World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(RESEARCH ARTICLE)

Check for updates

Machine learning for earthquake engineering analysis: Comparing regression models to predict peak ground acceleration

Shima Pakniat [1,*], Jafar Najafizadeh [2] and Monavvareh Kadkhodaavval [3]

[1] *Department of Civil Engineering, The University of Memphis, Memphis, Tennessee, USA.*
[2] *Department of Civil Engineering, Islamic Azad University, Mashhad Branch, Mashhad, Iran.*
[3] *Housing & Urban Development Research Center, Tehran, Iran.*

World Journal of Advanced Research and Reviews, 2025, 26(02), 856-867

Publication history: Received on 28 March 2025; revised on 03 May 2025; accepted on 06 May 2025

Article DOI: https://doi.org/10.30574/wjarr.2025.26.2.1714

Abstract

Advancements in machine learning have introduced powerful tools for enhancing seismic hazard assessment, offering improved predictive capabilities compared to traditional regression models. This study leverages machine learning algorithms to develop data-driven ground-motion models (GMMs) for predicting peak ground acceleration (PGA), a key parameter in seismic hazard analysis. Both parametric and nonparametric regression techniques, including linear regression, polynomial regression with second-degree terms, decision tree, and random forest, are employed. The models are trained on a comprehensive dataset comprising over 10,000 ground-motion records from small-to-moderate earthquakes (magnitude 3.5 to 5.8) with hypocentral distances up to 200 km. Predictor variables such as moment magnitude (Mw), hypocentral distance (Hypo-D), average shear wave velocity in the upper 30 meters (VS30), and focal depth (Ztor) are utilized to capture the complex relationship. Performance evaluation reveals that the random forest model significantly outperforms traditional regression-based GMMs like linear regression, demonstrating its potential to enhance seismic hazard assessment, particularly for regions prone to similar earthquakes.

Keywords: Machine learning; Regression Model; Seismic hazard assessment; Decision Tree; Random Forest; Ground-motion models (GMMs); Peak ground acceleration (PGA); Predictive modeling; Earthquake engineering

1. Introduction

Ground-motion models (GMMs) are fundamental tools in seismic hazard analysis, contributing to the development of hazard maps, earthquake-resistant building codes, and risk reduction strategies [1, 2, 3]. Among their key applications, GMMs are widely used to predict Peak Ground Acceleration (PGA) - a crucial parameter in site response analysis and structural design. Extensive research has been conducted in seismic analysis to enhance site response modeling. For instance, Najafizadeh et al. investigated the site response of various geological formations, including 2D triangular, irregular triangular, and rectangular alluvial deposits [4, 5, 6]. Similarly, Pakniat et al. developed SEISGRASP, a software package designed for signal processing, soil profile analysis, and comprehensive site response analysis results [7, 8]. Tools like SEISGRASP demonstrate the growing role of advanced computational methods in refining seismic hazard assessments. Earthquake records and their characteristics are also used in many structural analyses like seismic fragility assessment of buildings [9]. Ground motion models are based on factors such as earthquake magnitude, source-to-site distance, and site-specific conditions. Traditional empirical GMMs often utilize predefined functional forms to model ground motion parameters, as demonstrated by several seminal works [10, 11, 12]. While effective, Traditional ground motion prediction equations (GMPEs) rely on statistical regression with predefined functional forms to estimate intensity measures like PGA and pseudo-spectral acceleration (PSA). However, these models often face limitations in capturing nonlinear relationships and handling large datasets [13]. Recent advances in machine learning (ML) have

* Corresponding author: Shima Pakniat

introduced powerful nonparametric alternatives to classical regression techniques. Unlike traditional approaches, ML methods do not require predefined functional forms, enabling them to capture complex nonlinear relationships and adapt to large, high-dimensional datasets. Recent advances in environmental and geotechnical modeling increasingly emphasize hybrid methods that combine machine learning with traditional physics-based approaches to enhance predictive performance and computational efficiency. In recent work, machine learning models such as regression trees have been effectively used in flood forecasting, particularly in data-scarce regions [14]. Similarly, recent research has applied data-driven techniques to track and simulate the environmental transport of emerging contaminants like microplastics, integrating experimental insights with computational [15, 16, 17]. These approaches highlight the power of machine learning in capturing complex, nonlinear relationships in physical systems. Building on this concept, the current study applies machine learning algorithms to seismic hazard analysis, focusing specifically on predicting peak ground acceleration (PGA). By leveraging models such as genetic algorithms, simulated annealing, and regression-based techniques, this work demonstrates how machine learning can improve the estimation of PGA using seismic input parameters—especially in cases where traditional analytical models struggle with variability and limited data. This flexibility is particularly beneficial in regions with sparse earthquake data or complex geological conditions. Studies such as [18, 19, 20, 21, 22] have demonstrated the efficacy of ML techniques like artificial neural networks (ANN), random forest regressors (RFR) in developing GMMs. Induced earthquakes, often triggered by human activities such as fluid injection, present unique challenges due to their shallow depths and distinct attenuation characteristics [2, 8]. Traditional GMMs, which are primarily designed for tectonic earthquakes, may not fully capture these differences. Machine learning offers a flexible framework for modeling induced seismicity, leveraging historical records to better understand the behavior of small-to-moderate magnitude events. For instance, Alidadi et.al [18], developed a region-specific GMM for induced earthquakes in Central and Eastern North America (CENA), providing valuable insights into the unique attenuation patterns of these events [22].

