LightGBM-Driven Earthquake Magnitude Prediction: A Comparative Machine Learning Framework Using Global Seismic Data

Nima Khodadadi[1,*]
[1]Department of Civil and Architectural Engineering, University of Miami, Coral Gables, FL, USA
Emails: nima.khodadadi@miami.edu

Abstract

Earthquakes represent one of the most destructive natural hazards because they cause consequential destruction to entire communities and fatal consequences for people. Research has continued for decades because scientists aim to develop better forecasting tools for seismic events, which unpredictably strike society with massive economic losses. Research methods from classical earthquake science and statistical and physical earthquake models do not effectively demonstrate earthquake data's complex spatial and temporal characteristics. ML methods generated widespread interest in prediction work because they extract understanding from extensive data collections to produce accurate results independently of physical rules. The presented work examines various ML models that predict earthquake magnitudes by assessing an open-access global earthquake dataset from 2023. The evaluation consists of five predictive models, including Light Gradient Boosting Machine (LightGBM) and Support Vector Regression (SVR), as well as k-nearest Neighbors (KNN), Ridge Regression, along Extra Trees Regressor. The training process included stratified cross-validation and model optimization of hyperparameters for every model. The assessment included a mixture of statistical and mathematical performance indicators that measured Mean Squared Error (MSE) alongside Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Bias Error (MBE), Coefficient of Determination (R^2), Nash–Sutcliffe Efficiency (NSE), Willmott Index (WI), Pearson's Correlation Coefficient (r) and Relative Root Mean Squared Error (RRMSE). LightGBM outperformed all evaluation models by attaining a minimum MSE value of 0.0474 and a R^2 score of 0.9241. LightGBM's leaf-wise tree-building approach, robust scalability, and native regularization features enabled it to apply very well to unknown data samples without reducing computational speed. The experimental outcomes validate LightGBM as a powerful tool for recognizing delicate patterns within high-dimensional seismic data collections for potential use as a predictive modeling instrument in earthquake-prone zones. ML-based forecasting systems have displayed the capability to change earthquake prediction processes according to research outcomes. When used together, LightGBM and alternative advanced ML systems enhance real-time early warning systems, which leads to shortened emergency response time bet, better planning decisions, and lower numbers of human and economic losses from earthquakes. This approach, along with open-access datasets, allows the goal of seismic risk mitigation to achieve broader transparency and collaborative innovation through reproducible modeling strategies.

Keywords: Earthquake prediction; Seismic Data Analysis; Light Gradient Boosting Machine; Machine Learning Models

1 Introduction

Earthquakes remain among nature's most unpredictable destructive forces that seriously endanger human survival, infrastructure systems, and economic operations. Earthquakes strike suddenly while releasing enormous amounts of energy, which creates especially hazardous situations leading to destructive consequences, including fatalities and building destruction paired with extensive financial constraints to the economy. Seismic

DOI: https://doi.org/10.54216/JAIM.090205
Received: January 14, 2025 Revised: March 18, 2025 Accepted: May 10, 2025

disasters around the world have resulted in massive economic losses combined with hundreds of thousands of killed people throughout the past fifty years. Advanced forecasting methods are necessary for improving disaster preparedness because of extensive damage seen during events such as the 2004 Indian Ocean tsunami, the 2011 earthquake in Japan, and the 2023 Turkey-Syria earthquake [1, 2]. Research on earthquake predictive modeling has become essential because scientists now examine earthquakes as global-scale phenomena. The societal need for early warning systems increased exponentially as urban areas continued expanding into zones threatened by tectonic activity. Accurate predictions of magnitude combined with depth and peak ground acceleration (PGA) determine which emergency decisions must be made to protect citizens and infrastructure. Forecast data accuracy allows emergency responders to be deployed quickly and nuclear plants alongside transport systems to be shut down before accidents occur while vulnerable groups get alerted. Accurate modeling helps governments plan city development sustainably while establishing essential priorities for infrastructure increases and enhancing logistical efficiency following disasters, strengthening societal resistance [3]. Earthquake forecasting methods in the past identified their success in physical models combined with statistical seismology and historical data analysis. The beneficial contributions of these approaches to seismic behavior understanding stem from their background research but insufficient ability to monitor the nonlinear complex spatial and temporal earthquake dynamics patterns. The accuracy of classical forecasting methods is restricted because seismic signals show irregular patterns and earthquakes have various origins, while real-time data access is inadequate for high-definition monitoring. The introduction of machine learning (ML) changed seismic analysis because it enables models to extract complicated data-based representations of seismic motion along with no predefined assumptions or physical constraints. ML techniques show superior performance when they detect hidden relationships within data. They adjust their patterns to numerous data kinds and deploy predictive power across many geologic zones and seismic situations [4, 5]. The promising nature of ML does not eliminate the significant hurdles in using these methods for earthquake prediction purposes. The primary obstacle to tackle is the excessive number of dimensions. Such datasets contain many features that incorporate geospatial data, seismic wave characteristics, distances to fault lines, and environmental parameters. Many features in the system result in greater demands for computational resources, elevated memory storage requirements, and model overfitting, which minimizes generalization potential for new events. The dimensionality curse creates issues because it needs specific methods for choosing relevant features with appropriate dimension reduction techniques and optimized algorithms to preserve model accuracy and efficiency [6]. Performance typically degrades when multiple features repeat information through redundant features because strict data validation techniques are required to handle this issue [7]. A leading complex issue consists of executing hyperparameters correctly. The performance of machine learning models depends significantly on adjustable learning rates, tree depths, regularization terms, and kernel configurations. The best model generalization depends on these essential model parameters since they control the learning process of data patterns while revealing under or overfit results. The success of automated hyperparameter optimization techniques, including randomized search and Bayesian optimization, becomes essential because they help find the best model performances in varying configurations and data split setups [8]. The ability to reach high generalization across multiple datasets and under noisy conditions along with limited ground-truth labels remains an active challenge for ML-based seismic forecasting [9]. The main goal of this research is to build a reliable data-based prediction system for earthquake attributes through analysis of an extensive worldwide earthquake database with open access. This research evaluates advanced machine learning approaches to determine their performance capabilities regarding prediction accuracy com, rotational speed, and time-based and geographic model generalization. Light Gradient Boosting Machine (LightGBM) emerges as an efficient framework that provides exceptional capabilities to process extensive tabular data. LightGBM stands out due to its capabilities to manage missing data points while executing fast training procedures that leverage histogram-based decision tree learning for optimized feature selection [10]. Randomized hyperparameter tuning through cross-validation optimizes model learning with the benefit of minimizing overfitting, which enables the selection of best configurations that maximize accuracy and operational feasibility [11]. This research systematically analyzes four machine learning algorithms where Support Vector Regression (SVR) matches k-nearest Neighbors (KNN) and Ridge Regression with the Extra Trees ensemble method [12]. The researchers evaluate each algorithm by applying standardized evaluation metrics comprising Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE) alongside the coefficient of determination (R^2). This evaluation facilitates complete comprehension of how different models perform under different seismic prediction conditions. Data preprocessing, feature selection, and model tuning techniques substantially influence prediction performance and reliability through the comparative approach [13]. This research contributes primarily by constructing an open-access predictive framework that researchers can reproduce. This research fulfills its mission by utilizing public seismic dataset information while documenting modeling steps that help progress open science practices within earthquake science. Researchers and policy-making institutions can effortlessly transform the developed model framework into operational tools for regional prediction studies, stress anomaly detection, and real-time early warning systems. In summary, this work delivers four primary contributions: (i) the development of an optimized LightGBM-based predictive model enhanced through randomized hyperparameter tuning; (ii) systematic comparative analysis of multiple machine learning algorithms applied to earthquake prediction; (iii) insights into the role of dimensionality, feature redundancy, and generalization in seismic data modeling; and (iv) the promotion of open-access data and transparent modeling practices to support collaborative research and informed decision-making in seismic risk management. This study enhances scientific comprehension of earthquake forecasting through its findings while enabling better strategies to deal with seismic hazards in a world increasingly exposed to risk.

