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, severe destructive power, and extensive impact. The prediction of flood disasters holds great significance. This paper proposes a flood disaster prediction model based on a multi-layer perceptron (MLP). Firstly, the model employs the Spearman correlation coefficient and random forest feature importance algorithm to identify the most influential feature indicators. Subsequently, an MLP neural network is established, trained, and optimized. Experimental findings demonstrate that the model accurately forecasts the likelihood of flood disasters through sample processing, achieving a coefficient of determination of approximately 85.27%.

1. Introduction

Floods, caused by intense or sustained precipitation that inundates low-lying regions, are among the most significant natural disasters worldwide [1]. Flood disasters have wrought havoc on the safety of residents in affected areas, underscoring the need for enhanced scientific understanding and improved response strategies. 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]. 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 the Analytic Hierarchy Process (AHP), Frequency Ratio models, and machine learning approaches.

In 2016, Khosravi et al. [4] utilized the 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 floods. The AHP model was used to compute the flood potential index along the mountain flood valleys to ascertain the potential for flooding caused by the propagation of mountain floods. However, the AHP method primarily relies on the subjective judgment of scholars, who assign weights to each indicator feature based on experience. This reliance on subjective input can introduce bias, potentially affecting the objectivity of the final results.

Based on the frequency ratio model, Youssef et al. [6] applied an ensemble method of Frequency Ratio (FR) and Logistic Regression (LR) in 2015. This combined approach 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 essential to acknowledge that the frequency ratio model has limitations due to its reliance on historical data, which may not effectively account for the interactions 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) to address disagreements among decision-makers by providing an alternative to traditional weighting methods used in 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 MLP. The approach involves conducting Spearman correlation analysis to examine the relationships between variables. Subsequently, random forest feature importance analysis is used to assess the significance of various feature indicators. Finally, an MLP model is trained, and L2 is regularized to predict the likelihood 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, as detailed in Reference [9]. The dataset includes various indicators such as flood occurrence probability, infrastructure deterioration, terrain drainage, and monsoon intensity. Since the original observation data contains over 700000 flood-related information, we began with a preprocessing phase.

To ensure the integrity and consistency of flood indicator data while removing potential noise and redundancy, 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.

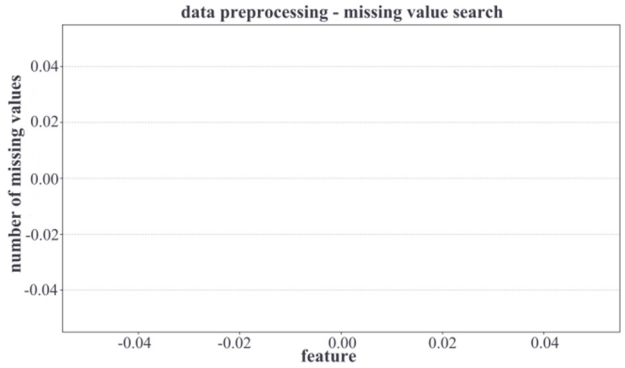

Figure 1. Data preprocessing - missing value search.

2.2. Spearman correlation analysis

We used the Spearman correlation coefficient for correlation research, which measures the non-parametric correlation between variables based on their ranks. This method is particularly effective for identifying relationships 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 suggests the rank difference between paired variables. The formula for calculating the Spearman correlation coefficient is presented in Formula (1):

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

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.

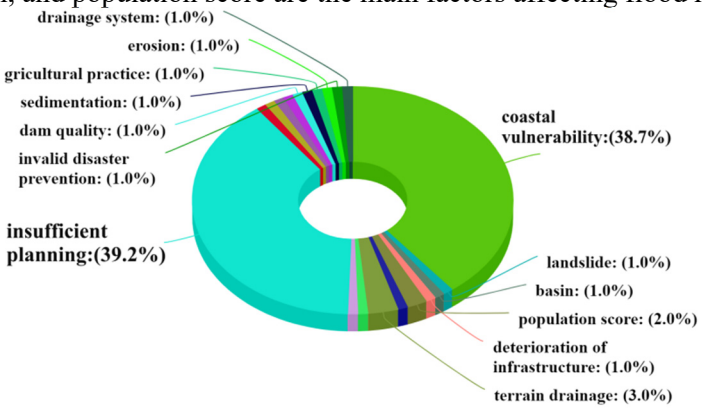

Figure 2. Analysis results of the importance of random forest features.

2.4. Sensitivity analysis

Sensitivity analysis is a vital research technique to examine how changes in model variables or environmental conditions affect the model's state or output. The analysis is achieved by adjusting specific parameters within the model and observing the resultant impact on its outputs, thereby revealing the influence of these parameters on the model's performance. The findings are detailed in Figure 3, Table 2, and Table 3.