In this study, we employ several supervised ML algorithms, including linear regression, polynomial regression, decision tree and random forest, to develop GMMs tailored to small-to-moderate induced earthquakes. After evaluating the accuracy of each model, the Random Forest Regressor was identified as the most reliable model, providing the best performance in terms of predictive accuracy. This research enhances our understanding of ground motion prediction for induced seismicity and contributes to more accurate seismic hazard assessments. Utilizing moment magnitude (Mw), hypocentral distance (Hypo-D), and VS30 as predictor variables. Our models aim to forecast PGA with improved reliability. This research contributes to advancing data-driven GMMs, addressing critical challenges in seismic hazard assessments for induced seismicity.

2. Material and methods

2.1. Data collection and Preprocessing

In this study, the data used for training the machine learning models comes from the NGA-West2 database, an expansion of the NGA-West1 database. The dataset includes ground-motion data from small-to-moderate magnitude earthquakes in California, as well as global strong ground motion recordings from shallow crustal earthquakes in active tectonic regions, such as Japan, New Zealand, and Italy, recorded after the year 2000. This updated version of the database, compiled by the Pacific Earthquake Engineering Research Center (PEER) [23], contains 21,538 recordings and is specifically designed for developing GMMs for shallow crustal earthquakes.

The recordings consist of uniformly processed time series and response spectral data, which include instrument-corrected, median, orientation-independent horizontal components (RotD50-d030) of ground-motion intensity measures (GMIMs). These GMIMs represent the 50th percentile of the response spectra across all nonredundant rotation angles [11]. The database covers a broad range of regions and includes detailed metadata on earthquake characteristics and site conditions.

There are numerous parameters recorded during an earthquake. To simplify the model, a selection of parameters is made based on findings from previous studies in this area. Additionally, a correlation matrix is presented to illustrate the relevance of each parameter to the observed PGA, helping to identify the most influential factors for prediction (figure 1). The input parameters chosen for the machine learning models include:

- Moment Magnitude (M): A unitless measure of earthquake size.
- Hypocenter Distance (Hypo-D): The distance (in km) from the earthquake's hypocenter to the station.
- Depth to the Top of the Rupture Plane (Ztor): Depth (in km) from the ground surface to the top of the fault rupture plane.

- VS30: The time-averaged shear wave velocity (in m/s) in the top 30 meters of the soil.

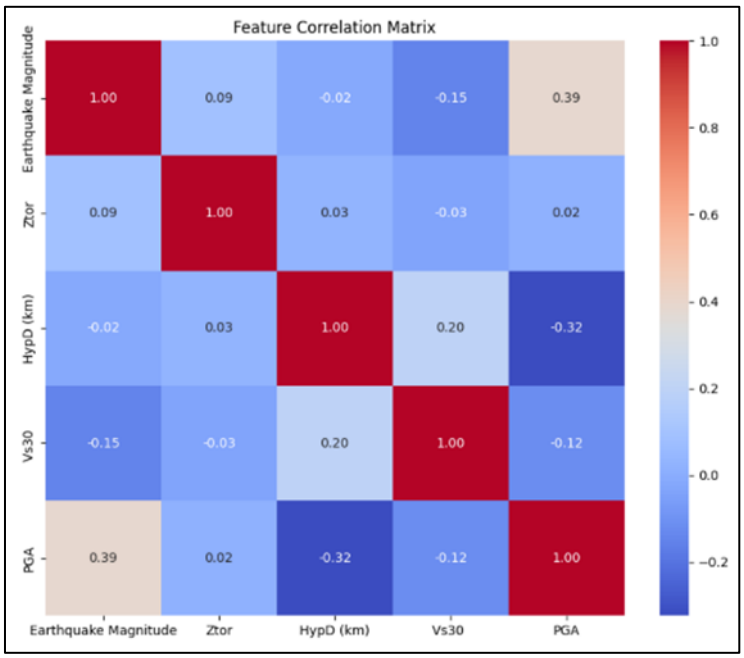

Figure 1 Correlation matrix of features and target parameter before transformation and standardization

2.2. Feature Engineering

Feature engineering in this study involves the transformation and scaling of input data to optimize model performance and improve predictive accuracy. Preliminary data analysis revealed that the logarithmic transformation of certain input variables significantly improves model convergence and performance, particularly for variables that exhibit strong skewness. Figure 2 illustrates the correlation between the transformed features. It is evident that the increase in correlation values highlights the impact of the feature transformations and standardization process.

The key features of transformations are as follows:

- Logarithmic Transformation: The input variables Hypo-D and VS30, along with the output PGA, are transformed using the natural logarithm. This transformation addresses the positive skewness in these variables, making the distribution more suitable for machine learning algorithms, particularly those that assume normality in the data.
- Standardization: Given that most machine learning algorithms are sensitive to the scale of numerical features, all parameters are standardized to have a mean of 0 and a standard deviation of 1. Standardization is particularly important for algorithms like linear regression and support vector machines, which rely on gradient-based optimization methods. Without standardization, features with larger numerical ranges would dominate the model training, leading to biased or inefficient learning [24]. Standardizing the features ensures that the model can treat all features equally, improving convergence speed and overall performance.

