International Journal of Computer Science and Information Technology
ISSN: 3005-9690 (Print), ISSN: 3005-7140 (Online) | Volume 2, Number 2, Year 2024
DOI: https://doi.org/10.62051/ijcsit.v2n2.31
Journal homepage: https://wepub.org/index.php/IJCSIT/index

# Evaluation of Earthquake Hazard Risk Level Based on Random Forest

Qiqi Song[1,*], Xinyue Wu[1], Yanan Lv[2]

[1]Institute of Disaster Prevention, Beijing, China
[2]Zhongke Pengyang(Hangzhou) Technology Co., Ltd, Hangzhou, China
*Corresponding Author: Qiqi Song

## ABSTRACT

Earthquakes, highly destructive natural disasters, pose significant challenges to China, a country situated at the intersection of two major seismic belts. The risk of earthquake disasters has been escalating due to population density and rapid economic development, threatening the safety of people's lives and the sustainability of the national economy. Therefore, rapid and accurate earthquake risk assessment is crucial for disaster prevention and mitigation. With the advancement of artificial intelligence technology, machine learning algorithms have become vital tools in earthquake science research. This study aims to develop a nationwide earthquake risk assessment model, utilizing earthquake data from 2005 to 2020. Through preprocessing techniques such as normalization, discretization, and type conversion, combined with Spearman correlation analysis to select key indicators. After training and testing with BP neural network, SVM, and Random Forest models, the Random Forest model demonstrated the best performance in key metrics such as accuracy, precision, recall, and F1 Score, proving its superior classification capability. After a comprehensive evaluation, the Random Forest model was chosen as the preferred model for earthquake risk assessment, ensuring the accuracy and reliability of the assessment.

## KEYWORDS

Earthquake Disaster; Risk Level Assessment; Machine Learning; RF

## 1. INTRODUCTION

Earthquakes are devastating natural disasters that can cause widespread destruction, such as the collapse of buildings and the fracturing of roads, leading to significant loss of property and life. Moreover, due to the sudden nature of earthquakes, their duration is typically only a few seconds to several tens of seconds, often catching people off guard and resulting in severe damage. China is among the countries with a high level of seismic activity and a high rate of destruction worldwide.[1] Over the past century, the country has experienced numerous catastrophic earthquakes, causing serious casualties and damage. Therefore, accurately assessing and forecasting earthquake disaster losses is crucial when an earthquake occurs. These assessments and analyses not only provide a basis for government disaster risk management and emergency management but also serve as key numerical support for government rescue operations. Through scientific methods and technical means, we can better respond to earthquake disasters, reduce losses, and ensure the safety of people's lives and property.

Research on earthquake disaster loss assessment originated in the United States. In the 1960s, Freeman [2] was among the first to delve into the field of regional loss assessment, thoroughly exploring the interrelationship between earthquake damage and earthquake insurance. In the 1970s,

American expert Algermissen [3] pioneered research on the extent of building damage under varying earthquake intensities. Subsequently, the U.S. government conducted comprehensive field surveys in four domestic cities, from which the NOAA/USGS earthquake disaster assessment methodology emerged. By the end of the 1980s, Algermissen introduced an innovative technique for damage assessment, which meticulously categorized buildings based on their materials and structural characteristics. This technique, known as the vulnerability classification inventory method, quickly gained global recognition and application. Towards the close of the 20th century, agencies such as the U.S. Federal Emergency Management Agency collaboratively developed the HAZUS earthquake technology. This technology primarily serves state, regional, and community governments, aiming to scientifically assess potential losses in earthquake scenarios, thereby providing a basis for disaster preparedness and emergency response planning. Additionally, HAZUS [4] earthquake technology is committed to facilitating planning efforts to mitigate potential future earthquake losses.

