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

Ensemble machine learning models for predictive analysis: Application to seismic ground motion data

Parya Dolatyabi *

Department of Computer Science, The University of Tulsa, Tulsa, Oklahoma, USA.

World Journal of Advanced Research and Reviews, 2025, 27(01), 558-568

Publication history: Received on 07 May 2025; revised on 25 June 2025; accepted on 27 June 2025

Article DOI: https://doi.org/10.30574/wjarr.2025.27.1.2474

Abstract

Computer science has become an essential discipline for solving complex, data-intensive problems across the natural sciences. This study demonstrates how machine learning algorithms—especially ensemble methods such as stacking, random forest, and gradient boosting—can be used to build data-driven ground-motion models (GMMs) for predicting peak ground acceleration (PGA), a key parameter in seismic hazard assessment. The stacking approach integrates multiple base learners (linear regression, polynomial regression, decision tree, and random forest) with meta-models (linear regression, decision tree, or random forest) to enhance prediction accuracy. Random forest constructs an ensemble of decision trees, while gradient boosting sequentially refines residuals to minimize errors. Models are trained on over 10000 records from small-to-moderate earthquakes (Mw 3.5–5.8) with hypocentral distances up to 200 km. Predictor variables include moment magnitude (Mw), hypocentral distance (Hypo-D), rupture-top depth (Ztor), and average shear-wave velocity in the upper 30 m (VS30). Performance evaluation reveals that the stacked model with a linear-regression meta-model achieves the highest accuracy, underscoring the potential of ensemble learning for seismic hazard modeling.

Keywords: Machine learning; Regression; Ensemble methods; Gradient boosting; Random forest; Stacking; Seismic hazard assessment; Peak ground acceleration (PGA)

1. Introduction

In seismic hazard analysis, machine learning (ML) techniques have emerged as powerful tools for developing data-driven ground-motion models that can capture complex relationships among seismic parameters more effectively than traditional empirical methods. Ground-motion models (GMMs) are essential tools in seismic hazard analysis, contributing to the development of hazard maps, earthquake-resistant building codes, and strategies for risk reduction [1, 2, 3]. Among their primary applications, GMMs are widely used to predict peak ground acceleration (PGA), a crucial parameter for site response analysis and structural design. Significant research has been conducted to improve site response modeling in seismic analysis. For example, Najafizadeh et al. examined 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 [7, 8]. Tools like SEISGRASP highlight the growing role of advanced computational methods in refining seismic hazard assessments. Additionally, earthquake records and their characteristics are widely used in structural analysis, including seismic fragility assessments of buildings [9]. Ensemble methods have driven breakthroughs in intelligent transportation—evidenced by traffic-scene understanding [10]—and in biomedical engineering, from scaffold-based drug delivery to regenerative-medicine outcome prediction [11, 12, 13,14, 15, 16, 17, 18]. These cross-domain successes highlight ensemble learning's power to extract robust patterns from noisy, high-dimensional data, motivating its use here to improve seismic ground-motion modeling and PGA forecasting.

* Corresponding author: Parya Dolatyabi

GMMs are based on factors such as earthquake magnitude, source-to-site distance, and site-specific conditions. Traditional empirical GMMs often rely on existing functional forms to model ground motion parameters, as demonstrated by several seminal works [19, 20, 21]. While these models are effective, they rely on simplifying assumptions and may struggle to capture the complexity of ground motion phenomena, particularly in regions with limited data. Recent advances in ML have introduced powerful nonparametric alternatives to classical regression techniques. In previous work, ML models such as regression trees have been effectively used in flood forecasting, particularly in data-scarce regions [22]. 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 [23, 24, 25]. Unlike traditional empirical GMMs that rely on predefined functional forms [26], ML approaches such as artificial neural networks and random forest regressors can flexibly capture complex nonlinear relationships in high-dimensional seismic data. This flexibility is especially valuable for regions with sparse or heterogeneous earthquake records. Recent studies have demonstrated the effectiveness of these models in improving ground motion predictions [27, 28, 29, 30, 31, 32, 33, 34, 35].

ML techniques offer significant advantages for modeling complex systems in earthquake engineering. In seismic hazard analysis, these methods can process large, heterogeneous datasets to uncover nonlinear relationships that conventional empirical ground-motion models often fail to capture. For example, decision trees, support vector machines, and deep learning architectures [10] flexibly model diverse geological conditions and maintain robust performance when historical data are sparse or non-uniform. By continuously integrating new seismic records, these approaches refine ground motion predictions over time, supporting more reliable hazard assessments and risk mitigation strategies. Induced earthquakes pose additional challenges due to their shallow depths and distinct attenuation characteristics [2, 21], which traditional GMMs designed for tectonic events may not fully capture. ML provides a flexible framework for modeling induced seismicity by leveraging historical records to improve understanding of small-to-moderate magnitude events. For instance, Alidadi et al. and Farajpour et al. developed region-specific GMMs for induced earthquakes in Central and Eastern North America (CENA), offering valuable insights into their unique attenuation patterns [26, 28].