The dataset, after these preprocessing steps, is now ready to be fed into machine learning algorithms for training. These steps ensure that the data is both clean and properly scaled, enabling the models to learn effectively from the input features.

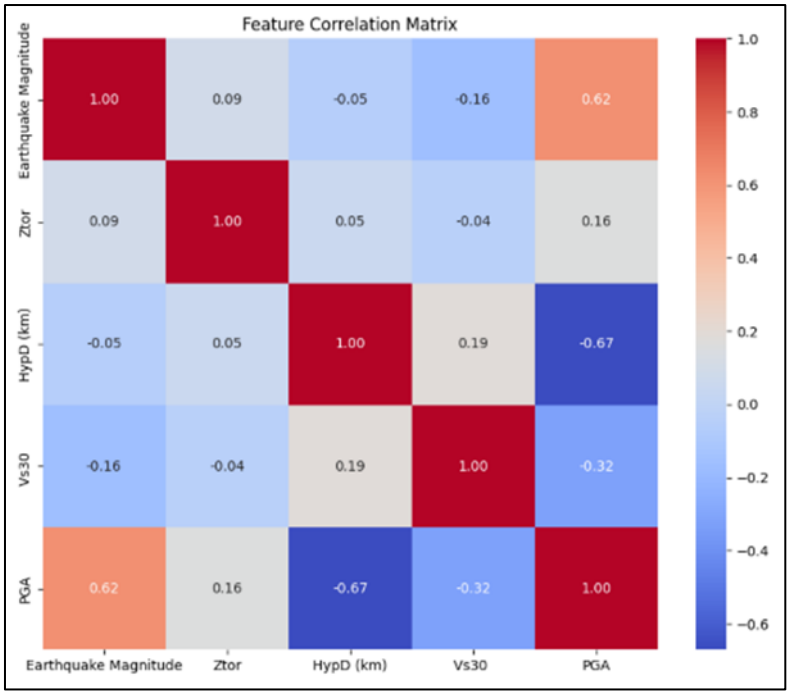

Figure 2 Correlation matrix of features and target parameter after transformation and standardization

2.3. Machine Learning Models

The estimation of GMIM is approached as a regression problem, where the Ground Motion Model (GMM) is represented as follows:

$$\ln(\mathrm{PGA}) = \mathrm{function}[M, \ln(\mathrm{Hypo} - D), Z\_tor, \ln(V\_s30)] \quad \ldots\ldots\ldots\ldots \quad (1)$$

In this equation, PGA represents peak ground acceleration, M is the moment magnitude, Hypo-D is the hypocentral distance, Ztor denotes the depth to the top of the ruptured plane, and VS30 refers to the average shear wave velocity in the upper 30 meters of soil. The data used in this study is based on real earthquake observations and is not synthetic or artificially generated. The relationships between the parameters in the data are unknown, which makes it challenging to model. To address this, we began with the simplest approach, using linear regression, which serves as a baseline model. From there, we progressively enhanced the model by incorporating more advanced techniques and selecting additional parameters to capture more complex relationships. Five machine learning methods were implemented for modeling: linear regression, lasso regression, polynomial regression, decision tree, and random forest. These models were chosen to compare the performance of both linear and nonlinear regression techniques, offering insights into how different model architectures handle seismic data. While linear and lasso regression act as baseline models, the polynomial regression, decision tree, and random forest models provide more flexible frameworks that can better capture nonlinear, complex relationships within the data.

The machine learning algorithms were implemented in Python, and several libraries were used to facilitate the modeling process, including:

- Scikit-learn for model implementation, training, and evaluation.
- NumPy for numerical operations and data manipulation.
- Pandas for handling and preprocessing the dataset.
- Matplotlib and Seaborn for visualizing data and results.

To evaluate model performance on both the training and unseen data, Mean Squared Error (MSE) and R-squared (R^2) metrics were used. These metrics provide a comprehensive assessment of each model's ability to generalize and fit the observed data.

- *Mean Squared Error (MSE)*

MSE represents the variance of the residuals. In general, the smaller variance the better, however, MSE is not expressed on the same scale as the depended variable, making this metric somewhat difficult to interpret. N, y and $\hat{y}$ represent the number of values, actual values and predicted value, respectively.

$$Mean\ Squared\ Error = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{2}$$

- *R-squared*

The coefficient of determination, or R^2, indicates the proportion of the variance in the dependent variable that is accounted for by the linear regression model. It ranges from 0 to 1, with a value of 1 meaning the model perfectly explains all the variance, and a value of 0 indicating that the model explains none of the variance. Negative R^2 values suggest that the model performs worse than simply predicting the mean value each time. In the formula, y, $\bar{y}$ and $\hat{y}$ represent actual value and average of actual values and predicted value, respectively.

$$R^2\_score = 1 - \frac{(y_i - \hat{y}_i)^2}{(y_i - \bar{y}_i)^2} \tag{3}$$

To visually assess model performance, predicted values are plotted against actual values on a 2D scatter plot. An x = y reference line is included to indicate perfect predictions. Data points clustering closely around this line reflect stronger model performance. The closer the data points are to the line, the better the model performs. This indicates that the model accurately capturing the relationship between the variables, highlighting instances of underfitting, overfitting, or systematic error patterns that may warrant further investigation. Each model is explained in the following sections.

