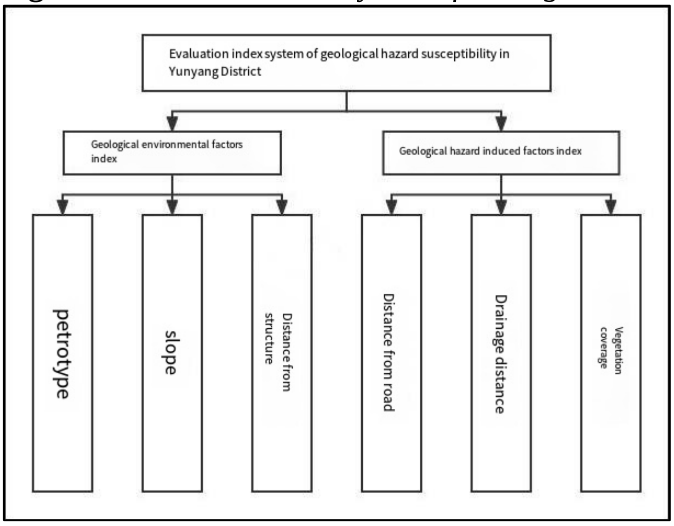
Table 1. *Kendall Correlation Analysis Coefficients Table*

correlation coefficient	roads	geomorphology	tectonic (geology)	elevation	slope direction	rainy season	plant cover	rock group
Distance from road	1							
Landform type	0.006	1						
Distance from structure	0.052	0.001	1					
elevation	0.125	0.512	-0.022	1				
slope direction	-0.013	0.404	-0.004	0.614	1			
Distance to water system	0.080	0.06	0.054	0.038	0.011	1		
vegetation cover	0.205	0.007	0.011	0.293	-0.038	0.063	1	
Rock group type	0.063	-0.044	-0.068	0.057	-0.017	-0.12	0.157	1

The results of Kendall correlation analysis show that there is a high correlation between slope gradient and slope direction, geomorphology and slope gradient and slope direction, with correlation coefficients of 0.614, 0.512, and 0.404, respectively, which may be due to the fact that the slope gradient, slope direction, and geomorphology are all analyzed according to the elevation data by the ArcGIS software, and therefore the correlation is high. The Kendall correlation coefficients between the remaining geohazard evaluation indicators were all ≤|0.3|. Therefore, the geomorphology and slope direction indicators with high Kendall correlation coefficients were excluded.

After Kendall analysis of the indicators in Utopia, it was determined that the geohazard susceptibility evaluation index system of the dissertation finally consists of the following six indicators: ① distance from roads, ② distance from tectonics, ③ slope, ④ distance from water system, ⑤ vegetation coverage, and ⑥ rock group category. The final established evaluation index system of geologic disaster susceptibility in Utopia is shown in Figure 2.

Figure 2. *Evaluation Index System of Geologic Disaster Susceptibility in Utopia*



Data Processing

Evaluation system, data processing of evaluation indicator layers in ArcGIS software. The element class files of each indicator layer were converted into raster files, and then the reclassification function in the ArcGIS toolbox was used to classify each indicator according to its defined category. Subsequently, the multi-value extraction to point function was used to extract the categorized attributes of the evaluation indicators into the evaluation grid cells of the study area, and each cell had a corresponding number FID, so that the attributes of the evaluation cells had been given.

There are 892 disaster points in the study area, and the disaster points are divided into the training and validation sets of the evaluation model in the ratio of 7:3, i.e., 627 disaster points are used for training and evaluation of the model, and 265 disaster points are used for the subsequent testing of the accuracy of the model evaluation results.