2 Literature Review

Scientists focus on earthquake prediction as a vital field of disaster prevention because earthquakes occur abruptly with severe effects. Earthquake unpredictability and its harmful effects on humans and infrastructure require effective real-time predictive techniques for producing dependable results. Machine learning is a leading technological instrument to solve problems related to earthquake prediction. This research looked into how the application of Random Forest Regressor and Neural Networks models performed for earthquake magnitude and depth prediction. Historical seismic activity and location-based features and timestamps went into training algorithms that proved the effectiveness of data-centric methodologies for seismic hazard reduction according to [14]. By combining statistical modeling with features from geographic information systems, this system detects patterns that traditional seismological analysis methods cannot detect. The authors applied transformer architecture in deep learning to predict earthquake magnitudes in the Horn of Africa region. Tests conducted on a benchmark model demonstrated that transformers outperformed LSTM, BiLSTM, and its attention counterpart, BiLSTM-AT. The transformer demonstrated superior performance by generating the lowest metric values of mean absolute error (MAE) as well as mean squared error (MSE), root mean squared error (RMSE), and mean absolute percentage error (MAPE) while showing its capability to capture complex temporal and spatial relationships in seismic data [15]. Multiple research investigations propose using joint systems that merge geological domain knowledge with ML methods. Researchers employed Random Forest algorithms connected to geological models to predict earthquakes in the short-term duration along the Pacific Ring of Fire in their study. The study demonstrated that pure machine learning and geological techniques cannot deliver adequate forecasting results separately. The integrated model combining skills from both domains produced better predictive accuracy together with high reliability [16]. Research investigators established two ML hybrid models for Southern California analysis by integrating FPA-ELM and FPA-LS-SVM. The research merged the Flower Pollination Algorithm (FPA) with Extreme Learning Machines (ELM) as well as FPA with Least Squares Support Vector Machines (LS-SVM). Seven seismic indicators allowed the predictive models to forecast seismic events up to fifteen days in advance. An FPA-LS-SVM model demonstrated better performance than alternatives when measuring RMSE, MAE, SMAPE and PMRE and proved hybrid evolutionary algorithms beneficial in seismic forecasting [17]. The breakthrough arrives from the SeisEML (Seismological Ensemble Machine Learning) model. The PGA prediction system SeisEML combines multiple hybrid models with kernel-based algorithms and regression trees for its design. The model successfully processed over 20,000 Japanese accelerograms with outstanding prediction capabilities, outperforming traditional attenuation models. Studies revealed that applying the model to Iranian earthquake data and other geographical areas confirmed its usefulness in creating seismic hazard maps across different regions [18]. The model used Bayesian ML with wavelet transform analysis to identify seismic cycles in five primary worldwide zones. Researchers inspected earthquake patterns from 1900 to 2021 that presented periodicity from 3.7 years to 55 years across different zones. The forecasting model indicates periods with high earthquake potential from Japan's 2024 to 2026 and South America's 2026 to 2031. The model provided a retrospective view of why the California Parkfield earthquake occurred late by explaining the anomaly [19]. The research investigated earthquake severity level classification through seven machine learning approaches, including Random Forest Bayes Net and Logistic Regression. The Simple Logistic and Logistic Model Tree (LMT) classifiers provided the best results when analyzing six Indian earthquake datasets and regional data. The classification system enables detailed planning resources in territories affected by earthquakes [20]. GPS satellite Total Electron Content (TEC) data has been researched to study ionosphere changes that could precede earthquakes. Deep learning models based on LSTM analysis TEC measurement data from past days to determine whether an earthquake would occur. The ionospheric anomalies proved useful as earthquake precursors when used with the LSTM model because it reached a 0.82 accuracy rate, which exceeded SVM, LDA, and Random Forest classifiers [21].

DOI: https://doi.org/10.54216/JAIM.090205
Received: January 14, 2025 Revised: March 18, 2025 Accepted: May 10, 2025

Science analysts used Himalayan earthquake data to apply artificial neural networks (ANNs) and Long-Short Term Memory (LSTM) networks for earthquake detection predictions. Results proved acceptable for up to medium-scale seismic events, but deep learning approaches demonstrated difficulty detecting high-magnitude earthquakes according to the research findings presented in [22]. A research study in Iran utilized Fault Density as a new variable computed using Kernel Density Estimation and Bivariate Moran's I methods. SVM, Decision Tree, Shallow Neural Network, and Deep Neural Network served as the models under evaluation in this study. Spatial parameters in SVM and DNN models brought about substantial performance enhancements for large earthquake magnitude patterns according to the authors' research results [23]. The earthquake-induced landslides (EQILs) prediction technique benefited from high-resolution satellite imagery and digital elevation models. The extracted abstract features through stacked autoencoder (SAE) achieved better results than regular models, including logistic regression and Random Forest. Research proved the necessity of deep learning algorithms to detect essential spatial signatures that help forecast earthquake-induced landslides [24]. Scientists have developed Inverse Boosting Pruning Trees (IBPT), which makes satellite data-based earthquake predictions possible. The multivariate model analyzed infrared and hyperspectral data from 1371 powerful earthquakes, which proved more effective than six alternative machine learning implementations. The results show how remote sensing technologies work best with sophisticated ensemble methods to achieve accurate short-term predictions according to [25]. A comprehensive study of 2,271 documents led to an analysis of 98 research papers to determine eight fundamental categories of ML applications in earthquake prevention efforts. The research focused on anti-seismic structures, forecasting methods, emergency response techniques, multimodal data processing and quality management. The review documented enhanced proficiency in ML applications, which branch out throughout the complete sequence of earthquake disaster management processes [26]. A research-based in Indonesia investigated ways to enhance Earthquake Early Warning Systems (EEWS). The research used raw seismogram data that lasted four seconds to conduct training for both regression and classification deep neural networks (DNN). Better pattern recognition occurred because the classification DNN led to lower final loss than Random Forest models in data processing. This makes it useful for real-time magnitude estimation, as reported in [27]. The research demonstrates that machine learning and hybrid methods will boost earthquake prediction accuracy. Deep neural architectures, ionospheric anomalies, remote sensing elements, spatial feature analysis, social seismological knowledge, and ensemble modeling form prospects for seismic risk reduction research.

3 Dataset

3.1 Dataset Description

This research uses Earthquakes 2023 Global as its dataset, which contains a detailed collection of global seismic occurrences during the year 2023 [28]. This dataset delivers essential information about earthquake spatiotemporal and geophysical features 2023 through an open-access seismological database. Through this database, researchers can perform studies about seismic safety assessments linked with geographical analysis and use machine learning models for prediction. There are 26,642 earthquake events organized into distinct rows with 22 descriptive columns for each record. Each seismic incident contains a broad spectrum of 22 parameters across the 26,642 rows, which makes this database both expansive and deep. The parameters within the dataset can be grouped into temporal elements alongside geospatial information, seismological numbers and quantified measurement errors. Specifically, the dataset includes:

- **Time**: The timestamp of the earthquake event.
- **Latitude**, **Longitude**: Geographic coordinates specifying the epicenter.
- **Depth**: The depth at which the earthquake originated, measured in kilometers.
- **Mag**: The magnitude of the earthquake, which serves as the primary **target variable** in this study.
- **MagType**: Type of magnitude scale used (e.g., Mw, ML).
- **Nst**: Number of seismic stations that recorded the event.
- **Gap**: Azimuthal gap between stations, indicating sensor distribution quality.