Earthquake-induced secondary disasters exhibit an extremely high level of complexity, posing numerous challenges for the precise assessment and prediction of direct and indirect losses. In contrast, systematically classifying, evaluating, and predicting the levels of earthquake disaster losses is a more feasible approach. For solving such multi-class problems, we have a variety of mature algorithms at our disposal, including decision trees, Naive Bayes algorithms, Support Vector Machines (SVM), and Back Propagation (BP) neural networks. Notably, the Random Forest classification model, as a rising star in the field of machine learning, has garnered widespread attention and been applied to a diverse array of issues in recent years. In this study, we employ the Random Forest classification model to construct a predictive model for earthquake disaster loss levels and, through comparative analysis with traditional BP neural network and SVM models, have uncovered some valuable insights. The results demonstrate that the Random Forest classification model exhibits significant advantages in terms of predictive accuracy and multiple relevant evaluation metrics, surpassing the traditional BP neural network and SVM models. This finding not only provides a new perspective and methodology for earthquake disaster loss assessment but also offers important references for researchers and practitioners in related fields.

# 2. DATA COLLECTION AND PROCESSING

## 2.1. Data Acquisition

The earthquake disaster data utilized in this paper spans a period of 15 years, from 2005 to 2020, encompassing a total of 141 case studies. These data meticulously document various aspects of earthquake disasters, including the seismic magnitude, focal depth, and epicentral intensity within the causative factors. Additionally, they capture key information from the disaster-prone environment, such as the seismic fortification intensity and the design basic earthquake acceleration. Data pertaining to the disaster-bearing capacity, including the affected area's per capita GDP, population, and the area impacted by the disaster, have also been compiled to provide a comprehensive reflection of the impacts and consequences of earthquake disasters. The seismic magnitude and focal depth data are sourced from the China Earthquake Networks Center, while the epicentral intensity is obtained from the China Earthquake Administration. The seismic fortification intensity and design basic earthquake acceleration are derived from the China Academy of Building Research's Seismic Design Code (2016 edition), and the per capita GDP and population data of the disaster-affected areas are sourced from the China Statistical Yearbook.

## 2.2. Data Pre-processing

Data transformation is a critical step in the data preprocessing process, involving the conversion or mapping of the original dataset into a new form to facilitate better analysis and processing. The purpose of data transformation is to enhance the quality and usability of data, making it more suitable

for subsequent analytical methods such as machine learning or statistical modeling. Data transformation encompasses a variety of methods, including normalization, standardization, discretization, binarization, type conversion, logarithmic transformation, Box-Cox transformation, principal component analysis, and handling of discrete values. The choice of data transformation method depends on the characteristics of the data and the requirements of the subsequent analysis. In practice, to find the most suitable transformation method for the current dataset and analytical objectives, it may be necessary to a variety of different transformation techniques. In this paper, three main data transformation methods were applied during the data processing: normalization, discretization, and type conversion. The application of these methods provided an essential data foundation for subsequent data analysis and modeling.

Normalization, widely employed as a technical means in data transformation, primarily aims to standardize data onto a uniform scale, typically within the range [0,1]. By implementing normalization, not only can the convergence rate of algorithms be effectively accelerated, enhancing computational efficiency, but it also eliminates the differences in units among various features, thereby making the features comparable with one another. These advantages position normalization as an important step in data processing. Normalization methods include several forms such as min-max normalization, Z-score normalization, decimal scaling normalization, and regularization normalization. Given the variance in the distribution of parameter values across each training sample, which leads to significant changes in sample spacing, this can introduce considerable bias when models fit the data. Moreover, even if a model performs well on the training set, its generalization capability may still be limited. Therefore, min-max normalization was selected as the processing method, with the formula as follows:

$$y = (y_{\max} - y_{\min}) \cdot \frac{x - x_{\min}}{x_{\max} - x_{\min}} + y_{\min} \tag{1}$$

Under the given circumstances, $y_{\max}$ and $y_{\min}$ are preset fixed values, set to 1 and 0 respectively. The value $x_{\max}$ represents the maximum value in the sequence of $x$, while the value $x_{\min}$ represents the maximum value of that attribute parameter among all the training samples.

The direct economic losses from earthquakes discussed in this paper constitute a vast set of continuous numerical values. Given the practical circumstances post-earthquake, rapid assessment of the magnitude of losses is often more valuable for rescue operations than precise numerical values. In the data processing phase, when the precision of specific values is less critical and the focus is more on the distribution range of the data, a discretization strategy is commonly employed. This involves converting continuous data into discrete interval data to facilitate more effective subsequent analysis and application. For the measurement of natural disaster losses, there are both absolute and relative methods. As stated in the paper 'Research on the Assessment Indicator System for Natural Disaster Losses,' natural disasters can be categorized into five degrees of severity: catastrophic, major, moderate, minor, and minimal. The relative measurement approach, on the other hand, is gauged by the damage ratio, which specifically refers to the proportion of the loss in various types of property in the affected area to their pre-disaster values. The classification of disaster losses is presented in Table 1 below.

