International Journal of Computer Sciences and Mathematics Engineering
Journal homepage: www.ijecom.org

Early Earthquake Prediction Using a Hybrid Feature Selection and Ensemble Learning Approach

Abdul Khaliq[1] , Muhammad Muttaqin[2]
[1,2]Computer Science Department, Universitas Pembangunan Panca Budi

Article Info

Article history:

Accepted Sep 27, 2024
Revised Sep 29, 2024
Accepted Oct 01, 2024

Keywords:

Earthquakes,
Voting Classifier,
CatBoost,
XGBoost,
Machine Learning

ABSTRACT

Earthquakes are one of the most destructive and unpredictable natural disasters, causing huge losses to infrastructure, economy, and lives. Therefore, an accurate earthquake prediction system is crucial in mitigating the impact of this disaster. Rapidly developing data processing and artificial intelligence technologies have opened up new opportunities to improve earthquake prediction capabilities, with machine learning-based approaches that can handle large and complex data. However, a major challenge in developing earthquake prediction models is the selection of relevant features from large and noisy datasets. Irrelevant or redundant features can reduce model accuracy and increase computational complexity. This study proposes a hybrid approach in feature selection and ensemble learning for early earthquake prediction. This approach combines several feature selection techniques to identify the most relevant data attributes and integrates an ensemble learning model to improve prediction accuracy. By adopting this strategy, it is expected that the model can produce faster and more reliable predictions, while reducing the risk of errors. This approach not only provides a technical solution but also makes a significant contribution to disaster risk mitigation efforts. With proper implementation, an early earthquake prediction system based on this hybrid method has the potential to save many lives and reduce losses due to earthquakes.

Corresponding Author:

Abdul Khaliq,
computer science,
Universitas Pembangunan Panca Budi,
Email: abdulkhaliq@pancabudi.ac.id

1. INTRODUCTION

Earthquakes are one of the most destructive and unpredictable natural disasters, causing significant losses to infrastructure, economy, and loss of life. The importance of an accurate earthquake prediction system is increasing as global efforts to mitigate the devastating impacts of these disasters increase. Early prediction of earthquakes can provide sufficient time for communities and authorities to take preventive measures, such as evacuation or protection of vital infrastructure.

E-ISSN 2962-4274

In the last decade, advances in data processing technology and artificial intelligence have opened up new opportunities to improve earthquake prediction capabilities. Traditional methods such as seismic analysis are now being augmented by machine learning-based approaches that are capable of handling large amounts of complex data. These approaches leverage data such as seismic activity, geology, and other physical parameters to detect patterns that precede earthquakes.

However, a major challenge in developing earthquake prediction models is selecting the most relevant features from large, often heterogeneous and noisy datasets. Irrelevant or redundant features can reduce the accuracy of the prediction model and increase the computational complexity. Therefore, the combination of efficient feature selection techniques with sophisticated machine learning algorithms is essential to improve prediction performance.

This study introduces a hybrid approach in feature selection and ensemble learning for early earthquake prediction. The hybrid approach combines several feature selection techniques to identify the most relevant data attributes and integrates ensemble learning models to improve prediction accuracy. By adopting this strategy, the model is expected to be able to produce faster and more reliable predictions, while reducing the risk of errors.

This approach not only offers a technical solution but also makes a significant contribution to disaster risk mitigation efforts. With proper implementation, an early earthquake prediction system based on this hybrid method has the potential to save many lives and reduce losses caused by earthquakes.

2. RELATED WORK

Earthquake prediction research has grown rapidly in the past few decades by utilizing various approaches, including traditional statistical methods, signal processing, and machine learning. In this section, we review some previous studies relevant to the topic of earthquake prediction, especially those related to the use of feature selection and ensemble learning algorithms.

Traditional Approaches to Earthquake Prediction

Early methods for earthquake prediction were mostly based on statistical analysis of historical seismic activity data. These studies focused on identifying earthquake patterns by utilizing parameters such as magnitude, frequency, and location. However, these approaches often suffer from limitations in handling non-linear and complex data, which are common characteristics of earthquake phenomena.

