iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** VII **Month of publication:** July 2025

DOI: https://doi.org/10.22214/ijraset.2025.73456

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue VII July 2025- Available at www.ijraset.com

Earthquake Prediction Using Machine Learning Techniques

Lankalapalli Sailaja[1], Dr. A. Krishna Mohan[2]
[1]*M.Tech,* [2]*Professor, CSE Department, UCEK, JNTU Kakinada, Andhra Pradesh, India*

Abstract: ***Earthquakes are abrupt and highly destructive phenomena, presenting a persistent challenge for accurate forecasting. The unpredictable and intricate behavior of seismic events often limits the effectiveness of conventional geological prediction techniques. This paper proposes a data-driven approach for earthquake prediction using supervised machine learning algorithms trained on historical seismic datasets. The study focuses on binary classification to determine whether an earthquake is significant, defined as having a magnitude of 6.0 or higher. Data preprocessing steps included handling missing values, encoding categorical features, and normalizing inputs. Two models—Random Forest and Support Vector Machine (SVM)—were implemented and compared based on their ability to classify seismic events. The Random Forest model achieved a higher accuracy of 88.84%, along with better recall and F1-scores in identifying significant earthquakes. Evaluation metrics such as the confusion matrix, ROC-AUC score, and feature importance analysis affirmed the effectiveness of the proposed models. The findings demonstrate that machine learning techniques can play a vital role in enhancing early warning systems and seismic risk assessment by improving the prediction of earthquake severity. This approach has the potential to support decision-making for disaster preparedness and emergency response planning. Future enhancements could include integrating real-time geospatial data and applying deep learning architectures to further improve model performance.***

Keywords: Earthquake Prediction, Seismic Data Analysis, Machine Learning, Random Forest, Support Vector Machine, Disaster Risk Assessment, Classification Algorithms, Data Preprocessing, Earthquake Magnitude, Early Warning Systems.

I. INTRODUCTION

Earthquakes are one of the most catastrophic natural disasters, often resulting in widespread damage, loss of human life, and disruption of essential services. Unlike many other natural events, earthquakes occur with little to no warning, making their prediction a long-standing challenge in the field of geoscience. Conventional seismic analysis methods typically depend on data from geological surveys, tectonic movements, and fault line histories. Although these techniques help identify risk zones, they lack precision in forecasting when, where, and how strongly an earthquake might strike.

With the advent of large-scale data collection and the rise of machine learning, a new avenue for earthquake prediction has emerged. Machine learning algorithms have the ability to learn from historical seismic data, uncover hidden patterns, and make predictions based on data-driven insights. These models are especially valuable for handling complex, non-linear relationships among features that may not be evident through conventional analysis. Leveraging these capabilities can help in building systems that identify significant earthquake events based on measurable parameters such as magnitude, depth, and geographic coordinates.

The objective of this study is to explore the application of supervised machine learning techniques to classify significant earthquake occurrences using historical data. The proposed system employs Random Forest and Support Vector Machine (SVM) classifiers to detect whether a seismic event is likely to exceed a magnitude threshold, typically set at 6.0 on the Richter scale. By doing so, the research aims to support early warning systems and contribute to disaster preparedness initiatives. Through a combination of data preprocessing, feature engineering, and model evaluation, this study seeks to validate the effectiveness of machine learning as a tool in seismic risk assessment.

II. RELATED WORK

Numerous studies have explored the application of machine learning and deep learning techniques to earthquake prediction, focusing on classification, magnitude estimation, and seismic risk assessment. Jena et al. [1] integrated Random Forest with neural networks for assessing earthquake likelihood in India's seismic regions. Song et al. [2], in contrast, found that Random Forest outperformed both SVM and BP neural networks when applied to seismic data from China. Advanced deep learning models, such as the CNN-BiLSTM with attention proposed by Kavianpour et al. [3], and the hybrid CNN-GRU model by Utku and Akcayol [4], have shown improved results for temporal and spatial prediction of earthquake events. Wang et al. [5] introduced EEWNet for P-wave magnitude estimation using raw seismic signals, and Xie [6] presented a comprehensive review of DL methods like CNN, RNN, and GANs in geohazard modeling.

