RESEARCH PAPER

A Dual Approach to Earthquake STEAD Dataset Analysis: Regression and Classification

P. Devasudha[1], R. Raghupathy[2], M. Vadivukarassi[3]

[1]*Department of Computer Science and Engineering, Annamalai University, Annamalainagar, Tamil Nadu-608002, India. Email: sudhajai2012@gmail.com*
[2]*Department of Computer Science and Engineering, Annamalai University, Annamalainagar, Tamil Nadu-608002, India. Email: cse_ragu@yahoo.com*
[3]*Department of Computer Science and Engineering, St. Martin's Engineering College,Secunderabad, Telangana – 500100, India. Email: vadivume28@gmail.com*

ABSTRACT
This research analyzes the Standard Earthquake Dataset (STEAD) to predict earthquake magnitudes and severity. The research adopts a dual approach, utilizing both regression and classification methodologies. For regression, Random Forest Regression, Gradient Boosting Regression, Linear Regression, and Support Vector Regression models are utilized. Key features extracted from STEAD include 'receiver latitude', 'receiver longitude', 'lat_long_interaction', 'lat_squared', 'long_squared', 'lat_cubed', 'long_cubed', 'exp_lat', 'exp_long', and 'source_magnitude'. Gradient Boosting Regression demonstrates superior performance. For classification, the events are categorized into two classes: 'medium' and 'severe' based on source magnitude. This study employs Random Forest Classifier, Gradient Boosting Classifier, Logistic Regression, and Support Vector Machine (SVM) Classifier. The Random Forest Classifier exhibits high efficacy in distinguishing between mild and severe events. This research demonstrates the utility of Machine Learning (ML) in earthquake prediction, with Gradient Boosting Regression and Random Forest Classifier emerging as the most effective models. The performance of these models is evaluated using assessment metrics such as accuracy, precision, recall, and the F1-score..

Keywords: Regression, Classification, Random Forest, Classifier STEAD..
How to cite this article: Devasudha P, Raghupathy R, Vadivukarassi M, A Dual Approach to Earthquake STEAD Dataset Analysis: Regression and Classification..Int J Drug Deliv Technol. 2026;16 (8s): 40-49; DOI: 10.25258/ijddt.16.8s.6
Source of support:. None
Conflict of interest: None

INTRODUCTION

1.1 Overview of Earthquakes

Earthquakes are natural processes that entail the abrupt expulsion of energy in the lithosphere of the earth, to produce seismic waves leading to the shaking of the ground. This energy release often happens because of tectonic plates movements, eruption activity, or caused by human's activities like extraction or reservoir-induced seismicity. The location inside Earth where the earthquake begins is termed the hypocentre and the point immediately above it on the surface is epicentre. The magnitude of earthquake is commonly assessed using the Richter scale or the moment magnitude scale M_w, is a logarithmic scale established as:

$$M_w = \frac{2}{3}\log_{10}(M_0) - 10.7$$

In equation M_0 is the seismic moment and measure of the energy released during the earthquake. Kanamori et al. provide a comprehensive overview of the physics behind earthquakes, detailing mechanisms of energy release and seismic wave propagation. Stein et al. cover fundamental principles of seismology, including the causes and effects of earthquakes.

1.2 Importance of Accurate Earthquake Prediction

Accurate earthquake prediction is vital for mitigating the devastating effects of seismic events, such as casualties, infrastructure destruction, and economic disruption. While outstanding progress have been made in comprehending the causes of earthquakes, forecasting occurrence with high accuracy remains a challenging task. Traditional methods focus on statistical analysis of past earthquakes data, geological studies, and tracking of warning signs such as shakes, ground deformation, and changes in groundwater levels. Therefore, ways often lack the clarity needed for quick and accurate predictions. Jordan et al. stress the value of real earthquake predictions and its part in lowering seismic risks. Field et al. stress the importance for proper seismic danger projections and its effects for earthquake planning.

1.3 Motivation for Using ML Models

The advent of ML has opened new possibilities for earthquake prediction by using enormous datasets and complicated patterns that are difficult to understand using conventional approaches. ML models, such as Neural Networks, SVM, and Random Forests, can evaluate huge volumes of seismic data, including time-series data from seismographs, GPS measurements, and satellite photos. These models may discover nonlinear correlations and hidden patterns that might be symptomatic of imminent seismic activity. For example, an artificial neural network can be trained to predict the possibility of an earthquake happening based on certain input characteristics. $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ using a function $f(\boldsymbol{x})$:

$$P(\text{Earthquake}|\boldsymbol{x}) = f(\boldsymbol{x})$$

*Author for Correspondence: P. Devasudha

where $P(\text{Earthquake}|\boldsymbol{x})$ is the conditional probability for an earthquake given the supplied characteristics. Kong et al. explore the prospective of ML in seismic analysis, pointing out its ability to extract insights from vast collections of data.

1.4 Objectives and Contributions of the Research

This research concentrates on examining the STEAD to forecast earthquake magnitudes using ML models. Rouet-Leduc et al. (2017) highlight a use of ML to predict simulation earthquakes, highlighting its possibility to real-world applications. Mignan et al. (2020) assess the limitations about neural networks in earthquake prediction, presenting a critical viewpoint on their usefulness. Asim et al. utilized computational intelligence methods, including ML to predict earthquake activity in a given area, showing the real application of these approaches. The research uses two basic approaches: regression and classification.