Feature Selection for Earthquake Prediction

Feature selection is an important step in building an accurate prediction model. Several studies have used techniques such as Principal Component Analysis (PCA), redundant feature removal, and mutual information-based algorithms to filter relevant data. For example, a study by Wang et al. (2018) showed that the use of a hybrid feature selection algorithm can significantly improve the accuracy of earthquake prediction models. This technique combines statistical-based selection with heuristic methods to select the best feature subset from seismic data.

Machine Learning in Earthquake Prediction

Machine learning algorithms have made significant contributions to earthquake prediction. Models such as Support Vector Machines (SVM), Random Forest, and Neural Networks are used to capture non-linear relationships in earthquake data. For example, Li et al. (2020) used the Random Forest model to predict earthquakes with a fairly high degree of accuracy based on seismic sensor data and geophysical parameters. However, one of the main challenges is to avoid overfitting, especially when the data used is limited or imbalanced.

Ensemble Learning Approach

Ensemble learning, such as Random Forest, Gradient Boosting Machines (GBM), and Extreme Gradient Boosting (XGBoost), has been widely used to improve the performance of prediction models. This

E-ISSN 2962-4274

method combines predictions from multiple base models to produce more reliable results. A study by Zhang et al. (2021) showed that XGBoost is able to provide more accurate results than individual models on a large earthquake dataset.

Hybrid Approaches in Earthquake Prediction

Hybrid approaches integrate multiple techniques to improve prediction accuracy. For example, Sharma et al. (2022) proposed an earthquake prediction system using a combination of correlation-based feature selection with ensemble learning algorithms such as AdaBoost. Their results showed that the hybrid approach not only improves accuracy but also reduces computational time.

3. MATERIAL AND METHOD

This section will display classification models in machine learning that used in this study. With some details:

3.1 Data Collection

The dataset used in this study was obtained from [earthquake data sources], including geophysical parameters such as magnitude, epicenter location, depth, and occurrence time. Additional data, such as pre-earthquake seismic activity, soil characteristics, and geological information, were incorporated to enrich the analysis. The dataset covers a specific time period to ensure pattern diversity and reduce data bias.

3.2 Data Preprocessing

To ensure the dataset is suitable for modeling, several preprocessing steps were applied:

1. **Handling Missing Data:** Missing values were imputed using interpolation or mean-based imputation methods to prevent their negative impact on the model.
2. **Normalization:** All features were scaled to the range [0,1] to ensure uniformity during model training.
3. **Dimensionality Reduction:** Principal Component Analysis (PCA) was employed to reduce the complexity of the data without significant information loss.

3.3 Feature Selection

A hybrid feature selection approach was implemented to identify the most relevant data attributes:

1. **Filter-Based Selection:** Statistical methods like chi-square ($\chi 2 \backslash chi^2 \chi 2$) and information gain were applied to evaluate feature relevance.
2. **Wrapper-Based Selection:** Recursive Feature Elimination (RFE) was used to validate features based on the predictive performance of the model.
3. **Hybrid Selection:** The combination of filter- and wrapper-based approaches was used to derive the optimal feature subset, balancing accuracy and computational efficiency.

3.4 Model Development

An ensemble learning-based model was developed using the following process:

E-ISSN 2962-4274

1. **Base Models:** Algorithms such as Decision Tree, Random Forest, and Gradient Boosting were used as the foundation for ensemble modeling.
2. **Ensemble Construction:** Bagging and Boosting techniques were applied to aggregate predictions from base models, enhancing prediction accuracy.
3. **Hyperparameter Tuning:** Parameters were optimized using Grid Search and Bayesian Optimization to achieve optimal model performance.

3.5 Mathematical Formulation

a. Feature Selection

The chi-square score for feature relevance is calculated as:

χ2=∑(Oi−Ei)2Ei\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}χ2=∑Ei(Oi−Ei)2

Where:

- OiO_iOi: Observed frequency
- EiE_iEi: Expected frequency

RFE selects features iteratively by ranking their importance through a predictive model.

b. Ensemble Prediction

Ensemble predictions combine outputs of individual models using techniques such as weighted voting:

Pensemble(x)=∑i=1nwiPi(x)P_{ensemble}(x) = \sum_{i=1}^{n} w_i P_i(x)Pensemble(x)=i=1∑nwiPi(x)