Received: January 14, 2025 Revised: March 18, 2025 Accepted: May 10, 2025

- **Dmin**: Minimum distance between the earthquake epicenter and the nearest station.
- **Rms**: Root Mean Square of the earthquake's amplitude spectrum.
- **Net**, **Id**, **Updated**, **Place**, **Type**: Meta-descriptive columns providing event identifiers and contextual descriptions.
- **HorizontalError**, **DepthError**, **MagError**: Estimation errors in the location, depth, and magnitude respectively.
- **MagNst**: Number of stations used specifically to calculate the magnitude.
- **Status**, **LocationSource**, **MagSource**: Information regarding review status and data origin.

The multiple analysis approaches result from diverse and detailed features. By having geospatial and temporal features, researchers can perform clustering analysis on seismic events create spatial and temporal maps, and build predictive models using measurements of magnitude and depth. A study was performed to determine the accuracy and reliability of dataset measurements in addition to key measurement precision. Figure 1 displays the correlation matrix demonstrating connections between the error variables and their relationship to both main features (magnitude and depth). The horizontal error measurement demonstrates a moderate correlation between how intense events are and how deep they occur below the surface.

Figure 1: Correlation Matrix of Measurement Errors

A review of measurement error distributions helped quantify data accuracy levels and variability. The measurement reliability for depth reaches high standards based on Figure 2, showing positive skewness that centers around low error values.

DOI: https://doi.org/10.54216/JAIM.090205

Received: January 14, 2025 Revised: March 18, 2025 Accepted: May 10, 2025

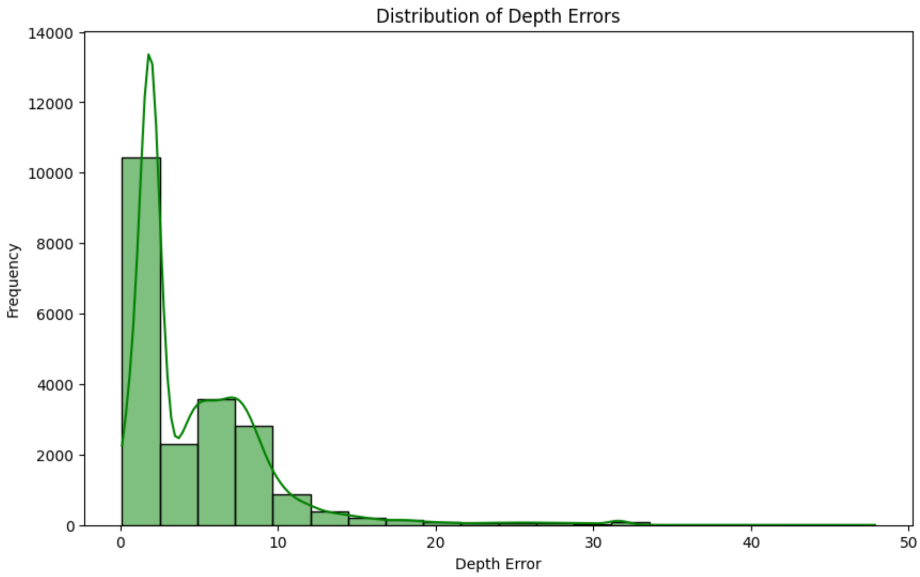

Figure 2: Distribution of Depth Errors

The majority of magnitude estimates shown in Figure 3 fit inside a restricted error range, thus strengthening the dataset's suitability as a regression target.

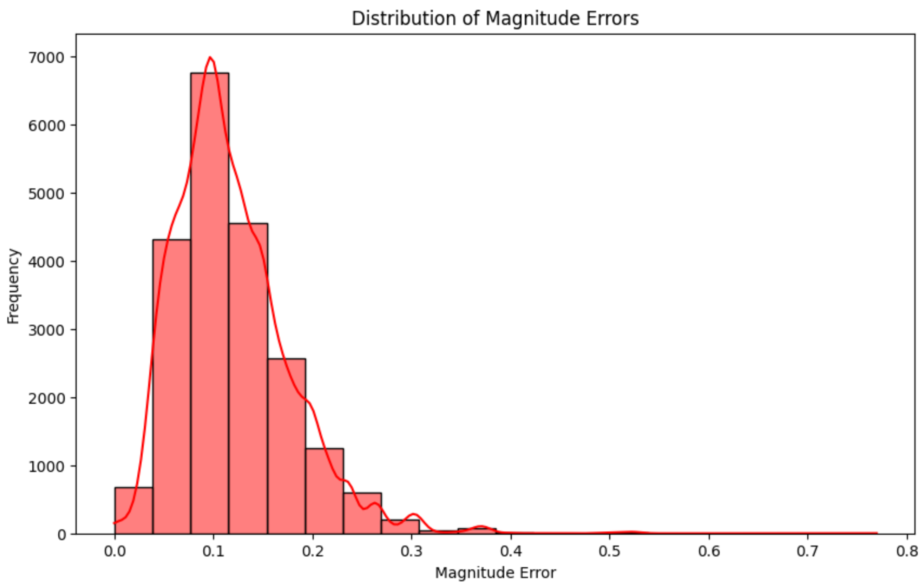

Figure 3: Distribution of Magnitude Errors

The figure identified as Figure 4 provides information about the distribution of HorizontalError. Most locations fall within permitted error zones, yet their dispersion pattern exhibits wider ranges than other measurement errors, which sensor distribution patterns or geographical terrain conditions could cause.

DOI: https://doi.org/10.54216/JAIM.090205

Received: January 14, 2025 Revised: March 18, 2025 Accepted: May 10, 2025

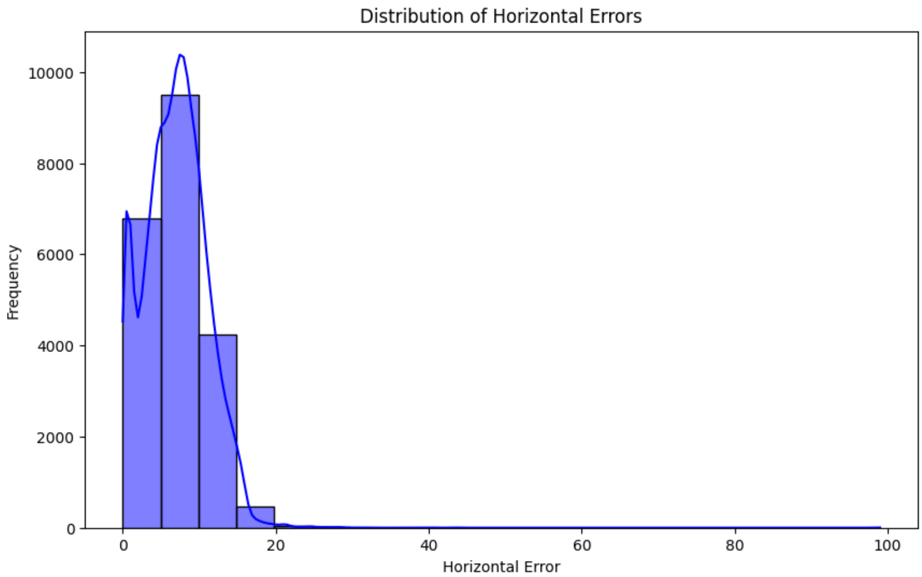

Figure 4: Distribution of Horizontal Errors

Visual representations and descriptive elements confirm the analytical strength of Earthquakes 2023 Global dataset data. The database forms an essential groundwork for creating dependable earthquake magnitude and depth prediction models and facilitating precise statistical tests with error measurement capabilities.