Previous studies have demonstrated correlations among the model parameters. A suite of regression techniques—linear regression, polynomial regression, decision trees, and random forests—was applied to develop parametric and non-parametric ground-motion models, providing a comprehensive comparison of their performance in earthquake engineering and PGA prediction. Building on these results, additional ensemble methods were evaluated to further enhance model accuracy and assess the impact of more advanced machine-learning approaches [29]. Unlike prior work by Alidadi and Pezeshk [26,28] and Pakniat et al. [29], which focused on single ensemble techniques or limited model comparisons, this study systematically evaluates multiple ensemble approaches—random forest, gradient boosting, and various stacking configurations—using the comprehensive and diverse NGA-West2 dataset. By comparing these methods side-by-side, this study provides new practical insights into selecting optimal ML techniques for predicting PGA in induced seismicity contexts. This is, to the best of the author's knowledge, the first study to demonstrate how stacking configurations can outperform single ensemble methods for data-driven ground-motion models.

In this study, according to the previous research, after comparing the different methods, the decision was made to focus on the one that performed best. The random forest regressor (RFR) was identified as the most reliable approach based on its consistent results and overall performance [29]. To further enhance GMM accuracy, the gradient boosting and stacking ensemble methods were implemented. This research enhances our understanding of ground motion prediction for induced seismicity and contributes to more accurate seismic hazard assessments. By utilizing moment magnitude (Mw), hypocentral distance (Hypo-D), and VS30 as predictor variables, the models aim to forecast PGA with improved reliability. This work contributes to the advancement of data-driven GMMs, addressing critical challenges in seismic hazard assessments for induced seismicity.

2. Material and methods

2.1. Ground Motion Database

The NGA-West2 database is a comprehensive and meticulously curated ground motion database developed as part of the Next Generation Attenuation (NGA) project [32]. This initiative aims to enhance ground motion prediction models specifically for shallow crustal earthquakes occurring in active tectonic regions, with a focus on Western North America, as well as other seismically active areas around the world. The database was compiled by the Pacific Earthquake Engineering Research Center (PEER), and it represents one of the most extensive and detailed datasets available for seismic hazard assessment. The NGA-West2 dataset includes instrument-corrected, median, and orientation-independent horizontal components of ground-motion intensity measures (GMIMs), specifically the RotD30 metric. The

RotD30 represents the 30th percentile of the response spectra across all nonredundant rotation angles, providing a robust and stable representation of ground motion intensity across various directions [29, 30].

In developing accurate ground motion prediction models, selecting appropriate input parameters and defining their ranges is crucial. The quality and reliability of these models depend significantly on these variables, as they influence both performance and the model's ability to represent the complex physical processes that govern seismic events. In this study, moment magnitude (Mw) (unitless), hypocentral distance (Hypo-D) (km), depth to the top of the rupture plane (Ztor) (km), and the time-averaged shear wave velocity in the top 30 meters (VS30) (m/s) are used as input features. These variables capture key aspects of the earthquake source and local site conditions, which affect the amplification of seismic waves at the surface. Together, they characterize the physical properties that are critical for predicting ground motion. The model's output is defined as the horizontal component of peak ground acceleration (PGA), expressed in units of g, which directly correlates with the potential for structural damage during an earthquake and is a key parameter in seismic hazard analysis. Moment magnitude (Mw) is a unitless measure that quantifies the size of an earthquake, representing the total energy released during the seismic event. Hypocenter distance (Hypo-D), measured in kilometers, indicates the distance from the seismic source to the observation site, which directly affects the intensity of the ground motion. Depth to the top of the rupture plane (Ztor), also measured in kilometers, captures the depth of the earthquake source, which influences the attenuation of seismic waves as they propagate through the Earth's crust. VS30 is the time-averaged shear wave velocity in the upper 30 meters of the Earth's surface, measured in meters per second. It is a key site-specific parameter that affects the amplification of ground shaking, with higher velocities generally indicating stiffer soil conditions that attenuate seismic waves more effectively. PGA serves as the observed data and the output in our model. PGA is a critical parameter in seismic hazard assessment as it represents the maximum acceleration experienced at the Earth's surface during an earthquake, directly correlating with the potential for damage to structures and infrastructure.

Utilizing these parameters as input features enables the development of a robust model capable of accurately predicting PGA values under various seismic conditions. By integrating the NGA-West2 database with these selected input parameters, ground motion prediction models are refined and enhanced, leading to improved seismic hazard assessments and risk mitigation strategies in regions prone to shallow crustal earthquakes.

2.2. Data Processing

Preliminary data analysis of the recorded ground motions suggests that the logarithms of GMIMs (ground motion intensity measures) are more effectively captured when using the logarithms of both Hypo-D and VS30 as input variables. This observation is based on the fact that these variables exhibit significant positive skewness, with distributions that are not symmetric and have longer right tails. The output variable, PGA, shows similar skewness. To address this and better capture the relationships among variables, Hypo-D, VS30 and PGA are transformed using the natural logarithm—yielding ln(Hypo-D), ln(VS30) and ln(PGA)—which helps normalize the data, reduce the impact of outliers and extreme values, and improve model performance [29].

The input parameters are continuous variables, meaning they are numerical features that can take on a wide range of values. Most ML algorithms, with the exception of tree-based models like decision trees and random forests, are sensitive to unscaled numerical features. When data is not properly scaled, it can result in significantly slower training times and hinder the convergence of gradient-based algorithms. This is particularly true for models such as linear regression and support vector machines, which rely on optimization methods that are sensitive to the scale of the input data [37]. In the case of unscaled data, features with larger numerical values could disproportionately influence the model, leading to biased or inefficient models. For example, if one feature, such as Hypo-D, has a much larger numerical range than another feature, such as M, the model may give excessive importance to Hypo-D simply because its numerical range is wider, even if M is just as important for prediction. To mitigate these issues and ensure the ML models are trained efficiently, all input parameters were standardized before feeding them into the ML algorithm. Standardization is the process of rescaling the features of a dataset so that each feature has approximately a mean of 0 and a standard deviation of 1. This ensures that all the input features are on a similar scale, preventing any one feature from dominating the learning process due to its larger numerical values [29]. Standardizing the data using Equation (1) ensures that optimization algorithms converge more quickly and avoids the pitfalls of unscaled inputs. Overall, feature transformation and standardization are critical preprocessing steps that enable ML models to learn effectively from the data. To implement this, each feature is standardized according to the following equation:

$$x' = \frac{\mathrm{x} - \mu}{\sigma}, \qquad \ldots\ldots (1)$$

where x′ is the standardized value, x is the original value, μ is the mean of the data, and σ is the standard deviation of the data.

In machine learning, dividing a dataset into training and test sets is essential for developing robust, reliable models. The training set enables the model to learn underlying patterns, while the test set evaluates its performance on unseen data—revealing issues such as overfitting or underfitting. To prevent inadvertent bias, the data should be shuffled prior to splitting, ensuring that any inherent ordering or grouping does not influence learning. In this study, the shuffled dataset was partitioned into an 80 percent training set and a 20 percent test set, providing ample data for model development while retaining sufficient hold-out samples for validation. By combining proper shuffling with an appropriate split ratio, the resulting models achieve more accurate, generalizable predictions of PGA, thereby strengthening seismic hazard assessments.

2.3. Machine Learning Models

The estimation of GMIM is approached as a regression problem, where GMM is represented as follows:

$$\ln(\text{PGA}) = \text{function}[\text{M}, \ln(\text{Hypo} - \text{D}), \text{Ztor}, \ln(\text{V}_{\text{s30}})] \ldots\ldots \quad (2)$$

A combination of linear and polynomial regression models, in conjunction with the bagging (bootstrap aggregation) technique, was employed to investigate the relationships between the dependent and independent variables. Regression analysis facilitated the evaluation of these relationships, while bagging enhanced model robustness and predictive accuracy by reducing variance [38, 39]. It is important to emphasize that the earthquake data used in this study are real earthquake data records rather than synthetic or simulated data. Since these earthquake records are naturally occurring and not artificially generated, the true relationship between the input parameters and PGA is unknown. This inherent uncertainty necessitates the exploration of multiple ML techniques to effectively capture the complex interactions among these parameters. By testing different models, we aim to identify the most effective approach for predicting PGA and understanding the underlying data patterns. In a recent study, linear regression, lasso regression, polynomial regression, decision tree and random forest were implemented, and their results showed a consistent improvement in the results [29]. To follow their procedure and develop better models, the random forest and two other ensemble models were developed. ML algorithms were implemented in Python using NumPy for numerical processing, Pandas for data preprocessing, scikit-learn for model development, training, testing, and evaluation, and Matplotlib and Seaborn for visualizing and plotting results.

2.3.1. Ensemble Machine Learning Models

Ensemble models are powerful ML techniques that combine the predictions of multiple individual models to improve overall performance and robustness. By aggregating the strengths of various models, ensemble methods can effectively reduce overfitting, improve prediction accuracy, and enhance model stability. Techniques such as bagging, boosting, and stacking are widely used ensemble approaches, each employing distinct strategies to refine predictions.

In this study, three ensemble models, random forest, gradient boosting, and stacking, were employed to assess their effectiveness in predicting PGA and to compare their performance with individual models. By combining multiple models, ensemble methods often outperform standalone models, particularly when dealing with complex, high-dimensional data like seismic parameters. Comparing individual and ensemble models provides valuable insights into the trade-offs between model simplicity, interpretability, and predictive performance, guiding the selection of the most effective approach for seismic hazard assessment.

2.3.2. Model Evaluation

To rigorously assess predictive performance, quantitative metrics—Mean Squared Error (MSE) to measure average squared prediction error and R^2 to quantify explained variance on the ln(PGA) scale—are combined with a graphical diagnostic. MSE measures the average squared difference between the actual and predicted values. Smaller MSE values indicate better model performance. N, y and $\hat{y}$ represent the number of values, actual values and predicted value, respectively (equation 3).

$$Mean\ Squared\ Error = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \ \ldots\ldots\ldots \quad (3)$$

R-squared (R^2) quantifies the proportion of variance in the dependent variable which is explained by the model. An R^2 value close to 1 indicates strong predictive performance, while negative values suggest the model performs worse than simply predicting the mean value. In the formula, y, $\bar{y}$ and $\hat{y}$ represent actual value and average of actual values and predicted value, respectively (equation 4).

$$R^2_{score} = 1 - \frac{(y_i - \hat{y}_i)^2}{(y_i - \bar{y}_i)^2} \dots\dots\dots (4)$$

Since the model is high dimensional with 4 variables, to reduce the dimensionality and further assess model performance, natural logarithm of predicted PGA values are plotted against natural logarithm of observed values on a 2D scatter plot. An x=y reference line is included to represent perfect predictions. Data points clustering closely around this line indicate stronger model performance, suggesting that the model is effectively capturing the relationship between input variables and PGA. The scatter plot provides a simplified yet insightful way to evaluate model accuracy, albeit in reduced dimensions. By combining these evaluation metrics and visual analysis, we aim to compare model performance and identify the most suitable method for predicting PGA based on real earthquake records. The following sections provide a detailed explanation of each ML model and its implementation.

3. Analysis and Results

3.1. Random Forest

The random forest model is a robust and highly effective ML 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. By training each tree on a random subset of data, the model benefits from the diversity of the different trees. Combining these different perspectives results in 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. This is particularly useful for feature selection and model interpretation. To assess the model's generalization capability and reduce overfitting, cross-validation was applied.

The random forest model demonstrated impressive performance, achieving an MSE of 0.1409 and an R-squared value of 0.8573. These results indicate that the model not only provides 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 85% of the variation in PGA values, marking a significant improvement compared to simpler models like linear regression or decision tree regression. This performance underscores the model's ability to capture complex patterns in the data, which is crucial in seismic hazard analysis, where relationships between features can be highly nonlinear and intricate.

As shown in Figure 1, the scatter plot of predicted versus actual PGA values further reinforces the model's performance. The points closely align with the diagonal line, indicating that the predictions are well-aligned with the true values. This visual alignment highlights the model's accuracy and its effectiveness in forecasting PGA values.

In comparison to individual decision tree models, random forests offer a clear advantage. Single decision trees tend to perform well on training data but struggle to generalize to new data due to overfitting. By combining multiple trees, random forests overcome this limitation, providing a more powerful and reliable approach, especially for complex tasks like seismic hazard analysis. Leveraging the strength of multiple decision trees, the random forest model offers improved predictive power, making it an ideal choice for high-dimensional and noisy datasets, which are common in earthquake engineering studies [29].

3.2. Gradient Boosting

Gradient boosting regression is a powerful technique that builds predictive models through an iterative process. It begins by fitting a simple model to provide an initial prediction. The algorithm then calculates the residuals — the differences between the actual values and the predicted values — which represent the errors the model needs to correct. In the next step, a new model is trained specifically to predict these residuals. This new model's predictions are then combined with the previous model's predictions, with the combination controlled by a learning rate parameter to prevent drastic changes. This process of sequentially training new models to minimize previous errors continues until

the model achieves satisfactory performance. The final prediction is obtained by aggregating the outputs of all the individual models, effectively improving accuracy and capturing complex patterns in the data.

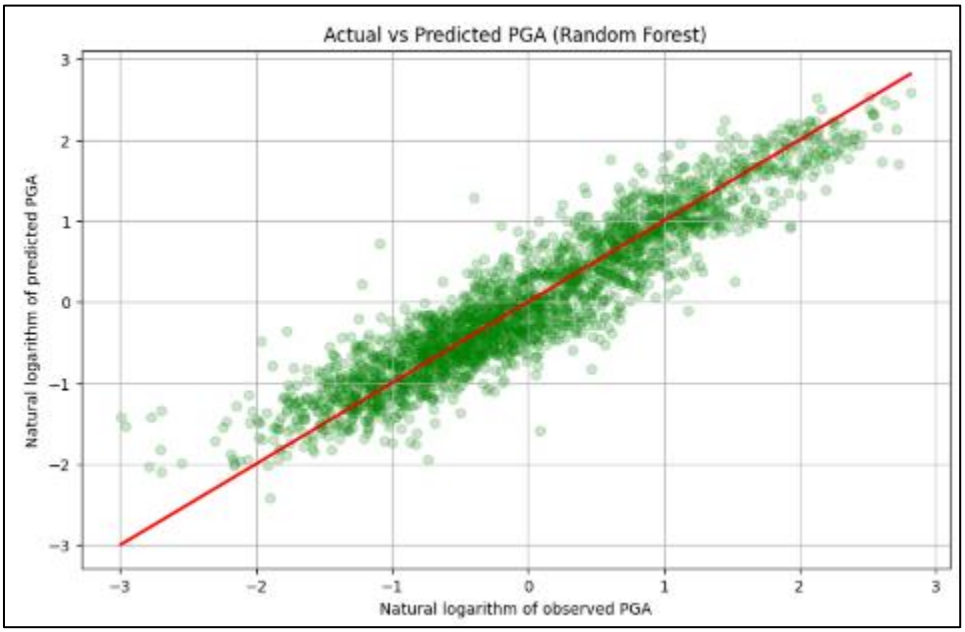

Figure 1 Random forest scatter plot of the observed ln(PGA) versus the predicted ln(PGA)

In this study, gradient boosting regression from the scikit-learn library in Python was employed to model the relationship between seismic parameters and PGA values. By progressively correcting prediction errors, this method effectively captured nonlinear patterns and complex dependencies. The model's performance was compared against individual models and other ensemble techniques to evaluate its effectiveness in improving predictive accuracy for seismic hazard analysis.

The gradient boosting regression model demonstrated strong performance, achieving an MSE of 0.138 and an R-squared value of 0.857. These results indicate that the model effectively captures the relationship between seismic parameters and PGA values, providing high predictive accuracy. The R-squared value suggests that the model explains approximately 85% of the variation in PGA values, marking a notable improvement compared to simpler models like linear regression and decision tree regression. This enhanced performance highlights the model's ability to capture complex, nonlinear patterns, making it particularly valuable in seismic hazard analysis where intricate dependencies between variables are common.

As shown in Figure 2, the scatter plot of predicted versus actual PGA values illustrates the model's effectiveness. Data points almost follow the diagonal reference line, demonstrating that the predicted values align well with the true observations. This visual alignment further confirms the model's reliability in forecasting PGA values and underscores the advantages of gradient boosting regression in improving predictive accuracy for seismic hazard assessment.

3.3. Stacking

Stacking, or stacked generalization, is an ensemble learning method that improves prediction accuracy by combining the outputs of multiple base models. In stacking, each base model is trained on the same dataset, and their predictions are used as features for a meta-model, which learns how to best combine these predictions to generate a final output. The advantage of stacking lies in its ability to leverage the strengths of multiple models, capturing different patterns and relationships in the data, ultimately leading to more robust and reliable predictions.

In this study, we applied stacking to model the relationship between seismic parameters and Peak Ground Acceleration PGA values. The base models selected for this study included linear regression, decision tree, random forest, and polynomial regression. Initially, a decision tree was used as the meta-model, resulting in an MSE of 0.277 and an R-squared of 0.715. Although this model captured some of the underlying trends, its performance was further enhanced by replacing the decision tree with a random forest as the meta-model. This change improved the model's performance significantly, achieving an MSE of 0.143 and an R-squared of 0.852. The final meta-model used was linear regression,

which achieved an impressive MSE of 0.128 and an R-squared of 0.868. These results highlight the progression in model performance as the meta-model evolves, illustrating how stacking allows for the combination of different base models to enhance predictive accuracy.

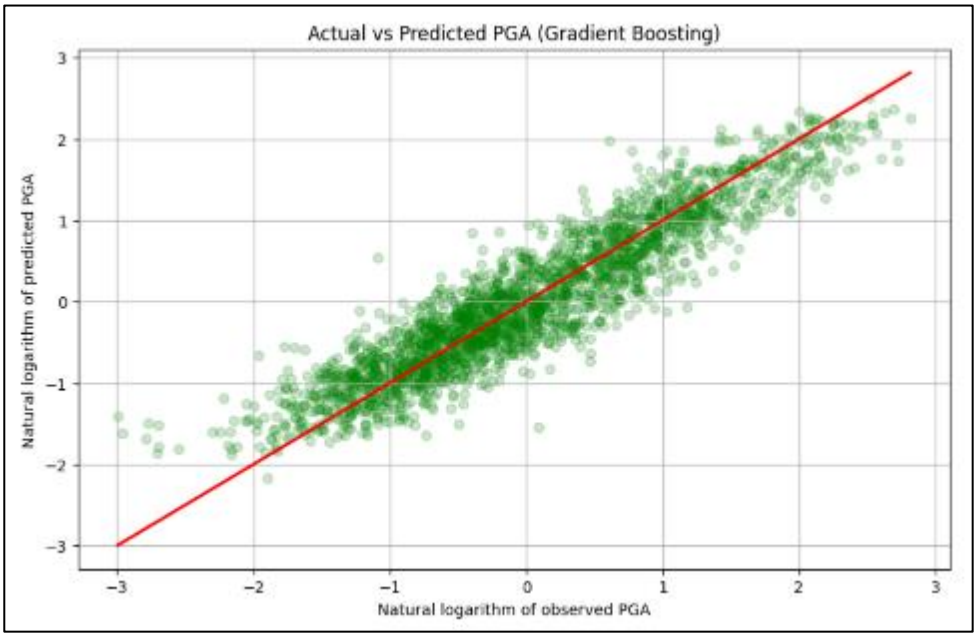

Figure 2 Gradient Boosting scatter plot of the actual ln(PGA) and the predicted ln(PGA)

Figures 3, 4, and 5 show the scatter plots for each of these analyses, providing a visual representation of the model's performance. These visualizations emphasize the improvements in model accuracy as stacking is applied, confirming the value of using an ensemble approach in predicting seismic hazard data. By using different meta-models, this study demonstrates how stacking can refine predictions and improve model performance. The varying results from each meta-model—from decision tree to random forest to linear regression—showcase how stacking can boost predictive accuracy in seismic hazard analysis.

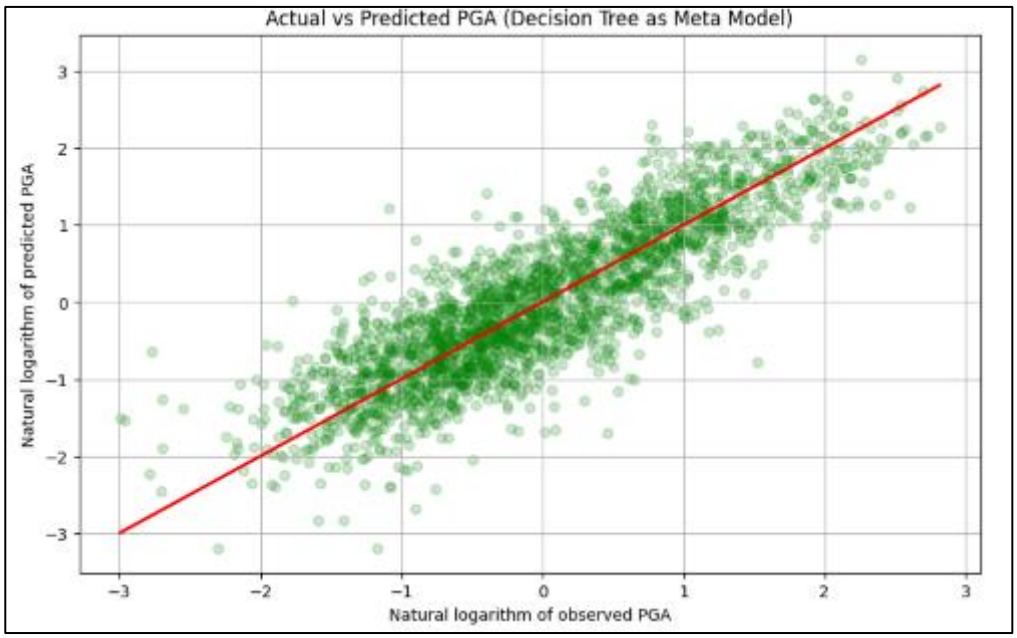

Figure 3 Stacking-Decision Tree as meta model scatter plot of the actual ln(PGA) and the predicted ln(PGA)

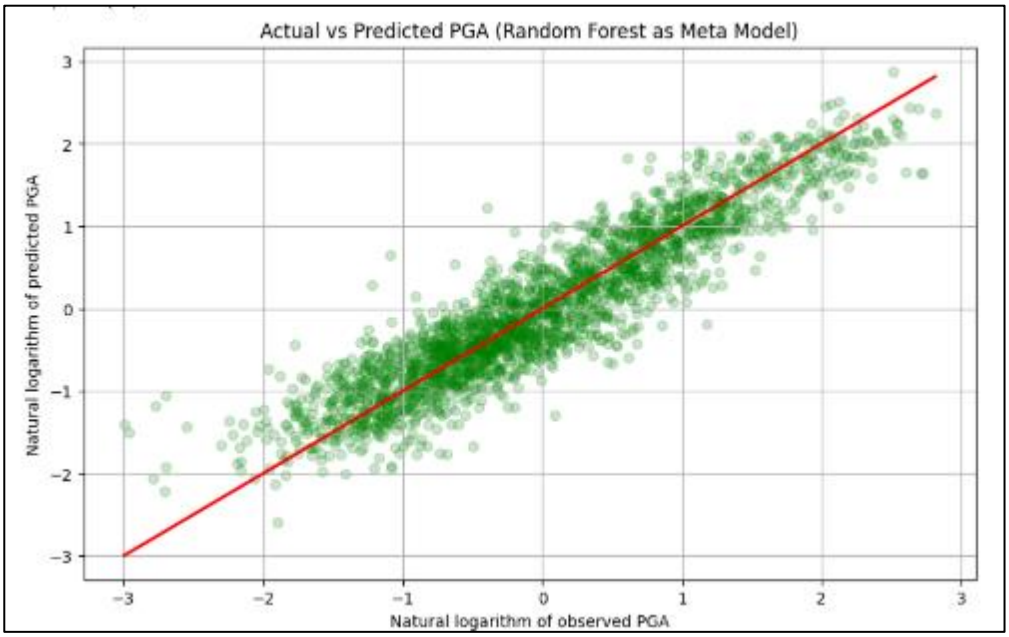

Figure 4 Stacking-Random Forest as meta model scatter plot of the actual ln(PGA) and the predicted ln(PGA)

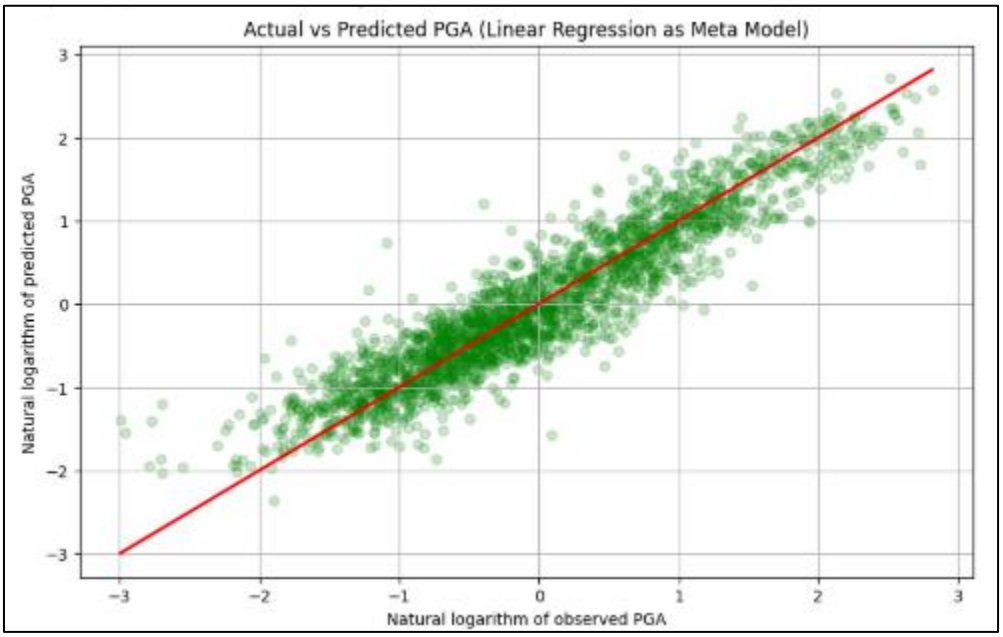

Figure 5 Stacking-Linear Regression as meta model scatter plot of the actual ln(PGA) and the predicted ln(PGA)

4. Discussion

The results from all the models evaluated in this study reveal varying levels of performance in predicting PGA values, showcasing the strengths and limitations of each approach. By comparing individual ML models [29] to ensemble methods like stacking and boosting, we can better understand the trade-offs between simplicity, interpretability, and predictive accuracy.

Individual Models: Linear regression, while simple and interpretable, only captured linear relationships between seismic parameters and PGA, leaving a significant portion of the variability in the data unexplained. Polynomial regression, by adding nonlinear terms, improved the model's performance but still couldn't account for all the

complexity in the relationships. Decision trees, though able to capture nonlinear patterns, suffered from overfitting, performing well on the training set but poorly on unseen data [29].

Ensemble Methods: Random forests, an ensemble of decision trees, performed better by reducing overfitting and capturing complex patterns, providing a more generalized model that was well-suited for seismic hazard analysis. Gradient boosting, an ensemble method that builds trees sequentially to correct errors made by previous trees, demonstrated strong performance by progressively improving its predictions. It was particularly effective in capturing complex, nonlinear relationships in the data. Stacking, which combined multiple base models to make predictions, showed further improvements. Using a decision tree as the meta-model yielded moderate performance, while switching to random forest as the meta-model significantly enhanced predictive power. The best results were obtained when linear regression was used as the meta-model, suggesting that combining base models with an appropriate meta-model can significantly boost performance.

A comparison of the results with previous studies reveals a correlation between the implemented methods and the findings of earlier research [26, 29, 31]. Alidadi and Pezeshk [26, 28] identified gradient boosting regression as the most effective model for predicting PGA, which aligns with the strong performance observed in this study. Similarly, Sedaghati and Pezeshk [31] found that random forest regression yielded the highest R-squared value among their individual models, a result consistent with the findings here. Also, Pakniat et al. found that the random forest has the best performance on the modeling of studied dataset. Additionally, Pakniat et al. [29] suggestion about improving model accuracy was employed in this study and the results indicated an improvement of R-squared evaluation. This comparison highlights that the outcomes of this study are in line with previous research, further validating the effectiveness of these ML techniques for seismic hazard analysis. Compared to Alidadi and Pezeshk [26, 28] and Pakniat et al. [29], who focused primarily on single ensemble methods or smaller, region-specific datasets, this study demonstrates that systematically comparing multiple ensemble techniques on a larger, more diverse ground motion database can reveal important performance differences. This finding confirms that stacking with different meta-models can achieve higher predictive accuracy than random forest or gradient boosting alone, offering practical insights for selecting ensemble strategies. This extends previous work and, to the best of the author's knowledge, represents the first demonstration of this advantage for induced seismicity hazard assessment.

For future studies, it is recommended to incorporate additional seismic and geological features, explore deep learning models for capturing complex relationships, and investigate feature engineering techniques like dimensionality reduction and feature selection. Further, experimenting with different ensemble methods and applying the models to diverse earthquake datasets could improve model performance and assess their generalizability across different seismic environments.