Where:

- Pensemble(x)P_{ensemble}(x)Pensemble(x): Final ensemble prediction
- Pi(x)P_i(x)Pi(x): Prediction from the iii-th base model
- wiw_iwi: Weight assigned to the iii-th base model

c. Model Evaluation

The performance metrics include:

- **Accuracy**:

Accuracy=TP+TNTP+TN+FP+FN\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}Accuracy=TP+TN+FP+FNTP+TN

- **Precision**:

Precision=TPTP+FP\text{Precision} = \frac{TP}{TP + FP}Precision=TP+FPTP

- **Recall**:

E-ISSN 2962-4274

Recall=TPTP+FN\text{Recall} = \frac{TP}{TP + FN}Recall=TP+FNTP

- **F1-Score**:

F1-Score=2·Precision·RecallPrecision+Recall\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}F1-Score=2·Precision+RecallPrecision·Recall

3.6 Evaluation and Validation

The model was evaluated using a k-fold cross-validation technique to ensure robust generalization across data subsets. Additionally, Receiver Operating Characteristic (ROC) and Area Under Curve (AUC) were analyzed to assess the model's discriminative capability.

3.7 Tools and Environment

Experiments were conducted using Python libraries such as Scikit-learn, TensorFlow, and XGBoost. The computational environment included a processor, memory, and GPU for efficient model training and testing.

This methodology aims to deliver a reliable and efficient hybrid model for early earthquake prediction, contributing to disaster risk mitigation strategies.

4. METHDOLOGY

4.1 Overview

The methodology for this research is designed to develop an early earthquake prediction system using a hybrid feature selection and ensemble learning approach. The framework consists of the following stages: data preparation, feature selection, model training, and evaluation. Each step is methodically designed to enhance the accuracy and reliability of earthquake predictions.

4.2 Framework

Step 1: Data Preparation

- **Data Acquisition:** Collect earthquake-related data, including seismic activity, geological properties, and soil characteristics, from reliable sources.
- **Preprocessing:** Address missing values, normalize the dataset, and reduce dimensionality using techniques like Principal Component Analysis (PCA).
- **Data Splitting:** Split the dataset into training and testing sets with a ratio of 80:20 to ensure sufficient data for both model training and validation.

Step 2: Feature Selection

The hybrid feature selection process is employed to enhance the model's efficiency:

1. **Filter-Based Methods:** Statistical measures like chi-square ($\chi 2$\chi^2$\chi 2$) and mutual information are used to rank feature importance based on their correlation with the target variable.

E-ISSN 2962-4274

2. **Wrapper-Based Methods:** Recursive Feature Elimination (RFE) iteratively selects features by evaluating model performance.
3. **Hybrid Selection:** The results from filter- and wrapper-based methods are combined to select the most relevant features for prediction.

Step 3: Model Training and Ensemble Learning

- **Base Models:** Develop individual base models using Decision Tree, Random Forest, and Gradient Boosting techniques.
- **Ensemble Construction:** Combine base models using ensemble methods:
  - **Bagging:** Aggregates predictions from multiple independent models to reduce variance.
  - **Boosting:** Focuses on reducing bias by sequentially training models to correct errors from previous iterations.
- **Hyperparameter Optimization:** Fine-tune parameters for each model using Grid Search and Bayesian Optimization to maximize performance.

Step 4: Model Evaluation

The predictive performance of the ensemble model is evaluated using several metrics:

1. **Accuracy:** Proportion of correct predictions to total predictions.
2. **Precision, Recall, and F1-Score:** Evaluates the balance between false positives and false negatives.
3. **ROC-AUC:** Assesses the model's ability to distinguish between positive and negative classes.
4. **Error Metrics:** Mean Absolute Error (MAE) is calculated to measure the average prediction error.

4.3 Mathematical Representation

Feature Selection

- **Chi-Square Test:**

χ2=∑(Oi−Ei)2Ei\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}χ2=∑Ei(Oi−Ei)2

- **Recursive Feature Elimination (RFE):**

  At each iteration, eliminate the least important features based on the model’s weight or importance ranking.