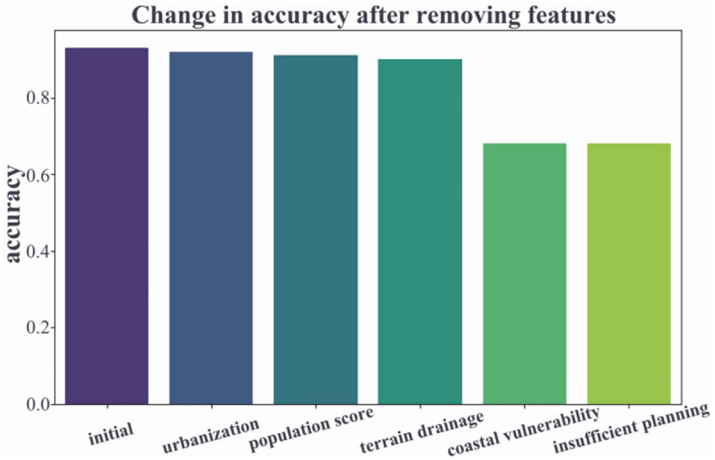

Figure 3. Sensitivity analysis result chart.

Table 2. Sensitivity analysis of the 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

This analysis shows inadequate planning and coastal vulnerability are the most significant factors affecting flood risk prediction. Eliminating these factors would substantially impair the model's performance. Terrain drainage, urbanization, and population scores are also crucial features, and their removal would also result in a notable decline in the model's effectiveness.

3. Establishment of multi-layer perceptron modelreliminaries

The MLP 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 an MLP consists of three layers: the input layer, hidden layers, and the 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 contain 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.

The MLP model iteratively updates its weights to enhance the accuracy of its predictions, adjusting based on the learning algorithm's feedback. The principle of the MLP model is shown in Figure 4.

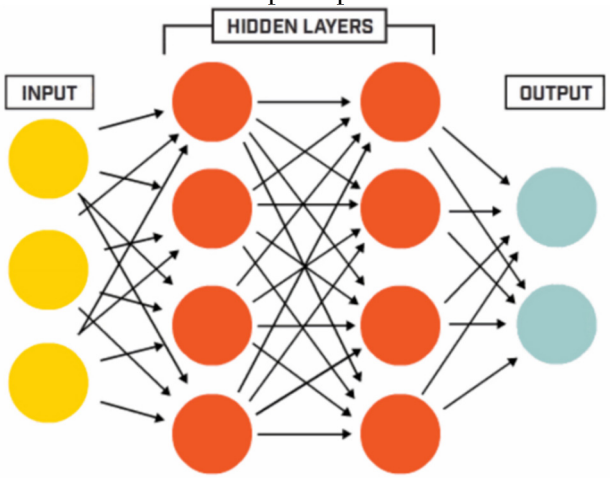

Figure 4. Schematic diagram of multi-layer perceptron model principle.

This study employed a dataset of 800,000 samples split into training and validation sets in a 7:3 ratio. Given the substantial size and diversity of the dataset, data augmentation techniques were deemed unnecessary. Then, we used the MLPRegressor function from sklearn for model training, applying the Adam optimizer to speed up the training process and mitigate 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 and adaptive learning rate algorithms, resulting in expedited convergence and enhanced generalization ability by calculating distinct adaptive learning rates for each parameter. The update rules for the Adam optimizer are as follows:

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

Among them, g_t is the gradient of the parameter, β_1 and β_2 are the attenuation coefficients of two exponential weighted averages, $\hat{m}_t$ and $\hat{v}_t$ are the moving averages of the gradient after deviation correction, θ_{t+1} is the updated parameter, η is the learning rate, and ϵ is a tiny constant used to avoid dividing by zero.

The training loss is shown in Figure 5.

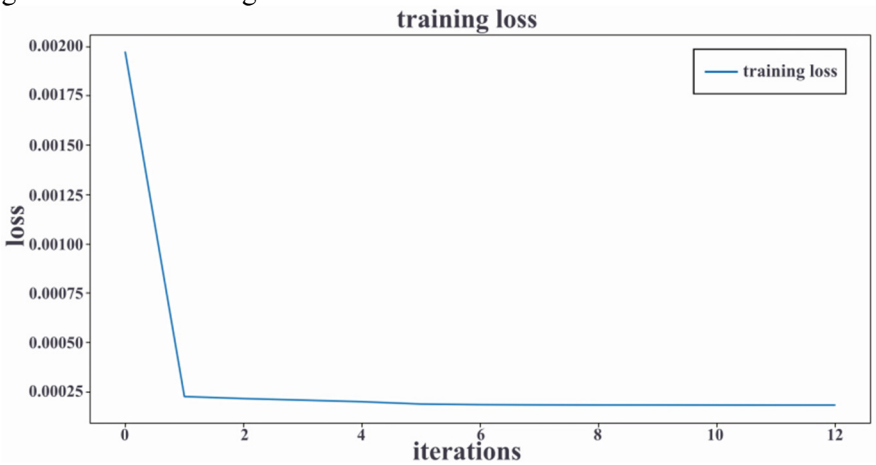

Figure 5. Training loss.

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 measures how well the regression model fits the sample data. A R^2value closer to 1 indicates a better fit of the model to the data. The formula is given as Formula (3):

$$R^2 = 1 - \frac{\sum_{k=1}^{n}(y_k - \hat{y}_k)^2}{\sum_{k=1}^{n}(y_k - \bar{y}_k)^2} \tag{3}$$

where y_k represents the true target value for the i-th observation and $\hat{y}_k$ is the predicted value of the model for the same observation, $\bar{y}$ denotes the average target value of all observations, 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 values, while the red line represents the predicted values.