- *linear and lasso regression*

In this study, linear regression was selected due to its simplicity and ease of interpretation, particularly in scenarios where a linear relationship between the input features and peak ground acceleration (PGA) can be reasonably assumed. Linear regression's straightforward approach makes it ideal for establishing baseline models and gaining insights into the impact of individual features on the target variable.

In contrast to linear regression, lasso regression was employed to enhance feature selection through the application of regularization. Lasso regression introduces a penalty term that shrinks the coefficients of less relevant predictors, effectively reducing their influence and, in some cases, setting them to zero. This technique streamlines the model by focusing only on the most impactful features. By doing so, lasso regression mitigates the risk of overfitting and enhances the model's generalization capability on unseen data.

To ensure optimal performance, the regularization parameter (alpha) in the lasso regression model was tuned using cross-validation. Cross-validation helps balance the trade-off between bias and variance by partitioning the data into multiple folds and ensuring consistent model evaluation across subsets. This process optimizes prediction accuracy by identifying the alpha value that minimizes error while maintaining model stability.

While linear regression retains all input features, lasso regression refines the model by excluding less significant predictors. This targeted selection of impactful features often results in improved model performance and better interpretability, particularly in datasets with numerous correlated variables.

To evaluate the performance of all models, the dataset was divided into 80% training and 20% test sets. Cross-validation was applied specifically in the lasso regression model to address potential issues related to data skewness or variance inconsistencies across folds and find the best alpha.

Performance metrics, including Mean Squared Error (MSE) and R-squared (R^2), were calculated for both models. Each model achieved an MSE value of 0.167 and an R-squared value of 0.827. These results indicate that the models collectively explain approximately 83% of the variability in PGA values. However, the remaining 17% of the variability remains unexplained, suggesting potential room for improvement. Incorporating additional features or adopting more advanced machine learning models, such as polynomial regression or ensemble methods, may enhance predictive accuracy. To visually assess model fit, scatter plots were generated to compare actual versus predicted PGA values

(Figure 3-4). Also, table 1 represents the factor of each parameter in linear regression and Lasso regression obtained from Scikit learn library of python.

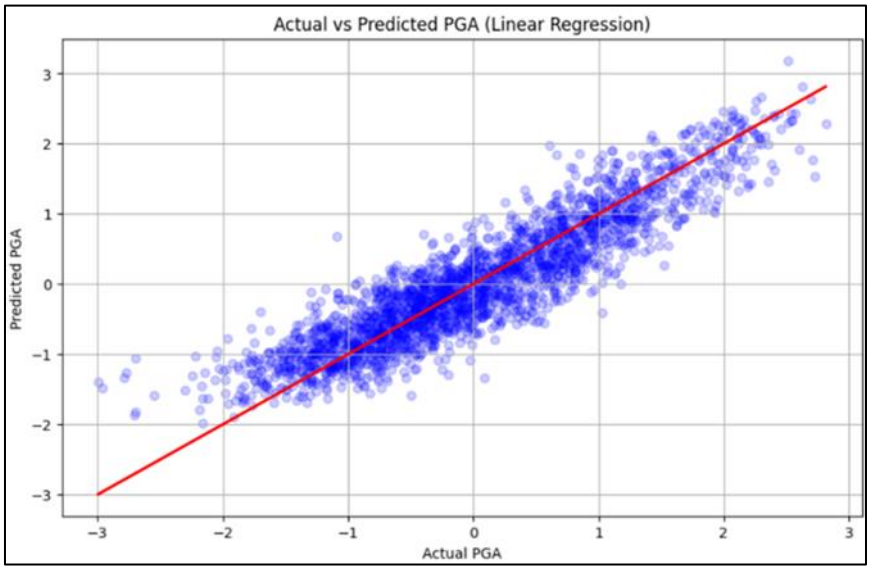

Figure 3 Linear regression scatter plot of actual ln (PGA) and predicted ln (PGA)

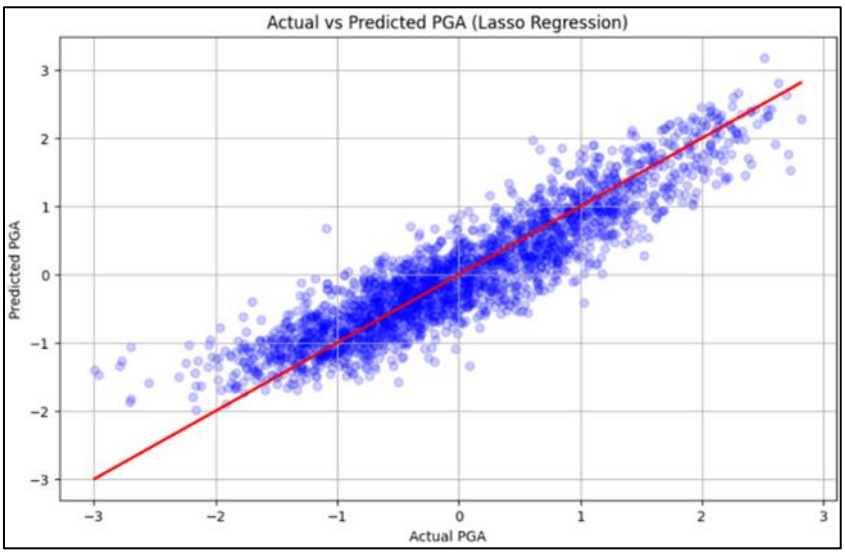

Figure 4 Lasso regression scatter plot of actual ln (PGA) and predicted ln (PGA)