Ensemble Learning

- **Bagging Prediction:**

Pbagging(x)=1n∑i=1nPi(x)P_{bagging}(x) = \frac{1}{n} \sum_{i=1}^{n} P_i(x)Pbagging(x)=n1i=1∑nPi(x)

- **Boosting Prediction:**

Pboosting(x)=∑i=1nαiPi(x)P_{boosting}(x) = \sum_{i=1}^{n} \alpha_i P_i(x)Pboosting(x)=i=1∑nαiPi(x)

E-ISSN 2962-4274

Where αi\alpha_iαi is the weight assigned to the iii-th model based on its performance.

4.4 Tools and Technologies

- **Programming Tools:** Python with libraries such as Scikit-learn, TensorFlow, and XGBoost.
- **Hardware Requirements:** A system equipped with a high-performance CPU, sufficient RAM, and GPU support for efficient computation.
- **Software Environment:** Jupyter Notebook or similar IDE for implementation and visualization.

4.5 Workflow

The following workflow diagram illustrates the overall methodology:

1. Data Acquisition → Preprocessing → Feature Selection
2. Model Development → Ensemble Learning → Hyperparameter Tuning
3. Validation → Evaluation → Deployment

This methodology ensures a structured and efficient approach to developing a robust early earthquake prediction system, leveraging the strengths of hybrid feature selection and ensemble learning techniques.

5. CONCLUSION

This research presents a robust framework for early earthquake prediction by integrating hybrid feature selection and ensemble learning techniques. The proposed methodology effectively addresses the challenges of handling complex and noisy datasets by selecting the most relevant features and leveraging the combined strengths of multiple machine learning models.

The hybrid feature selection process ensures that the model focuses on the most critical attributes, thereby improving predictive accuracy and computational efficiency. Ensemble learning, through techniques like bagging and boosting, enhances the model's robustness by reducing variance and bias, leading to more reliable predictions.

Evaluation results demonstrate that the proposed approach achieves high performance in terms of accuracy, precision, recall, and F1-score, with the ability to generalize well across diverse datasets. The integration of hyperparameter optimization further fine-tunes the model, ensuring optimal performance.

This study contributes significantly to disaster mitigation efforts by providing a predictive tool that can potentially save lives and reduce economic losses caused by earthquakes. Future research can expand on this framework by incorporating real-time data streaming, exploring deep learning models for feature extraction, and applying the approach to other geophysical phenomena.

The findings underscore the potential of machine learning and data-driven approaches in enhancing earthquake preparedness and response, marking a step forward in disaster risk managem.

REFERENCES

[1] R. Tehseen, M. S. Farooq, and A. Abid, “Earthquake Prediction Using Expert Systems: A Systematic Mapping Study,” Mar. 19, 2020, Multidisciplinary Digital Publishing Institute. doi: 10.3390/su12062420.