3.2 Data Preprocessing

A comprehensive series of data preprocessing methods was executed first on the Earthquakes 2023 Global dataset before moving into the modeling stage. The developed procedures include methods for treating missing values and anomaly elimination, encoding approaches, scale normalization, and temporal data modification techniques [29]. **Handling Missing Values** The dataset underwent an evaluation process to determine if it contained empty values. Multiple numerical features, MagError, DepthError, HorizontalError and MagNst, showed missing data. The median imputation method was used on these fields because it proved robust to outliers, and categorical fields received mode-imputation, or their data was discarded if sparsity made information useless. **Outlier Detection and Removal** An interquartile range (IQR) analysis and boxplot inspection improved the reliability of subsequent analyses. Records of earthquakes with impossible values, such as negative depths or magnitudes below 1.0, were excluded to isolate seismically essential data points. **Categorical Encoding and Field Reduction** We applied a one-hot encoding technique to convert the categorical features MagType, Type and Status into numerical values since it enables machine learning models to work effectively. The analysis removed several non-predictive attributes consisting of Id, Updated and Net because they served as unique identifiers while providing no explanatory value to the model. **Feature Scaling and Geospatial Normalization** The models that displayed sensitivity to input scales received standardized numerical data through z-score normalization of Depth, Mag, Gap and Dmin features. A normalization process for geospatial coordinates consisting of Latitude and Longitude values normalized all geographic elements to achieve uniform scale representation. **Temporal Feature Engineering** When converting datasets, the Time field timestamps were transformed into cyclic patterns by applying sine and cosine methods. Daily and seasonal earthquake patterns found naturally in seismic data remain observable through the successful implementation of encoding methodology. **Feature Selection and Final Preparation** Pearson correlation analysis determined a strong interrelation between Rms and MagNst features, so they were both eliminated to decrease multicollinearity. The analyzed dataset was split into training and testing sections with an 80:20 split ratio before preprocessing. In later stages, the predictive model received support through k-fold cross-validation procedures, so predictive performance became reliable without train-test bias. The preprocessing phase developed a stable foundation for seismic magnitude prediction through its work to resolve inconsistencies and convert unstructured data to a standardized format according to [30].

DOI: https://doi.org/10.54216/JAIM.090205
Received: January 14, 2025 Revised: March 18, 2025 Accepted: May 10, 2025

4 Results and Discussion

The section presents the predictive results of applying machine learning models to the earthquake prediction dataset. The research is divided into three essential sections where it evaluates chosen machine learning models together with evaluation approaches and performance measurements followed by model outcome comparison. The combined elements provide a complete model performance evaluation regarding predictive accuracy, robustness, and practical seismic forecasting potential.

4.1 Machine Learning Models

The subsequent part examines machine learning models designed to analyze earthquake prediction information. The predictive analytics methods use unique modeling approaches with their accompanying mathematical approaches. Multiple model types in the selection provide diverse behavior models to help effectively evaluate prediction accuracy and processing efficiency for seismic data. Each machine learning algorithm used in the research functions to predict earthquake magnitudes because they show unique strengths in managing structured high-dimensional data analysis. The study employs LightGBM and three other predictive models consisting of SVR and KNN and Ridge Regression used together with Extra Trees Regressor. Researchers selected these predictive models because they successfully perform regression tasks and manage linear and non-linear connections with high clarity when analyzing geospatial and temporal seismic data. Accuracy prediction measurements and generalization tests showed that LightGBM stood as the most effective model among competitors. LightGBM builds gradient-boosting systems through leaf-oriented tree development which delivers high performance instead of using traditional level-wise methods. The method facilitates the model to produce deep trees because it makes its loss reduction abilities more powerful. This framework enables quick processing of large datasets because it conducts parallel processing of tuples simultaneously with GPU learning. LightGBM delivers superior outcomes for working with extensive features and missing data and discrete features. LightGBM accommodates built-in regularization features to combat overfitting issues that enable its perfect application to earthquake prediction analysis involving uneven and unpredictable data distributions. The LightGBM model implements a loss function minimization routine by adding weak learner trees one after the other:

$$X_t(f) = X_{t-1}(f) + \eta \cdot y_t(f)$$

At iteration t the boosted prediction becomes $X_t(f)$ while the learning rate equals η and $y_t(f)$ represents the newly added decision tree. This study employed a Randomized Search method to tune LightGBM through various hyperparameter tests that underwent five-fold cross-validation to ensure reliable results. The table in Table 1 highlights the final set of optimal parameters used in modeling. The chosen parameters resulted in top-notch predictive performance for the model, which exceeded traditional models according to every evaluation metric measured.

Table 1: Optimized Hyperparameters for LightGBM Model

Parameter	Value	Description
`num_leaves`	31	Controls the complexity of the individual trees; more leaves allow more complex models.
`learning_rate`	0.1	Determines the step size at each iteration while minimizing the loss function.
`n_estimators`	100	Number of boosting iterations (trees) used to fit the model.
`subsample`	0.8	Fraction of observations randomly selected for each tree to prevent overfitting.
`colsample_bytree`	0.8	Fraction of features randomly selected for constructing each tree, improving generalization.
`min_child_samples`	20	Minimum number of samples required in a leaf node to control overfitting.
`random_state`	42	Random seed to ensure reproducibility of results.

DOI: https://doi.org/10.54216/JAIM.090205

Received: January 14, 2025 Revised: March 18, 2025 Accepted: May 10, 2025

Support Vector Regression (SVR) The researchers selected SVR because this model successfully analyzes intricate non-linear relationships. SVR seeks to discover functions representing target data points using a specified tolerance margin, which is known as ϵ.

$$q(x) = \langle h, x \rangle + a$$

The optimization objective is:

$$\min_{h,b,\xi,\xi^*} \frac{1}{2}\|h\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*)$$

Subject to:

$$\begin{cases} y_i - \langle h, x_i \rangle - a \leq \epsilon + \xi_i \\ \langle h, x_i \rangle + a - y_i \leq \epsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{cases}$$

k-Nearest Neighbors (KNN) This method uses non-parametric techniques to determine predictions through the average output calculation of the k nearest data points.

$$\hat{E}(y) = \frac{1}{k}\sum_{i \in \mathcal{N}_k(y)} E_i$$

Ridge Regression The model includes L2 regularization for OLS to reduce coefficient sizes and handle issues of multicollinearity.

$$\hat{\beta} = \arg\min_{\beta}\left\{\sum_{i=1}^{n}(y_i - x_i^\top\beta)^2 + \lambda\|\beta\|^2\right\}$$

Extra Trees Regressor Random Decision Forests and Extremely Randomized Trees function as ensemble learning techniques to generate various unpruned decision trees. A randomized cut-point selection process through random selection determines the points of execution where randomness will be introduced.

$$\hat{v}(i) = \frac{1}{T}\sum_{t=1}^{T} b_t(i)$$

The comparison process involves running the algorithms simultaneously on one datasets to determine performance levels. This research provides direct comparison of earthquake prediction modeling strategies to obtain beneficial practices for academic applications and real-time seismic monitoring.

