# A probability prediction model for flood disasters based on Multi-layer Perceptron

**Yiquan Wang*[1], Jialin Zhang[1] and Yuhan Chang[1]**

[1]College of Mathematics and System Science, Xinjiang University, Urumqi, Xinjiang,830046,China

*Corresponding author's e-mail: ethan@stu.xju.edu.cn

**Abstract**. Flood disasters are characterized by high frequency, strong destructive power, and wide impact range. The prediction of flood disasters holds great significance. This article proposes a flood disaster prediction model based on multi-layer perceptron. Firstly, the Spearman correlation coefficient and random forest feature importance algorithm are utilized to identify the feature indicators that have the greatest impact on the model. Secondly, an MLP neural network is established and trained and optimized. Experimental findings demonstrate that the model accurately forecasts the likelihood of flood disasters through sample processing. The coefficient of determination of this model can reach around 85.27%.

## 1. Introduction

Floods, resulting from intense or sustained precipitation that results in the submergence of low-lying regions, stand as one of the most impactful natural calamities globally. [1] Flood disasters have wrought havoc on the safety of residents in affected areas, a harsh reality that compels us to deepen our scientific comprehension and bolster our effective response to such disasters. Consequently, predicting the probability of flood disasters is of great significance. Flooding is a complex process affected by precipitation events, basin characteristics, and natural geographical conditions. The flood process exhibits strong nonlinearity, non-stationarity, and stochastic characteristics. [2] At present, the utilization of remote sensing and Geographic Information System technology for flood risk delineation has progressively emerged as the primary approach for identifying flood risks. [3] Commonly used methods include Analytic Hierarchy Process, Frequency Ratio models, and machine learning models.

In 2016, Khosravi et al. [4] utilized the Analytic Hierarchy Process (AHP) to evaluate flood risk by determining the relative importance of various factors influencing flood susceptibility.In 2020, Costache et al. [5] utilized the Fuzzy Analytic Hierarchy Process (FAHP) to identify and categorize the valleys within the study area based on their susceptibility to flash floodsThe AHP model was utilized to compute the flood potential index along the mountain flood valleys, in order to ascertain the potential for flooding caused by the propagation of mountain floods.However, the AHP method largely relies on the subjective judgment of scholars, which assigns weights to each indicator feature based on experience. This may lead to subjective bias in the results, thereby affecting the objectivity of the final outcome.

Based on the frequency ratio model, Youssef et al. [6] applied an ensemble method of Frequency Ratio (FR) and Logistic Regression (LR) in 2015. The combination of these two statistical methods can generate a comprehensive model that assesses the impact of various conditional factors and the influence of different classes of each conditional factor on landslide occurrence, providing accurate assessments for disaster management and decision-making.In 2016, Khosravi et al.[4] conducted a binary statistical analysis (BSA) to analyze the impact of various factors on floods, creating receiver operating characteristic curves and the area under the curve (AUC) for different flood sensitivity maps. However, it is important to note that the frequency ratio model is limited by the time scale of historical data and may overlook the interaction between factors.

With the emergence of artificial intelligence, machine learning models have become widely utilized. In 2014, Radmehr et al. [7] employed an Artificial Neural Network (ANN) as an alternative approach to the weighting process used by decision makers in order to address disagreements among them during decision-making analysis. In 2018, Khosravi et al. [8] conducted a study testing four machine learning models based on decision trees for flash flood susceptibility mapping.

This article establishes a flood disaster probability prediction model based on multi-layer perceptron. Firstly, conduct Spearman correlation analysis to determine the relationship between variables. Then, random forest feature importance analysis will be used to evaluate the importance of feature indicators. Finally, a multi-layer perceptron model will be trained and L2 regularized to predict the probability of flood disasters.