5. Conclusion

In conclusion, this research demonstrated the effectiveness of various ML models, including ensemble methods like random forest, gradient boosting, and stacking. Among these, linear regression as a meta-model in the stacking approach provided the best performance. The results of the ensemble models outperformed individual models that were studied in a recent study [28], particularly in terms of capturing complex patterns and improving predictive accuracy for seismic hazard analysis. Our findings aligned with previous studies highlighted the success of gradient boosting regression and regression for predicting PGA. Future studies could explore incorporating additional features, using more advanced ensemble techniques, or further refining the meta-model to enhance performance. Overall, this research demonstrates that systematically comparing multiple ensemble ML techniques can yield valuable insights for seismic hazard analysis. In particular, the findings show that stacking with carefully selected meta-models can outperform single ensemble methods like regression or gradient boosting, improving the predictive accuracy of data-driven ground-motion models. To the best of the author's knowledge, this is the first study to demonstrate this advantage using the comprehensive NGA-West2 dataset for induced seismicity hazard assessment.

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] Dolatyabi P, Regan J, Khodayar M. Deep learning for traffic scene understanding: A review. *IEEE Access*. 2025;13:13187–237. doi:10.1109/ACCESS.2025.3529289.

[11] Saberian E, et al. Application of scaffold-based drug delivery in oral cancer treatment: A novel approach. *Pharmaceutics*. 2024 Jun 1;16(6):802. Available from: https://doi.org/10.3390/pharmaceutics16060802

[12] Mohammadinezhad F, et al. Preparation, characterization and cytotoxic studies of cisplatin-containing nanoliposomes on breast cancer cell lines. *Asian Pac J Cancer Biol*. 2023 Jul 30;8(2):155–9. Available from: https://doi.org/10.31557/apjcb.2023.8.2.155-159

[13] Jalili Sadrabad M, et al. Success in tooth bud regeneration: A short communication. *J Endod*. 2024 Mar 1;50(3):351–4. Available from: https://doi.org/10.1016/j.joen.2023.12.005

[14] Arabmoorchegani M, Abbasi M, Asadalizadeh M, Motavaf F. Integrative cancer care: Leveraging nutrition and positive psychology for optimal outcomes. *Asian Pac J Cancer Nutr*. 2025. Available from: https://doi.org/10.31557/apjcn.1796.20250504

[15] Moravedeh R, Samadnezhad MZ, Asadalizadeh M, Abbasi M, Nadaki A. Enhanced anticancer potential of curcumin-loaded liposomal nanoparticles in oral cancer treatment. *Asian Pac J Cancer Biol*. 2025 May 7;10(2):293–9. Available from: https://doi.org/10.31557/apjcb.2025.10.2.293-299

[16] Saberian E, et al. Applications of artificial intelligence in regenerative dentistry: Promoting stem cell therapy and the scaffold development. *Front Cell Dev Biol*. 2024 Dec 6;12:1497457. Available from: https://doi.org/10.3389/fcell.2024.1497457

[17] Saberian E, et al. Combination therapy of curcumin and cisplatin encapsulated in niosome nanoparticles for enhanced oral cancer treatment. *Indian J Clin Biochem*. 2024 Nov 11. Available from: https://doi.org/10.1007/s12291-024-01279-9

[18] Hoseinifar, M.J., Aghaz, F., Asadi, Z. et al. Facilitating DNAzyme transport across the blood-brain barrier with nanoliposome technology. Sci Rep 15, 18914 (2025). https://doi.org/10.1038/s41598-025-04433-2

[19] 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.

[20] 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.

[21] 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.

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

[23] 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

[24] 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.

[25] 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.

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

[27] 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.

[28] 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.

[29] Pakniat S, Najafizadeh J, Kadkhodaavval M. Machine learning for earthquake engineering analysis: Comparing regression models to predict peak ground acceleration. *World J Adv Res Rev*. 2025;26(2):856–67. doi:10.30574/wjarr.2025.26.2.1714

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

[31] 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.

[32] Soltani A, Imani MA, Overcoming implementation barriers to renewable energy in developing nations: A case study of Iran using MCDM techniques and Monte Carlo simulation. *Results Eng*. 2024;24:103213. doi:10.1016/j.rineng.2024.103213.

[33] Safary A, Shafieasl H, Mitani J. Parameterized folded state shape modeling of David Huffman's ellipse. J Geom Graph. 2025;28(2):199-212

[34] Safary A, Shafieasl H, Mitani J. Geometric design tool for One-Fold, a curved origami with a single fold. J Geom Graph. 2024;28(1):89-101. Copyright Heldermann Verlag; 2024

[35] Pourmahmood-Aghababa H, Sattari MH, Shafieasl H. Bounded pseudo-amenability and contractibility of certain Banach algebras. Filomat. 2020;34(5):1701-1712

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

[37] 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/

[38] Soltani A. Exploring the interplay of foreign direct investment, digitalization, and green finance in renewable energy: Advanced analytical methods and machine learning insights. *Energy Conversion and Management: X*. 2024;24:100802. doi:10.1016/j.ecmx.2024.100802

[39] Naseri Baygi SM, Rezazadeh Eidgahi D. A Highly Efficient Optimized Artificial Neural Network for Predicting Deformation of Geogrid-Reinforced Soil Walls. In: *Proceedings of the 15th International Conference on Transportation and Traffic with an Approach to Artificial Intelligence in Civil Engineering*; 2023. p. 8. CIVILICA.