**Table 1.** Disaster loss classification

| Disaster Level | Property Damage | Disaster Loss Rate Index |
|---|---|---|
| **mega-disaster** | $>10^6$ | >0.5 |
| **Large-scale disaster** | $10^4$~$10^6$ | 0.4~0.5 |
| **Moderate disaster** | $10^3$~$10^4$ | 0.3~0.4 |
| **Minor disaster** | $10^2$~$10^3$ | 0.2~0.3 |
| **Minimum disaster** | $10^2$ | <0.2 |

Data types can generally be divided into numerical and non-numerical categories. For non-numerical data, a categorization conversion operation is necessary to meet the subsequent processing requirements of machine learning algorithms. Type conversion specifically refers to the process of transforming textual data into numerical data. In this paper, a relatively straightforward conversion method is employed, mapping the categories of catastrophic, major, moderate, minor, and minimal disasters to the numerical values of 1, 2, 3, 4, and 5, respectively.

## 2.3. Correlation Analysis

Correlation analysis [5], a vital method in statistics, is primarily used to quantify the degree and direction of association between two or more variables. This method plays a pivotal role in exploratory data analysis, aiding researchers in uncovering potential correlations and their strength among variables. The outcome of correlation analysis, known as the correlation coefficient, is a value that ranges between -1 and +1, describing the linear relationship between variables. A correlation coefficient of +1 indicates a perfect positive linear correlation between two variables, while a coefficient of -1 signifies a perfect negative linear correlation; a coefficient of 0 implies no linear association between the variables in question. In the realm of correlation analysis, the Pearson product-moment correlation coefficient and the Spearman rank correlation coefficient are the two most commonly utilized methods. This study employs the Pearson product-moment correlation coefficient as its analytical tool, aiming to delve deeper into the extent of the association between variables and to provide a more accurate and reliable foundation for subsequent statistical analyses.

The Pearson product-moment correlation coefficient, commonly referred to as the Pearson correlation coefficient or Pearson's r, is a statistical measure used to assess the degree of linear association between two variables. Named after the statistician Karl Pearson, this coefficient ranges from -1 to 1, inclusive. An r value of 1 signifies a perfect positive linear relationship, where an increase in one variable results in a proportional increase in the other variable. Conversely, an r value of -1 indicates a perfect negative linear relationship, where an increase in one variable leads to a proportional decrease in the other. An r value of 0 suggests that there is no linear relationship between the two variables. The formula for calculating the Pearson correlation coefficient is as follows:

$$r = \frac{\sum_{i=1}^{n}(X-\bar{X})(Y-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X-\bar{X})^2(Y-\bar{Y})^2}} \tag{2}$$

In the formula, $X$ and $Y$ are the sample values of two variables, while $\bar{X}$ and $\bar{Y}$ are their respective sample means. The correlation between the variables strengthens as the value of $r$ increases. The specific degree of correlation is shown in Table 2.

**Table 2.** Pearson correlation coefficient r specific meaning

| Pearson's correlation coefficient | degree of correlation |
|---|---|
| **[0.6,1]** | Strong positive correlation |
| **[0.4,0.6)** | Medium positive correlation |
| **[0.2,0.4)** | Weak positive correlation |
| **[-0.2,0.2)** | Uncorrelated |
| **(-0.4,-0.2]** | Weak negative correlation |
| **(-0.6,0.4]** | Medium negative correlation |

After a detailed data analysis, eight key indicators were selected as the training samples for this study. These indicators encompass the seismic magnitude, focal depth, and epicentral intensity from the causative factors of the disaster; the seismic fortification intensity and design basic earthquake acceleration from the disaster-prone environment; and the per capita GDP of the disaster area, population, and the area affected by the disaster from the disaster-bearing capacity. Rigorous correlation analysis was conducted on these eight indicators, yielding corresponding results as illustrated in Figure 1.