Table 1 Linear and Lasso Regression Parameters

Parameters	Linear Regression	Lasso Regression
Magnitude	0.56059374	0.559839
Hypo_D	-0.62959772	-0.628711
Ztor	0.13716594	0.136225
Vs30	-0.10435275	-0.103690
Bias	0.00077	0.00077

- *Polynomial regression*

Polynomial regression is an extension of linear regression designed to capture more complex relationships between independent variables (features) and the dependent variable (PGA). Unlike linear regression, which assumes a straight-line relationship, polynomial regression introduces higher-degree terms to account for potential non-linear patterns in the data. By incorporating these additional terms, the model gains flexibility to fit curved trends, which may better reflect the behavior of PGA in response to various seismic parameters.

In this study, second-degree polynomial terms were added to the model to improve its ability to capture non-linear dependencies between the features and PGA. This decision was guided by the need to balance model complexity and performance. It is important to mention that including excessively high-degree terms can increase the risk of overfitting, where the model fits the training data too closely and performs poorly on unseen data. To ensure the model's robustness and generalizability, cross-validation was applied.

The performance of the polynomial regression model was evaluated using two key metrics: Mean Squared Error (MSE) and R-squared (R^2). The polynomial regression model achieved an MSE of 0.164 and an R-squared value of 0.830. These results indicate improved predictive performance compared to the linear regression models. The improved R-squared value suggests that the polynomial regression model explains a greater proportion of the variability in PGA values. Despite this improvement, the R-squared value of 0.83 indicates that approximately 17% of the PGA variability remains unexplained. This highlights the potential need for incorporating additional influential features or exploring more sophisticated machine learning algorithms to further enhance model performance.

Figure 5 visualizes the model's effectiveness. The parametric model of second-degree terms which is a combination of main parameters, and a bias term is also shown in table 2. In the following table x_1 , x_2, x_3, x_4 stand for Magnitude, Ztor, Hypo-D and Vs30, respectively.

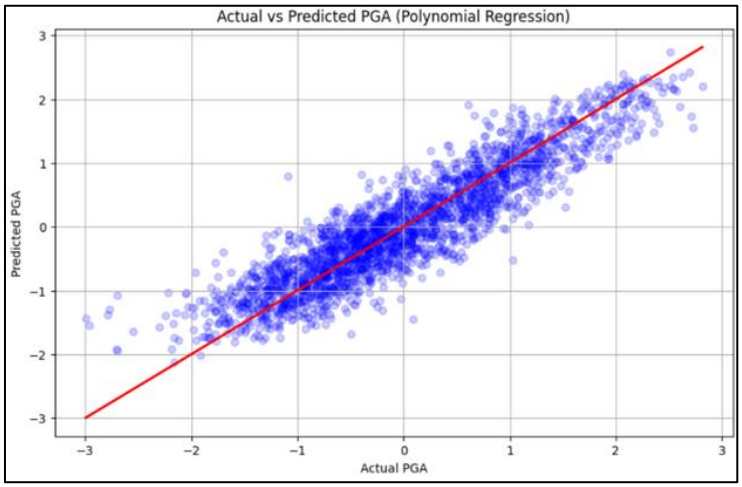

Figure 5 The polynomial scatter plot of actual ln (PGA) and predicted ln (PGA)

Table 2 Polynomial Regression Parameters

Parameter	Value	Parameter	Value
x_1	0.59000	x_2	0.131121
x_3	-0.65780	x_4	-0.09610
${x_1}^2$	-0.03550	${x_2}^2$	0.01094
${x_3}^2$	0.023779	${x_4}^2$	-0.039178
x_1x_2	0.000361	x_1x_3	-0.01481
x_1x_4	0.00391	x_2x_3	-0.024385
x_2x_4	-0.027506	$x_3\ x_4$	-0.00706
Biase		0.06798	

- *Decision tree*

A decision tree is a versatile machine learning model that predicts outcomes by recursively partitioning the data into smaller subsets based on specific conditions. Each split is chosen to maximize the homogeneity of the resulting groups concerning the target variable. Decision trees are particularly effective in capturing complex, nonlinear relationships between input features and the dependent variable, making them a suitable choice for earthquake engineering data where interactions between seismic parameters can be intricate. In this study, a decision tree regression model was developed and evaluated using cross-validation to ensure reliable performance. The decision tree model achieved the following performance metrics:

Mean Squared Error (MSE): 0.25

R-squared (R^2) on the Test Set: 0.74

Mean Squared Error (MSE): 0.00

R-squared (R^2) on the Training Set: 1.00

While the training set R^2 value of 1.00 indicates that the model fits the training data perfectly, this is a strong indication of overfitting. Overfitting occurs when the model memorizes the training data rather than learning its underlying patterns, causing poor generalization to new data. The test set R^2 of 0.74, which is notably lower than the training set score, further confirms this issue. Figure 6 illustrates the actual versus predicted PGA values of the training set for the decision tree model. The alignment of points precisely along the x = y line suggests that the model has memorized the training data. However, this level of precision typically fails to translate to unseen data, explaining the model's weaker test performance.