## 2. Preliminaries

### 2.1. *Data acquisition and preprocessing*

This study is mainly based on flood data from the Asia and Pacific Mathematical Contest in Modeling, provided by reference [9], including flood occurrence probability, infrastructure deterioration, terrain drainage, monsoon intensity and other flood indicator data. Due to the fact that the original observation data contains over 700000 flood related information, this study first preprocessed the data.

To ensure the integrity and consistency of flood indicator data, and to eliminate potential noise and redundant information, we first conducted a series of screenings and processing for potential missing and outlier values in the dataset.After review, there are no missing values in the dataset, and the results are plotted in Figure 1.

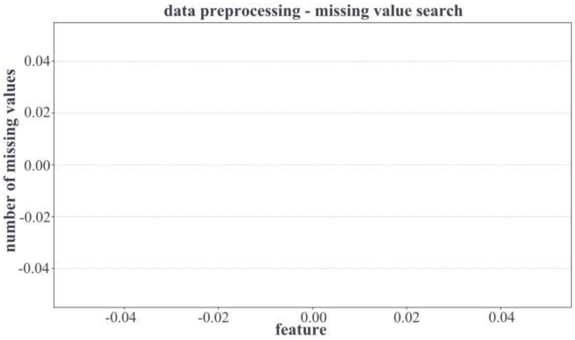

**Figure 1.** Data preprocessing - missing value search

### 2.2. *Spearman correlation analysis*

We used spearman correlation coefficient for correlation research, which measures the non parametric correlation between variables based on rank and is mainly suitable for revealing the relationship between two continuous variables. [10] The absolute value of the Spearman correlation coefficient approaches 1, indicating a stronger correlation.N represents the number of samples, while d indicates the difference in position of the paired variables after sorting them individually.The spearman correlation coefficient calculation formula is:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \tag{1}$$

The Spearman correlation coefficients between various features and flood probability are presented in Table 1.:

**Table 1.** Correlation coefficients between different features and the probability of flooding

| feature | coefficient | feature | coefficient |
|---|---|---|---|
| flood probability | 1.000000 | population score | 0.188808 |
| deterioration of infrastructure | 0.192852 | landslide | 0.187898 |
| Terrain drainage | 0.191362 | climate change | 0.187465 |
| monsoon intensity | 0.191353 | deforestation | 0.187441 |
| Dam quality | 0.189720 | invalid disaster prevention | 0.186922 |
| sedimentation | 0.189705 | agricultural practice | 0.186685 |
| river management | 0.189569 | others | <0.18614 |

### *2.3. Importance analysis of random forest features*

Random forest is an ensemble machine learning algorithm that is a classifier composed of multiple decision trees that aggregate their prediction results to improve the accuracy and stability of classification. The results of Figure 2 indicate that inadequate planning, coastal vulnerability, terrain drainage, urbanization, and population score are the main factors affecting flood risk.

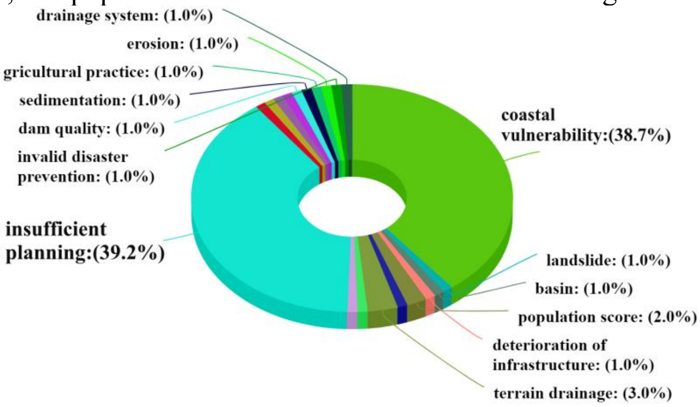

**Figure 2.** Analysis results of the importance of random forest features

### *2.4. Sensitivity analysis*

Sensitivity analysis is a crucial research technique aimed at investigating the responsiveness of changes in the state or output of a model to variations in system variables or environmental conditions. This is achieved by adjusting a certain parameter within the model and observing its impact on the model results, which can effectively unveil the extent of influence that specific parameters exert on model output. The findings are presented in Figure 3, as well as Table 2 and Table 3.