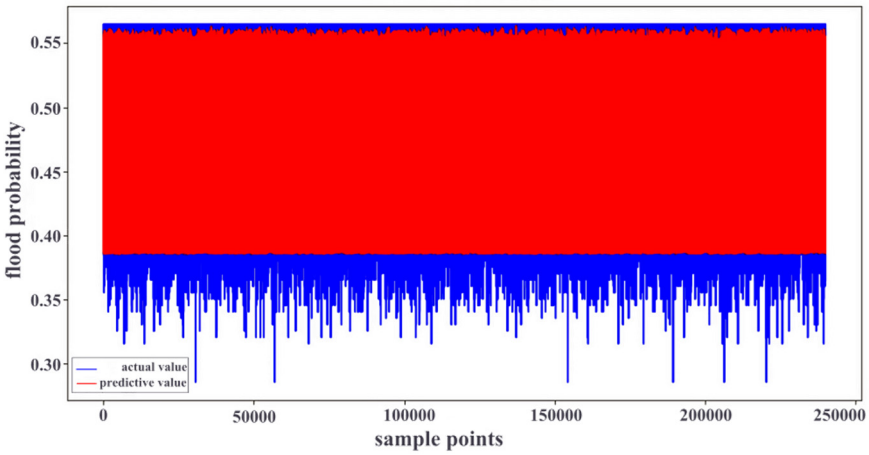

Figure 6. Line graph of flood probability prediction.

Table 4. Model prediction results.

Mean square error of the 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 applying the L2 regularization to alleviate the problem of overfitting and improve the model's generalization ability. The updated code introduces an alpha parameter of 0.001 through continuous optimization and iteration, improving the model's prediction accuracy and practical utility. This regularization technique introduces an additional regularization term to the loss function 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 presented in Formula (4):

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{4}$$

L2 regularization introduces a regularization term into the original model's loss function to penalize the parameters' magnitude. The form of the L2 regularization loss function is as Formula (5):

$$\mathrm{Loss} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda\sum_{j=1}^{p}\omega_j^2 \tag{5}$$

where λ 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.

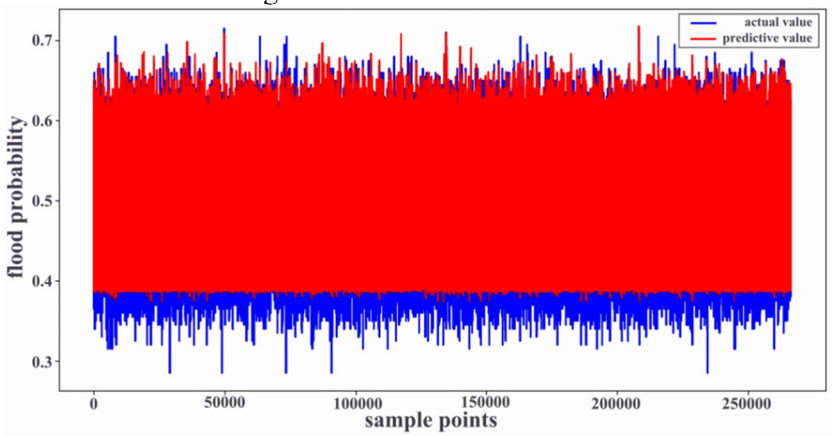

Figure 7. Comparison between actual and predicted flood probability values.

Table 5. Optimized model prediction results.

Mean square error of the 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 illustrated in Figure 7 and Table 5, showcasing a comparison between actual disaster data and predicted outcomes. The mean square errors of the training and testing sets are minimal, at 0.000369 and 0.000368, respectively. The R^2 values for the test and training sets are 0.8533 and 0.8527, respectively. These results indicate that although the number of indicators in the model has decreased, the R^2 has significantly improved, demonstrating a strengthened model effectiveness. The model's prediction results are more accurate.

This study introduces a novel mathematical modeling approach based on a multi-layer perceptron. We have identified key features that significantly impact flood probability by integrating correlation analysis with the traditional random forest algorithm and conducting sensitivity analysis. This approach also eliminates extraneous factors, reduces the subjectivity in parameter weighting, and enhances model interpretability. The experimental results confirm that the proposed model shows robust predictive performance and can be a valuable tool for forecasting flood risks and addressing other natural disasters.

4. Conclusion

This study introduces a novel mathematical modeling approach based on a multi-layer perceptron. By combining correlation analysis with the traditional random forest algorithm and conducting sensitivity analysis, we have identified features that significantly impact flood probability. This approach eliminates extraneous factors, reduces the subjectivity in parameter weighting, and improves model interpretability. The experimental results demonstrate that the proposed model exhibits robust predictive performance and can serve as a valuable reference for forecasting flood risks 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] Xu Yuanhao, Hu Caihong, Wu Qiang, 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 modeling based on 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 the 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.