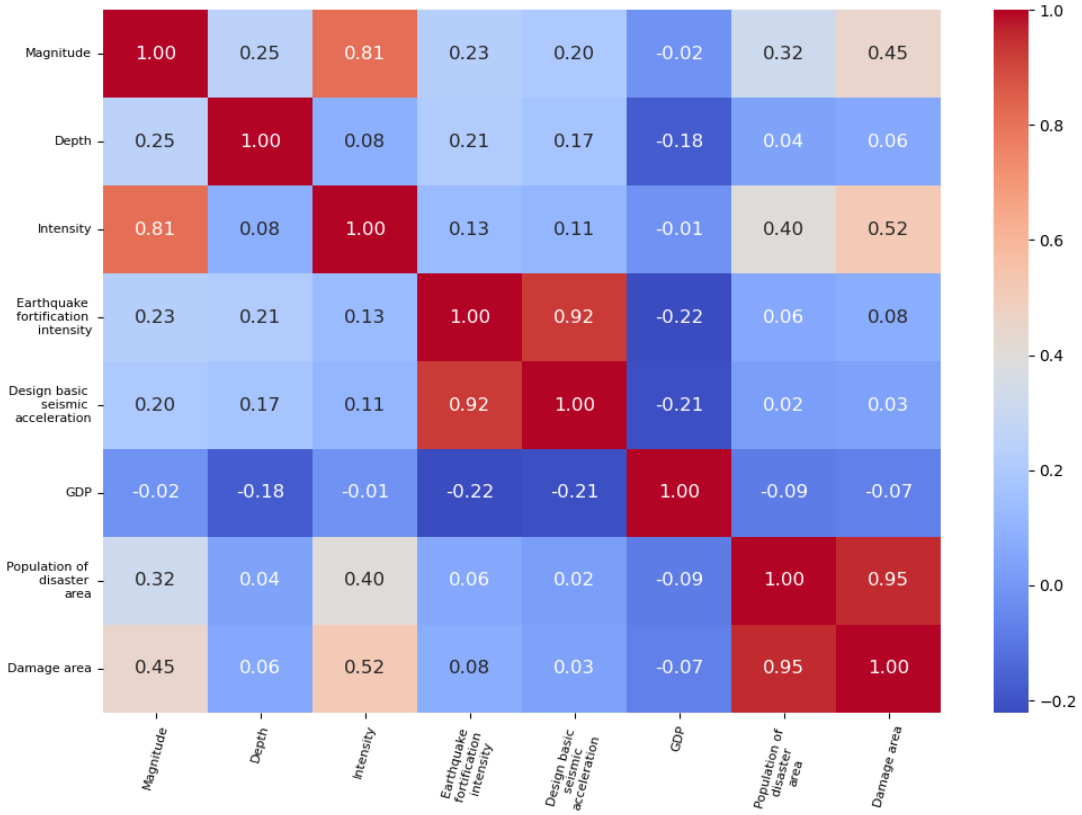


**Figure 1.** Correlation analysis heat map

After an in-depth analysis of Table 2 and Figure 1, the following conclusions were drawn: The correlation between focal depth and per capita GDP of the disaster area with other indicators is relatively weak, thus both indicators can be retained. There is a strong positive correlation between seismic magnitude and epicentral intensity, and a moderate correlation with the area affected by the disaster. Further observation reveals a significant positive correlation between the area affected and the population of the disaster area. Based on the above analysis, the indicators of seismic magnitude and population of the disaster area were chosen for retention to more accurately assess the impact and situation of the earthquake. Additionally, a comparison shows a strong positive correlation between seismic fortification intensity and design basic earthquake acceleration, with weaker correlations to other indicators. Therefore, in subsequent analyses, one of these indicators can be selected for retention to enhance the accuracy and efficiency of the assessment. After the aforementioned analysis, the final decision was to retain the following five key indicators: seismic magnitude, focal depth, design basic acceleration, per capita GDP of the disaster area, and population of the disaster area. The indicators of epicentral intensity, seismic fortification intensity, and the area affected by the disaster were excluded. This decision aims to ensure the accuracy and effectiveness of the assessment work, in order to provide robust support for the recovery and reconstruction efforts in the disaster area.