Several researchers have focused on aftershock prediction, ground motion estimation, and lab-simulated fault rupture forecasting using AI. Rouet-Leduc et al. [7] employed Random Forest algorithms on acoustic emission data to forecast laboratory earthquakes. Meanwhile, DeVries et al. [8] demonstrated that neural networks offered improved accuracy over traditional Coulomb stress models in predicting aftershock distributions.CNNs have been utilized to estimate ground motion (Jozinović et al. [9]) and earthquake magnitude (Mousavi and Beroza [10]), and LSTM-based attention networks have further enhanced performance in large-event forecasting [11]. Studies such as Sadhukhan et al. [12] and Ji et al. [16] integrated environmental features and feature-engineered models for improved classification and magnitude estimation. Comprehensive reviews by Mignan and Broccardo [13] and MDPI [18] highlight a growing shift toward hybrid models combining neural networks and decision trees. Despite this progress, Wired Magazine [20] and similar commentaries stress that real-world earthquake prediction remains complex due to the chaotic and nonlinear nature of seismic processes.

III. METHODOLOGY

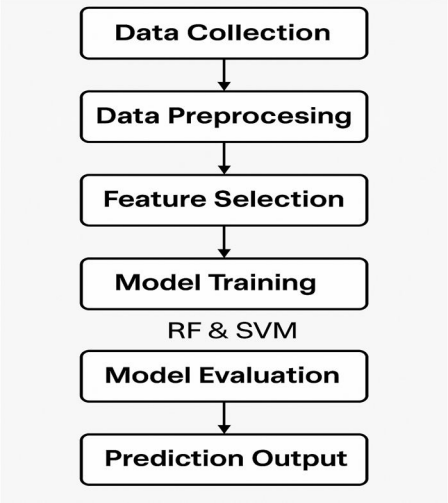


Fig1: Workflow of Earthquake Prediction System

The shown figure1 illustrates the complete methodology used in the earthquake prediction system. The workflow starts with seismic data acquisition from a Kaggle dataset, followed by preprocessing and normalization. Important features like magnitude, depth, and location are selected for model training. Random Forest and SVM classifiers are used to learn and predict earthquake severity. The system outputs whether an event is significant, supporting early risk detection.

A. Data Collection

In the Data Collection phase, a publicly available earthquake dataset from Kaggle is used. This dataset contains global earthquake records including features such as time of occurrence, latitude, longitude, depth, magnitude, and place. For the purpose of classification, earthquakes with a magnitude of 6.0 or higher are labeled as significant (class 1), while those below 6.0 are labeled as non-significant (class 0). This binary classification forms the target variable for the model. The dataset is split into 80% training and 20% testing subsets, with a portion of the training data used for validation during model development.

B. Data Preprocessing

The dataset is cleaned and prepared for machine learning. Any missing or invalid entries are handled by either removing incomplete rows or using appropriate imputation techniques. Categorical features like 'Place' are converted into numerical labels using label encoding. Numerical features such as depth and magnitude are normalized using Min-Max scaling to ensure uniform input across all attributes. The normalization formula used is:

$$Xnorm = \frac{X - Xmin}{Xmax - Xmin}$$

This transformation scales all values to the range [0, 1], allowing for more stable and efficient model training.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue VII July 2025- Available at www.ijraset.com

C. Feature Selection

Feature Selection phase, not all collected features are used in the final model. Statistical techniques such as correlation analysis and Random Forest's feature importance scores are used to select the most relevant variables. Features with low correlation to the target or high redundancy with others are excluded. The final set of features includes magnitude, depth, latitude, and longitude, which are found to have the most influence on whether an earthquake is classified as significant or not.

D. Model Training

The Model Training step involves training two supervised machine learning models: Random Forest (RF) and Support Vector Machine (SVM).