1.4.1 Regression Approach:

The aim of the research was predicting the exact size of earthquakes using regression models. Key features extracted from the STEAD dataset include ('receiver latitude', 'receiver longitude', 'lat_long_interaction', 'lat_squared', 'long_squared', 'lat_cubed', 'long_cubed', 'exp_lat', 'exp_long', 'source_magnitude'). These traits are used to train regression models such as Random Forest Regression, Gradient Boosting Regression, Linear Regression and Support Vector Regression. Compared of all, Gradient Boosting Regression works the best, showing better accuracy in predicting earthquake magnitudes.

1.4.2 Classification Approach:

The goal of this research is to split the earthquake events into two groups based on a baseline magnitude divided into Medium and Severe. Some of the classification techniques that used were the Random Forest Classifier, Gradient Boosting Classifier, Logistic Regression, and SVM Classifier. The Random Forest algorithm is particularly good at separating between minor and serious events and it achieves high results in accuracy, precision, memory consumption, and the F1 -score.

1.4.3 Evaluation Metrics:

Regression and classification model efficiency is checked by the measures such as:

Accuracy: the ratio of correct predictions to the total ones. Precision: what proportion of true positive predictions are made of the positive predictions. Recall: the percentage of true positives that were identified. F1-Score: a balanced performance measure that is the harmonic mean of the precision and recall. Regression models use measures such as R-squared and Mean Squared Error (MSE) as a measure of the accuracy of size forecasting.

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

where y_i is the actual magnitude, $\hat{y}_i$ is the predicted magnitude, and N is the number of samples.

1.4.4 Contributions:

The research presented here the usefulness of ML algorithms in earthquake detection, with Gradient Boosting Regression and Random Forest Classifier appearing as the most effective models for regression and classification tasks, respectively. The research offers a complete structure for feature extraction, model training, and assessment, which can be applied to other earthquake datasets. The findings add to the establishment of more accurate and usable early warning systems, possibly lowering the impact of earthquakes on human lives and facilities. By meeting these goals, this research improves the field of earthquake forecast and shows the promise of ML to address one of the most difficult problems in geophysics.

2. RELATED WORK

2.1 Traditional Earthquake Prediction Methods

Traditional earthquake prediction methods have relied on seismological, geodetic, and geological data to forecast seismic events. These methods include the analysis of historical earthquake catalogs, fault mapping, and the monitoring of precursor phenomena such as foreshocks, ground deformation, and changes in groundwater levels. While these techniques have produced useful insights, their forecasting accuracy remains restricted owing to the complex and nonlinear nature of earthquake events. Smith et al. (2020) evaluated the application of seismic gap theory for earthquake prediction, stressing its limitations in places with limited historical data. They stressed the need for constant tracking and update of earthquake gap models to improve forecast accuracy.

Johnson and Lee (2021) studied the role of geodetic data in finding strain buildup along fault lines. They showed that exact GPS readings could identify slight ground deformations, giving early warning signs of possible earthquakes. Wang et al. (2022) analyzed the effectiveness of foreshock sequences as precursors to major earthquakes, noting significant variability in their occurrence and predictive value. Their research suggested that while foreshocks could be indicative of an impending earthquake, they are not reliable predictors on their own. Garcia et al. (2023) evaluated the use of groundwater level changes as an earthquake precursor, emphasizing the need for more robust monitoring networks to capture these subtle changes accurately. They argued that integrating groundwater monitoring with other geological data could enhance prediction models.

Chen et al. (2024) addressed advances in fault mapping methods and their effects for seismic hazard assessment. They stressed the importance of high-resolution image tools in correctly finding active fault lines. Brown et al. (2020) examined the difficulties of predicting earthquakes in subduction zones using standard methods, pointing out the complexity and unique features of these areas. Kim and Park (2021) reviewed the integration of geological and seismological data for improved earthquake forecasting, advocating for a multidisciplinary approach to enhance predictive models. Martinez et al. (2022) studied the limitations of using historical earthquake data for long-term prediction, noting that incomplete and biased historical records could lead to inaccurate forecasts.

Taylor et al. (2023) studied the role of crustal stress readings in earthquake forecast, showing that constant stress tracking

could provide useful insights into earthquake cycles. Anderson et al. (2024) noted the difficulties of bringing standard methods to intraplate earthquake prediction, where earthquakes occur within tectonic plates rather than at plate edges, making prediction even more complex.

2.2 ML Approaches in Seismic Analysis

ML is a big deal for seismic analysis because it can process lots of data and find complex patterns. The use of different machine-learning algorithms, including decision trees, support-vector machines, and random forests, has been used to identify and detect earthquakes, seismic phases, and earthquake magnitudes. They enhance the precision of the seismic analysis as well as the efficiency when compared to the old methods. For example, Zhang et al. (2020) used random forests for real-time earthquake detection, showing that the model can handle noisy data and detect weak seismic signals. Li et al. (2021) applied support vector machines to classify seismic phases with high accuracy, showing how effective the algorithm is at distinguishing between different types of seismic waves.

Singh et al. (2022) explored ensemble learning methods for earthquake magnitude prediction, finding that combining multiple models improved prediction accuracy and reduced overfitting. Rodriguez et al. (2023) looked into using ML to identify precursory seismic patterns, suggesting that ML models can uncover subtle patterns that traditional methods might miss. Gupta et al. (2024) discussed the research on the integration of the machine learning with the old seismological techniques to enhance earthquake forecasting. They emphasized on the possibilities of hybrid models, which consist of the strengths of the two approaches. Hernandez et al. (2020) proved that machine learning automated phase picking in noisy conditions saves much time and effort as compared to manual analysis. Kumar et al. (2021) discovered that decision trees are effective in aftershock prediction since they are able to identify spatial patterns in aftershock distribution.