## 3. STOCHASTIC FOREST THEORY METHOD

Random Forest (RF), in full, is an ensemble learning algorithm that is widely used for both classification and regression tasks. The algorithm was proposed by Breiman [6] in 2001, with its core concept combining Breiman's bootstrap sampling strategy introduced in 1996, as well as the feature random selection methods independently proposed by Ho [7] in 1995 and 1998, and Amit and Geman [8] in 1997. The integration of these elements aims to construct a collection of decision trees with controllable diversity, thereby enhancing the predictive accuracy and stability of the model. Utilizing the bootstrap technique, alternative samples derived from the training data are used to build each decision tree within the ensemble. Statistically, approximately 64% of the instances in the samples are expected to appear at least once. These instances from the samples are termed as "in-bag" instances, while the remaining instances (about 36%) are referred to as "out-of-bag" instances. Each tree in the ensemble serves as a basic classifier to determine the class labels of unmarked instances. This process is implemented through a majority voting mechanism, where each classifier casts a vote for its predicted class label, and the class label with the most votes is then used to categorize the instance.

Breiman optimized the application of Classification and Regression Trees (CART) by introducing additional randomness in the construction of decision trees. In this technique, the subset of features selected at each internal node is evaluated using the Gini index heuristic algorithm. The feature with the highest Gini coefficient is chosen as the splitting feature for that node. The concept of the Gini index was initially proposed by the Italian statistician Corrado Gini in 1912 and was later further popularized and applied by Breiman, Friedman, Olshen, and Stone [9]. The Gini index, as a function to measure the impurity of data, reflects the degree of uncertainty when an event occurs. In classification tasks, this event is closely related to the determination of category labels. The Gini index is typically calculated as follows:

$$Gini(t) = 1 - \sum_{i=1}^{N} P(C_i \mid t)^2 \tag{3}$$

Where $t$ is the condition, $N$ is the number of classes in the dataset, and $C_i$ is the i class label in the dataset.

In Breiman's seminal paper on Random Forests (RF), it is explicitly stated that the error rate of RF is influenced by two core factors: the correlation between trees and the strength of individual trees. Specifically, increasing the correlation between any two trees in the RF can lead to an increase in the overall error rate. Trees with lower error rates, referred to as strong classifiers, play a significant role within the forest. On the other hand, enhancing the strength of individual trees helps to reduce the overall error rate of the RF. These findings align with the research results of Bernard, Heutte, and Adam [10], who demonstrated through empirical research that by jointly maximizing the strength of individual trees and minimizing the correlation between trees, the error rate can be effectively reduced, leading to statistically optimized performance.

Compared to AdaBoost, the primary advantages of Random Forests (RF) lie in their robustness against noise and their ability to mitigate overfitting phenomena [11]. Overfitting typically occurs when the model construction process excessively fits the training data, causing the model to fit the data beyond a reasonable range. Such overfitted models often exhibit poor predictive performance because they struggle to generalize effectively to new situations. The ability to generalize, that is, the model's capacity to make predictions on data not involved in training, is a crucial metric for evaluating model performance. Hawkins [12] points out that overfitting not only increases the complexity of the model without any performance improvement but also, more seriously, may lead to a significant degradation in model performance. Overfit classifiers may show a lower error rate on training

instances (in-bag instances) but may have a higher error rate on unknown instances (out-of-bag instances).

According to Breiman's understanding, Random Forests (RF) possess significant advantages. In terms of accuracy, RF is capable of matching, and even surpassing, Adaboost in certain scenarios. Compared to bagging or boosting methods, RF offers faster computation speeds, enhancing efficiency. Moreover, RF provides internal estimates regarding error, strength, correlation, and variable importance, offering users a deeper understanding of the data. Lastly, the simplicity and ease of implementing RF, along with its amenability to parallel processing, further augment its flexibility and practicality in real-world applications. These strengths position the random forest as a powerful machine learning tool, widely applied across various data analysis tasks.