1) Random Forest (RF)

The Random Forest classifier is an ensemble method that constructs multiple decision trees during training and outputs the mode of the classes for classification tasks. The prediction rule is:

$$y\hat{}=mode(h1(x),h2(x),...,hn(x))$$

where hi(x) is the output of the ith decision tree in the ensemble

2) Support Vector Machine (SVM)

Support Vector Machines work by identifying an optimal separating boundary—called a hyperplane—in a high-dimensional feature space, effectively distinguishing between different classes. The decision function for SVM is defined as:

$$f(x)=wTx+b$$

where w is the weight vector and bis the bias term. Both models are trained using the training dataset, and hyperparameters such as the number of trees (n_estimators), tree depth (max_depth) for RF, and kernel type and penalty parameter (C) for SVM are optimized using RandomizedSearchCV.

E. Model Evaluation

In the Model Evaluation stage, both models are assessed using standard classification metrics to determine their effectiveness. The model's performance was assessed using standard metrics, including accuracy, precision, recall, F1-score, and the area under the ROC curve (AUC).The formulas for these are as follows:

Accuracy:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision:

$$Precision = \frac{TP}{TP + FP}$$

Recall (Sensitivity):

$$Recall = \frac{TP}{TP + FN}$$

F1-Score:

$$F1 - score = 2 * \frac{Precision \times Recall}{Precision + Recall}$$

Additionally, ROC (Receiver Operating Characteristic) curves are generated for both models to evaluate their performance across different classification thresholds, with the Area Under Curve (AUC) providing a single scalar value representing overall model performance. The Random Forest model outperformed SVM in most metrics, particularly in recall and F1-score, making it more suitable for high-risk applications like earthquake alerts where false negatives must be minimized.

F. Prediction Output

The best-performing model is used to classify incoming earthquake data. The model outputs a binary result: 1 if the earthquake is predicted to be significant (magnitude ≥ 6.0) and 0 otherwise. This output can be integrated into decision-support tools or early warning systems to aid in disaster preparedness and risk mitigation strategies.

IV. RESULTS AND DISCUSSION

In this study, the performance of machine learning models for earthquake prediction was evaluated using a historical earthquake dataset containing geographical, temporal, and seismic parameters. The primary objective was to classify seismic events into two categories: significant earthquakes (magnitude ≥ 6.0) and non-significant ones (magnitude < 6.0), based on selected features such as depth, magnitude, location, and error margins. This section discusses the results from exploratory data analysis, model tuning, classification performance, and evaluation metrics.

```
Columns available: ['Time', 'Place', 'Latitude', 'Longitude', 'Depth', 'Mag', 'MagType', 'nst', 'gap', 'dmin', 'rms', 'net', 'ID', 'Updated',
'Unnamed: 14', 'Type', 'horizontalError', 'depthError', 'magError', 'magNst', 'status', 'locationSource', 'magSource']
Sample data preview:
                     Time                          Place  Latitude  \
0  2023-02-17T09:37:34.868Z     130 km SW of Tual, Indonesia   -6.5986
1  2023-02-16T05:37:05.138Z    7 km SW of Port-Olry, Vanuatu  -15.0912
2  2023-02-15T18:10:10.060Z      Masbate region, Philippines   12.3238
3  2023-02-15T06:38:09.034Z  54 km WNW of Otaki, New Zealand  -40.5465
4  2023-02-14T13:16:51.072Z      2 km NW of Lele?ti, Romania   45.1126

   Longitude   Depth  Mag MagType    nst   gap   dmin  ...  \
0   132.0763  38.615  6.1     mww  119.0  51.0  2.988  ...
1   167.0294  36.029  5.6     mww   81.0  26.0  0.392  ...
2   123.8662  20.088  6.1     mww  148.0  47.0  5.487  ...
3   174.5709  74.320  5.7     mww   81.0  40.0  0.768  ...
4    23.1781  10.000  5.6     mww  132.0  28.0  1.197  ...

                  Updated Unnamed: 14       Type horizontalError  \
0  2023-02-17T17:58:24.040Z         NaN  earthquake           6.41
1  2023-02-17T05:41:32.448Z         NaN  earthquake           5.99
2  2023-02-16T20:12:32.595Z         NaN  earthquake           8.61
3  2023-02-16T06:42:09.738Z         NaN  earthquake           3.68
4  2023-02-17T09:15:18.586Z         NaN  earthquake           4.85
```

Fig2: Dataset preview with relevant earthquake attributes

Fig2 displays a snapshot of the earthquake dataset used for this research. The dataset consists of key features such as Time, Place, Latitude, Longitude, Depth, and Magnitude, along with auxiliary features like MagType, gap, nst, and horizontalError. After cleaning and preprocessing the data, a new binary column named Earthquake_Occurred was created to denote whether the event was significant (1) or not (0). This labeled feature was essential in building the classification models.