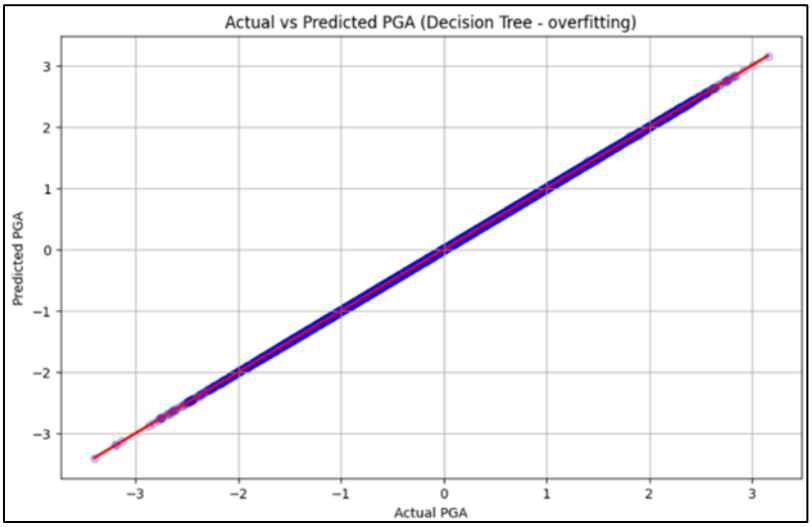

Figure 6 The decision tree training data scatter plot of actual ln (PGA) and predicted ln (PGA)

To enhance model performance and mitigate overfitting, several strategies can be employed:

- Pruning the Decision Tree

Pruning simplifies the model by trimming branches that contribute little to predictive accuracy. By removing these less impactful branches, the model becomes less specialized to the training data, improving its ability to generalize.

Techniques such as cost complexity pruning effectively control the trade-off between model accuracy and complexity by introducing a regularization parameter.

- Limiting Tree Depth

Restricting the maximum depth of the tree helps prevent it from becoming overly complex. Setting a reasonable depth limit (e.g., 3–7 levels) can improve the model's ability to generalize to new data.

- Adjusting Minimum Sample Requirements

Increasing the minimum number of samples required for node splits or leaf formation helps reduce the risk of overly specific, narrow splits that contribute to overfitting.

- Using Random Forests

Random forests mitigate overfitting by combining multiple decision trees, each trained on different subsets of the data and features. This ensemble method effectively reduces variance and enhances model stability.

In this study, to overcome the overfitting problem, Random Forest algorithm is implemented which is described in next section.

- *Random Forest algorithm*

The Random Forest model is a robust and highly effective machine learning technique that improves prediction accuracy by aggregating the outputs of multiple decision trees. Each individual tree in the random forest makes an independent prediction based on a subset of the data and features. The final prediction is determined by averaging the individual predictions from all the trees. This ensemble learning approach helps mitigate the overfitting problem often associated with individual decision trees, where a model becomes too tailored to the training data, leading to poor performance on new, unseen data.

Random forests also enhance model stability by reducing variance, as the model benefits from the diversity of the different trees. Each tree is trained on a random subset of the data, and by combining these different perspectives, the random forest creates a more generalized model that captures a broader range of patterns and reduces the likelihood of error caused by noise or outliers in the data. Additionally, random forests naturally handle feature importance, providing insights into which features contribute most to the predictions, which is useful for feature selection and model interpretation. To assess the model's generalization capability and prevent it from overfitting, cross-validation was applied

The random forest model demonstrated impressive performance, achieving an MSE of 0.13 and an R-squared value of 0.86. These results indicate that the model not only has high accuracy but also explains a substantial portion of the variability in the data. Specifically, the R-squared value suggests that the model accounts for 86% of the variation in PGA values, which is a significant improvement compared to simpler models like linear or decision tree regression. This is a strong indicator of the model's ability to capture complex patterns in the data, especially in the context of seismic hazard analysis where relationships between features can be highly nonlinear and intricate.

As depicted in Figure 7, the scatter plot for predicted versus actual PGA values further reinforces the model's performance. The points almost closely follow the diagonal line, indicating that the predictions are closely aligned with the true values. This visual alignment highlights the model's better accuracy and reinforces its effectiveness in forecasting PGA values.

In comparison to individual decision tree models, random forests offer a clear advantage. A single decision tree tends to perform well on training data but can struggle to generalize new data due to overfitting. Random forests, by combining multiple trees, overcome this limitation and provide a more powerful and reliable approach, particularly for handling complex relationships in different tasks such as seismic hazard analysis. By leveraging the strength of multiple decision trees, the random forest model provides improved predictive power, making it an ideal choice for high-dimensional and noisy datasets typically encountered in earthquake engineering studies.