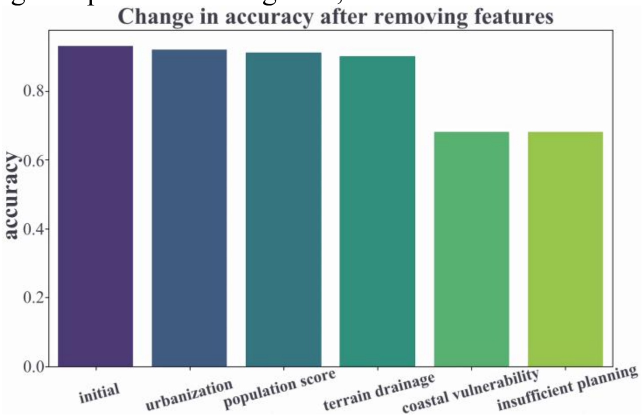

**Figure 3.** Sensitivity analysis result chart

**Table 2.** Sensitivity analysis of initial situation

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.94 | 0.94 | 0.94 | 261683 |
| 1 | 0.93 | 0.93 | 0.93 | 272199 |
| 2 | 0.93 | 0.93 | 0.93 | 354086 |
| accuracy | | | 0.93 | 887968 |
| macro avg | 0.93 | 0.93 | 0.93 | 887968 |
| weighted | 0.93 | 0.93 | 0.93 | 887968 |

**Table 3.** Model accuracy after removing a certain feature

| condition | accuracy（%） |
|---|---|
| initial | 0.93 |
| remove urbanization | 0.92 |
| remove population score | 0.91 |
| remove terrain drainage | 0.90 |
| remove coastal vulnerability | 0.68 |
| remove planning deficiencies | 0.68 |

From this, it can be seen that inadequate planning and coastal vulnerability are the two characteristics that have the greatest impact on flood risk prediction, and removing them will significantly reduce model performance. Terrain drainage, urbanization, and population scores are also important features, and removing them can lead to a decrease in model performance.

## 3. Establishment of Multi Layer Perceptron Modelreliminaries

The multi-layer perceptron model is a fully connected feedforward neural network model that continuously modifies weight values during training iterations to optimize various training parameters.[11]The basic structure of MLP includes three layers: input layer, hidden layer, and output layer. The input layer receives data features as its input, with each feature corresponding to an individual input neuron. The hidden layer is situated between the input layer and the output layer. It may consist of one or more hidden layers, each containing multiple neurons. The output layer produces the model's predicted results, with each output corresponding to an output neuron. These neurons are typically used to represent classification categories or regression values in academic papers.