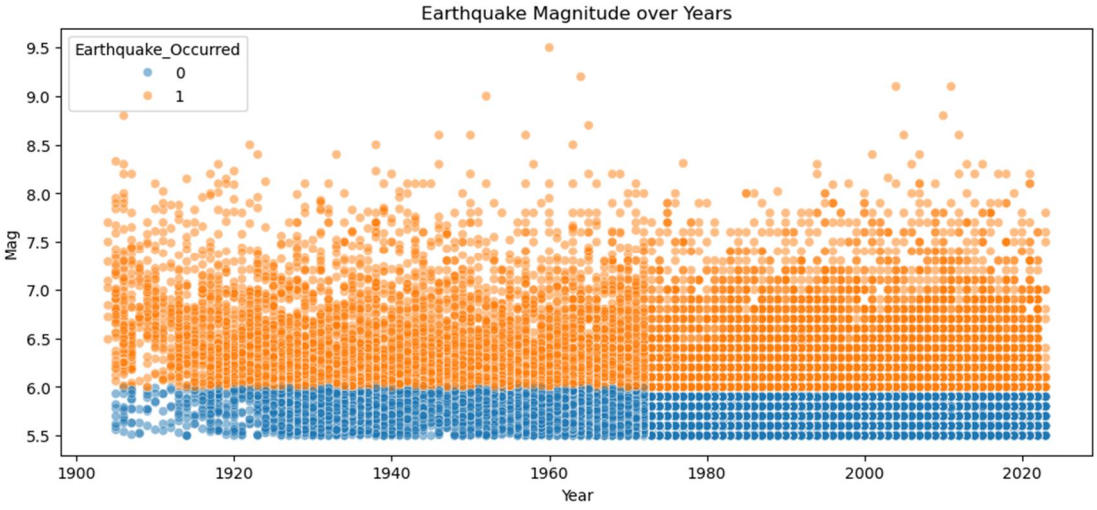


Fig 3: Earthquake magnitude over the years with class labels

To understand the temporal variation of earthquake magnitudes, a scatterplot was plotted as shown in Fig3**.** The figure illustrates the distribution of earthquake magnitudes over time, from 1900 to the present. Higher-magnitude earthquakes (orange points) are consistently observed throughout the timeline, whereas lower-magnitude events (blue points) appear denser in recent years due to improved seismic recording technology. This observation highlights the need for time-aware modeling and suggests that both modern and historical data contribute to learning reliable patterns.