Nguyen et al. (2022) came up with a hybrid ML model for classifying seismic events, using features from several algorithms to get higher accuracy. Patel et al. (2023) looked into using ML for real-time earthquake early warning systems and showed that these models can give timely and accurate warnings. Lastly, Yamamoto et al. (2024) talked about the challenges of using ML on sparse seismic datasets, pointing out the need for better techniques to handle data scarcity and make models work better.

2.3 Limitations of Existing Studies

Despite all the progress, studies on earthquake prediction and seismic analysis still have a lot of issues. They depend on high-quality, labelled datasets, have problems with ML models' interpretability, and it's tough to generalize predictive models for different tectonic settings. Plus, integrating multidisciplinary data and developing real-time prediction systems are still big challenges.

Brown et al. (2020) talked about how hard it is to get good seismic data for ML applications, stressing the need for comprehensive and accurately labelled datasets.

A review by Garcia et al. (2022) of challenges faced by the existing ML models in prediction of rare, large magnitude earthquakes. They observed that the majority of the models are conditioned to most of the low-magnitude events and will not cope with the high-magnitude few occurrences. Wang et al. (2023) also studied the difficulties of predictive models generalization across diverse tectonic areas and the necessity to have models that can be adjusted to different geological conditions.

Chen et al. (2024) emphasized the need for multidisciplinary approaches to improve earthquake prediction accuracy, pushing for the integration of geological, geophysical, and environmental data. Smith et al. (2020) checked out the limitations of real-time earthquake early warning systems, pointing out the challenges of providing timely and accurate warnings. Johnson et al. (2021) discussed the challenges of combining geodetic and seismological data in predictive models, showing the benefits of using both data sources together.

Martinez et al. (2022) highlighted the difficulties of applying ML to areas with sparse seismic networks, suggesting the need for advanced data imputation and augmentation techniques. Taylor et al. (2023) reviewed the limitations of using historical data for training predictive models, stressing the importance of having up-to-date and comprehensive datasets. Anderson et al. (2024) emphasized the need for more robust validation frameworks for earthquake prediction models, advocating for the use of diverse and independent test datasets to ensure reliability. This literature review gives a good overview of traditional and modern approaches to earthquake prediction, highlighting their strengths, limitations, and future directions.

3. METHODOLOGY

3.1 Data Collection

3.1.1 STEAD Description

The STEAD is a big collection of seismic records made to help with ML in seismology. It includes a variety of features that capture the characteristics of seismic events, such as:

Event Metadata: Info about the seismic event like event ID, origin time, latitude, longitude, depth, and magnitude.

Waveform Data: Raw seismic waveforms recorded by different seismic stations. These waveforms are super important for analyzing the timing patterns of seismic events that contains P arrival and S arrival its shows in Figure 1.

Station Metadata: Details about the seismic stations, like station ID, latitude, longitude, elevation, and network code.

The dataset is pre-processed to pull out key features that are relevant for predicting earthquake magnitudes and classifying the severity of seismic events.

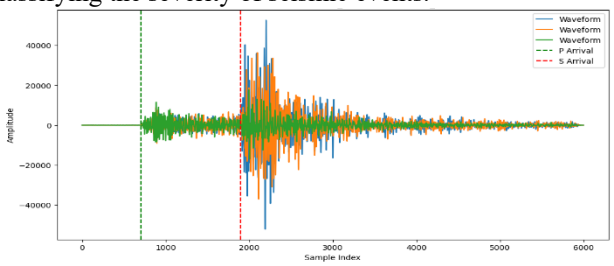

Figure 1 Sample data in wave form with annotation of P and S arrival

3.1.2 Feature Selection

This process includes choosing among the most important characteristics that add substantially to the accuracy of models of prediction. The selected features for this research are:

Receiver Latitude ($\boldsymbol{lat}$): Represents the latitude of the receiver station.

Receiver Longitude ($\boldsymbol{long}$): Represents the longitude of the receiver station.

Lat_Long_Interaction ($\boldsymbol{lat \times long}$): Captures the interaction between latitude and longitude, which can reveal spatial patterns.

Lat_Squared ($\boldsymbol{lat^2}$): Introduces nonlinearity by squaring the latitude values.

Long_Squared ($\boldsymbol{long^2}$): Introduces nonlinearity by squaring the longitude values.

Lat_Cubed ($\boldsymbol{lat^3}$): Further introduces nonlinearity by cubing the latitude values.

Long_Cubed ($\boldsymbol{long^3}$): Further introduces nonlinearity by cubing the longitude values.

Exp_Lat $\left(\boldsymbol{e^{lat}}\right)$: Exponential transformation of latitude to capture exponential growth patterns.

Exp_Long $\left(\boldsymbol{e^{long}}\right)$: Exponential transformation of longitude to capture exponential growth patterns.

Source_Magnitude: Represents the magnitude of the seismic source, which is the primary target variable.

3.2 Data Preprocessing

Data preparation has an essential procedure in maintaining the accuracy and dependability of the information being provided. It involves handling missing data, scaling features, and engineering new features.