The machine learning algorithm needs sample set to train the model, which consists of input indicators and output indicators, where the input indicators are the indicator attributes of the evaluation cells, and the output indicators are the results of the susceptibility partition. Considering the number of evaluation grid cells in the study area, the paper selects 627 disaster grid cells as the sample set of disaster points, and then randomly selects twice the number of grid cells as the sample set of non-disasters from the grid cell area of non-disasters, i.e., 1254 non-disasters, to form the sample set, and then according to the "Shiyan City Geological Disasters Refined Meteorological Risk Early Warning Forecast Project" project research in the partitioning area, the sample set is composed of input indicators and output indicators. The sample set is composed of 1254 non-hazardous point samples, and then the susceptibility zoning results of the sample set are extracted from the zoning results of the "Shiyan City Geological Hazard Refined Meteorological Risk Early Warning and Forecasting Project".

Methodologies

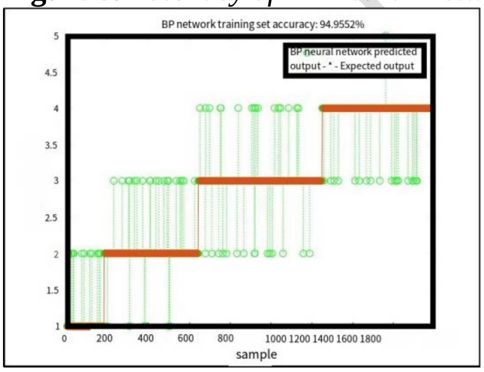
BP Neural Network

Modeling of BP Neural Networks

When solving problems, it is crucial to construct a reasonable model. In this study, we used a 3-layer neural network, with the input layer containing 6 nodes corresponding to landslide susceptibility evaluation indexes and the output layer containing 1 node. Among them, the hidden layer is 1 layer. In practice, although the number of nodes in the hidden layer can be chosen arbitrarily, we found that decreasing the number of nodes in the hidden layer increases the model output error, while increasing the number of nodes in the hidden layer reduces the model output error. However, increasing the number of hidden layer nodes leads to an increase in the number of weight matrices. Therefore, weighing the accuracy and efficiency, this study chooses a moderate number of hidden layer nodes. Through extensive debugging and training, the number of hidden layer nodes of this BP neural network is set to 15.

The parameters of the BP neural network are set as follows, the maximum number of training times is set to 1000, the learning rate is set to 0.01, and the learning accuracy is set to $1e^{-8}$, in order to achieve the desired desired value of the output results, it is necessary to repeatedly train the model until the error reaches the requirements before stopping the training. The model training process is shown in Figures 4.3, 4.4 and 4.5. Eventually, the highest model training set accuracy of the BP neural network model over the multiple training process is 94.96% as shown in Figure 3.

Figure 3. *Accuracy of BP Neural Network Training Set*



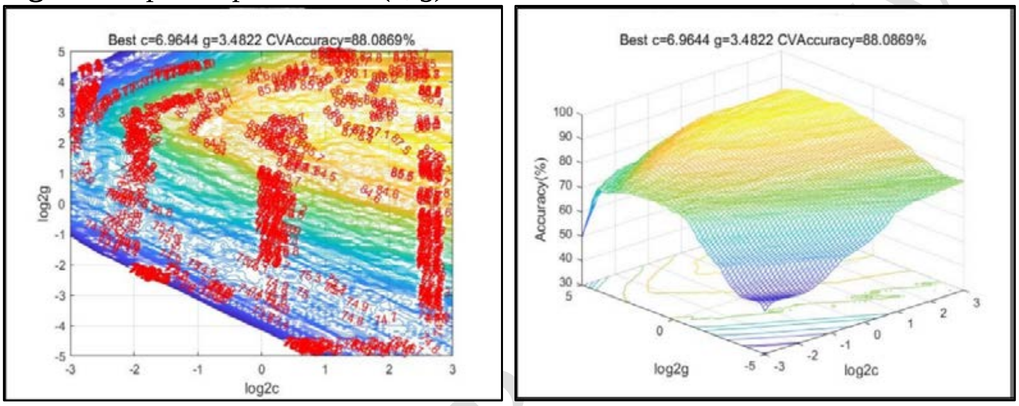
Support Vector Machine

Support Vector Machine Modeling

The paper uses support vector machines with four kinds of kernel functions as evaluation models to carry out the evaluation of geohazard susceptibility in the study area, respectively. The LN-SVM, PL-SVM, RBF-SVM, and Sigmoid-SVM evaluation models were established by MATLAB platform and LIBSVM software package respectively. In the support vector machine evaluation model, the selection of appropriate kernel function parameter g and error penalty parameter c is crucial to the model performance of SVM.