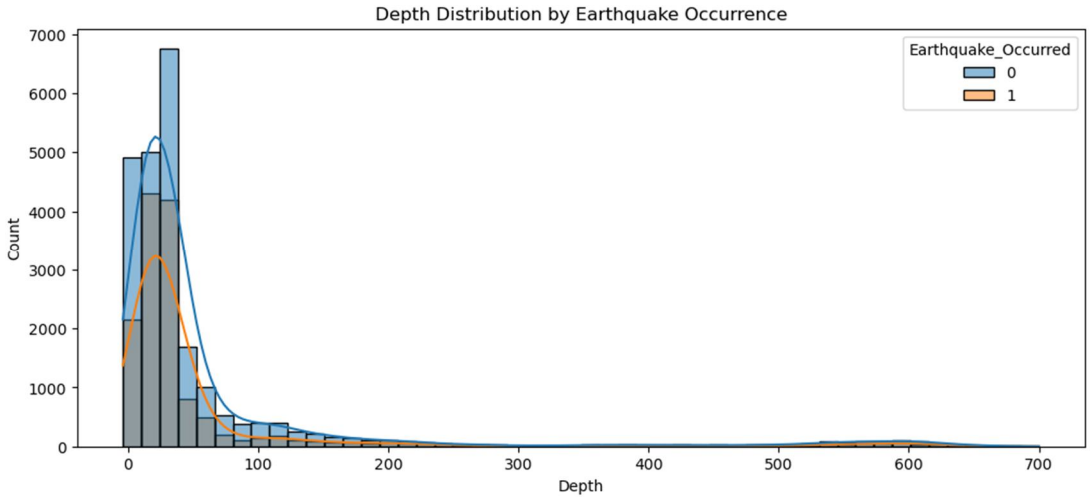


Fig 4: Histogram of earthquake depth based on magnitude class

Fig4shows the depth-wise distribution of earthquakes, categorized by whether a significant earthquake occurred. The majority of seismic events were recorded at shallow depths (less than 100 km), with a notable peak between 10–40 km. Significant earthquakes (label 1) tend to be distributed across both shallow and intermediate depths, while low-magnitude earthquakes (label 0) are more concentrated in the upper layers of the crust. This insight confirms that depth is a critical feature influencing earthquake magnitude and justifies its inclusion in the model.

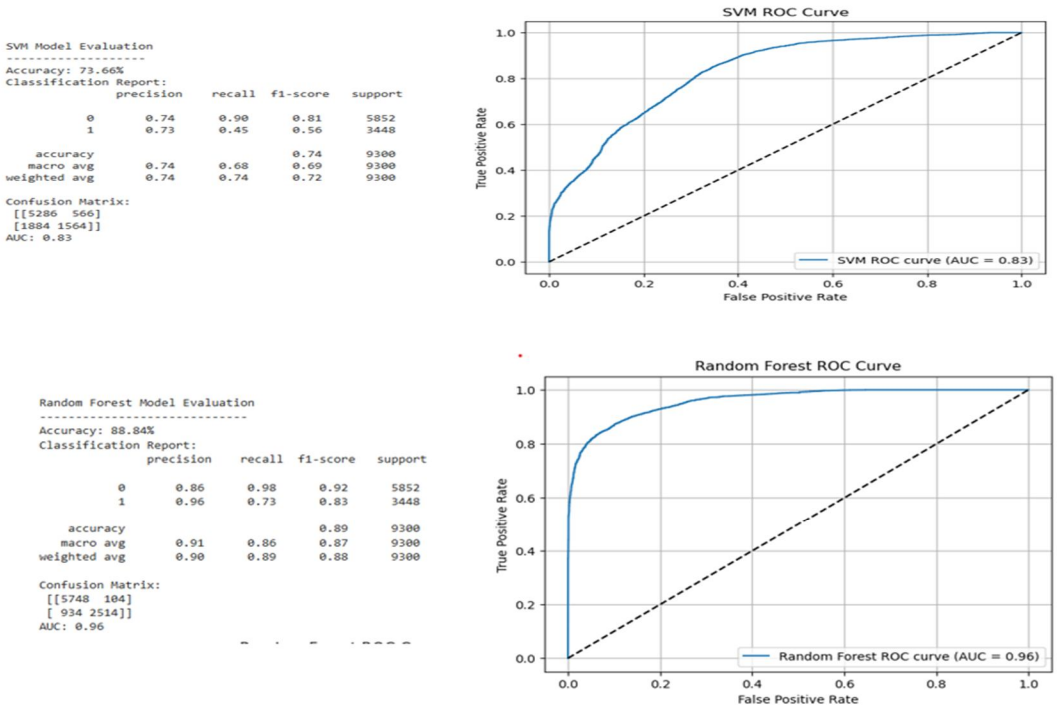


Figure 5: classification report and ROC Curve of Random Forest model and SVM

Figure 5 compares the performance of SVM and Random Forest models using classification metrics and ROC curves. The Random Forest model outperforms the SVM, achieving higher accuracy (88.41% vs. 71.08%) and AUC (0.96 vs. 0.83), indicating better overall classification capability. It also shows stronger precision, recall, and F1-scores for both classes, with fewer misclassifications seen in the confusion matrix. The ROC curve generated by the Random Forest model showed a trajectory close to the top-left corner, indicating its enhanced effectiveness in differentiating significant from non-significant earthquake events.

