Received: 24 July 2025 | Revised: 25 November 2025 | Accepted: 5 December 2025 | Published online: 17 December 2025

RESEARCH ARTICLE

Archives of Advanced Engineering Science
2025, Vol. 00(00) 1–9
DOI: 10.47852/bonviewAAES52026930

BON VIEW PUBLISHING

Accurate Prediction of Peak Ground Acceleration Using Random Forests: A Data-Driven Approach with PEER Updated NGA-WEST2 Ground Motion Records

Asadullah Ziar[1,*] and Ender Basari[2]

[1]*Department of Civil Engineering, Ghazni Technical University, Afghanistan*

[2]*Department of Civil Engineering, Manisa Celal Bayar University, Türkiye*

Abstract: Accurate prediction of peak ground acceleration (PGA) is crucial for seismic hazard assessment and earthquake-resistant design. Traditional regression-based ground motion prediction equations often fall short in capturing the complex, nonlinear interactions among earthquake parameters. This study proposes a machine learning approach using a random forest (RF) model to predict PGA based on five key input variables: moment magnitude (Mw), closest distance to the rupture plane (ClstD), hypocentral depth, rake angle, and average shear wave velocity in the top 30 meters (Vs30). A comprehensive dataset of 16,211 ground motion records from the Pacific Earthquake Engineering Research Center (PEER), Updated NGA-WEST2 Flatfile was used. The RF model was optimized through Grid Search and 5-fold cross-validation, achieving high predictive performance with R^2 scores of 0.948 on training data and 0.928 on test data. Feature importance analysis indicated Mw and ClstD as the most influential parameters. The study demonstrates the robustness, accuracy, and generalization capability of the RF model, confirming its potential as a valuable tool in seismic risk analysis and providing a foundation for future development of more adaptable, data-driven models in earthquake engineering.

Keywords: peak ground acceleration, random forest, Pacific Earthquake Engineering Research (PEER), machine learning, ground motion prediction equations (GMPEs)

1. Introduction

An earthquake is a natural phenomenon that cannot be prevented by human intervention and represents a significant natural hazard, posing substantial risks to human life and property. The intensity and potential for damage become more critical in proximity to populated regions, particularly those situated within tectonically active zones. When an earthquake occurs, waves of energy rapidly travel from the earthquake's center through the ground. When these waves reach the surface, they cause the ground to shake, sometimes for just a few seconds, and other times for several minutes. The strength and duration of the shaking depend on the earthquake's magnitude, the distance from the source, and the local ground conditions. If the area is close to the hypocenter or the rupture plane of the fault, the shaking can be extremely strong and cause severe damage. Designing earthquake-resistant structures requires accurate prediction of strong ground motions and their impact on the built environment. One of the fundamental components of seismic risk assessment is estimating the expected ground motion parameters at a given distance from an earthquake of known magnitude [1]. Generally, these parameters are typically grouped into two broad categories: time-domain and response-domain measures. The time-domain set encompasses quantities such as peak ground acceleration (PGA), peak ground velocity (PGV), and peak ground displacement (PGD), which describe the maximum amplitudes of motion irrespective of structural behavior. Owing to their independence from the dynamic properties of structures, these parameters remain the most commonly adopted indicators in practical earthquake engineering applications [2].

Among these parameters, peak ground acceleration is the most widely employed indicator in seismic hazard assessments [3]. PGA indicates the maximum acceleration of the ground during an earthquake, which is the most widely used parameter in seismic hazard assessments [4].

Accurate prediction of peak ground parameters, especially PGA, is a cornerstone of seismic hazard assessment and earthquake engineering, as it directly informs structural design, risk mitigation, and disaster preparedness [5]. It is typically estimated using attenuation relationships; these relationships predict PGA based on factors such as earthquake magnitude, the distance between the earthquake source and the site, local ground conditions, the nature of the fault (e.g., strike-slip, normal, or reverse), and how seismic waves travel through the Earth [1, 6].

The aforementioned attenuation relationships, known as ground motion prediction equations (GMPEs), are empirically

***Corresponding author:** Asadullah Ziar, Department of Civil Engineering, Ghazni Technical University, Afghanistan. Email: asadullah.ziar@teug.edu.af

derived through regression analyses and have served as the primary predictive framework in seismic hazard studies for several decades [7, 8].

Despite their widespread use, GMPEs are inherently constrained by their reliance on log-linear functional forms and simplifying statistical assumptions, which often fail to capture the highly nonlinear and multidimensional relationships among seismic predictors. These models average data across regions and assume normally distributed residuals, thereby overlooking critical local variations in soil stratification, damping, and geological heterogeneity. As a result, GMPEs frequently yield biased predictions when applied outside their calibration regions, particularly in areas with strong site effects or complex stratigraphy, leading to systematic under- or overestimation of PGA values [9, 10]. Early regression-based efforts [7, 11–12] contributed to the development of GMPEs, but their limited flexibility andoversimplification of earthquake processes highlight the need for more advanced approaches.

In response to these limitations, machine learning (ML) algorithms have gained increasing attention in earthquake engineering. ML methods are capable of modeling nonlinear, high-dimensional relationships in seismic data, making them suitable for capturing the complex interactions between earthquake source, path, and site parameters. They have been successfully applied in diverse earthquake-related tasks, including seismic signal processing, earthquake detection, structural damage assessment, and ground motion modeling [13–22]. Recent studies emphasize ML's potential to overcome the limitations of GMPEs, providing more accurate and flexible predictions of ground motions, particularly PGA, when combined with large ground motion datasets [2, 6, 23].