4.2 Evaluation Metrics

The reliability, accuracy, and performance of predictive systems require thorough evaluation when implementing models for earthquake prediction since operational results depend on model capabilities. The assessment of machine learning models with their linked feature selection methods in this application needs strict evaluation to produce dependable and generalizable outcomes. The evaluation framework must integrate several statistical and mathematical assessment methods to provide comprehensive evaluations of model conduct while allowing for proper algorithm and optimization strategy comparisons. This research evaluated machine learning model predictions through a created framework that measured three performance aspects, including predictive accuracy and error magnitude, together with agreement with observed actual data. Society places great importance on understanding model capabilities for predicting data beyond what has already been shown because it helps them forecast earthquakes more effectively under real-world conditions. Error-based metrics form the core metric category within this framework as they determine the size of prediction discrepancies

DOI: https://doi.org/10.54216/JAIM.090205

Received: January 14, 2025 Revised: March 18, 2025 Accepted: May 10, 2025

relative to observed results. The evaluation techniques employ MSE together with RMSE and MAE for assessment. The squared nature of MSE makes it sensitive to critical prediction errors so it functions well for primary deviation assessment. The square root calculation of MSE transforms its result back into original unit measurement to provide practitioners with an easier-to-understand interpretation format. MAE gives predicted data versus actual data comparisons by averaging their absolute value differences that perform identically for small and large errors without considering differences in size. These absolute error metrics are supported by correlation-based statistical indicators in the framework, which determine the relationship between predicted and observed value strengths. The coefficient of determination (R^2) represents the percentage of target variable variance within the model's predictions. The data distribution matches better when the R-squared statistic increases. The Nash-Sutcliffe Efficiency evaluates system performance by comparing model residuals to observed data variance levels to determine the model's ability to capture original variations. The WI provides a normalized model performance evaluation by measuring prediction-to-observation value concordance. It establishes perfect agreement at WI = 1. To identify consistent prediction biases the Mean Bias Error (MBE) detects systematic errors to reveal positive and negative deviations from actual measurement values. The linear correlation between predicted value estimations and actual data results can be evaluated using Pearson's correlation coefficient (r). The collection of assessment measures delivers an extensive technique to evaluate model performance against its actual behavior. The Relative Root Mean Squared Error (RRMSE) presents itself as a metric allowing satisfactory performance analysis between datasets with different scale sizes. The RRMSE metric enables fair benchmarking through normalization of RMSE with observed value range because it works across datasets with different magnitudes or distributions of features. The paper provides an extensive summary of evaluation metrics used throughout the study, containing their formal definitions and mathematical expressions in Table 2. The evaluation metrics produce systematic assessment methods that lead to transparent model validation in earthquake modeling applications with high stakes.

4.3 Results

Empirical evidence showed how machine learning algorithms perform functions to detect earthquake magnitude levels in this section. The necessary objective is to establish which model achieves the highest predictive accuracy alongside superior generalization abilities and accurate seismic recognition. Various mathematical and statistical evaluation metrics evaluate prediction mistakes by measuring their dimensions together with their contour shapes for the assessment to take place. This analysis examined five regression models which included LightGBM (Tuned) together with Support Vector Regression (SVR) and k-Nearest Neighbors (KNN) and Ridge Regression and Extra Trees Regressor. The assessment metrics used to evaluate model performance appeared in Table 3 through nine indicators that include Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Bias Error (MBE) and Pearson Correlation Coefficient (r), Coefficient of Determination (R^2), Relative Root Mean Squared Error (RRMSE), Nash–Sutcliffe Efficiency (NSE), and Willmott Index (WI). The model assessment consists of nine metrics which determine the ability to maintain original data relationships for precise predictions.

Table 3: Evaluation of Machine Learning Models for Predicting Earthquakes

Model	MSE	RMSE	MAE	MBE	r	R^2	RRMSE	NSE	WI
LightGBM (Tuned)	**0.0474**	**0.2176**	**0.1619**	**0.0018**	**0.9613**	**0.9241**	**0.0538**	**0.9241**	**0.9799**
SVR	0.0562	0.2371	0.1743	-0.0160	0.9542	0.9100	0.0586	0.9100	0.9761
KNN	0.0687	0.2621	0.1944	-0.0021	0.9434	0.8900	0.0648	0.8900	0.9703
Ridge Regression	0.0799	0.2827	0.2136	0.0083	0.9339	0.8720	0.0699	0.8720	0.9647
Extra Trees	0.0490	0.2213	0.1644	0.0076	0.9600	0.9216	0.0547	0.9216	0.9791

Among the models assessed in the study LightGBM (Tuned) proves to be the most efficient method for earthquake magnitude predictions according to all performance measures. The predicted output of LightGBM (Tuned) proved its superiority among all models by generating predictions with innovative 0.0474 MSE and 0.2176 RMSE and 0.1619 MAE values because it maintained consistent target value accuracy during assessment. According to the Mean Bias Error analysis, the model produces negligible 0.0018 scores that demonstrate its ability to prevent systematic over- or under-forecasting of final results. Two metrics demonstrated why LightGBM became the top-performing method by achieving an R^2 coefficient of 0.9241 and a Pearson correlation value of $r = 0.9613$ improving its description capability and prediction accuracy for observed

DOI: https://doi.org/10.54216/JAIM.090205

Received: January 14, 2025 Revised: March 18, 2025 Accepted: May 10, 2025

Table 2: Machine Learning Prediction Metrics

Metric	Description
Mean Squared Error (MSE)	Measures the average squared difference between predicted and actual values, penalizing larger errors more than smaller ones. $\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
Root Mean Squared Error (RMSE)	The square root of MSE, representing the standard deviation of prediction errors and giving a sense of the magnitude of prediction errors in the same units as the target variable. $\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$
Mean Absolute Error (MAE)	Measures the average of the absolute errors between predicted and actual values, offering an intuitive and direct understanding of prediction accuracy. $\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\lvert y_i - \hat{y}_i\rvert$
Mean Bias Error (MBE)	Measures the average bias in the predictions, indicating whether the model tends to underestimate or overestimate the target variable. A negative value indicates underestimation, while a positive value indicates overestimation. $\text{MBE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)$
Pearson's Correlation Coefficient (r)	Measures the linear relationship between predicted and actual values. A value of 1 indicates a perfect positive correlation, -1 indicates a perfect negative correlation, and 0 indicates no correlation. $r = \frac{\sum(y_i-\bar{y})(\hat{y}_i-\bar{\hat{y}})}{\sqrt{\sum(y_i-\bar{y})^2}\sqrt{\sum(\hat{y}_i-\bar{\hat{y}})^2}}$
R-squared (R^2)	Represents the proportion of variance in the target variable explained by the model, with higher values indicating better model fit. A value of 1 indicates perfect predictions. $R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$
Relative Root Mean Squared Error (RRMSE)	A normalized version of RMSE compares RMSE with the range of observed values, making it easier to compare errors across datasets with different scales. $\text{RRMSE} = \frac{\text{RMSE}}{\max(y)-\min(y)}$
Nash-Sutcliffe Efficiency (NSE)	Measures the model's predictive power by comparing the model's variance to the variance of observed data. A value of 1 indicates a perfect fit, while values below 0 suggest poor predictive ability. $\text{NSE} = 1 - \frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$
Willmott Index (WI)	Measures the agreement between predicted and observed values, with a value of 1 indicating perfect agreement. WI values closer to 1 suggest better model performance. $WI = 1 - \frac{\sum_{i=1}^{n}\lvert y_i-\hat{y}_i\rvert}{\sum_{i=1}^{n}(\lvert y_i-\bar{y}\rvert+\lvert\hat{y}_i-\bar{y}\rvert)}$