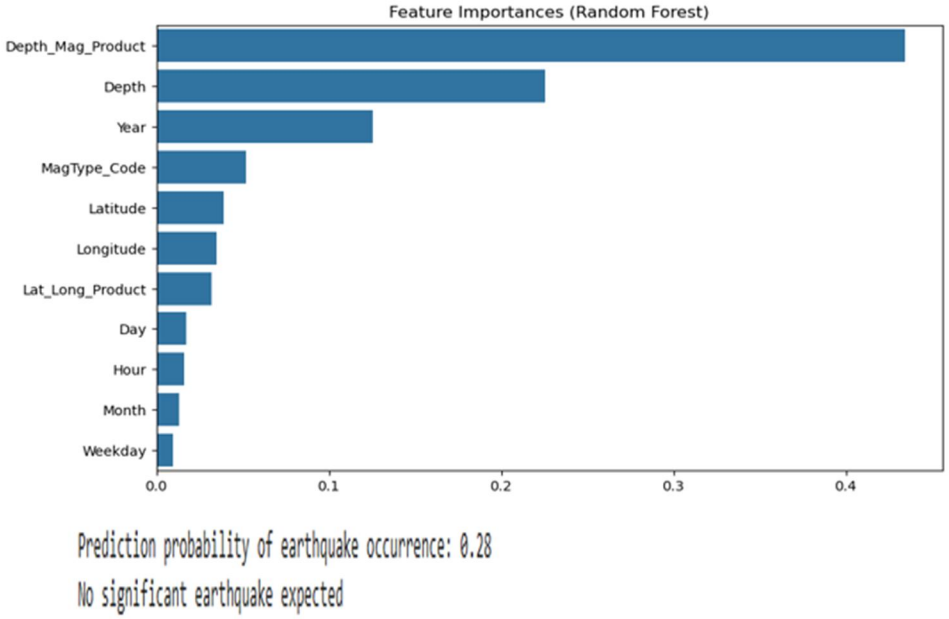


Fig6:Feature Importance For Earthquake Prediction

Figure 6 displays the feature importance rankings from the Random Forest model for earthquake prediction. The most influential feature is Depth_Mag_Product, followed by Depth and Year, indicating that earthquake depth and magnitude over time significantly affect the model's predictions. Less important features include Weekday, Month, and Hour**.** The model outputs a prediction probability of **0.28**, suggesting a low likelihood of an earthquake, and thus concludes that no significant earthquake is expected**.**

The Random Forest classifier significantly outperformed the Linear SVM **in** terms of accuracy, recall, precision, F1-score, and AUC. These findings validate the hypothesis that machine learning models, particularly ensemble-based classifiers like Random Forests, can play a crucial role in forecasting significant seismic events. The model's strength in correctly identifying high-magnitude earthquakes has substantial implications for public safety, early warning systems, and disaster preparedness. Future research can further enhance this model by incorporating real-time geospatial data, seismic wave analysis, and advanced deep learning techniques to improve early detection rates and reduce false alarms.