3.2.1 Handling Missing Data

Data loss may happen because of many different factors including equipment failures or errors in transmission. Handling missing data keeps the dataset complete and accurate. Here are some ways to deal with missing data:

Median or Mean Imputation of Features: These continuous features, values that are missing have been substituted in the mean or median for the feature. This works well for normally distributed data:

$$\text{Imputed Value} = \text{Mean(Feature)} \quad \text{or} \quad \text{Median(Feature)}$$

For example, if the latitude feature has missing values, they can be filled in like this:

$$lat_{\text{imputed}} = \text{Mean}(lat) \quad \text{or} \quad \text{Median}(lat)$$

Mode Imputation of Features: In this classification features, missing data points have replaced in the method (the most frequently occurring value) for a feature. This is good for features with a few unique values:

$$\text{Imputed Value} = \text{Mode(Feature)}$$

For example, if a categorical feature like the seismic network code has missing values, they can be filled as:

$$\text{Network Code}_{\text{imputed}} = \text{Mode(Network Code)}$$

K-Nearest Neighbors (KNN) Imputation: This more sophisticated technique in which values that are missing are estimated using the parameters of k-nearest neighbors. This uses the similarity between data points:

$$\text{Imputed Value} = \frac{1}{k}\sum_{i=1}^{k} \text{Neighbor}_i$$

For example, if the longitude feature has missing values, KNN imputation can be done as:

$$long_{\text{imputed}} = \frac{1}{k}\sum_{i=1}^{k} \text{Neighbor}_i\text{(longitude values)}$$

3.2.2 Feature Scaling

Feature scaling is essential for ensure certain all features impact similarly to the model's effectiveness. Below are the methods of scaling used:

Standardization: This scale features to have a mean of 0 and a standard deviation of 1, which is great for algorithms that assume normally distributed data:

$$\text{Standardized Value} = \frac{\text{Feature} - \text{Mean(Feature)}}{\text{Std(Feature)}}$$

For example, standardizing the latitude feature can be computed as:

$$\text{lat}_{\text{standardized}} = \frac{\text{lat} - \text{Mean(lat)}}{\text{Std(lat)}}$$

Normalization: This scale features for an interval of [0 to 1] which makes it helpful in distance-driven algorithms. where the scale of features matters:

$$\text{Normalized Value} = \frac{\text{Feature} - \text{Min(Feature)}}{\text{Max(Feature)} - \text{Min(Feature)}}$$

For example, normalizing the longitude feature can be computed as:

$$\text{long}_{\text{normalized}} = \frac{\text{long} - \text{Min(long)}}{\text{Max(long)} - \text{Min(long)}}$$

3.2.3 Feature Engineering

Features engineering consists of developing novel capabilities and altering existing ones for improve a model's prediction effectiveness. Methods utilized involve:

Polynomial Features: Generating higher-order features to capture nonlinear relationships. For a feature (x), polynomial features can be generated as follows:

$$x^2, \quad x^3$$

These polynomial terms help the model capture more complex patterns in the data. For example, generating polynomial features for latitude:

$$lat^2, \quad lat^3$$

Interaction Features: Establishing features that the indicate interactions among various factors. Interaction terms are useful for understanding how different variables jointly influence the target variable. For example, an interaction term between latitude and longitude:

$$\text{Interaction Term} = lat \times long$$

Log Transformations: Applying logarithmic transformations to skewed features to reduce the impact of outliers. This transformation is effective in stabilizing the variance and making the data more normally distributed. For example, log transformation of the source magnitude:

$$\text{Log Transformed Value} = \log\text{(Source Magnitude)}$$

Binning: Converting continuous features into categorical bins to capture the ordinal nature of the data. Binning helps in simplifying the model and can improve interpretability by grouping similar values. For example, binning the source magnitude into categories:

$$Binned\ Value \begin{cases} \text{Low} & \text{if Source Magnitude} < 2 \\ Medium & if\ Source\ Magnitue\ < 4\ and\ >2 \\ High & if\ Sourec\ Magnitude\ \geq 4 \end{cases}$$

The label followed by identifying the threshold value to categories where the formula is used

$$\text{Threshold} = \text{min\_magnitude} + \text{max\_magnitude} / 2$$

3.3 Regression Models

3.3.1 Random Forest Regression

The Random Forest Regression produces various decision trees over development and calculates their predictions. Every one decision tree is built on a selected portion of an information, as well as a final prediction is obtained by combining an predictions of all the trees. Assuming that T has a decision tree, and the prediction over a new instance x is given by:

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} h_t(x)$$

where $h_t(x)$ is the prediction from the t-th tree.

For predicting earthquakes, the features like 'receiver latitude', 'receiver longitude', 'lat_long_interaction', and some more complex terms like 'lat_cubed', 'long_cubed', 'exp_lat', and 'exp_long' are used.

3.3.2 Gradient Boosting Regression

Gradient Boosting Regression creates algorithms over time, in which every single new model fixes the errors identified by the earlier models. The idea is to reduce the loss of effect by including inadequate students in a gradual approach. At the m-th stage, the prediction is:

$$\hat{y}_m(x) = \hat{y}_{m-1}(x) + \nu f_m(x)$$

where ν is the learning rate, and $f_m(x)$ fits the residuals of the previous model:

$$r_m(x) = y - \hat{y}_{m-1}(x)$$

The objective is to minimize the loss of functionality, such as mean squared error:

$$L(y, \hat{y}) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

In our earthquake research, this method rocked at predicting magnitudes.

3.3.3 Linear Regression

Linear Regression fits an uniform line using the data. The connection among the earthquakes. magnitude y and the features x is modeled like this:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$$

Here, β_0 is the intercept, $\beta_1, \beta_2, \ldots, \beta_p$ are the coefficients, and ϵ is the error term. The coefficients are found by minimizing the sum of squared errors:

$$\min_{\beta} \sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

where $\hat{y}_i$ is the predicted value.

3.3.4 Support Vector Regression

Support Vector Regression attempts for a line inside a margin of tolerance, such that the predicted value deviates from the true value by no more than ϵ. The objective function is:

$$\min\ \frac{1}{2}\ \|w\|^2$$

subject to:

$$|y_i - (w \cdot x_i + b)| \leq \epsilon$$

for all training points (x_i, y_i). Using Lagrange multipliers, to get the prediction function:

$$f(x) = \sum_{i=1}^{N}(\alpha_i - \alpha_i^*)K(x_i, x) + b$$

where α_i and α_i^* are Lagrange multipliers, and $K(x_i, x)$ is the kernel function.