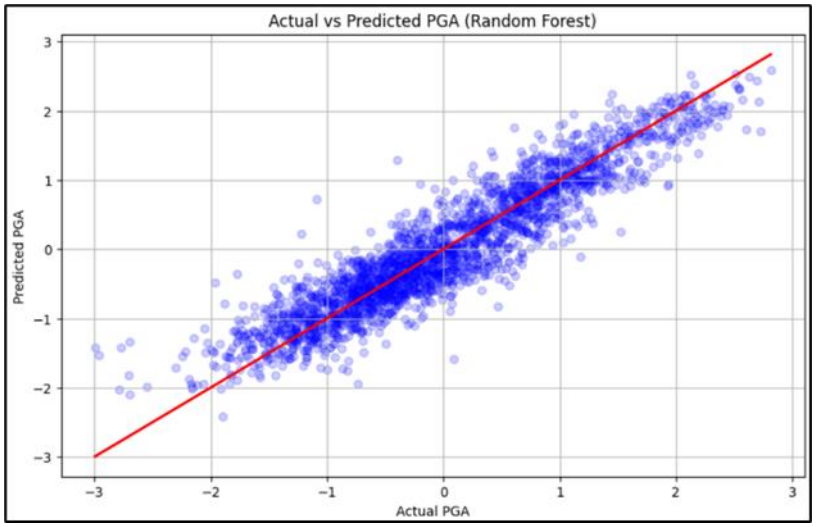

Figure 7 The Random Forest scatter plot of actual ln (PGA) and predicted ln (PGA)

3. Result and Discussion

In this study, machine learning techniques were implemented using Python, leveraging libraries such as scikit-learn for model creation and evaluation. Before creating the models, data preprocessing was carefully conducted to handle missing values, scale the features, and encode categorical variables, ensuring that the data was ready for effective model training.

To conclude, five different models—linear regression, lasso regression, polynomial regression, decision tree, and random forest—were assessed for PGA. Linear regression and Lasso regression, while simple and easy to interpret, had the lowest accuracy. Lasso regression, by adding regularization, performed slightly better than linear regression but still fell short in capturing the complexity of the data.

Polynomial regression improved the model by handling non-linear relationships, providing a better fit compared to the linear models. However, its performance suggests that there's still room for improvement. Fine-tuning higher-degree terms or using more sophisticated methods could help refine this approach.

The decision tree model, which is more flexible, did well initially but suffered from overfitting, meaning it performed excellently on training data but struggled with new, unseen data. This is a common issue where the model essentially memorizes the training data, which limits its ability to generalize. To overcome this, the random forest model was used, combining multiple decision trees to provide a more stable and accurate prediction.

Among all the models, random forest delivered the best results with the highest R^2 and the lowest MSE, making it the most reliable for predicting PGA. However, there is still room for improvement in refining the model further. While random forest performed the best overall, additional work like fine-tuning the parameters, adding more features, or exploring other ensemble methods like boosting could further enhance its accuracy.

In earthquake engineering, where seismic hazard assessments are crucial for designing safe structures and infrastructure, improving model accuracy is key. It is important to mention that the model used in this study was simplified by choosing limited parameters. To improve the accuracy of the model in future studies, several strategies can be considered. First, adding more relevant features, such as fault mechanism, and geo-graphical data, can enhance the model's predictive power. Feature engineering, such as creating interaction terms or polynomial features, may also help capture more complex relationships in the data. Leveraging domain-specific knowledge of earthquake engineering will guide the selection of critical features. Increasing the size of the dataset, either by collecting more real-world data or generating synthetic data, can improve model generalization. Hyperparameter tuning using techniques like grid search or random search will help optimize model performance. Additionally, exploring more complex models, such as gradient boosting methods (XGBoost, LightGBM, CatBoost) or neural networks, can capture intricate patterns that simpler models might miss. Ensemble methods, such as stacking or boosting, could also combine the strengths of various models for more accurate predictions. Implementing k-fold cross-validation ensures robust performance evaluation and prevents overfitting. Addressing data issues, such as handling missing values more effectively and detecting

outliers, will also contribute to improving model reliability. By applying these strategies, the model's accuracy and robustness in predicting peak ground acceleration can be significantly improved, leading to more reliable seismic hazard assessments in earthquake engineering. Better predictions lead to more effective site response analysis, ultimately helping to create buildings and infrastructures that can withstand earthquakes.

4. Conclusion

This study demonstrated the effectiveness of machine learning regression models in predicting peak ground acceleration (PGA) for seismic hazard analysis of earthquake engineering. By comparing both traditional and advanced algorithms, it was evident that ensemble methods, particularly the random forest model, significantly outperformed linear and polynomial regressions in terms of predictive accuracy and generalization. While simpler models offered interpretability, they struggled to capture the complex, nonlinear relationships inherent in seismic data. The random forest model, by leveraging the power of multiple decision trees, showed strong potential in improving the reliability of ground-motion models, especially for regions characterized by small-to-moderate earthquakes.

However, this work also highlights that further enhancements are possible. Expanding the feature set, applying more advanced ensemble techniques, and integrating domain-specific knowledge could yield even more robust models. As seismic hazard assessments form the foundation of earthquake-resistant design, improving model precision is not just a technical endeavor but a critical step toward safer infrastructure and resilient communities.

Compliance with ethical standards

Disclosure of Conflict of Interest:

The authors declare no conflict of interest.

Data Availability

All data used in this study were obtained from publicly available seismic databases and are properly cited within the manuscript.