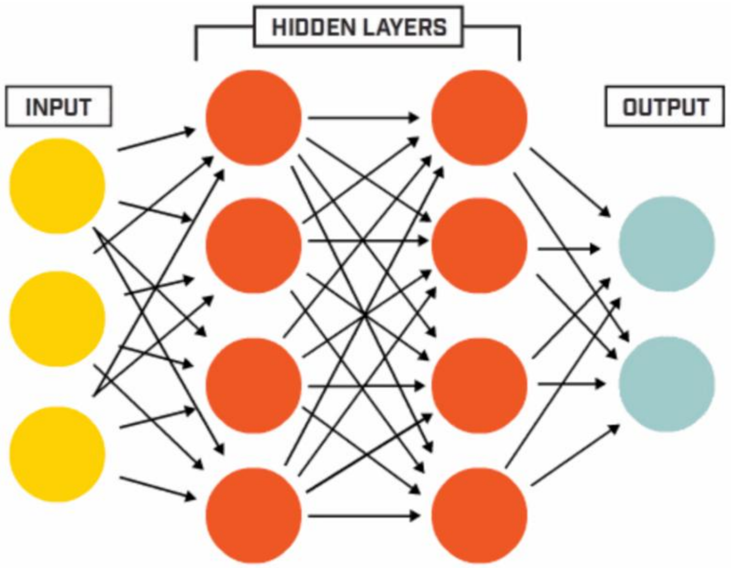

**Figure 4.** Schematic diagram of multi-layer perceptron model principle

This article utilized a total of 800,000 samples, which were divided into training and validation sets in a 7:3 ratio. Owing to the substantial sample size and diversity of the dataset, there is no need to use data enhancement techniques. Then, in the MLPRegressor function of sklearn, we use the Adam optimizer to accelerate the model training process and reduce the risk of overfitting. We set the hidden layer sizes to 64 and 32, the activation function to the corrected linear unit function (RELU), and the maximum iteration count to 500.

The Adam optimizer is a gradient descent algorithm utilized for training neural networks. It integrates the momentum algorithm and adaptive learning rate algorithm, resulting in expedited convergence and enhanced generalization ability through the calculation of distinct adaptive learning rates for each parameter. The update rules for the Adam optimizer are as follows:

$$\begin{gathered} m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t \\ v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2 \\ \widehat{m}_t = \frac{m_t}{1-\beta_1^t} \\ \widehat{v}_t = \frac{v_t}{1-\beta_2^t} \\ \theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{\widehat{v}_t + \epsilon}} \widehat{m}_t \end{gathered} \tag{2}$$

Among them, $g_t$ is the gradient of the parameter, $\beta_1$ and $\beta_2$ are the attenuation coefficients of two exponential weighted averages, $\widehat{m}_t$ and $\hat{v}_t$ are the moving averages of the gradient after deviation correction, $\theta_{t+1}$ is the updated parameter, $\eta$ is the learning rate, and $\epsilon$ is a very small constant used to avoid dividing by zero.

The training loss is shown in Figure 5.

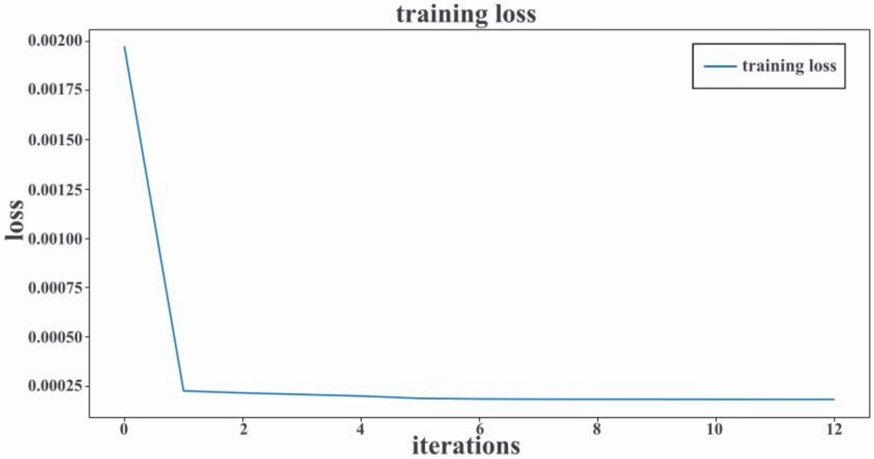

**Figure 5.** Training loss

The measurement standard we use to verify the accuracy of the model is $R^2$(R − squared ,also known as the coefficient of determination), which indicates the extent to which the regression model fits the sample data.The closer the absolute value of $R^2$ is to 1, the better the model's fit to the data. The formula is as follows:

$$R^2 = 1 - \frac{\sum_{k=1}^{n} (y_k - \hat{y}_k)^2}{\sum_{k=1}^{n} (y_k - \bar{y}_k)^2} \tag{5}$$

Among them, $y_k$ is the true target value of the i-th observation value, and $\hat{y}_k$ is the predicted value of the model for the i-th observation value; $\bar{y}$ is the average target value of all observed values; n is the number of samples.

The final results we obtained are shown in Figure 6 and Table 4,The blue line depicts the actual value, while the red line represents the predicted value.

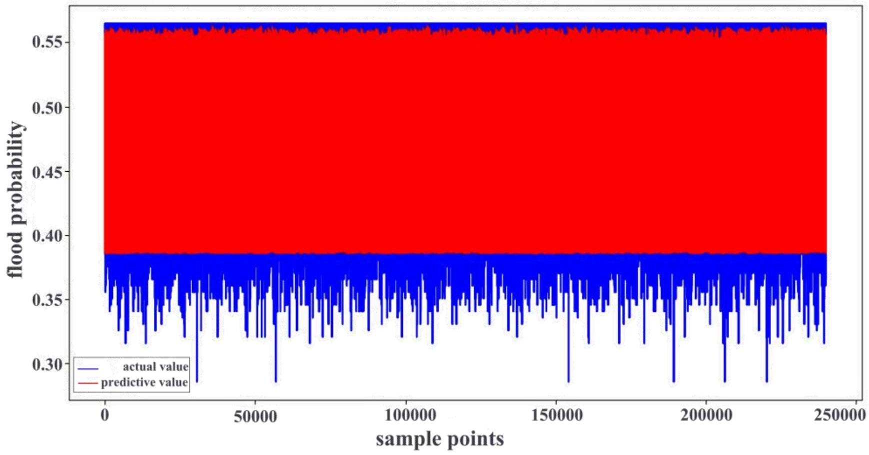

**Figure 6.** Line graph of flood probability prediction

**Table 4.** Model prediction results

| | |
|---|---|
| mean square error of training set | 0.00036385 |
| mean square error of the test set | 0.00036374 |
| training set $R^2$ | 0.79732624 |
| test set $R^2$ | 0.79626938 |

Based on the above results, we optimized the neural network model by using L2 regularization strategy to alleviate the problem of overfitting and improve the model's generalization ability. Through continuous optimization and iteration, the optimized code has added the alpha=0.001 parameter, the prediction accuracy and practicability of the model can be further enhanced. It introduces an additional regularization term in the loss function of the model by applying a penalty term to the square of the model coefficients.

Assuming we have a linear regression model, the mean square error loss function is:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{3}$$

L2 regularization introduces a regularization term into the original model's loss function in order to penalize the magnitude of the parameters.The form of the L2 regularization loss function is as follows:

$$\mathrm{Loss} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda\sum_{j=1}^{p}\omega_j^2 \tag{4}$$

Among them, $\lambda$ is a non-negative regularization parameter used to control the strength of the regularization term, which needs to be selected through methods such as cross validation.

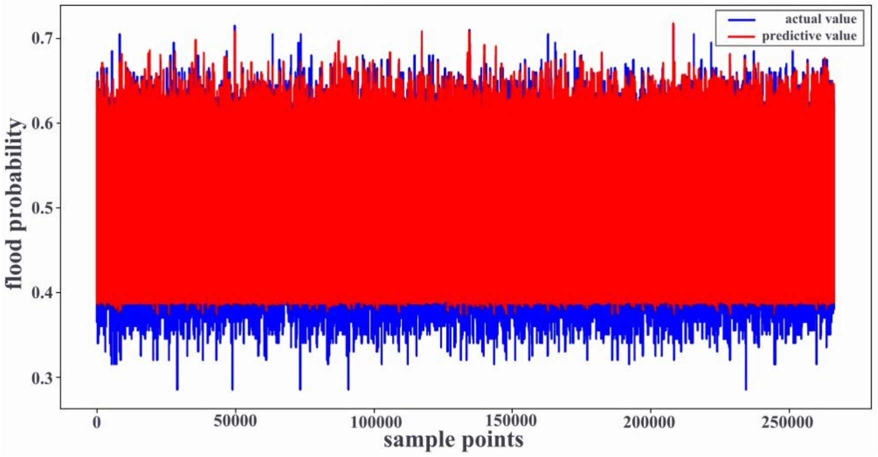

**Figure 7.** Comparison between actual and predicted flood probability values

**Table 5.** Optimized model prediction results

| | |
|---|---|
| mean square error of training set | 0.00036912 |
| mean square error of the test set | 0.00036816 |
| training set $R^2$ | 0.85331226 |
| test set $R^2$ | 0.85271417 |

The findings of the enhanced model are presented in Figure 7 and Table 5, which demonstrate the comparison between actual disaster data and predicted results. The mean square errors of both the training and testing sets are exceptionally small, at 0.000369 and 0.000368, respectively. The $R^2$ values of the test set and training set are 0.8533 and 0.8527, respectively. This indicates that although the number of indicators in the model has decreased, our $R^2$ has significantly improved, indicating that the effectiveness of the model has been further strengthened. The model's prediction results are more accurate.

## 4. Conclusion

In this study, a novel mathematical modeling method based on multi-layer perceptron is proposed. By combining correlation analysis with traditional random forest algorithm and conducting sensitivity analysis, we have identified features that have a significant impact on flood probability, eliminated other interference factors, reduced the subjectivity of parameter weights, and improved the interpretability of the model. The experimental results demonstrate that the proposed model exhibits strong predictive performance and can serve as a valuable reference for predicting the likelihood of flood disasters and mitigating other natural calamities.

## References

[1] Shi Peijun, Yuan Yi. (2014) Integrated Assessment of Large-Scale Natural Disasters in China. Progress in Geography[J], 33 (9):1145-1151.

[2] Yuanhao Xu, Caihong Hu, Qiang Wu, et al. (2022) Research on particle swarm optimization in LSTM neural networks for rainfall-runoff simulation.Journal of Hydrology[J],Volume 608.

[3] Zeng Ziyue, Xu Jijun, Wang Yongqiang, et al. (2020) Advances in flood risk identification and dynamic modelling basedon remote sensing spatial information[J].Advances in Water Science,31(3):463-472.

[4] Khosravi, K., Nohani, E., Maroufinia, E. et al. (2016) A GIS-based flood susceptibility assessment and its mapping in Iran: a comparison between frequency ratio and weights-of-evidence bivariate statistical models with multi-criteria decision-making technique. Nat Hazards 83, 947-987.

[5] Costache, R., Barbulescu, A., Pham, Q.B. (2021) Integrated Framework for Detecting the Areas Prone to Flooding Generated by Flash-Floods in Small River Catchments. Water , 13, 758.

[6] Youssef, A.M., Pradhan, B., Jebur, M.N. et al. (2015) Landslide susceptibility mapping using ensemble bivariate and multivariate statistical models in Fayfa area, Saudi Arabia. Environ Earth Sci 73, 3745–3761.

[7] Radmehr, A., and Shahab A. (2014) Developing strategies for urban flood management of Tehran city using SMCDM and ANN. Journal of Computing in Civil Engineering 28.6 : 05014006.

[8] Khosravi, K. et al. (2018) A comparative assessment of decision trees algorithms for flash flood susceptibility modeling at Haraz watershed, northern Iran. Science of the Total Environment 627: 744-755.

[9] Asia and Pacific Mathematical Contest in Modeling.2024.http://www.apmcm.org/detail/2478.

[10] Taoyu Z., Penghua S. (2024) Predicting Surface Roughness of Parts Manufactured by the Fused Deposition Modeling Based on Coupled Machine Learning models[J].China Plastics Industry,52(05):116-123.

[11] Chong Z., Mo C., Yuanyuan L. et al. (2024)Research on Coupled Prediction Model of Meteorological and Water Quality Based on Machine Learning[J/OL].Journal of China Hydrology,1-9.