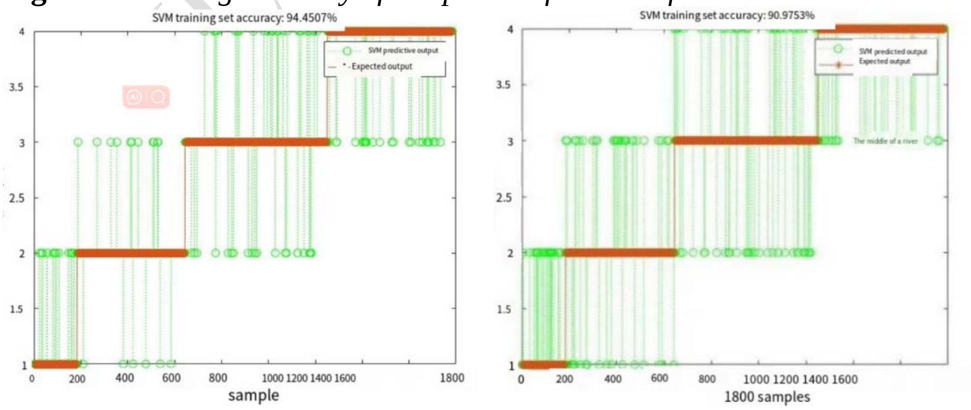
In this paper, the K-fold cross-validation method is used to verify the training performance of the optimal parameters of the SVM model.K-fold cross-validation, that is, the data are randomly and evenly divided into K parts, of which (K-1) parts are used to build the model, and the validation is carried out in the remaining part of the data. In this paper, the value of K is chosen as 5, and the sample set is imported into the MATLAB platform, and the optimal penalty parameter c of the SVM model is sought by the K-fold cross-validation method as 6.9644 , and the parameter g is 3.4822, and the optimal accuracy of cross-validation is 88.09%, as shown in Fig. 4.

Figure 4. *Optimal parameters (c, g)*

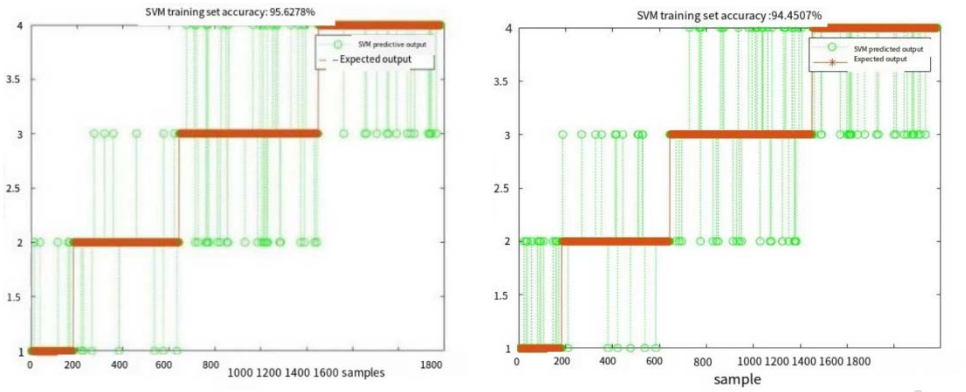


The LN-SVM, PL-SVM, RBF-SVM, Sigmoid-SVM models and the optimal parameters (c, g) are trained on the sample set to obtain the LN-SVM, PL -SVM, RBF-SVM, and Sigmoid-SVM optimal models, and the accuracy of the trained sample set with different kernel function models are 94.4507%, 90.9753%, 95.6278%, and 94.4507%, respectively, as shown in Fig. 5.