The Random Forest (RF) algorithm, with its exceptional performance, has been widely applied across various fields, including ecology, medicine, astronomy, forensics, traffic and transportation regulations, agriculture, bioinformatics, and computational biology. These practical applications fully demonstrate the significant advantages of random forests in handling complex datasets, feature selection, classification accuracy, and model interpretability. Its inherent diversity and robustness make it a powerful tool for addressing a range of practical issues, providing robust support for decision-making in various domains.

# 4. MODEL CONSTRUCTION AND RESULT ANALYSIS

## 4.1. Model Construction

After preprocessing and meticulous correlation analysis, this paper ultimately identified five key indicators—earthquake magnitude, focal depth, design basic acceleration, per capita GDP of the disaster area, and population of the disaster area—as the foundation for constructing the training sample set. Based on disaster severity theory, earthquake disasters are categorized into four distinct levels, which serve as the true labels for predictive analysis. Subsequently, this paper utilizes the Random Forest classification model from machine learning and compares it with the BP neural network and SVM classification models for an in-depth prediction analysis of earthquake disaster risk levels. A comparison of the results from the three models is presented in Table 3.

**Table 3.** The results of the three models were compared

| | BP | SVM | RF |
|---|---|---|---|
| **Accuracy** | 0.59 | 0.62 | 0.77 |
| **Precision** | 0.25 | 0.15 | 0.37 |
| **Recall** | 0.27 | 0.25 | 0.39 |
| **F1 Score** | 0.26 | 0.19 | 0.38 |

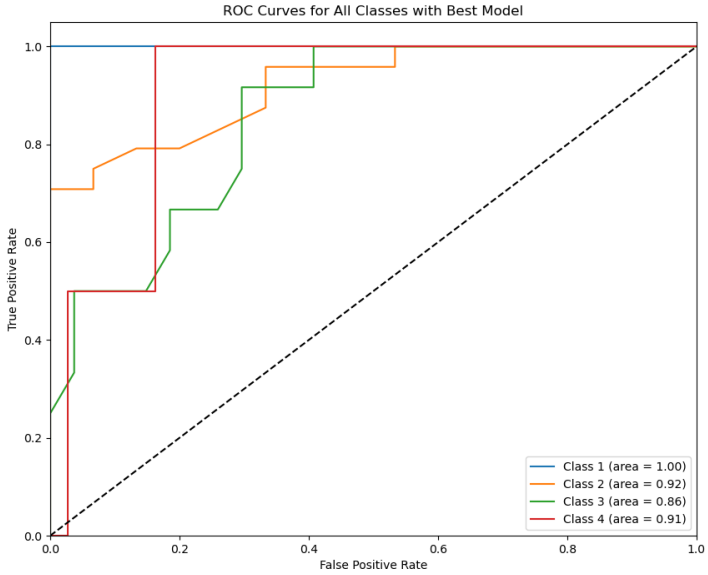


**Figure 2.** ROC curve under random forest model

## 4.2. Result Analysis

Table 3 presents the performance metrics of three distinct machine learning models in assessing the risk levels of earthquake disasters: the Back Propagation (BP) neural network, Support Vector Machine (SVM), and Random Forest. The performance of each model is measured by accuracy, precision, recall, and F1 score.

Accuracy is a fundamental performance metric that represents the proportion of samples correctly predicted by the model relative to the total number of samples. In the assessment of earthquake disaster risk levels, the Random Forest achieved the highest accuracy at 0.77, meaning it correctly classified 77% of the samples. The BP neural network had an accuracy of 0.59, while the SVM scored 0.62, indicating that the Random Forest performed best in overall classification accuracy. Precision focuses on the proportion of samples that were actually positive among those the model predicted as positive. Here, the BP neural network had the highest precision at 0.25, indicating that 25% of its positive predictions were indeed positive. The Random Forest had a precision of 0.37, while the SVM had the lowest precision at 0.15, suggesting that the SVM generated a higher number of false positives. Recall measures the model's ability to correctly identify all positive samples. In the assessment of earthquake disaster risk levels, the Random Forest had the highest recall at 0.39, indicating it could identify 39% of the actual positive samples. The BP neural network's recall was 0.27, and the SVM's was 0.25, showing that the Random Forest was more effective in detecting positive samples. The F1 score is the harmonic mean of precision and recall, attempting to balance the two. In the assessment of earthquake disaster risk levels, the Random Forest achieved the highest F1 score at 0.38, indicating a better balance between precision and recall. The BP neural network's F1 score was 0.26, while the SVM's, the lowest, was 0.19, which may imply that the SVM performed poorly in balancing false positives and false negatives.