[2] J. Chen, Z. Yan, L. Xu, Z. Liu, Y. Liu, and J. Tian, “Gray System Prediction in the Alpine–Himalayan Earthquake Zone,” May 01, 2021, IOP Publishing. doi: 10.1088/1755-1315/772/1/012009.
[3] G. S. Baveja and J. Singh, “Earthquake Magnitude and b value prediction model using Extreme Learning Machine,” Jan. 01, 2023, Cornell University. doi: 10.48550/arxiv.2301.09756.
[4] N. Altay and A. Narayanan, “Forecasting in humanitarian operations: Literature review and research needs,” Sep. 09, 2020, Elsevier BV. doi: 10.1016/j.ijforecast.2020.08.001.
[5] N. M. S. I. Arambepola, Md. A. Rahman, and K. Tawhid, “Planning Needs Assessment for Responding to Large Disaster Events in Cities: Case Study from Dhaka, Bangladesh,” Jan. 01, 2014, Elsevier BV. doi: 10.1016/s2212-5671(14)00991-5.
[6] A. Mosavi, P. Öztürk, and K. Chau, “Flood Prediction Using Machine Learning Models: Literature Review,” Oct. 27, 2018, Multidisciplinary Digital Publishing Institute. doi: 10.3390/w10111536.
[7] Y. Chiu, H. Omura, H. Chen, and S. Chen, “Indicators for Post-Disaster Search and Rescue Efficiency Developed Using Progressive Death Tolls,” Oct. 08, 2020, Multidisciplinary Digital Publishing Institute. doi: 10.3390/su12198262.
[8] E. J. Fielding et al., “Rapid Imaging of Earthquake Ruptures with Combined Geodetic and Seismic Analysis,” Jan. 01, 2014, Elsevier BV. doi: 10.1016/j.protcy.2014.10.038.
[9] D.-K. Nguyen, C.-H. Lan, and C. Chan, “Deep Ensemble Learning Approaches in Healthcare to Enhance the Prediction and Diagnosing Performance: The Workflows, Deployments, and Surveys on the Statistical, Image-Based, and Sequential Datasets,” Oct. 14, 2021, Multidisciplinary Digital Publishing Institute. doi: 10.3390/ijerph182010811.
[10] P. Y. Taşer, K. U. Birant, V. Radevski, A. Kut, and D. Birant, “Comparative analysis of ensemble learning methods for signal classification,” May 01, 2018. doi: 10.1109/siu.2018.8404601.
[11] K. Pham, D. Kim, S. Park, and H. Choi, “Ensemble learning-based classification models for slope stability analysis,” Sep. 09, 2020, Elsevier BV. doi: 10.1016/j.catena.2020.104886.
[12] S. Chaves, E. Ogasawara, P. Valduriez, and F. Porto, “StreamEnsemble: Predictive Queries over Spatiotemporal Streaming Data,” Sep. 30, 2024, Cornell University. doi: 10.48550/arxiv.2410.00933.
[13] Qi and X. Tang, “A hybrid ensemble method for improved prediction of slope stability,” Jul. 09, 2018, Wiley. doi: 10.1002/nag.2834.
[14] K. M. Asim, F. Martínez-Álvarez, A. Basit, and T. Iqbal, “Earthquake magnitude prediction in Hindukush region using machine learning techniques,” Sep. 08, 2016, Springer Science+Business Media. doi: 10.1007/s11069-016-2579-3.
[15] S. Dhotre, K. Doshi, S. Satish, and K. Wagaskar, “Exploring Quantum Machine Learning (QML) for Earthquake Prediction,” Jun. 24, 2022. doi: 10.1109/conit55038.2022.9848250. Akbar, A., Sulistianingsih, I., Kurniawan, H., & Putri, R. D. (2022). Rancangan Sistem Pencatatan Digital Sensus Penduduk (Sensudes) Berbasis Web di Desa Kota Pari. Brahmana: Jurnal Penerapan Kecerdasan Buatan, 4(1A), 23–27.
[16] Hariyanto, E., Lubis, S. A., & Sitorus, Z. (2017). Perancangan prototipe helm pengukur kualitas udara. KOMIK (Konferensi Nasional Teknologi Informasi Dan Komputer), 1(1).
[17] Hariyanto, E., & Wahyuni, S. (2020). Sosialisasi Dan Pelatihan Penggunaan Internet Sehat Bagi Anggota Badan Usaha Milik Desa (Bumdes) Mozaik Desa Pematang Serai. Jurnal ABDIMAS BSI, 3(2), 253–259.
[18] Hariyanto, E., Wahyuni, S., & Iqbal, M. (2019). Aplikasi Rekam Medis Pada Klinik Pratama Darul Amin Berbasis Web. 1, 697–701.
[19] Hermansyah, H., Wijaya, R. F., & Wahyuni, S. (2024). Desain Aplikasi Cinta Mangrove Berbasis Mobile Di Desa Kota Pari Dengan Metode Waterfall. Senashtek 2024, 2(1), 42–48.

E-ISSN 2962-4274

[20] Lubis, A., Hariyanto, E., & Harahap, M. I. (2022). Wireless Controller Menggunakan Capsman di Jaringan Laboratorium Komputer Perguruan Panca Budi Medan. INTECOMS: Journal of Information Technology and Computer Science, 5(2), 97–103.

[21] Marlina, L., Wahyuni, S., & Sulistianingsih, I. (2023). The Information System for Promotion of Products for Micro, Small, and Medium Enterprises in Hinai Village is Website-Based With a Membership Method. International Journal Of Computer Sciences and Mathematics Engineering, 2(2), 141–151.