Figure 5. *Training accuracy of sample set of 4 kernel function models*



LN-SVM model PL-SVM model



RBF-SVM model Sigmoid-SVM model

Among them, RBF-SVM model has the highest training accuracy, followed by Sigmod-SVM model and LN-SVM model, and PL-SVM model has the lowest accuracy. It can be seen that the training effect of RBF-SVM model is the best, so the RBF-SVM model was finally selected as the training model for geohazard susceptibility assessment in the study area to predict the results of susceptibility zoning in the study area.

Information Quantity Evaluation Model

<u>Results of single-factor informativeness calculations</u>

The grading of each evaluation factor and the distribution of disaster points in Utopia have been briefly counted above, and the information quantity of each evaluation factor was calculated according to the formula, and the information quantity of a single factor was brought into the attribute statistical table of the study area to calculate the total information quantity Ii of the evaluation grid in the study area, and subsequently, the total information quantity value was imported into ArcGIS software, and according to the method of natural breakpoints, it was classified into geohazard low susceptibility zone, medium susceptibility zone, medium high susceptibility zone and high susceptibility zone.

Results

Visualization of the results of the three model evaluations

The results of the three model evaluations were imported into the evaluation grid attributes of the study area in ArcGIS software according to the corresponding grid number for visualization and analysis, and the study area was classified into low susceptibility, medium susceptibility, medium-high susceptibility and high susceptibility according to the respective evaluation results, and the susceptibility zones of the evaluation results of the three models are shown in Figures 6, 7 and 8.

Figure 6. *Evaluation result of the susceptibility of information model in Yunyang District*

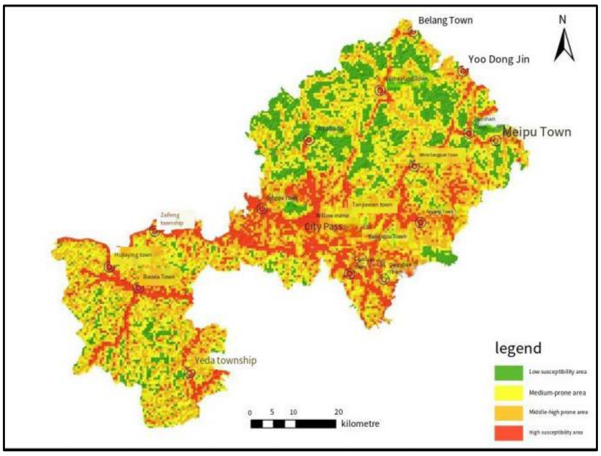


Figure 7. *Utopia BP neural network model susceptibility evaluation results*

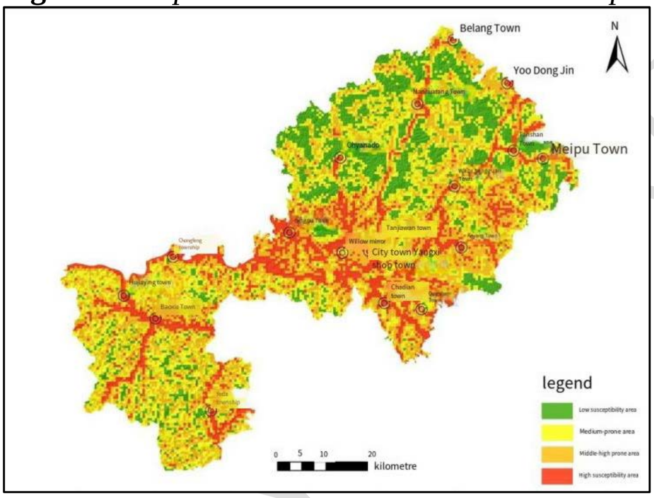
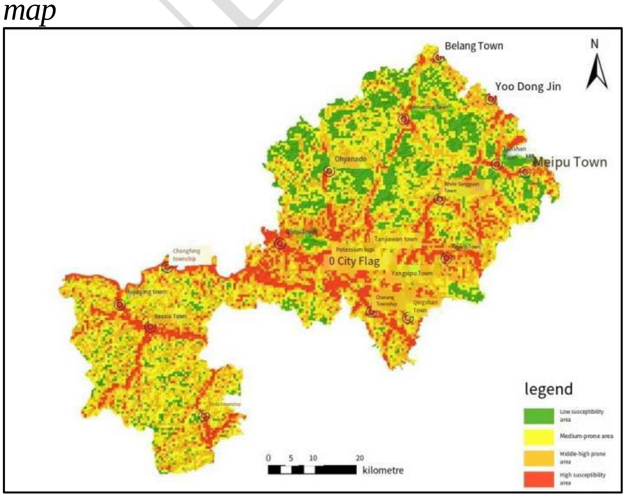