The present study addresses this gap by integrating modern ML with the Pacific Earthquake Engineering Research Center (PEER), Updated NGA-WEST2 database [24]. A random forest (RF) algorithm is developed to predict PGA using five key input parameters: moment magnitude (Mw), closest distance to the rupture plane (ClstD), hypocenter depth (D, km), rake angle (RA, °), and average shear wave velocity in the upper 30 meters (Vs30, m/s). By leveraging both the nonlinear modeling capabilities of RF and the comprehensive NGA-West2 dataset, this study aims to establish a more reliable framework for PGA prediction.

The remainder of this paper is organized as follows: Section 2 reviews related studies, Section 2.1 outlines the dataset and the preprocessing steps used for model development, while Section 2.1.1 describes the ground motion parameters employed in this study. Section 3 introduces the overall research methodology, with Sections 3.1 and 3.2 focusing on the development and implementation of the RF ML model. Section 4 presents the results and discussion of the model's performance, and finally, Section 5 concludes the study with a summary of key findings and recommendations for future research.

2. Related Studies

Numerous studies have been carried out to predict peak ground motion parameters during earthquakes using both traditional and advanced computational approaches. Güllü and Erçelebi (2007) employed an artificial neural network (ANN) model based on the Fletcher–Reeves conjugate gradient back-propagation algorithm to estimate PGA values from Turkish strong-motion records. The model utilized earthquake magnitude, source-to-site distance, and local site conditions as inputs, with PGA as the sole output parameter. The obtained correlation coefficients of 0.9801, 0.9923, and 0.9637 across three datasets confirmed the ANN's ability to capture nonlinear dependencies among seismic predictors. Compared with traditional regression methods, the ANN yielded superior predictive accuracy, and the use of a single hidden layer with the Fletcher–Reeves method further improved its performance. These results demonstrate that ANN-based models can effectively represent attenuation characteristics and enhance the understanding of seismic parameter influence on PGA [3].

Similarly, Derakhshani and Foruzan (2019) developed deep neural network (DNN) models to estimate the three key time-domain parameters of ground motion such as PGA, PGV, and PGD using data from the comprehensive NGA-West2 database of the PEER. By incorporating multiple hidden layers and neurons, the DNN framework captured more complex nonlinear relationships than conventional ANNs. The models achieved correlation coefficients of 0.902, 0.899, and 0.911 for PGA, PGV, and PGD, respectively, outperforming earlier soft computing techniques. Moreover, lower root mean square error (RMSE), mean absolute error (MAE), and mean absolute percentage error values confirmed their enhanced accuracy and generalization capability [25].

In a more recent study, Mandal and Mandal (2024) utilized the XGBoost algorithm, a supervised ML technique, to predict PGA in the Kachchh region of Gujarat, India. Using inputs such as earthquake magnitude, focal depth, epicentral distance, and Vs30, the model, which was trained on 244 data points, achieved excellent performance (R = 0.994 for the full dataset and R = 0.844 for the test set) [26].

Furthermore, Javan-Emrooz et al. (2018) focused on vertical ground motion components, addressing a major research gap in Iranian seismic studies, which had previously concentrated primarily on horizontal parameters. Analyzing 463 three-component records from 107 events (Mw 4.5–7.4, ≤100 km) recorded between 1976 and 2016, they applied Prefix Gene Expression Programming to derive GMPEs for PGA, PGV, and PGD without assuming predefined regression forms. Their findings highlighted the potential of symbolic ML in developing flexible and data-driven attenuation models [4].

In summary, the reviewed literature illustrates a clear evolution from traditional regression-based GMPEs toward ML and deep learning methodologies. Techniques such as ANN, DNN, and XGBoost have consistently shown superior capability in modeling the nonlinear and multidimensional relationships among seismic predictors

2.1. Dataset description

The PEER, Updated NGA-WEST2 Flatfile Vertical 5% damping dataset [24] is employed in this study. The entire database is filtered based on the closest distance to the rupture plane (ClstD in km) up to 500 km, fault type identified by the RA, Mw ranging from 3.40 to 7.90, and hypocentral depth from 1.8 km to 81.0 km. After filtering, the employed dataset comprises a total of 16211 data points.

The input variables used for the RF model included Mw, RA, hypocentral depth, and Vs30. The target variable was the logarithm of peak ground acceleration (log(PGA)). Prior to model training, all records containing missing or incomplete entries were removed. Outliers were identified through visual inspection of variable distributions for all input and output parameters and excluded to minimize their influence on the dataset. Continuous variables were normalized using the StandardScaler() function from the Scikit-learn library to achieve consistent scaling across all parameters. As the dataset contained no categorical variables, additional encoding was not required. These preprocessing steps ensured that the final dataset was complete, statistically balanced, and free from

anomalies, making it suitable for reliable model development and validation.

The statistical distribution of each variable is illustrated in Figure 1. The horizontal axis of each plot represents the range of values (or bins) for a specific variable, while the vertical axis (frequency) quantifies the number of data points falling within that range. Table 1 provides the basic statistical summary for the five input parameters and the single output parameter used in the RF model, summarizing key measures, minimum, maximum, mean, median, and standard deviation values across the entire dataset.

The relationships between the input variables and the PGA are presented in Figure 2 as a correlation heatmap based on Pearson correlation coefficients, which range from −1 to +1. These coefficients quantify the strength and direction of the linear association between each pair of variables. In this heatmap, darker blue colors indicate stronger positive correlations, while lighter shades represent weaker or even negative correlations.

Among the input variables, Mw exhibits the strongest positive correlation with the logarithm of peak ground acceleration (Log_{10}(PGA)), with a correlation coefficient of r = 0.66, suggesting that larger magnitude earthquakes are associated with higher ground motion intensities. This is followed by RA and hypocentral depth, which show weaker positive correlations with Log10(PGA) (r = 0.15 and r = 0.13, respectively), indicating modest contributions to ground motion variability.

In contrast, ClstD and Vs30 exhibit negative correlations with Log10(PGA), with r = −0.55 and r = −0.33, respectively. This suggests that as the rupture distance increases or the ground becomes stiffer, the PGA tends to decrease. Notably, Vs30 also has a negative correlation with Mw (r = −0.21), potentially reflecting regional geological trends.

The weakest correlations are observed among depth, RA, and ClstD, with several near-zero or slightly negative values (e.g., RA and ClstD, r = −0.0058), indicating minimal direct linear relationships. These results highlight that while some variables like Mw and ClstD strongly influence PGA, others may interact in more complex, nonlinear ways that require advanced modeling techniques such as ML models for accurate prediction.

Figure 1

Histogram distributions of input and output variables: Closest distance to the rupture plane (ClstD), hypocentral depth, shear wave velocity in the top 30 meters (Vs30), rake angle (RA), moment magnitude (Mw), and the logarithm of peak ground acceleration (Log_{10}(PGA))

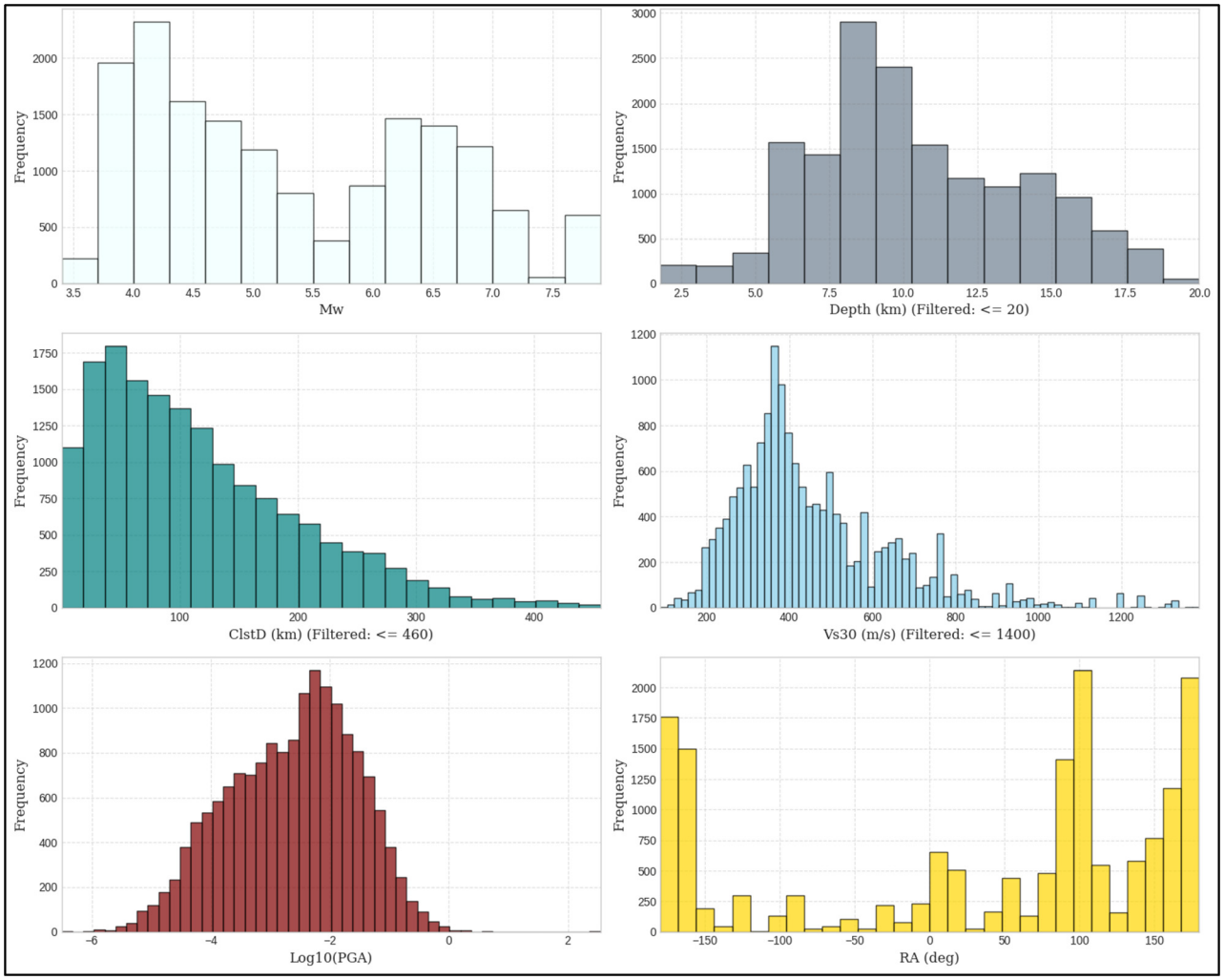
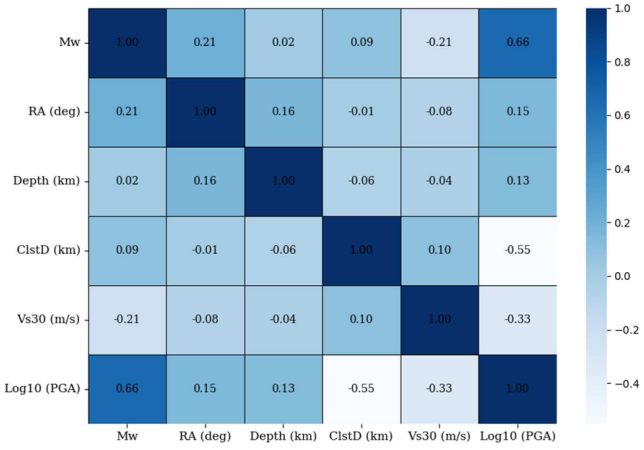


Table 1
Basic statistical summary: Minimum, maximum, range, mean, median, and standard deviation values of input and output parameters

Parameters	Mw	RA (deg)	Depth (km)	ClstD (km)	Vs30 (m/s)	Log_{10} (PGA)
Minimum	3.40	–180.00	1.80	0.05	89.32	–6.51
Maximum	7.90	180.00	81.00	1532.66	2100.00	2.55
Range	4.50	360.00	79.20	1532.61	2010.68	9.06
Mean	5.31	35.27	10.53	117.74	460.19	–2.63
Median	5.03	90.00	10.00	97.93	405.00	–2.50
Standard deviation	1.19	126.54	3.77	87.94	204.48	1.07

Figure 2
Correlation heatmap illustrates the linear relationships between input variables and PGA



2.1.1. Ground motion parameters used in the study

In this research, the variables selected as predictors are Mw, RA, hypocentral depth, and Vs30, while the PGA is considered the response variable. The Mw expresses the total seismic energy released at the source. Although several other magnitude scales such as the surface wave magnitude (Ms), the body wave magnitude (Mb), the local magnitude (ML), and the duration magnitude (Md) are often reported, Mw was used exclusively in this study. When other magnitude types were encountered, they were converted to Mw through empirical correlations available in previous research [27]. The RA represents the direction of slip along the fault and assists in identifying the type of fault mechanism involved. The hypocentral depth describes the point beneath the surface where the earthquake begins, expressed in kilometers. The parameter Vs30, measured in meters per second, reflects the mean shear wave velocity within the upper 30 meters of soil and is commonly applied to define site conditions. Finally, PGA corresponds to the maximum ground acceleration recorded during a seismic event and serves as an important measure of shaking severity and structural response.

3. Research Methodology

3.1. Methodology and model development

This study employed the RF ML algorithm to predict the horizontal component of PGA. A dataset comprising 16,211 data points was utilized for the model development. The data was split into two subsets: 80% (12968 data points) were used for training, and 20% (3243 data points) were used for testing the final model.

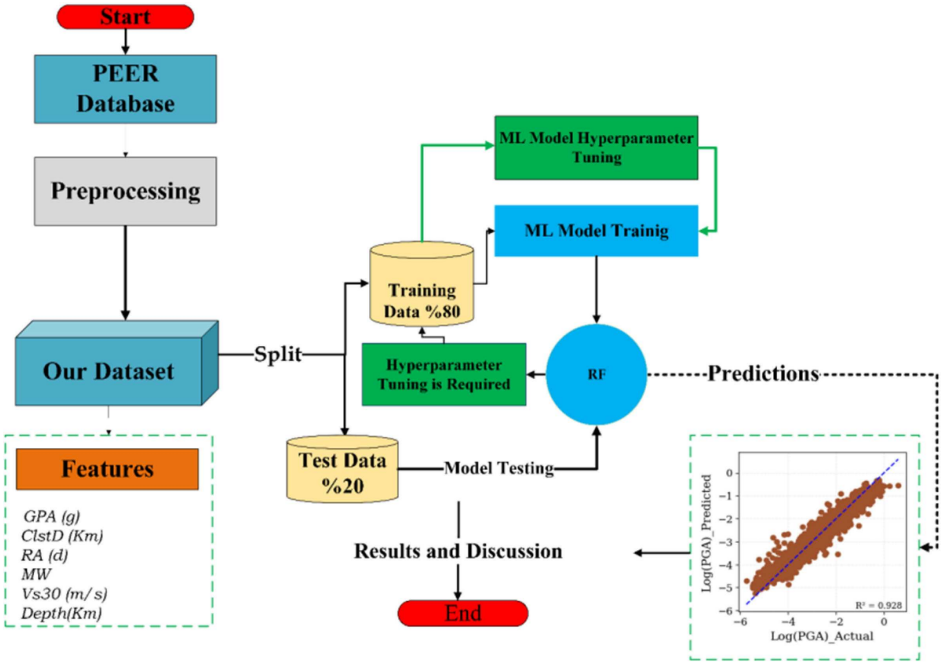
To enhance the model's predictive performance and generalization ability, hyperparameter tuning was performed using grid search in combination with 5-fold cross-validation. Instead of relying on the default parameters provided by the Scikit-learn library, the optimization process systematically evaluated 120 parameter combinations, resulting in a total of 600 model fits. The grid search explored four main parameters: maximum tree depth (max_depth) in the range of 3–10, number of trees (n_estimators) from 100 to 500, number of features considered at each split (max_features) as "sqrt" and "log2," and minimum samples required at a leaf node (min_samples_leaf) between 1 and 4.

The best-performing combination of parameters for the model was max_depth = 10, *n*_estimators = 500, max_features = "sqrt," and min_samples_leaf = 1. The optimized model achieved a cross-validated coefficient of determination (R^2) score of 0.9216, indicating strong predictive performance during training. When evaluated on the test dataset, it produced an R^2 of 0.9284 and an RMSE of 0.2875, confirming its excellent generalization capability. The close agreement between the cross-validated and test set R^2 values indicates that the model is neither underfitted nor overfitted and that the adopted 5-fold cross-validation approach ensured stable performance across different data partitions.

The hyperparameters used in the model are detailed as follows. The max_depth is set to 10, limiting the maximum depth of each decision tree to prevent overfitting by avoiding overly complex trees. The *n*_estimators parameter is set to 500, indicating the number of trees in the forest. While increasing this value generally enhances performance, it also raises computational time. The max_features parameter is set to "sqrt," meaning that at each split, only the square root of the total number of features is considered, that is, an approach commonly used in classification problems. Lastly, min_samples_leaf is set to 1, allowing leaf nodes to contain a minimum of one sample. This setting enables fine-grained splits that can better capture training patterns but may also lead to overfitting if not properly managed [28]. The methodological chart of the study is shown in Figure 3.

At the end, various statistical metrics were used to evaluate the performance of the developed ML models, including RMSE [Eq. (1)], MAE [Eq. (2)], and the R^2 [Eq. (3)]. Each metric highlights a different aspect of model accuracy. As noted by Kumar et al. [29], RMSE reflects the standard deviation of prediction errors and penalizes larger errors more heavily. MAE provides the average magnitude of errors between predicted and actual values, offering a clear measure of overall accuracy. R^2 indicates how well the model explains the variability in observed data. Collectively, these

Figure 3
Flowchart illustrating the methodology followed in this study



metrics enable a robust and comparative assessment of the models' predictive performance and reliability [30].

$$\mathrm{RMSE} = \frac{1}{n}\sqrt{\sum_{i=1}^{n}\left(y_{Actual\_PGA} - y_{Predicted\_PGA}\right)^2} \quad (1)$$

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(y_{Actual\_PGA} - y_{Predicted\_PGA}\right)^2 \quad (2)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_{Actual\_PGA} - y_{predicted\_PGA}\right)^2}{\sum_{i=1}^{n}\left(y_{Actual\_PGA} - \bar{y}\right)^2} \quad (3)$$

In these formulas, n represents the number of samples, while $\bar{y}$ denotes the mean of the actual values in the dataset.

3.2. Random forest

ML has been widely adopted in earthquake engineering over recent years, driven by its success in modeling nonlinear systems across many scientific domains. Techniques such as ANNs, support vector machines, gradient boosting machines, and RF have been explored for seismic hazard applications, often demonstrating better predictive performance than GMPEs. ML methods are particularly well suited for handling large datasets, where multicollinearity and nonlinear dependencies can obscure relationships in conventional regression models. Recent research has demonstrated that ML models can reduce residual variance and provide more accurate estimates of PGA and spectral accelerations when trained on large-scale ground motion datasets [31–33].

Among ML models, ensemble approaches such as RF have emerged as especially powerful tools, which is an ensemble-based ML algorithm that builds multiple decision trees by utilizing bootstrapped subsets of the dataset and randomly selecting features at each split [28, 34]. In regression tasks, the final output is determined by averaging the predictions from all individual trees, while in classification tasks, it relies on majority voting. Each tree in the forest is trained on a unique subset of data created through sampling with replacement, a technique known as bagging [13, 35]. This process introduces variation among the trees, as not all data points are used in training each tree.

The data points that are left out during the training of a specific tree, referred to as out-of-bag (OOB) samples, serve as a built-in validation set [36]. These OOB samples are instrumental in estimating the prediction error without the need for a separate validation dataset.

Moreover, instead of evaluating all features at every split, RF selects the best splitting feature from a randomly selected subset. This random feature selection reduces inter-tree correlation and boosts the ensemble's generalization ability, even if it means individual trees may be slightly less accurate.

The architecture of the RF model is illustrated in Figure 4.

4. Results and Discussion

The results of the developed RF model are presented in Figures 5 and 6. Figure 5 shows the model's prediction performance on the training dataset. Performance metrics such as $R^2 = 0.948$, RMSE = 0.244, and MSE = 0.060 indicate that the model fits the training data well, demonstrating a high prediction capability.

Figure 6 illustrates the model's prediction results on the test dataset. Similar to its performance on the training data, the model performs well on the test data, as shown by the evaluation metrics: $R^2 = 0.928$, RMSE = 0.288, and MSE = 0.083. These results confirm that the developed model generalizes effectively and maintains strong predictive performance on unseen data.

Additionally, the performance of the trained model was evaluated against previous studies that predicted PGA. As summarized in Table 2, the proposed RF model demonstrates performance levels consistent with other state-of-the-art approaches, such as ANN,

Figure 4
Schematic representation of the random forest algorithm architecture, illustrating the ensemble of decision trees and the voting mechanism

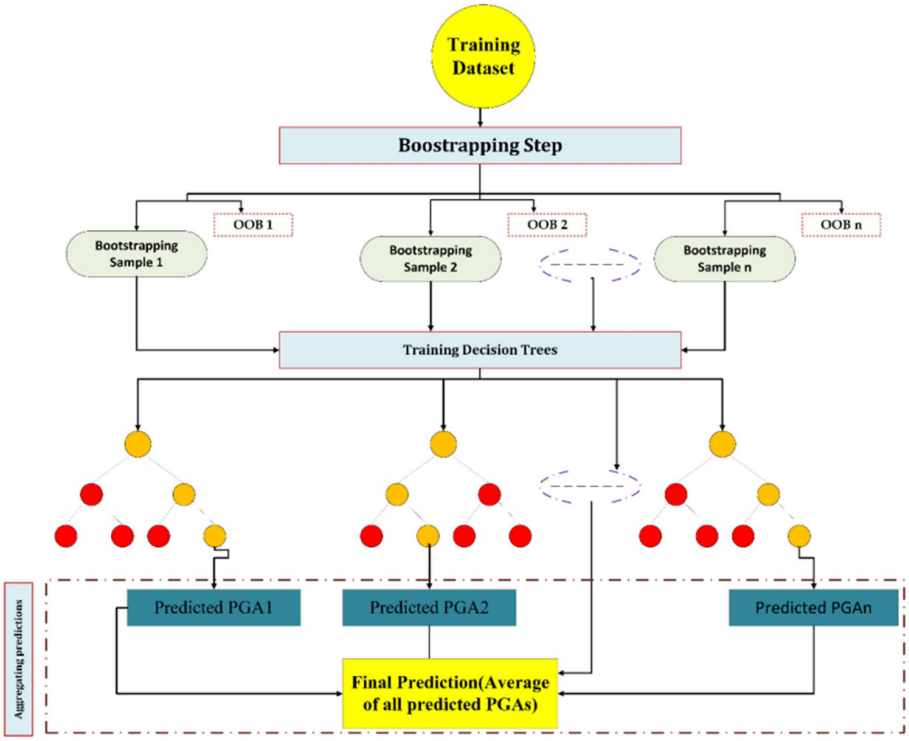


Figure 5
Prediction performance of the random forest model on the training dataset

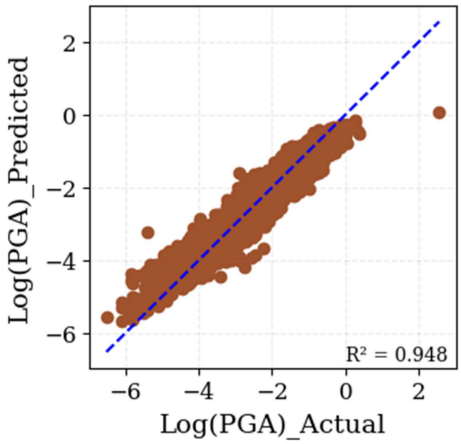
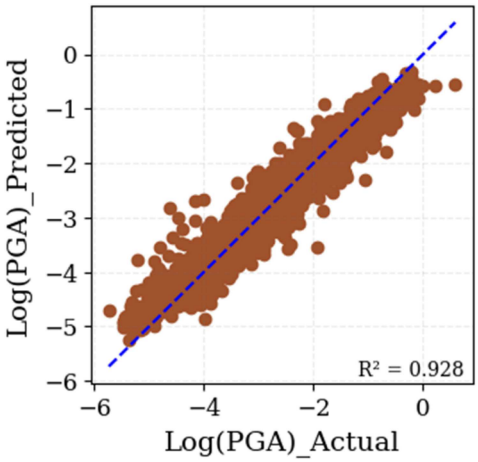


Figure 6
Prediction performance of the random forest model on the test dataset



DNN, and XGBoost, confirming its reliability and robustness across diverse datasets.

The feature-important results of the developed RF model are presented in Figure 7. Feature importance provides insight into how much each input variable contributes to the prediction made by the model. In this context, the importance values are normalized and indicate the relative influence of each feature in determining the model's output.

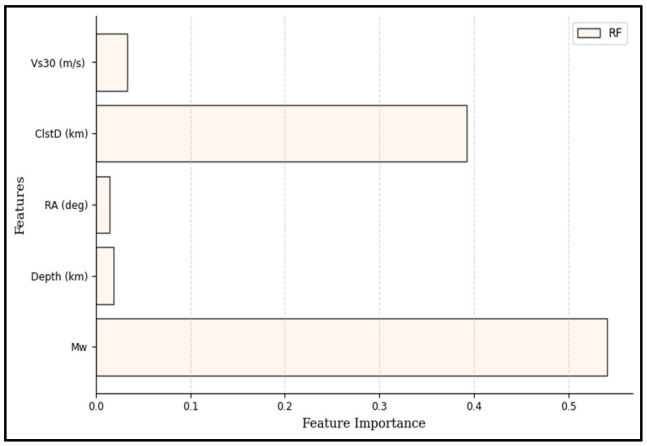
As shown in Figure 7, the most influential features are Mw and closest distance to the rupture plane (ClstD), with importance scores of approximately 0.52 and 0.39, respectively. These two variables contribute the most to the model's prediction accuracy. In contrast, depth, Vs30, and rake angle (RA) have significantly lower values, each contributing less than 0.05, indicating that their impact on the model's prediction is minimal.

This result suggests that seismic magnitude and proximity to the site are the dominant factors influencing the target variable in this model. These findings are consistent with established geotechnical understanding, where ground motion parameters are more strongly affected by the magnitude and distance of seismic events.

Table 2
Comparison of the predictive performance of the proposed RF model with previously published ANN, DNN, and XGBoost models for PGA prediction

Study	Model type	Data	R or R^2 (test)
Güllü and Erçelebi [3]	Artificial neural network	Turkish Strong Motion	0.964
Derakhshani and Foruzan [25]	Deep neural network	NGA-West2	0.902
Mandal and Mandal [26]	XGBoost	CSIR-NGRI, Hyderabad	0.844
This Study	Random forest	NGA-West2	0.928

Figure 7
Importance levels of input variables in the random forest model for PGA prediction



5. Conclusion

This study successfully developed a robust and reliable random ML model to predict the horizontal component of PGA using a comprehensive dataset of 16,211 ground motion records. The model was optimized through grid search with 5-fold cross-validation, ensuring strong performance and generalization capabilities.

The optimized model achieved an R^2 score of 0.948 on the training data and 0.928 on the test data, with corresponding RMSE values of 0.244 and 0.288 and MSE values of 0.060 and 0.083, respectively. These results indicate that the model is well-fitted, generalizes effectively, and maintains high predictive accuracy on unseen data.

Feature importance analysis revealed that Mw and closest distance to the rupture plane are the most influential predictors of PGA, which is consistent with the physical principles of earthquake ground motion. Larger magnitudes release greater seismic energy, resulting in stronger shaking amplitudes and higher PGA values, while PGA decreases with increasing distance from the rupture due to geometric spreading and energy dissipation through surrounding materials. The combined influence of Mw and ClstD reflects the balance between source energy and attenuation effects, where high-magnitude earthquakes occurring at short distances generate the most intense ground motions. These findings align well with the observations reported in recent literature [37, 38]. Other features such as Depth, Vs30, and RA contributed marginally to the prediction outcome.

As discussed in the introduction, PGA is a key parameter in seismic hazard assessment and earthquake-resistant design because it directly influences structural performance and risk evaluation. The RF model developed in this study provides an efficient and data-driven approach for estimating PGA across a broad range of geological and seismic conditions. Trained on a large and diverse dataset, the model can predict PGA more accurately than traditional GMPEs. It can be used to estimate PGA values and perform sensitivity analyses within the scope and limitations of this work.

Overall, the study demonstrates the strong potential of ML, particularly RF algorithms, for accurately predicting seismic ground motion parameters. The findings highlight the central role of Mw and rupture distance in controlling PGA and offer a valuable framework for seismic hazard evaluation and earthquake-resistant design applications.

Recommendations

Although the primary focus of this study was to develop an RF model for predicting PGA, there are numerous other ML models capable of predicting peak ground motion parameters. Future researchers are encouraged to explore and compare the performance of models such as gradient boosting, support vector regression, and DNNs. Additionally, statistical models can be developed and benchmarked against ML models to evaluate their relative prediction accuracy and interpretability.

Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

Data are available from the corresponding author upon reasonable request.

Author Contribution Statement

Asadullah Ziar: Conceptualization, Methodology, Software, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization, Project administration. **Ender Basari:** Conceptualization, Validation, Writing – review & editing, Supervision.

References

[1] Kramer, S. L., & Stewart, J. P. (2024). *Geotechnical earthquake engineering*. USA: CRC Press.

[2] Gandomi, A. H., Alavi, A. H., Mousavi, M., & Tabatabaei, S. M. (2011). A hybrid computational approach to derive new ground-motion prediction equations. *Engineering Applications of Artificial Intelligence*, *24*(4), 717–732. https://doi.org/10.1016/j.engappai.2011.01.005

[3] Güllü, H., & Erçelebi, E. (2007). A neural network approach for attenuation relationships: An application using strong ground motion data from Turkey. *Engineering Geology*, *93*(3-4), 65–81. https://doi.org/10.1016/j.enggeo.2007.05.004

[4] Javan-emrooz, H., Eskandari-Ghadi, M., & Mirzaei, N. (2018). Prediction equations for horizontal and vertical PGA, PGV, and PGD in northern Iran using prefix gene expression programming. *Bulletin of the Seismological Society of America*, *108*(4), 2305–2332. https://doi.org/10.1785/0120170155

[5] Khalid, F., & Razbin, M. (2024). Modeling peak ground acceleration for earthquake hazard safety evaluation. *Scientific Reports*, *14*(1), 31032.

[6] Douglas, J. (2003). Earthquake ground motion estimation using strong-motion records: A review of equations for the estimation of peak ground acceleration and response spectral ordinates. *Earth-Science Reviews*, *61*(1-2), 43–104. https://doi.org/10.1016/S0012-8252(02)00112-5

[7] Boore, D. M., Joyner, W. B., & Fumal, T. E. (1997). Equations for estimating horizontal response spectra and peak acceleration from western North American earthquakes: A summary of recent work. *Seismological research letters*, *68*(1), 128–153. https://doi.org/10.1785/gssrl.68.1.128

[8] Douglas, J. (2021). *Ground motion prediction equations 1964-2021*. University of Strathclyde.

[9] Sreenath, V., Podili, B., & Raghukanth, S. T. G. (2023). A hybrid non-parametric ground motion model for shallow crustal earthquakes in Europe. *Earthquake Engineering & Structural Dynamics*, *52*(8), 2303–2322. https://doi.org/10.1002/eqe.3845

[10] Kohrangi, M., Kotha, S. R., & Bazzurro, P. (2021). Impact of partially non-ergodic site-specific probabilistic seismic hazard on risk assessment of single buildings. *Earthquake Spectra*, *37*(1), 409–427. https://doi.org/10.1177/8755293020938813

[11] Ambraseys, N. N., Simpson, K. A., & Bommer, J. J. (1996). Prediction of horizontal response spectra in Europe. *Earth quake Engineering & Structural Dynamics*, *25*(4), 371–400. https://doi.org/10.1002/(SICI)1096-9845(199604)25:4<371::AID-EQE550>3.0.CO;2-A

[12] Campbell, K. W., & Bozorgnia, Y. (2007). *Campbell-Bozorgnia NGA ground motion relations for the geometric mean horizontal component of peak and spectral ground motion parameters*. USA: Pacific Earthquake Engineering Research Center.

[13] Yang, C., Pan, Y., Zhang, K., Yue, M., Wen, H., & Wang, F. (2025). Machine learning-based ground peak acceleration attenuation prediction model. *Journal of Earthquake Engineering*, *29*(2), 324–338. https://doi.org/10.1080/13632469.2024.2443638

[14] Thaler, D., Stoffel, M., Markert, B., & Bamer, F. (2021). Machine-learning-enhanced tail end prediction of structural response statistics in earthquake engineering. *Earthquake Engineering & Structural Dynamics*, *50*(8), 2098–2114. https://doi.org/10.1002/eqe.3432

[15] Mousavi, S. M., & Beroza, G. C. (2020). A machine-learning approach for earthquake magnitude estimation. *Geophysical Research Letters*, *47*(1), e2019GL085976. https://doi.org/10.1029/2019GL085976

[16] Khosravikia, F., & Clayton, P. (2021). Machine learning in ground motion prediction. *Computers & Geosciences*, *148*, 104700. https://doi.org/10.1016/j.cageo.2021.104700

[17] Joshi, A., Raman, B., Mohan, C. K., & Cenkeramaddi, L. R. (2024). Application of a new machine learning model to improve earthquake ground motion predictions. *Natural Hazards*, *120*(1), 729–753.