data points. The predictive capabilities of LightGBM are proven exceptional because the model delivers high predictive metrics during seismic data analysis of complex nonlinear patterns. The results from LightGBM exceeded other methods by obtaining superior Nash–Sutcliffe Efficiency (NSE = 0.9241) and Willmott Index (WI = 0.9799). The scoring metrics of LightGBM achieve near-perfect accuracy in matching predicted value series with actual sequences because these evaluation methods focus on time sequence predictions for geophysical data. The model shows resistance across different seismic levels due to its low relative RMSE value which amounts to 0.0538. LightGBM brings superior outcomes from its combination of designed structures and optimization functionalities. LightGBM implements a leaf-wise tree construction method which properly optimizes loss reduction efficiency as well as accelerates convergence. LightGBM excels in multi-dimensional dataset analysis because it incorporates built-in regularization and early stopping mechanisms that specialize in preventing overfitting when dealing with earthquake data structures which possess categorical variables. The algorithm delivers multi-threaded functionality through histogram-based learning processes which allow quick evaluation of big dataset point information. The experiment showed that Extra Trees produced comparable performance to LightGBM (MSE = 0.0490, $R^2 = 0.9216$, WI = 0.9791) although it obtained slightly worse results compared to LightGBM during evaluation. Measures of bias and magnitude during SVR resulted in underpredictions followed by significant deviations (MBE = -0.0160) from original measurement outputs. The combination of Ridge Regression and KNN produced inferior results than other models when analyzing non-linear earthquake patterns through RMSE and MAE metrics. The LightGBM (Tuned) model functions as the preferred model for earthquake magnitude prediction because it provides superior efficiency throughout such applications. LightGBM (Tuned) presents remarkable capability within all performance metrics which

Received: January 14, 2025 Revised: March 18, 2025 Accepted: May 10, 2025

establishes it as a essential component for seismic early warning systems alongside risk assessment tools that support disaster preparedness. LightGBM delivers crucial seismic risk management benefits that enable emergency decision-makers as well as engineers and policymakers to succeed in their response activities. The standard assessment of machine learning models' earthquake prediction abilities takes place via performance metrics evaluation of five algorithms. The performance evaluation combines Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), Coefficient of Determination (R^2), Nash–Sutcliffe Efficiency (NSE) as well as Willmott Index (WI) metrics for assessment. The different performance metrics provide unique information about model accuracy which combines its reliability capabilities with its ability to align with seismic observations. This visualization systems aim to uncover both superior and inferior aspects of the models while finding the optimal approach which yields minimal errors and maximizes generalizability in earthquake magnitude prediction. Mean Absolute Error (MAE) functions as the initial performance metric because it calculates performance accuracy by determining the average absolute predictive error measurement without directionality considerations. Users can interpret MAE as an approach to assessing predictive accuracy because of its clear functionality. As Figure 6 shows LightGBM (Tuned) emerges as the top performer because its 0.1619 MAE stands as the lowest among all models. The LightGBM (Tuned) algorithm produces the smallest error gap between its prediction values and actual results throughout. Extra Trees exhibits comparable performance while Ridge Regression lags behind significantly in MAE results thereby indicating unpredictable results for critical cases requiring minimal absolute errors.

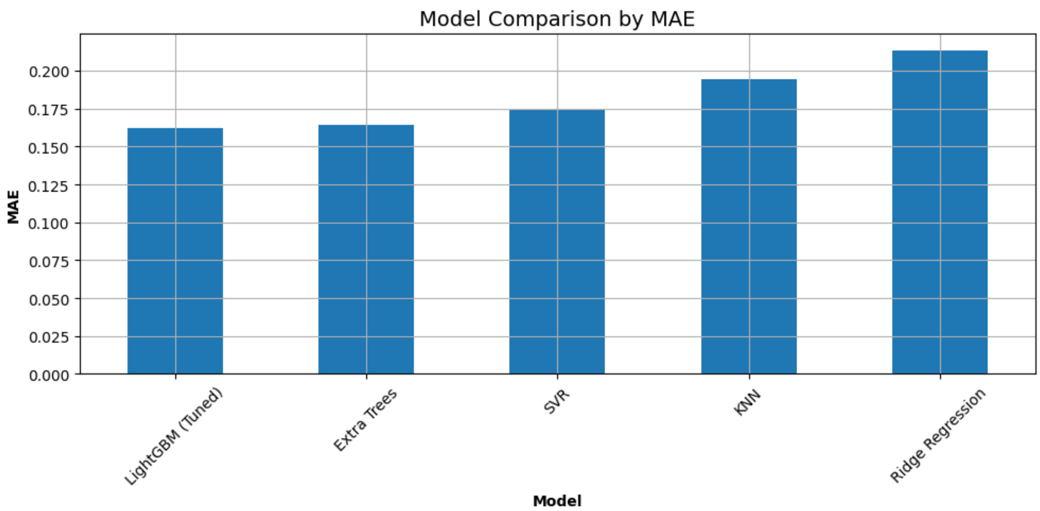

Figure 5: Comparing models based on average prediction error, LightGBM (Tuned) has the lowest average prediction error.

Our evaluation of model capabilities involved calculation of the Root Mean Squared Error (RMSE) metric because this metric shows high sensitivity to large prediction mismatches with actual values. The RMSE plot in Figure 6 confirms that LightGBM (Tuned) produced the smallest error value of 0.2176 while SVR and Extra Trees occupied the next positions. The results show the strong ability of LightGBM to minimize severe errors since this capability remains fundamental in earthquake prediction when wrong prediction results might trigger unneeded emergency responses or cause wrongful alarms.

DOI: https://doi.org/10.54216/JAIM.090205

Received: January 14, 2025 Revised: March 18, 2025 Accepted: May 10, 2025

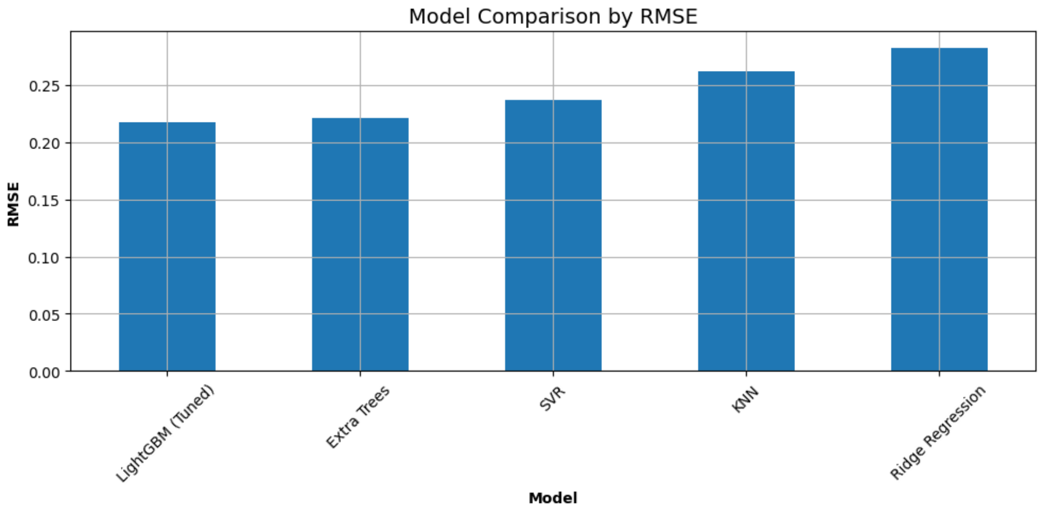

Figure 6: Model Comparison by RMSE: LightGBM (Tuned) minimizes large prediction errors, showcasing robustness.