References

[1] Abrahamson N, Silva W. Summary of the Abrahamson and Silva NGA ground-motion relations. Earthq Spectra. 2008;24(1):67–97.

[2] Boore DM, Joyner WB, Fumal TE. Equations for estimating horizontal response spectra and peak acceleration from western North American earthquakes: A summary of recent work. Seismol Res Lett. 1997;68(1):128–153.

[3] Douglas J. Earthquake ground motion estimation using strong-motion records: a review of equations for the estimation of peak ground acceleration and response spectral ordinates. Earth Sci Rev. 2003;61(1-2):43–104.

[4] Aminpour P, Najafizadeh J, Kamalian M, Jafari MK. Seismic Response of 2D Triangular-Shaped Alluvial Valleys to Vertically Propagating Incident SV Waves. J Seismol Earthq Eng. 2015;17(2):89–101.

[5] Najafizadeh J, Kamalian M, Jafari MK, Aminpour P. Seismic nonlinear behaviour of rectangular alluvial valleys subjected to vertically propagating incident SV waves using the spectral finite element method. In: Proceedings of the 7th International Conference on Seismology & Earthquake Engineering; 2015 May 18–21; Tehran, Iran.

[6] Najafizadeh J, Kamalian M, Jafari MK, Khaji N. Seismic analysis of rectangular alluvial valleys subjected to incident SV waves by using the spectral finite element method. Int J Civ Eng. 2014;12(3), Transaction B: Geotechnical Engineering.

[7] Jalili J, Moosavi M, Pakniat S. A newly generated seismic ground response analysis software package - SeisGRASP - by International Institute of Earthquake Engineering and Seismology. Iran J Sci Technol Trans Civ Eng. 2024;48:1467–1482.

[8] Pakniat S, Moosavi M, Jalili J. Effect of Seismic Site Response on Damage Distribution in Sarpol-e Zahab City Caused by 12 November 2017 Mw 7.3 Strong Ground Motion: Fooladi area. J Seismol Earthq Eng. 2021;23(3):11–24.

[9] Hemmati Khollari MR, Asadi A, Tajammolian H. Seismic fragility assessment of SMRFs equipped with TMD considering cyclic deterioration of members and nonlinear geometry. Buildings. 2023;13(6):1364.

[10] Bommer JJ, Dost B, Edwards B, Stafford PJ, van Elk J, Doornhof D, Ntinalexis M. Developing an application-specific ground-motion model for induced seismicity. Bull Seismol Soc Am. 2016;106(1):158–173.

[11] Campbell KW, Bozorgnia Y. NGA ground motion model for the geometric mean horizontal component of PGA, PGV, PGD and 5% damped linear elastic response spectra for periods ranging from 0.01 to 10 s. Earthq Spectra. 2008;24(1):139–171.

[12] Pezeshk S, Zandieh A, Campbell KW, Tavakoli B. Ground-motion prediction equations for central and eastern North America using the hybrid empirical method and NGA-West2 empirical ground-motion models. Bull Seismol Soc Am. 2018;108(4):2278–2304.

[13] Alidadi N, Pezeshk S. State of the art: Application of machine learning in ground motion modeling. Eng Appl Artif Intell. 2025;149:110534.

[14] Tufail SA. Investigation of data-driven flood forecasting models performance applied to the Müglitz river basin: regression trees [Master's thesis]; 2016.

[15] Tufail SA, Jazaei F, Bakhshaee A, Ashiq MM, Hamza M. Assessment of microplastic contamination in biosolids from wastewater treatment plants and its implications for terrestrial environments. AGU24. 2024 Dec 11

[16] Bakhshaee A, Jazaei F, Ashiq MM, Tufail SA, Ali AS. Microplastic identification and quantification using combined fluorescence microscopy and hotplate techniques. AGU24. 2024 Dec 11.

[17] Ashiq MM, Jazaei F, Bakhshaee A, Ali AS, Tufail SA. Investigating the transport behavior of low-density polyethylene microplastics in sandy aquifers. AGU24. 2024 Dec 11.

[18] Alidadi N, Pezeshk S. Ground-Motion Model for Small-to-Moderate Potentially Induced Earthquakes Using an Ensemble Machine Learning Approach for CENA. In preparation or submitted; 2024.

[19] Khosravikia F, Clayton P. Machine learning in ground motion prediction. Comput Geosci. 2021;148:104700.

[20] Sedaghati F, Pezeshk S. Machine learning–based ground motion models for shallow crustal earthquakes in active tectonic regions. Earthq Spectra. 2023;39(4):2406–2435.

[21] Soltani A. Exploring the interplay of foreign direct investment, digitalization, and green finance in renewable energy: advanced analytical methods and machine learning insights. Energy Convers Manag X. 2024.

[22] Farajpour Z, Pezeshk S. A ground-motion prediction model for small-to-moderate induced earthquakes for central and eastern United States. Earthq Spectra. 2021;37:1440–1459.

[23] NGA-West2 Project Website [Internet]. Berkeley: NGA-West2 Project; [cited 2025 Apr 18]. Available from: https://ngawest2.berkeley.edu/

[24] Brownlee J. Machine Learning Algorithms from Scratch with Python [Internet]. Machine Learning Mastery; [cited 2023 Jun 30]. Available from: https://machinelearningmastery.com/machine-learning-algorithms-fromscratch/