[18] Jeddi, A. B., Shafieezadeh, A., Hur, J., Ha, J. G., Hahm, D., & Kim, M. K. (2022). Multi-hazard typhoon and earthquake collapse fragility models for transmission towers: An active learning reliability approach using gradient boosting classifiers. *Earthquake Engineering & Structural Dynamics*, *51*(15), 3552–3573. https://doi.org/10.1002/eqe.3735

[19] Impraimakis, M. (2024). A convolutional neural network deep learning method for model class selection. *Earthquake Engineering & Structural Dynamics*, *53*(2), 784–814. https://doi.org/10.1002/eqe.4045

[20] Hussaini, S. M. S., Caicedo, D., Mohammadi, A., Karimzadeh, S., & Lourenço, P. B. (2024). Nonparametric ground motion models of arias intensity and significant duration for the Italian dataset. *Journal of Physics: Conference Series*, *2647*(6), 062001. https://doi.org/10.1088/1742-6596/2647/6/062001

[21] Farahani, S., & Barari, A. (2023). A simplified procedure for the prediction of liquefaction-induced settlement of offshore wind turbines supported by suction caisson foundation based on effective stress analyses and an ML-based group method of data handling. *Earthquake Engineering & Structural Dynamics*, *52*(15), 5072–5098. https://doi.org/10.1002/eqe.4000

[22] Bhatta, S., & Dang, J. (2023). Seismic damage prediction of RC buildings using machine learning. *Earthquake Engineering & Structural Dynamics*, *52*(11), 3504–3527. https://doi.org/10.1002/eqe.3907

[23] Mousavi, S. M., Gandomi, A. H., Alavi, A. H., & Vesalimahmood, M. (2010). Modeling of compressive strength of HPC mixes using a combined algorithm of genetic programming and orthogonal least squares. *Structural Engineering and Mechanics: An International Journal*, *36*(2), 225–241. https://doi.org/10.12989/sem.2010.36.2.225

[24] PEER Ground Motion Database. (2013). *Pacific Earthquake Engineering Research Center (PEER)*. Retrieved from: https://ngawest2.berkeley.edu/

[25] Derakhshani, A., & Foruzan, A. H. (2019). Predicting the principal strong ground motion parameters: A deep learning approach. *Applied Soft Computing*, *80*, 192–201. https://doi.org/10.1016/j.asoc.2019.03.029

[26] Mandal, P., & Mandal, P. (2024). Peak ground acceleration prediction using supervised machine learning algorithm for the seismically hazardous Kachchh rift zone, Gujarat, India. *Natural Hazards*, *120*(2), 1821–1840. https://doi.org/10.1007/s11069-023-06257-7

[27] Al-Heety, E. A. M. S. (2014). A complete and homogeneous magnitude earthquake catalogue of Iraq. *Arabian Journal of Geosciences*, *7*(11), 4727–4732. https://doi.org/10.1007/s12517-013-1131-y

[28] Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/a:1010933404324

[29] Kumar, P., Pratap, B., Sharma, S., & Kumar, I. (2024). Compressive strength prediction of fly ash and blast furnace slag-based geopolymer concrete using convolutional neural network. *Asian Journal of Civil Engineering*, *25*(2), 1561–1569. https://doi.org/10.1007/s42107-023-00861-5

[30] Hodson, T. O. (2022). Root mean square error (RMSE) or mean absolute error (MAE): When to use them or not. *Geoscientific Model Development Discussions*, *15*, 5481–5487. https://doi.org/10.5194/gmd-15-5481-2022

[31] Kubo, H., Naoi, M., & Kano, M. (2024). Recent advances in earthquake seismology using machine learning. *Earth, Planets and Space*, *76*(1), 36.10.1186/s40623-024-01982-0.

[32] Noor, U. A. (2025). Machine learning innovations in revolutionizing earthquake engineering: A review: UA Noor. In *Archives of Computational Methods in Engineering* (pp. 1–57).

[33] Hu, Y., Wang, W., Li, L., & Wang, F. (2024). Applying machine learning to earthquake engineering: A scientometric analysis of world research. *Buildings*, *14*(5), 1393. https://doi.org/10.3390/buildings14051393

[34] Ebadi, A., Auger, A., & Gauthier, Y. (2024). Machine learning insights into hypersonics research evolution: A 21st century perspective. *Archives of Advanced Engineering Science*, *2*(2), 79–92. https://doi.org/10.47852/bonviewAAES32021471

[35] Hatwell, J., Gaber, M. M., & Azad, R. M. A. (2020). CHIRPS: Explaining random forest classification. *Artificial Intelligence Review*, *53*(8), 5747–5788. https://doi.org/10.1007/s10462-020-09833-6

[36] Peters, J., De Baets, B., Verhoest, N. E. C., Samson, R., Degroeve, S., De Becker, P., & Huybrechts, W. (2007). Random forests as a tool for ecohydrological distribution modelling. *Ecological Modelling*, *207*(2-4), 304–318. https://doi.org/10.1016/j.ecolmodel.2007.05.011

[37] Temiz, C., Hussaini, S. M. S., Karimzadeh, S., Askan, A., & Lourenço, P. B. (2025). Seismic scenario simulation and ANN-based ground motion model development on the North Tabriz Fault in Northwest Iran. *Journal of Seismology*, *29*(1), 147–169. https://doi.org/10.1007/s10950-024-10264-x

[38] Colavitti, L., Bindi, D., Tarchini, G., Scafidi, D., Picozzi, M., & Spallarossa, D. (2024). A high-quality data set for seismological studies in the east Anatolian fault zone, Türkiye. *Earth System Science Data Discussions*, *17*(6), 3089–3108. https://doi.org/10.5194/essd-17-3089-2025

How to Cite: Ziar, A., & Basari, E. (2025). Accurate Prediction of Peak Ground Acceleration Using Random Forests: A Data-Driven Approach with PEER Updated NGA-WEST2 Ground Motion Records. *Archives of Advanced Engineering Science*. https://doi.org/10.47852/bonviewAAES52026930