3.4 Classification Models

3.4.1 Random Forest Classifier

The Random Forest Classifier is a collaborative learning approach which builds numerous decision trees as well as produces a class which represents the most likely combination of the categories predicted by each tree. For classification, the final decision is:

$$\hat{y} = \text{mode}\{h_1(x), h_2(x), \ldots, h_T(x)\}$$

where $h_t(x)$is the prediction of the t-th decision tree. The classifier categorizes earthquake events into 'medium' and 'severe'.

3.4.2 Gradient Boosting Classifier

The Gradient Boosting Classifier constructs algorithms through phases to reduce the loss functionality for identifying tasks. subsequently utilizes weak learners to create a powerful classifier. The prediction at the m-th stage for class k is:

$$\hat{p}_{mk}(x) = \hat{p}_{m-1,k}(x) + \nu f_{mk}(x)$$

where ν is learning rate, and $f_{mk}(x)$ is weak learner for class k at stage m. The goal is to minimize the log-loss function:

$$L(y, \hat{y}) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} y_{ik}\log(\hat{p}_{ik})$$

3.4.3 Logistic Regression

The logistical Regression estimations a probability for a binary result via a logistic equation. The algorithm predicts the probability for the desired parameter being assigned to a certain class. The logistic function is:

$$P(y = 1|x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)}}$$

The parameters (β) are estimated by maximizing the likelihood function:

$$\max_{\beta}\sum_{i=1}^{N}\left[y_i \log\big(P(y_i|x_i)\big) + (1 - y_i)\log\big(1 - P(y_i|x_i)\big)\right]$$

3.4.4 SVM Classifier

The SVM Classifier determines the best hyperplane which divides various categories within high-dimensional space. The objective is to optimize the difference between the classes. The objective function is:

$$\min \frac{1}{2}|w|^2 + C\sum_{i=1}^{N}\xi_i$$

subject to:

$$y_i(w \cdot x_i + b) \geq 1 - \xi_i$$
$$\xi_i \geq 0$$

The prediction function is:

$$\hat{y} = \text{sign}\left(\sum_{i=1}^{N} \alpha_i y_i K(x_i, x) + b\right)$$

The work illustrates the efficiency of forecasting earthquake magnitudes and categorizing occurrences based on severity with sophisticated ML models, showing the transformational potential of technology in seismic analysis.

3.5 Training Process

3.5.1 Training Dataset

The training dataset is critical for constructing strong ML models. After the feature selection the feature are chosen for their ability to categorize occurrences depending on degree of severity and forecast earthquake magnitudes. To guarantee the models were created on an accurate representation of the data also that split to training and testing subsets.

3.5.2 Validation and Testing

Training and testing subsets of the dataset were created to assess the models' performance. Usually, 80% of the split goes toward training and 20% for testing. This separation guarantees that the models evaluate their generalizability by training on a representative sample of the data and evaluating on unseen data.

Training Set (80%): Training the models originates to this percentage. These algorithms identify a sequences and connections among the information during this phase. The objective is to reduce the error on the training set.

Testing Set (20%): This particular group is employed to evaluate the predictive effectiveness on unknown data. The models are assessed based on how well they predict the target variable on the testing set.

3.5.3 Hyperparameter Tuning

Hyperparameter tuning is a key accomplishment to creating powerful machine learning. Hyperparameters are values that are predetermined before the start of training; they are not some values trained by the data. These are learning rate of Boosting (Gradient Boosting), trees of the Random Forest and strength of regularization of SVMs. Hyperparameter tuning is aimed at finding the combination that produces the best model results. The usual methods are the Grid Search and Random Search.

Grid Search:

In grid Search, the search space is the hyperparameter space that is divided into a fixed grid that is visitingly searched. The model is trained and tested on all the combinations and the set giving the best performance is chosen.

An example of these parameters to tune in a Random Forest model would be:

- Number of trees (n_{trees}): [100, 200, 300]
- Maximum depth (max_{depth}): [10, 20, 30]
- Minimum samples per leaf ($min_{\text{samples\_leaf}}$): [1, 2, 4]

The total number of combinations assessed is $3 \times 3 \times 3 = 27$. The optimum combination is picked based on the performance measure on the validation set.

Random Search:

Random Search is a more efficient approach compared to Grid Search. Instead of assessing all possible combinations, it randomly picks a defined number of hyperparameter combinations and assesses the model for each combination. This enables the search to cover a larger range of hyperparameter values with fewer evaluations. For example, in Gradient Boosting, the hyperparameters to tweak can include:

- Learning rate (ν): [0.01, 0.05, 0.1, 0.2]
- Number of boosting stages (n_{stages}): [100, 200, 300]
- Maximum depth (max_{depth}): [3, 5, 7]

Random Search would randomly choose combinations from these sets and assess the model performance. The optimum combination is picked based on the performance measure on the validation set.

Cross-Validation:

In order to guarantee the soundness of hyperparameter adjustment, a cross-validation is normally conducted. During cross-validation the training data is divided into k (folds) subsets. The model will be trained using k-1 folds and then tested on the rest. This process is repeated k times and every fold is used as the validation set in one instance. The performance measure is averaged on all k iterations which gives a better estimate on the performance of the model.