The Coefficient of Determination (R^2) functions as a significant evaluation metric to determine the accuracy of suggested values in representing actual data point variation. These predictive models that include LightGBM (Tuned) and Extra Trees reach satisfactory matches according to Figures 0.9241 and 0.9216R^2 as shown in Figure 7. The high R-squared values indicate the models correctly find a large portion of relevant data patterns that are necessary to analyze spatially complex and temporally random events.

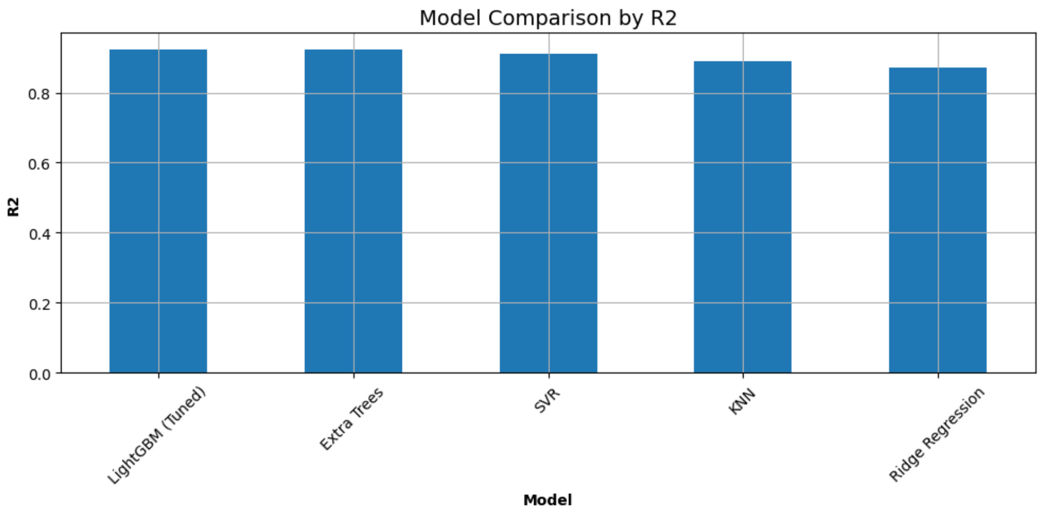

Figure 7: Comparing the models by R^2, LightGBM (Tuned) and Extra Trees show a high level of explanatory power.

The second validation approach used the Nash–Sutcliffe Efficiency to determine model predictive power through evaluation of predictive outcomes versus baseline mean forecast results. A model demonstrates better predictive ability with an NSE value that surpasses the average baseline mean observed value strength. The lightGBM (Tuned) and Extra Trees algorithms maintain their superior position by reaching nearly the perfect NSE score value of 1.0 as shown in Figure 8. An ensemble-based approach produces superior generalization abilities during high-precision measurements in comparison to the linear and local methods Ridge Regression and KNN.

DOI: https://doi.org/10.54216/JAIM.090205

Received: January 14, 2025 Revised: March 18, 2025 Accepted: May 10, 2025

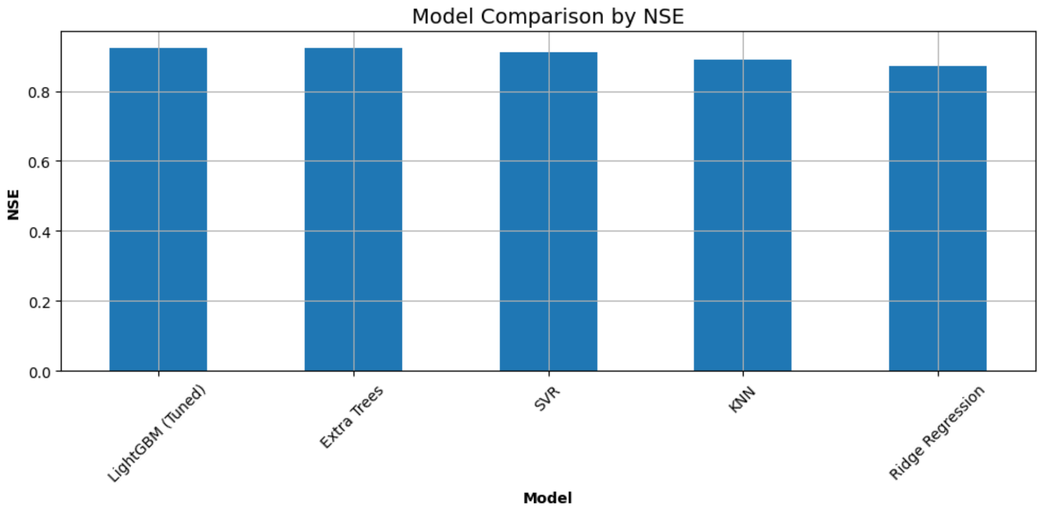

Figure 8: Model Comparison via NSE: In comparison to observed data, LightGBM (Tuned) obtains a good prediction efficiency.

We consider the Willmott Index (WI) because it provides a method for assessing prediction-actual value agreement while normalizing evaluation data for different range intervals. A normalized version of this measure provides useful insights into model consistency that operates across wide data ranges. The actual-data-based WI assessment confirmed that LightGBM (Tuned) achieved a score of 0.9799 superior to Extra Trees at 0.9791 although WI visual results are not provided. The scores validated absolute correspondence between field seismic data observed in the monitoring facility and model predictions developed for this purpose. Through every assessment including statistical and visual evaluation the LightGBM (Tuned) model demonstrated its position as the best prediction model. The model demonstrates outstanding prediction ability because its low error rating joins superior observed data correlation statistics with better generalization performance for earthquake magnitude prediction. Several components that create its strengths come directly from its ensemble architecture optimization and its usage of advanced regularization techniques and its robust dataset management abilities. The tuned version of LightGBM demonstrates better performance than all other models because it preserves both high accuracy and stability which facilitates necessary real-time earthquake detection alongside emergency readiness preparations. The vital importance of models and their parameters becomes clear through comprehensive evaluation of the selection process for geophysical prediction machine learning systems.

5 Conclusion and Future Work

Our research analyzed machine learning (ML) models for earthquake magnitude prediction by evaluating globally obtained year-2023 data. This evaluation of prominent regression algorithms included LightGBM and SVR, KNN and Ridge Regression, and Extra Trees by analyzing predictive features of geographic coordinates, seismic depth, and historical magnitude values. Throughout the evaluation LightGBM yielded superior performance on every statistical measurement. 0.9241 R^2 along with 0.0474 Mean Squared Error finalized the model, which showed exceptional performance and generalization potential for predicting data outcomes. The framework, combined with efficient regularization and scalability, permitted fast detection of seismic nonlinearity. According to the presented results, real-time earthquake prediction and early warning system operations benefit from ensemble-based ML approaches. The research supports the critical contribution of ML to seismic risk modeling that enhances traditional geophysical methods through data-based strategies to aid in evidence-based emergency planning. The framework provides researchers with a reproducible and extensible structure by combining open-access datasets and systematic evaluation metrics united with diverse model architectures. The first goal concentrates on improving dataset temporal frequency by incorporating additional earthquake records to strengthen long-term prediction approaches. Our research intends to merge auxiliary data sources consisting of satellite ionospheric measurements, tectonic plate velocity data, and records of past seismic faults to boost feature complexity and predictive performance metrics. The analysis will investigate explainable artificial intelligence (XAI) techniques for model-based decision interpretation to generate trust in operational

DOI: https://doi.org/10.54216/JAIM.090205

Received: January 14, 2025 Revised: March 18, 2025 Accepted: May 10, 2025

systems. The research will study real-time earthquake early warning framework deployment of the models to test their operational effectiveness within seismic events. The research adds essential information about machine learning capabilities for earthquake magnitude forecasting while enabling new ways to combine work between seismologists, artificial intelligence experts, and geoinformaticians.

References

[1] Artem Vesnin, Yury Yasyukevich, Natalia Perevalova, and Erman Şentürk. Ionospheric response to the 6 february 2023 turkey–syria earthquake. *Remote sensing*, 15(9):2336, 2023.

[2] Hua-Lun Huang. Natural hazards: Earthquakes. In *Encyclopedia of Security and Emergency Management*, pages 662–669. Springer, 2021.

[3] Mikhail B Gokhberg. *Earthquake prediction*. CRC Press, 2024.

[4] Munish Bhatia, Tariq Ahamed Ahanger, and Ankush Manocha. Artificial intelligence based real-time earthquake prediction. *Engineering Applications of Artificial Intelligence*, 120:105856, 2023.

[5] Assem Turarbek, Maktagali Bektemesov, Aliya Ongarbayeva, Assel Orazbayeva, Aizhan Koishybekova, and Yeldos Adetbekov. Deep convolutional neural network for accurate prediction of seismic events. *International Journal of Advanced Computer Science and Applications*, 14(10), 2023.

[6] Alexander N Safronov. Astronomical triggers as a cause of strong earthquakes. *International Journal of Geosciences*, 13(9):793–829, 2022.

[7] David Teh and Tehmina Khan. Types, definition and classification of natural disasters and threat level. In *Handbook of disaster risk reduction for resilience: new frameworks for building resilience to disasters*, pages 27–56. Springer, 2021.

[8] Mei-Yun Lin and Raluca Ilie. A review of observations of molecular ions in the earth's magnetosphere-ionosphere system. *Frontiers in Astronomy and Space Sciences*, 8:745357, 2022.

[9] Denis Anikiev, Claire Birnie, Umair bin Waheed, Tariq Alkhalifah, Chen Gu, Dirk J Verschuur, and Leo Eisner. Machine learning in microseismic monitoring. *Earth-Science Reviews*, 239:104371, 2023.

[10] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30, 2017.

[11] Luca Franceschi, Michele Donini, Valerio Perrone, Aaron Klein, Cédric Archambeau, Matthias Seeger, Massimiliano Pontil, and Paolo Frasconi. Hyperparameter optimization in machine learning. *arXiv preprint arXiv:2410.22854*, 2024.

[12] Gautam Kunapuli. *Ensemble methods for machine learning*. Simon and Schuster, 2023.

[13] Davide Chicco, Matthijs J Warrens, and Giuseppe Jurman. The coefficient of determination r-squared is more informative than smape, mae, mape, mse and rmse in regression analysis evaluation. *Peerj computer science*, 7:e623, 2021.

[14] Gaurav Singh Manral and Alka Chaudhary. Prediction of earthquake using machine learning algorithms. In *2023 4th International Conference on Intelligent Engineering and Management (ICIEM)*, pages 1–5. IEEE, 2023.

[15] Nada Badr Jarah, Abbas H Hassin Alasadi, and Kadhim Mahdi Hashim. Earthquake prediction technique: A comparative study. *IAES Int. J. Artif. Intell.(IJ-AI)*, 2252:1027, 2023.

[16] Ewnetu Abebe, Hailemichael Kebede, Mickus Kevin, and Zelalem Demissie. Earthquakes magnitude prediction using deep learning for the horn of africa. *Soil Dynamics and Earthquake Engineering*, 170:107913, 2023.

DOI: https://doi.org/10.54216/JAIM.090205
Received: January 14, 2025 Revised: March 18, 2025 Accepted: May 10, 2025

[17] Mustafa Abdul Salam, Lobna Ibrahim, and Diaa Salama Abdelminaam. Earthquake prediction using hybrid machine learning techniques. *International Journal of Advanced Computer Science and Applications*, 12(5):654–6652021, 2021.

[18] Anushka Joshi, Balasubramanian Raman, C Krishna Mohan, and Linga Reddy Cenkeramaddi. Application of a new machine learning model to improve earthquake ground motion predictions. *Natural Hazards*, 120(1):729–753, 2024.

[19] Victor Manuel Velasco Herrera, Eduardo Antonio Rossello, Maria Julia Orgeira, Lucas Arioni, Willie Soon, Graciela Velasco, Laura Rosique-de La Cruz, Emmanuel Zúñiga, and Carlos Vera. Long-term forecasting of strong earthquakes in north america, south america, japan, southern china and northern india with machine learning. *Frontiers in Earth Science*, 10:905792, 2022.

[20] Papiya Debnath, Pankaj Chittora, Tulika Chakrabarti, Prasun Chakrabarti, Zbigniew Leonowicz, Michal Jasinski, Radomir Gono, and Elżbieta Jasińska. Analysis of earthquake forecasting in india using supervised machine learning classifiers. *Sustainability*, 13(2):971, 2021.

[21] Rayan Abri and Harun Artuner. Lstm-based deep learning methods for prediction of earthquakes using ionospheric data. *Gazi University Journal of Science*, 35(4):1417–1431, 2022.

[22] Bharat Bhargava and Sumanta Pasari. Earthquake prediction using deep neural networks. In *2022 8th international conference on advanced computing and communication systems (ICACCS)*, volume 1, pages 476–479. IEEE, 2022.

[23] Mohsen Yousefzadeh, Seyyed Ahmad Hosseini, and Mahdi Farnaghi. Spatiotemporally explicit earthquake prediction using deep neural network. *Soil Dynamics and Earthquake Engineering*, 144:106663, 2021.

[24] Yao Li, Peng Cui, Chengming Ye, José Marcato Junior, Zhengtao Zhang, Jian Guo, and Jonathan Li. Accurate prediction of earthquake-induced landslides based on deep learning considering landslide source area. *Remote Sensing*, 13(17):3436, 2021.

[25] Pan Xiong, Lei Tong, Kun Zhang, Xuhui Shen, Roberto Battiston, Dimitar Ouzounov, Roberto Iuppa, Danny Crookes, Cheng Long, and Huiyu Zhou. Towards advancing the earthquake forecasting by machine learning of satellite data. *Science of the Total Environment*, 771:145256, 2021.

[26] Yiyi Zhao, Shuai Lv, and Pengfei Liu. Advances in earthquake prevention and reduction based on machine learning: a scoping review (july 2024). *IEEE Access*, 2024.

[27] M Apriani, SK Wijaya, et al. Earthquake magnitude estimation based on machine learning: Application to earthquake early warning system. In *Journal of Physics: Conference Series*, volume 1951, page 012057. IOP Publishing, 2021.

[28] Earthquakes 2023 global. `https://www.kaggle.com/datasets/mustafakeser4/earthquakes-2023-global`, 2023.

[29] Cheng Fan, Meiling Chen, Xinghua Wang, Jiayuan Wang, and Bufu Huang. A review on data preprocessing techniques toward efficient and reliable knowledge discovery from building operational data. *Frontiers in energy research*, 9:652801, 2021.

[30] Andrea Barucci, Stefano Diciotti, Marco Giannelli, and Chiara Marzi. Data preparation for ai analysis. In *Introduction to Artificial Intelligence*, pages 133–150. Springer, 2023.

DOI: https://doi.org/10.54216/JAIM.090205
Received: January 14, 2025 Revised: March 18, 2025 Accepted: May 10, 2025