Figure 8. *Utopia support vector machine model susceptibility evaluation result map*



It can be seen that the susceptibility evaluation zoning maps of the three models are very close to each other, and the general distribution is as follows: the high susceptibility zone is distributed in the central and southwestern part of the study area, where more geohazards have already occurred; the medium and high susceptibility zones are mainly distributed in the central and eastern part of the susceptibility zone, and most of them are distributed along the perimeter of the high susceptibility zone; the medium susceptibility zones are distributed in the northern and southwestern parts of the study area, and the low susceptibility zones are distributed in the northern and northeastern parts of the study area. The medium-prone areas are located in the north and southwest of the study area, and the low-prone areas are mainly located in the north and northeast of the study area.

Comparison of the accuracy of the evaluation results of the three models

Precision testing

Figure 9. *ROC curves for the three evaluation models*

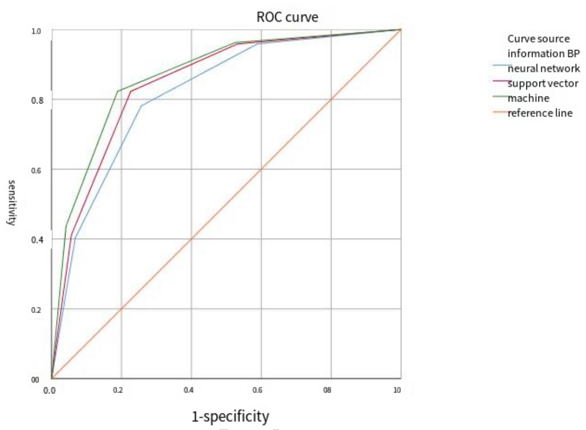


Table 2. *AUC values of the three evaluation models*

Evaluation Models	AUC Value
Information quantity model	0.817
BP Neural Network Model	0.847
Support Vector Machine Model	0.868

According to Figure 9 and Table 2, it can be seen that the AUC values of the three evaluation models are relatively similar, and the AUC value of the support vector machine model is the largest, 0.868, followed by the BP neural network model, 0.847, and the last is the informative model, 0.817, so the support vector machine model has the best prediction effect in the ROC curve test.

Table 3. *Distribution of disaster sites*

Evaluation Models	Low susceptibility zone	Medium susceptibility zone	Medium-high susceptibility zone	High susceptibility zone
Information quantity model	25/2.8%	166/18.61%	331/37.11%	370/41.48%
BP Neural Network Model	25/2.8%	163/18.27%	338/37.89%	366/41.03%
Support Vector Machine Model	23/2.58%	162/18.16%	326/36.55%	383/42.94%

According to the Table 3, it can be seen that the distribution of disaster points in the three models is also relatively similar, in the evaluation results of the informativeness model, the distribution of disaster points in the prone area accounted for 2.8%, in the medium-prone area accounted for 18.61%, in the medium-high prone area accounted for 37.11%, and in the high prone area accounted for 41.48%; the evaluation results of the BP neural network model, the disaster points in the prone area In the evaluation results of BP neural network model, the distribution of disaster points in prone area accounts for 2.8%, in medium prone area accounts for 18.27%, in medium-high prone area accounts for 37.89%, and in high prone area accounts for 41.03%; in the evaluation results of support vector machine model, the distribution of disaster points in prone area accounts for 2.58%, in medium prone area accounts for 18.16%, in medium-high prone area accounts for 36.55%, and in high prone area accounts for 42.94%; the distribution of disaster points in the three model evaluation results in the medium-high and high susceptibility zones accounted for 78.59%, 78.92% and 79.49%, respectively. Obviously, the distribution of disaster points in the support vector machine model is the most reasonable, combined with the accuracy test results of the three models, the evaluation results of the support vector machine model are selected as the results of the geohazard susceptibility zoning in the study area.

Conclusion and Discussion

Reach a Verdict

(1) Based on the geological environment condition of Utopia, six disaster-causing influence factors, slope, rock group type, distance to water system, distance to tectonics, distance to slope and vegetation cover, were selected to construct the sample dataset for the early warning model.

(2) Based on 1881 training samples, two machine learning algorithms, BP neural network model and support vector machine model, and an informativeness evaluation model were used to carry out the evaluation study of regional geohazard susceptibility.

(3) Based on the prediction results and accuracy verification of the BP neural

network algorithm model and the support vector machine algorithm model, the machine learning algorithms have excellent performance in regional geohazard susceptibility evaluation, and the prediction results are better than the traditional informativeness model.

Discussion

(1) Machine learning algorithm model in the process of model design to the visualization of prediction results, the selection of relevant parameters has a great impact on the model accuracy, and the selection of optimal parameters is one of the goals of model design.
(2) In the evaluation process, the steps of selection of evaluation indexes, grading of evaluation indexes, division of evaluation units and partitioning of susceptibility results will have an impact on the evaluation results, and there is no uniform specification in the current evaluation of regional geohazard susceptibility, and the phenomenon of strong subjectivity is common.

References

[1] Renneng B, Markus S, Joachim R, et al. Landslide susceptibility analysis based on ArcGIS and Artificial Neural Network for a large catchment in Three Gorges region, China[J]. Environmental Earth Sciences, 2014, 72(6).
[2] Paraskevas T, Andreas B. Estimating landslide susceptibility through a artificial neural network classifier[J]. Natural Hazards, 2014, 74(3).
[3] Christos P, Maria F, Christos C. A comparative study of landslide susceptibility mapping using landslide susceptibility index and artificial neural networks in the Krios River and Krathis River catchments (northern Peloponnesus, Greece)[J]. Bulletin of Engineering Geology and the Environment,2015,74(1)
[4] Joaquín A V O, Antonio M M. A neural network model applied to landslide susceptibility analysis (Capitanejo, Colombia)[J]. Geomatics Natural Hazards & Risk, 2018, 9(1):1106-1128.
[5] Andang S S, Tetsuya K, Hideaki M. Optimization of causative factors using logistic regression and artificial neural network models for landslide susceptibility assessment in Ujung Loe Watershed, South Sulawesi Indonesia[J]. Journal of Mountain Science,2019,16(2).
[6] Moayedi H, Mehrabi M, Mosallanezhad M, et al. Modification of landslide susceptibility mapping using optimized PSO-ANN technique[J]. Engineering with Computers, 2019, 35(3):967-984.
[7] Bragagnolo L, Silva R V, Grzybowski J M V. Artificial neural network ensembles applied to the mapping of landslide susceptibility[J]. Catena, 2020, 184.
[8] Dong V D, Abolfazl J, Mahmoud B, et al. A spatially explicit deep learning

neural network model for the prediction of landslide susceptibility[J]. Catena, 2020, 188.

[9] Liu R, Yang X, Xu C, et al. Landslide susceptibility mapping based on convolutional neural network and conventional machine learning methods. 2021.