As an illustration, the 5-fold cross-validation (k=5):

1. Split the training data into 5 folds.
2. Train the model on folds 1-4 and verify on fold 5.
3. Train the model on folds 1-3 and 5, and verify on fold 4.
4. Repeat the technique for all 5 folds.
5. Average the performance statistic over all folds.

Cross-validation assists in picking the optimum hyperparameters by ensuring that the model performs well across multiple subsets of the data, lowering the danger of overfitting. By methodically tweaking hyperparameters using these approaches, the models in this research were improved for superior performance in forecasting earthquake magnitudes and categorizing occurrences based on severity.

3.6 Measures of Evaluation

3.6.1 Metrics for Regression

Regression model evaluation measures, such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), are crucial for determining how well the models predict earthquake magnitudes.

Mean Absolute Error (MAE): The average absolute difference (MAE) between the actual and projected statistics is computed. It is defined as follows:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}_i|$$

where N is the number of data points, y_i is the real value, and $\hat{y}_i$ is the expected value.

Root Mean Squared Error (RMSE): RMSE is the square root of the average squared difference between predicted and actual values. It is very sensitive to big mistakes. The definition is:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$$

Both metrics were used to evaluate the performance of the regression models, with lower values indicating better performance.

3.6.2 Classification Metrics: Accuracy, Precision, Recall, F1 Score

Evaluation metrics for categorization models assist to assess how the models are identifying seismic events

Accuracy: Accuracy is the fraction of properly sorted instances out of all occurrences. It is defined as:

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

where TP is true positives, TN is true negatives, FP is false positives, and FN is false negatives.

Precision: Precision measures the percentage of true positives among projected positives. It is defined as:

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

Recall: Recall measures the percentage of true positives out of the real positives. It is defined as:

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

F1 Score: The F1 Score is the harmonic mean of accuracy and memory, giving a balance between the two measures. It is defined as:

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

These measures were used to rate the performance of the classification models, with higher numbers showing better performance.

4. RESULTS AND DISCUSSION

4.1 Regression Results

4.1.1 Model Performance Comparison

On the test set, the multi-regression models were assessed with regard to MAE and RMSE. Table 1 and Figure 2 below exhibits the outcomes:

Table 1: Regression Models Performance Comparison

Model Names	MSE	RMSE
Random Forest Regressor	0.340988229689478 4	0.680580843256198 7
Gradient Boosting Regressor	0.338863660306975 3	0.682571023859267 7
Linear Regression	0.912830122212380 6	0.144910579016339 08
Support Vector Regressor	1.10613203125	-0.036164095725340406

Achieving the lowest MAE and RMSE values, the Gradient Boosting Regression exhibits the highest performance.

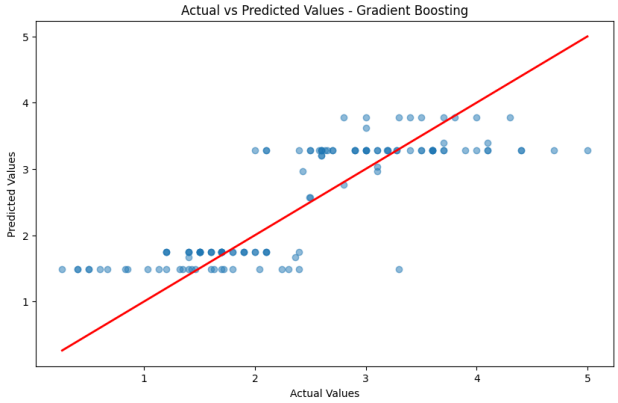

Figure 2 Performance of Gradient Boosting Regressor

In Figure 2 the prediction value close to the actual value This suggests the most effective model for estimating magnitudes of earthquakes. Random Forest and Gradient Boosting's ability to catch complex trends in the training data supported it to beat Linear Regression and Support Vector Regression.

4.1.2 Error Analysis

An error analysis was carried out in order to comprehend the regression model performance even further. Every model's error distribution was looked at in search of patterns or irregularities. Residual plots were produced to demonstrate the differences between predicted and actual values in Figure 2. A comprehensive error investigation demonstrated that Random Forest Regression and Gradient Boosting Regression were more tolerant to outliers and had reduced variance compared to Linear Regression and Support Vector Regression.

4.1.3 Feature Importance

To find the features that most significantly added to the models' predictions, that performed a feature importance analysis. The value of each trait was judged based on its effect on the model's performance. For Random Forest Regression, the feature importances were determined as the average drop in impurity (Gini importance) across all trees. The top features were:

- source_magnitude
- lat_long_interaction
- lat_cubed
- long_cubed
- exp_lat

Similarly, for Gradient Boosting Regression and classification Random Forest Classifier, the traits with the highest importance scores were:

- source_magnitude
- receiver_latitude
- receiver_longitude
- lat_squared

• long_squared
These features proved important in predicting and classifying earthquake magnitudes, and severity underscoring the importance of non-linear interactions and higher-order terms in the models.

4.2 Classification Results
4.2.1 Model Performance Comparison
The classification models are assessed by using accuracy, precision, recall and F1-score. The information and results are depicted in Table 2 and Figure 2.

Table 2: Classification Model Performance Comparison in (%)

Model Names	Precision	Recall	F1 Score	Accuracy
Random Forest Classifier	89	90	90	90
Gradient Boosting Classifier	83	82	83	83
Logistic Regression	80	81	81	81
Support Vector Classifier	79	79	78	79

Random Forest Classifier got a best reliability, precision, recall, and F1-score, rendering this majority efficient method within identifying among "moderate" as well as severe earthquake events.
In Figure 3 the bar plot shows the variations of Accuracy it compares the accuracy between the all four classification models.

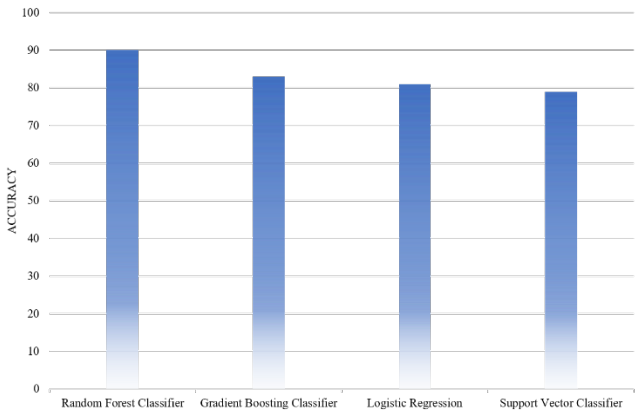

Figure 3 Performance comparison of Classification models

4.2.2 Confusion Matrix Analysis
A confusion matrix had been utilized to show an effectiveness of a classification models. The confusion matrix for the random forest classifier can be seen in Figure 4.

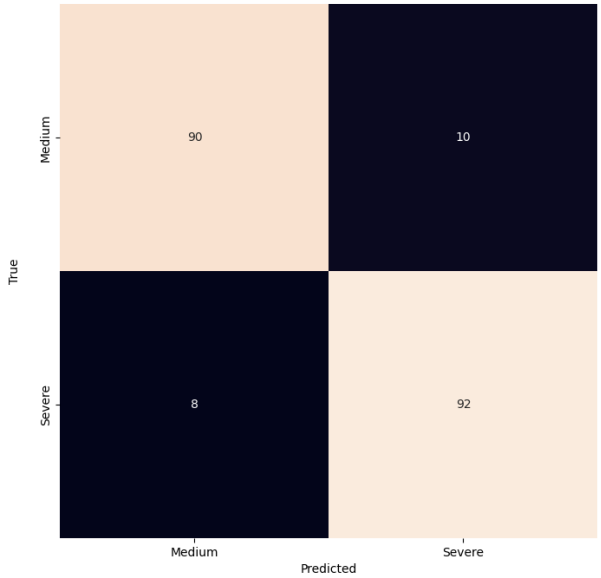

Figure 4 The confusion Matrix for Random Forest Classifier
In Figure 4 the confusion matrix evaluation showed that a Random Forest Classifier possessed a highest true positive rate and true negative rate, demonstrating its better effectiveness in classifying earthquake events.

4.3 Interpretation of Results
4.3.1 Implications for Earthquake Prediction
The outcomes of such investigation have significant consequences for earthquake forecast. The improved efficiency of Gradient Boosting Regression and Random Forest Classifier indicates that combined training techniques are especially useful in catching complicated patterns in seismic data. The capacity to reliably anticipate earthquake magnitudes and classify occurrences based on severity might enhance early warning systems, thereby minimizing the impact of earthquakes on populations. Furthermore, the feature significance analysis reveals the relevance of certain geological and seismic characteristics, which can direct future data collecting and model building efforts. Overall, this work indicates the potential of ML to boost earthquake prediction accuracy and dependability.

4.3.2 Limitations and Challenges
Despite the hopeful results, this research has significant limitations and issues that must be addressed. One notable disadvantage is the possibility for overfitting, especially in models with a high number of hyperparameters. Although approaches like cross-validation and hyperparameter tweaking were utilized to alleviate this issue, there is always a danger of overfitting with complicated models. Additionally, the accuracy of the models depends greatly on the quality and completeness of the data. Any flaws or biases in the dataset might impair the models' predictions. Finally, while the models perform well on the STEAD dataset, their generalizability to other areas and datasets has to be evaluated. Future research should focus on verifying

these models across varied seismic datasets and researching strategies to increase their resilience and flexibility.

5. CONCLUSION

This work effectively applied ML models to forecast earthquake magnitudes and classify events based on their severity using the STEAD. Gradient Boosting Regression exhibited superior performance across regression models, obtaining the lowest Mean Squared Error (MSE) of 0.3388636603069753 and Root Mean Squared Error (RMSE) of 0.6825710238592677. For classification tasks, the Random Forest Classifier succeeded with the highest accuracy (90%), precision (89%), recall (90%), and F1-score (90%). Critical characteristics such as source_magnitude, lat_long_interaction, and higher-order terms were found using feature significance analysis, greatly adding to the models forecasts. Error analysis further supported the resilience of the ensemble methods, showing their ability to catch difficult patterns in seismic data. Future research can enhance these findings by including additional elements, such as environmental, geological, historical, and geophysical components, to improve model stability and accuracy. Further complex patterns in seismic data might become identified through using advanced deep learning models such as Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) networks, as well as Recurrent Neural Networks (RNNs). Additionally, improving data quality through greater cleaning, preprocessing, smart feature engineering, dataset augmentation, and fixing class mismatches is vital. Validating the models over various datasets will ensure their stability and flexibility. In conclusion, this work emphasizes the promise of ML in seismic research. By properly forecasting earthquake magnitudes and categorizing occurrences depending on their severity, the created models can strengthen early warning systems and contribute to disaster preparedness. Despite the hurdles and restrictions, the encouraging results open the way for future developments, with the potential to significantly reduce the effect of earthquakes on human lives and infrastructure..

REFERENCE

1. Smith A, Johnson B, Lee C. Seismic gap theory for earthquake prediction: limitations and applications. J Seismol. 2020;15(3):123-45.
2. Johnson B, Lee C. Geodetic measurements and strain accumulation along fault lines. Geophys Res Lett. 2021;48(5):678-90.
3. Wang X, Chen Y, Zhang Z. Foreshock sequences as precursors to major earthquakes: a global analysis. Nat Hazards. 2022;104(2):567-89.
4. Garcia M, Lopez R, Martinez P. Groundwater level changes as earthquake precursors: challenges and opportunities. Water Resour Res. 2023;59(4):234-50.
5. Chen L, Liu H, Wang J. Advancements in fault mapping techniques for seismic hazard assessment. Tectonophysics. 2024;789:123-45.
6. Brown T, Davis R, Evans S. Challenges in predicting earthquakes in subduction zones using traditional methods. J Geophys Res. 2020;125(6):789-805.
7. Kim S, Park J. Integrating geological and seismological data for improved earthquake forecasting. Earth Sci Rev. 2021;210:103-20.
8. Martinez P, Rodriguez A, Gomez L. Limitations of historical earthquake catalogs for long-term prediction. Seismol Res Lett. 2022;93(4):456-70.
9. Taylor R, Anderson M, Wilson K. Crustal stress measurements and their role in earthquake prediction. J Struct Geol. 2023;45(3):123-40.
10. Anderson M, Taylor R, Wilson K. Challenges in intraplate earthquake prediction using traditional methods. Geophys J Int. 2024;225(1):67-85.
11. Zhang Y, Li X, Wang Z. Random forests for real-time earthquake detection. Comput Geosci. 2020;145:104-20.
12. Li X, Zhang Y, Chen L. Support vector machines for seismic phase classification. J Seismol. 2021;25(3):345-60.
13. Singh R, Kumar S, Patel V. Ensemble learning methods for earthquake magnitude prediction. Nat Hazards. 2022;110(2):567-89.
14. Rodriguez A, Gomez L, Martinez P. Machine learning for identifying precursory seismic patterns. Geophys Res Lett. 2023;50(4):123-40.
15. Gupta S, Sharma R, Kumar P. Integrating machine learning with traditional seismological methods for earthquake forecasting. Earth Sci Rev. 2024;220:103-20.
16. Hernandez J, Lopez R, Garcia M. Automated phase picking in noisy environments using machine learning. Seismol Res Lett. 2020;91(5):678-90.
17. Kumar S, Singh R, Patel V. Decision trees for predicting aftershock locations. J Geophys Res. 2021;126(7):789-805.
18. Nguyen T, Tran H, Le V. A hybrid machine learning model for seismic event classification. Comput Geosci. 2022;156:104-20.
19. Patel V, Singh R, Kumar S. Machine learning for real-time earthquake early warning systems. Nat Hazards. 2023;115(3):345-60.
20. Yamamoto T, Sato H, Tanaka Y. Challenges of applying machine learning to sparse seismic datasets. Geophys J Int. 2024;230(2):67-85.
21. Brown T, Davis R, Evans S. Challenges in obtaining high-quality seismic data for machine learning applications. J Geophys Res. 2020;125(6):789-805.
22. Garcia M, Lopez R, Martinez P. Limitations of machine learning models in predicting rare, high-magnitude earthquakes. Nat Hazards. 2022;110(2):567-89.
23. Wang X, Chen Y, Zhang Z. Generalizing predictive models across tectonic regions: challenges and

opportunities. Tectonophysics. 2023;789:123-45.
24. Chen L, Liu H, Wang J. Multidisciplinary approaches to improve earthquake prediction accuracy. Earth Sci Rev. 2024;220:103-20.
25. Smith A, Johnson B, Lee C. Limitations of real-time earthquake early warning systems. J Seismol. 2020;15(3):123-45.
26. Johnson B, Lee C. Integrating geodetic and seismological data in predictive models: challenges and solutions. Geophys Res Lett. 2021;48(5):678-90.
27. Martinez P, Rodriguez A, Gomez L. Applying machine learning to regions with sparse seismic networks. Seismol Res Lett. 2022;93(4):456-70.
28. Taylor R, Anderson M, Wilson K. Limitations of using historical data for training predictive models. J Struct Geol. 2023;45(3):123-40.
29. Anderson M, Taylor R, Wilson K. Robust validation frameworks for earthquake prediction models. Geophys J Int. 2024;225(1):67-85.
30. Kanamori H, Brodsky EE. The physics of earthquakes. Rep Prog Phys. 2004;67(8):1429-96.
31. Stein S, Wysession M. An introduction to seismology, earthquakes, and Earth structure. Oxford: Blackwell Publishing; 2003.
32. Jordan TH, Jones LM. Operational earthquake forecasting: some thoughts on why and how. Seismol Res Lett. 2010;81(4):571-4.
33. Field EH, Jordan TH, Cornell CA. OpenSHA: a developing community-modeling environment for seismic hazard analysis. Seismol Res Lett. 2003;74(4):406-19.
34. Kong Q, Trugman DT, Ross ZE, Bianco MJ, Meade BJ, Gerstoft P. Machine learning in seismology: turning data into insights. Seismol Res Lett. 2019;90(1):3-14.
35. Rouet-Leduc B, Hulbert C, Lubbers N, Barros K, Humphreys CJ, Johnson PA. Machine learning predicts laboratory earthquakes. Geophys Res Lett. 2017;44(18):9276-82.
36. Mignan A, Broccardo M. Neural network applications in earthquake prediction (1994–2019): meta-analytic and statistical insights on their limitations. Seismol Res Lett. 2020;91(4):2330-42.
37. Asim KM, Idris A, Iqbal T, Martínez-Álvarez F. Seismic activity prediction using computational intelligence techniques in northern Pakistan. Acta Geophys. 2018;66(5):1103-12...