Figure 2 displays the ROC curve for the Random Forest classifier, which demonstrates exceptional performance. The curve shows a distinct upward trend, indicating that as the threshold increases, the accuracy of predicting positive samples gradually improves. During the ascent of the curve, its curvature also increases, signifying that the classifier's discrimination ability is becoming stronger. Eventually, the curve levels off, suggesting that the classifier has reached its maximum performance capacity. In the graph, the line y=x represents a curve for random guessing; it is observable that the

ROC curve is almost entirely above the y=x line. The further to the right and upper the curve's position, the better the performance of the classifier. The AUC value of the classifier was also calculated, reflecting the average performance across all possible thresholds. A higher AUC value indicates better classifier performance.

## 5. SUMMARY

In the aftermath of an earthquake disaster, governments rely on scientific risk assessments to formulate rescue decisions and emergency plans. The accuracy and timeliness of these assessments are crucial for ensuring the effectiveness of relief efforts and the rational allocation of resources. An inadequate assessment may lead to insufficient rescue preparations and impaired relief outcomes, while excessive preparation could result in resource wastage. Therefore, scientific risk assessment is essential for the efficiency and safety of seismic rescue operations in affected areas. In recent years, machine learning-based models for earthquake risk assessment have offered new perspectives and methodologies for disaster response and management. This study employs machine learning classification algorithms, utilizing three distinct data preprocessing methods for feature selection on nationwide earthquake disaster data. After a comprehensive analysis of the final evaluation results and key performance indicators of the three classification models under optimal parameter settings, the Random Forest model was observed to exhibit the most outstanding overall performance. Specifically, this model achieved the highest levels in key evaluation metrics such as accuracy, precision, recall, and F1 score, demonstrating its robust classification capability. Consequently, considering all evaluation metrics and model performances, the Random Forest model was ultimately chosen as the classification model for earthquake disaster risk assessment to ensure the accuracy and reliability of the assessment outcomes.

## REFERENCES


[1] Jena, Ratiranjan, et al. "Integrated model for earthquake risk assessment using neural network and analytic hierarchy process: Aceh province, Indonesia." Geoscience Frontiers 11.2 (2020): 613-634.

[2] Freeman JR. 1932. Earthquake damage and earthquake insurance. New York: Mc Graw-Hill.

[3] Algermissen S T, Steinbrugge K V,1984: Seismic Hazard and risk assessment: Some case studies[J], The Geneva Papers on Risk and Insurance, Vol.9, PP8-26.

[4] Kircher, Charles A., Robert V. Whitman, and William T. Holmes. "HAZUS earthquake loss estimation methods." Natural Hazards Review 7.2 (2006): 45-59.

[5] Gogtay, Nithya J., and Urmila M. Thatte. "Principles of correlation analysis." Journal of the Association of Physicians of India 65.3 (2017): 78-81.

[6] Breiman, Leo. "Random forests." Machine learning 45 (2001): 5-32.

[7] Ho, Tin Kam. "The random subspace method for constructing decision forests." IEEE transactions on pattern analysis and machine intelligence 20.8 (1998): 832-844.

[8] Amit, Yali, and Donald Geman. "Shape quantization and recognition with randomized trees." Neural computation 9.7 (1997): 1545-1588.

[9] Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and regression trees (1st ed.). New York/Boca Raton, FL: Chapman and Hall/CRC.

[10] Bernard, Simon, Laurent Heutte, and Sébastien Adam. "A study of strength and correlation in random forests." Advanced Intelligent Computing Theories and Applications: 6th International Conference on Intelligent Computing, ICIC 2010, Changsha, China, August 18-21, 2010. Proceedings 6. Springer Berlin Heidelberg, 2010.

[11] Boinee, Praveen, Alessandro De Angelis, and Gian Luca Foresti. "Meta random forests." International Journal of Computational Intelligence 2.3 (2005): 138-147.

[12] Hawkins, Douglas M. "The problem of overfitting." Journal of chemical information and computer sciences 44.1 (2004): 1-12.