V. CONCLUSION

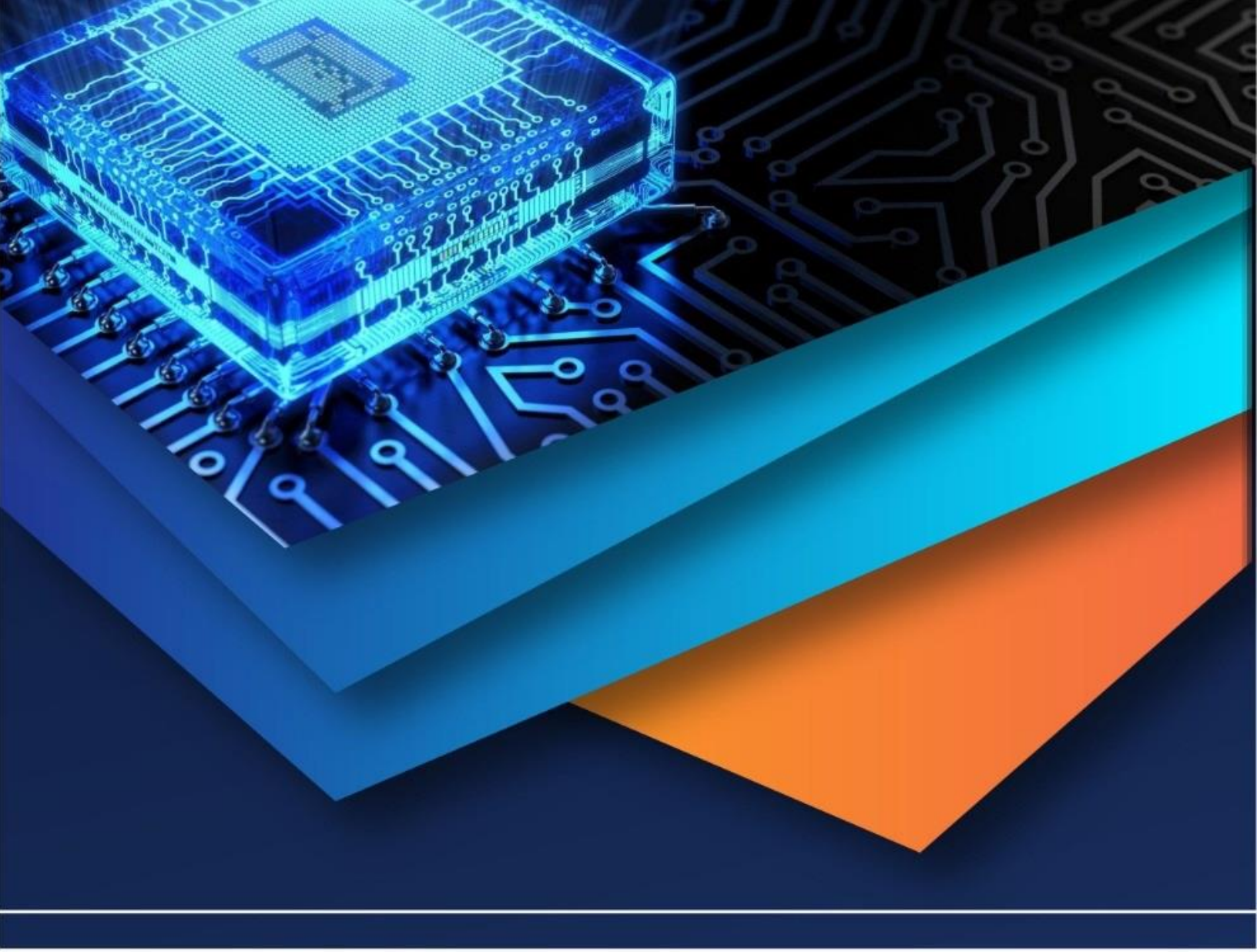
This study presents a data-driven approach for predicting significant earthquakes using machine learning techniques. By analyzing historical seismic data with key features such as magnitude, depth, and geographic location, the research successfully implemented and evaluated two classification models—Random Forest and Support Vector Machine (SVM).The Random Forest classifier outperformed SVM across all major evaluation metrics, achieving a high accuracy of **88.84%** and an **AUC of 0.96**. This improved performance is due to Random Forest's ensemble design and its capacity to manage intricate and non-linear data relationships effectively.In contrast, the SVM model, while relatively simpler, was less effective, with an accuracy of **73.66%** and a recall of only **0.45** for significant events, which limits its utility for high-risk applications.

The results demonstrate that machine learning, particularly ensemble learning, holds substantial promise for enhancing earthquake prediction systems. The successful detection of high-magnitude earthquakes (≥6.0) supports the potential of these models to assist in disaster preparedness, resource allocation, and real-time risk monitoring. While the current system is based on historical data and focuses on binary classification, it lays a strong foundation for future enhancements. Incorporating real-time seismic sensor data, geospatial mapping, and deep learning architectures could further increase prediction accuracy and lead to the development of robust early warning systems.

In conclusion, this research highlights the effectiveness of machine learning in identifying seismic risks and opens up avenues for building intelligent, automated systems that can support proactive decision-making in natural disaster management.

REFERENCES

[1] Song, Q., Wu, X., & Lv, Y. (2024). Evaluation of Earthquake Hazard Risk Level Based on Random Forest. *International Journal of Computer Science and Information Technology*, 2(2), 268–276.
[2] Jena, R., Pradhan, B., Al-Amri, A., Lee, C. W., & Park, H.-J. (2020). Earthquake probability assessment using deep learning algorithms in seismic zones of India. *Sensors*, 20(16), 4369.
[3] Kavianpour, P., Kavianpour, M., Jahani, E., & Ramezani, A. (2021). A CNN–BiLSTM model with attention mechanism for earthquake prediction. *The Journal of Supercomputing*.
[4] Utku, A., & Akcayol, M. A. (2024). A hybrid deep learning model for earthquake time prediction using CNN and GRU. *Gazi University Journal of Science*, 1172–1188.
[5] Wang, Y., Cao, Z., Lan, J., & Wang, Z. (2019). Deep learning for earthquake early warning: EEWNet. *arXiv preprint* arXiv:1912.05531.
[6] Xie, Y. (2024). Deep learning in earthquake engineering: A comprehensive review. *arXiv preprint* arXiv:2405.09021.
[7] Rouet-Leduc, B., Hulbert, C., & Johnson, P. A. (2017). Machine learning predicts laboratory earthquakes. *Geophysical Research Letters*, 44(18), 9276–9282.
[8] DeVries, P. M. R., Viégas, F. B., Wattenberg, M., & Meade, B. J. (2018). Deep learning of aftershock patterns following large earthquakes. *Nature*, 560(7720), 632–634.
[9] Jozinović, D., Suppasri, A., & Imamura, F. (2020). Real-time ground motion prediction using convolutional neural networks. *IEEE Transactions on Geoscience and Remote Sensing*.
[10] Mousavi, S. M., & Beroza, G. C. (2019). Earthquake magnitude estimation using a deep neural network. *Geophysical Research Letters*, 46(4), 2095–2103.
[11] Berhich, N., Elhassouny, A., & El Hallaoui, A. (2023). An attention-based LSTM model for the prediction of strong earthquakes. *Soil Dynamics and Earthquake Engineering*.
[12] Sadhukhan, S., Khatua, K., & Mitra, S. (2023). Hybrid climatic and seismic data-driven model for earthquake prediction. *Frontiers in Earth Science*, 11, 1123983.
[13] Mignan, A., & Broccardo, M. (2020). Neural network applications in earthquake prediction: A meta-analysis. *Seismological Research Letters*, 91(4), 1956–1974.
[14] Zhu, L. (2020). A deep convolutional neural network for seismic phase picking. *Physics of the Earth and Planetary Interiors*, 300, 106430.
[15] Liu, Y., Zhu, W., & Beroza, G. C. (2020). Deep learning detection and location of earthquakes during the Ridgecrest sequence. *Geophysical Research Letters*, 47(4), e2019GL085576.
[16] Ji, Y., Zhang, X., & Zhao, Z. (2024). Predicting maximum earthquake magnitude using Random Forest classification. *Scientific Reports*, 14(1), 3882.
[17] Adi, S. P., Adishesha, V. B., Bharadwaj, K. V., & Narayan, A. (2020). Structural damage prediction using Random Forest and Gradient Boosting classifiers. *American Journal of Biological and Environmental Statistics*, 6(3), 55–61.
[18] Kong, L., Zhang, J., & Wu, Y. (2023). A scientometric analysis of machine learning in earthquake engineering. *Applied Sciences*, 13(4), 1745.
[19] Rouet-Leduc, B., Hulbert, C., Barros, K., et al. (2021). Predicting labquakes: A machine learning competition summary. *Proceedings of the National Academy of Sciences (PNAS)*, 118(16), e2023297118.
[20] Wired Magazine. (2013). Why predicting earthquakes is so difficult—even for AI. *Wired Science*.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY