Natural Hazards (2024) 120:729–753
https://doi.org/10.1007/s11069-023-06230-4

ORIGINAL PAPER

Application of a new machine learning model to improve earthquake ground motion predictions

Anushka Joshi[1] · Balasubramanian Raman[1] · C. Krishna Mohan[2] · Linga Reddy Cenkeramaddi[3]

Received: 20 April 2023 / Accepted: 20 September 2023 / Published online: 11 October 2023

Abstract

A cross-region prediction model named SeisEML (an acronym for Seismological Ensemble Machine Learning) has been developed in this paper to predict the peak ground acceleration (PGA) at a given site during an earthquake. The SeisEML model consists of hybridized models, kernel-based algorithms, tree regression algorithms, and regression algorithms. The model ablation study is conducted to examine the performance and the selection of meta-machine learning models in the SeisEML. The training and testing dataset consists of 20852 and 6256 accelerograms recorded by the Kyoshin Network, Japan. The mean absolute error (*MAE*) and root mean square error (*RMSE*) have been utilized to compare the predicted peak ground acceleration (PGA) for the test data. The SeisEML model yields approximately half the *MAE* and *RMSE* values obtained with conventional attenuation relations. The SeisEML model has been used to compute Japan's iso acceleration contour map of three earthquakes (M_{JMA} 7.4, 6.6, and 6.1). The qualitative comparison of iso acceleration contours obtained from actual and predicted PGA using SeisEML clearly shows that the model can reliably predict the PGA distribution during an earthquake compared to the regional ground motion prediction equation (GMPE). The cross-region prediction was performed on the datasets of the Iranian earthquakes using SeisEML. The comparison of predicted and observed peak ground acceleration in terms of *MAE* and *RMSE* shows that the machine learning model's performance is superior to the regional attenuation relation. The predictions of PGA from the developed ML model indicate that this trained model has the potential for predicting regional and global scenarios with similar tectonic setups. The ML model developed in this paper can considerably enhance the reliability of PGA prediction for seismic hazard mapping of any region and can serve as a reliable substitute for GMPEs.

Keywords Machine learning · Ensemble model · PGA · Earthquake · Meta heuristic algorithms

Extended author information available on the last page of the article

1 Introduction

The design of earthquake-resistance structures has developed into a truly multidisciplinary area of engineering where artificial intelligence allows numerous exciting developments. The peak ground acceleration (PGA) recorded during an earthquake is a major controlling factor for the design of earthquake-resistant structures. The primary construction site in a seismically active environment seldom contains strong motion records. Therefore, the practical approach is to compute this parameter from the simulated strong motion records. The strong ground motion can be simulated using various simulation techniques (Hartzell 1978; Kanamori 1979; Hadley and Helmberger 1980; Mikumo et al. 1981; Irikura and Muramatu 1982; sheng Paul Lai 1982; Hartzell 1982; Boore 1983; Irikura 1983; Munguía and Brune 1984; Hutchings 1985; Saikia and Herrmann 1985; Irikura 1986; Boore and Atkinson 1987; Kamae and Irikura 1992; Midorikawa 1993; Saikia 1993; Zeng et al. 1994; Yu 1994; Yu et al. 1995; Motazedian and Atkinson 2005; Joshi and Midorikawa 2005). The simulation of strong-motion records using either of these techniques requires many parameters of the rupture plane responsible for an earthquake, which is difficult to determine, and several parameters related to the medium characteristics are not always available at the site where construction is required.

The magnitude of ground motions experienced during an earthquake is influenced by intricate interactions among the earthquake source, observation site, and the path connecting them Douglas (2003a). The relationship between these factors can be represented through ground motion prediction equations (GMPEs) or attenuation relationships (Si and Midorikawa 1999; Douglas 2003b; Morikawa and Fujiwara 2013; Abrahamson et al. 2014; Boore et al. 2014; Campbell and Bozorgnia 2014; Douglas and Edwards 2016; Kubo et al. 2020). The GMPE uses a simple regression model that relates the dependent parameter like PGA with the independent parameters like magnitude, hypocentral distance, site-specific parameters, fault type responsible for earthquake (Abrahamson and Litehiser 1989), and plate boundaries (Abrahamson and Litehiser 1989). The GMPEs are used in both the deterministic and probabilistic seismic hazard and risk assessments (Douglas and Edwards 2016) and earthquake early warning (Hoshiba et al. 2008). The GMPEs are developed using least square regression analysis, which results in errors in the ground motion predictability (Ambraseys et al. 1996; Campbell 1997; Boore and Atkinson 2008; Nath et al. 2009; Anbazhagan et al. 2013). Although GMPEs provide simple and ready answers to civil engineers, their reliability is always in question owing to the method of determination of coefficients used in the regression model together with the quantity and quality of the dataset.

Recently, machine learning (ML) approaches have been applied in place of GMPEs (Derras et al. 2012, 2013, 2016; Trugman and Shearer 2018; Kubo et al. 2020; Mandal and Mandal 2022). The primary benefit of employing this approach is its utilization of non-parametric ML methods. These methodologies can learn functions directly from data without necessitating the assumption of regression equations. Moreover, the machine learning model's inherent flexibility enables the incorporation of novel data or explanatory variables that are typically not utilized in traditional ground-motion prediction equations (GMPEs) (Kubo et al. 2020). The machine learning application in the field of GMPEs is based on a single ML model like an artificial neural network (Derras et al. 2012, 2013, 2016), random forest (Trugman and Shearer 2018), extremely randomized trees (Kubo et al. 2020) and XGboost (Mandal and Mandal 2022). It has been observed that the use of a single ML model may amount to the problem of excessive over or underestimation. Hence, this paper

adopts a phased approach, employing multiple machine learning models to predict the peak ground acceleration (PGA). This strategy harnesses the distinct advantages offered by various ML models within a unified framework. A five-phase stacked ensemble model, namely SeisEML, has been designed to predict the PGA, which can also be used for cross-region PGA prediction. The model consists of hybridized models such as Xgboost (XGB) (Chen and Guestrin 2016), Catboost (CatB) (Ostroumova et al. 2018), and Random Forest (RF) (Breiman 2004) coupled with meta-heuristic optimization algorithms such as Sanitized Gray Wolf Optimizer (SGWO) (Mirjalili et al. 2014) and Bayesian optimizer (BO) (Frazier 2018), a kernel-based algorithm such as Kernel Ridge Regression (K-RR), tree regression algorithms such as CatB, LightGBM (LGBM) (Ke et al. 2017), XGB, and RF, and regression-based algorithm such as Ridge Regression (RR).

In this paper, five phase-stacked ensemble model has been trained using the strong motion dataset of Kyoshin Network (K-NET) (Aoi et al. 2011) Japan, and its reliability has been checked by comparing the actual and predicted PGA values from the test data sets. RobustScalar, a feature scaling method, has been used in this work due to its robustness to outliers in the dataset. The comparison of the performance of the SeisEML model with other prominent machine learning models, such as the Gaussian Process Regressor (GPR) (Rasmussen and Williams 2009), Extra Tree Regressor (ExTR) (Geurts et al. 2006), Random Subspace Catboost (RSubCatB), AdaBoost (Freund and Schapire 1995), Decision Tree (DTree) (Breiman et al. 1984), and RR have been studied in this work to confirm the suitability of developed five-phase stacked ensemble model over single machine learning models.

2 Machine learning model

In this paper, a new ML model named SeisEML has been developed. The first step in developing the SeisEML model is to train multiple regression models and then apply them in stacks. Stacking is also called stack generalization. Stacking distinguishes itself from bagging and boosting methods by employing diverse base learning algorithms instead of combining homogeneous weak learners. Unlike bagging and boosting, which utilize the same type of weak learners, stacking takes advantage of different base learning algorithms. (Zhou and Jiao 2022).

The ML model used in the present work utilizes the logarithm of PGA in cm/s^2 as the target parameter. The PGA_{lg} used as a target by the model is the logarithm of the absolute value of the maximum of two horizontal components, which are calculated as:

$$PGA_{lg} = \log_{10} \max(PGA_{NS}, PGA_{EW}) \tag{1}$$

In this expression, PGA_{EW} and PGA_{NS} define the EW and NS components of the acceleration records, respectively. The following expression represents the functional dependency of PGA:

$$PGA = f(M, R_h, R_e, s_t, d_i, V_{S30}) \tag{2}$$

The above expression shows that PGA in our machine learning model is dependent on various features like the earthquake magnitude (M), the hypocentral distance in km (R_h), the epicentral distance in km (R_e), site characteristics (s_t), focal depth (d_i), magnitude (M) and average shear wave velocity at 30 m depth (V_{S30}). These features are obtained directly from the earthquake catalogue and other sources. The complete dataset, denoted as D, is

partitioned into training and testing sets, with a distribution ratio of 70 percent for training and 30 percent for testing. The validation set is 30 percent of the test dataset. The ML model developed in this work uses various common machine learning algorithms, hybrid, and kernel-based algorithms, which are discussed in the coming sections.

2.1 Common machine learning algorithms

The present work incorporates six popular machine learning algorithms, including Tree-based Regressors and Ridge Regression. The ensemble models give more accurate results when bagging (Breiman 1996) and boosting (Schapire 1990; Freund and Schapire 1996) are included in the model. These models are based on "resampling" techniques where different sets are built for training. Bagging is a "bootstrap" (Efron and Tibshirani 1994) ensemble method. The concept of bagging involves individual training models using resample sets and then combining the results. The bagging technique that has been used in the SeisEML model is RF. The other models, such as CatB, LGBM, and XGB, fall under the category of boosting ensemble models. Among the boosting models, we have three XGB (XGB-1, XGB-2, and XGB-3) models, two LGBM (LGBM-1 and LGBM-2), and one CatB model where different parameters are initialized.

One commonly employed parameter estimation technique, known as Ridge Regression (RR), effectively addresses the issue of collinearity that often arises in multiple linear regression (Hoerl and Kennard 2000). It is observed that few features are correlated. To mitigate the effects of multicollinearity, we employ ridge regression, which utilizes parameter-shrinking techniques. It also reduces model complexity by coefficient shrinking. The Ridge Regression is a form of simple linear regression along with a regularization term. The Ridge Regression model has been used in the current work in Layers 2, 3, and 4, named RR-2, RR-3, and RR-4.

2.2 Hybrid machine learning algorithms

The meta-heuristics algorithms and popular ML algorithms such as XGB, CatB, and RF have been applied in this work. The XGB, CatB, and RF are employed as objective functions in this study to employ hyperparameter optimization to optimize its fitness characteristics. Sanitized Gray Wolf Optimizer (GWO) and the Bayesian Optimizer (BO) methods are used to optimize the hyperparameters.

The Gray Wolf Optimization (GWO) algorithm has demonstrated successful results in diverse optimization tasks, such as optimal feature set selection (Hu et al. 2020) and kernel EML parameter tuning for bankruptcy prediction (Wang et al. 2017). The GWO algorithm has yielded high success rate in many applications than other meta-heuristic algorithms(Prasanth et al. 2020), such as particle swarm optimization, grid search, genetic algorithm, and other algorithms.

Animal behaviour served as inspiration for the SGWO. The use of the Chaos Theory produces a butterfly effect on GWO, thus naming it SGWO. The optimal parameter set obtained in SGWO to describe the social hierarchy of wolves is indicated by alpha, followed by beta and delta. Therefore, the GWO algorithm has been proposed in this work for the hyperparameter tuning of two models used in our study, such as XGB and RF. The fitness function (F_f) employed by SGWO in our study to minimize the error rate is represented by the following equation:

$$F_f = (1 - M_{cvAcc}) * 100 \tag{3}$$

$$M_{cvAcc} = \frac{c_1 + c_2 + c_3}{3} \tag{4}$$

In this expression, M_{cvAcc} is the mean of a threefold cross-validation score (c_1, c_2, and c_3) produced by XGB and RF models.

The other automatic hyperparameter optimization meta-heuristic algorithm used in the study is Bayesian Optimization (BO). It is superior to other evolutionary optimization methods because other optimization algorithms require a large number of training cycles and are noisy (Victoria and Maragatham 2021). Bayesian optimization overcomes this problem by unveiling global optima of the black box function (Victoria and Maragatham 2021) of the selected models, namely, XGB, RF, and CatB. Figure 1 depicts how the meta-heuristic algorithms are employed in our study to accurately do parameters optimization on various models such as XGB, RF, and CatB.

2.3 Kernel-based algorithm

The Kernel-based algorithm used in the SeisEML model is the Kernel-based Ridge Regression (K-RR). The following polynomial kernel has been applied in the K-RR model:

$$k(x, y) = (\gamma x^T y + c_0)^d \tag{5}$$

Where d is the degree selected as 4 in our algorithm, and c_0 is the coefficient set as one by default. The ablation study, as illustrated in Section 4, has been undertaken to determine the optimal degree.

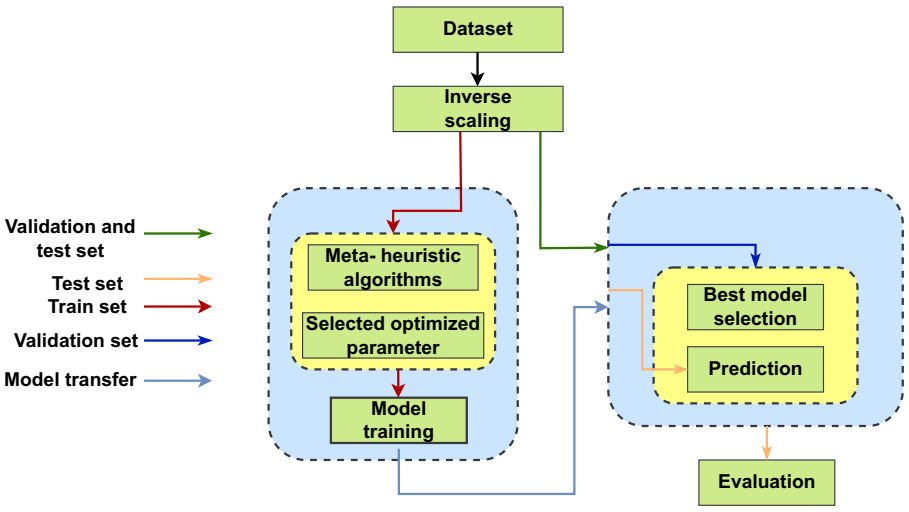

Fig. 1 Optimized model hyperparameter selection flowchart

2.4 SeisEML: a new ML model

It is evident that each of these ML models has its own advantages and disadvantages. A new ML model named SeisEML has been developed in this paper, which has the potential to incorporate the advantages of various ML models. The SeisEML is a five-level hybrid stacked ensemble model which includes various base learners in different layers, including XGB-1, XGB-2, XGB-3, LGBM-1, LGBM-2, CatB, SGWO-RF, SGWO-XGB, BO-CatB, BO-XGB, BO-RF, K-RR-1, K-RR-2, K-RR-3, RR-1, RR-2, and RR-3. The SeisEML takes six input features set denoted by $Input_{p1}$ that include R_e, R_h, d_i, s_t, V_{S30} and M. The predictions from individual models are passed to the following layer of models. The final predictions are then combined using the final layer algorithm. The performance of ensemble models is often more than individual machine learning models (Maclin and Opitz 1999).

The machine learning algorithms are impacted since different features do not have the same scale (Thiagarajan et al. 2017). Hence, feature scaling is necessary. The feature scaling method, RobustScalar, has been applied to reduce the inaccuracy in the final prediction by the following expression:

$$value = \frac{(X_i - X_{med})}{(p_{75} - p_{25})} \tag{6}$$

In the above equation, p_{75} and p_{25} are the $75th$ and $25th$ quantiles. X_{med} is the median of the dataset X, and X_i is the i^{th} instance. The input data is scaled using RobustScalar to reduce outliers and inaccuracy, as shown in Fig. 2a, which resulted in standardization. The SeisEML model passes the dataset through four phases, and following the fourth phase, an inverse RobustScaling is carried out on the predictions passed to the fifth phase. The evaluation is then completed using evaluation metrics such as coefficient of determination (R^2), mean absolute error (MAE), and mean error (ME).

Figure 2a presents the proposed model pipeline to predict PGA from input features. The parameters of the ensemble model have been selected using the grid search method and

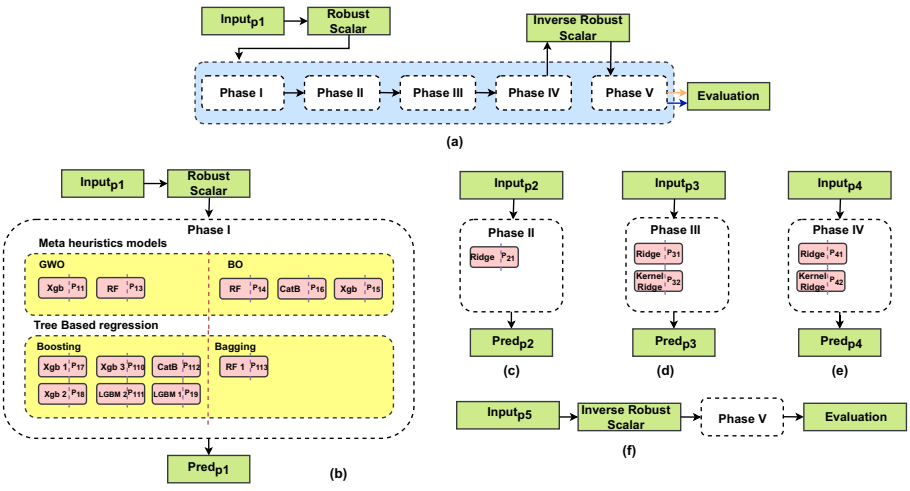

Fig. 2 SeisEML model architecture

meta-heuristic algorithm. The model is trained on 20852 records from the Japan dataset. The robust transformation is performed on these six features before passing them as input to Phase I. The PGA prediction pipeline consists of five major phases. The input set of Phase I ($Input_{p1}$) includes R_e, R_h, d_i, s_t, M, and V_{S30}. The model is tested on similar and cross-region data sets from regions like Iran. Figure 2b shows that Phase I consists of various meta-learners such as XGB-1, XGB-2, XGB-3, RF-1, RF-2, SGWO-XGB, SGWO-RF, BO-XGB, BO-RF, BO-CatB, LGBM-1, LGBM-2, and CatB.

The prediction set obtained from Phase I ($Pred_{p1}$) consists of P_{11}, P_{12}, P_{13}, P_{14}, P_{15}, P_{16}, P_{17}, P_{18}, P_{110}, P_{111}, P_{112}, and P_{113}. The input to Phase II set ($Input_{p1}$) that includes $Pred_{p1}$ and is passed to the Ridge layer as shown in Fig. 2c. The predictions from Phase II are denoted by $Pred_{p2}$. The input feature set in phase III ($Input_{p3}$) includes P_{19}, P_{111}, P_{17}, P_{110}, P_{112}, and P_{16} predictions from Phase I and $Pred_{p2}$ from Phase II, as presented in Fig. 2d. The output from Phase III consists of P_{31} and P_{32}, denoted by $Pred_{p3}$. Phase IV input set denoted by $Input_{p3}$ includes prediction P_{19}, P_{111}, P_{112}, P_{110}, and P_{16} from Phase I, $Pred_{p2}$ from Phase II, and $Pred_{p3}$ from Phase III passed as input to Phase IV as shown in Fig. 2e. The prediction from Phase IV is named as $Pred_{p4}$. Finally, the inverse transform is applied to the predictions and impactful predictions, including P_{16}, P_{112}, P_{110}, P_{111}, P_{18}, P_{19}, P_{21}, $Pred_{p3}$, and $Pred_{p4}$ from Phases I to IV are selected for averaging. Figure 2f represents that the result from Phase V is the final result that is used for evaluation.

3 Data

The effectiveness of the machine learning model relies on both the quality and quantity of the dataset used for its training. There are two prominent earthquake networks in Japan, viz., the Kiban Kyoshin network (KiK-net) and the Kyoshin Network (K-NET) (Kinoshita 1998; Okada 2004). The subsurface information available at each station of the K-NET is extended up to 20 ms depth from the surface but consists of more stations than the KiK-net, which has deeper subsurface information at each stations (Boore et al. 2011). As the model's training is highly dependent on the quantity of data, this study uses data from K-NET for training, testing, and validating the developed ML model. The K-NET data has been maintained by the Nation Research Institute for Earth Science and Disaster Resilience (NIED), Japan (Aoi et al. 2011). The K-NET dataset consists of high-quality near-source recordings in a considerable amount from the year 1996. The entire dataset used in this work consists of 29789 records from various earthquakes recorded at 1011 stations that are recorded between 1996 and 2022 with a magnitude range of 3 to 9 on the JMA scale. The location of the epicenter of earthquakes used for training, testing and validation is shown in Fig. 3a. The total distribution of records according to the magnitude of earthquakes utilized in the present study, as demonstrated in Fig. 3b, indicates that the dataset consists of a sufficient number of major destructive earthquakes.

The entire dataset is divided into training, validation, and testing datasets, consisting of 20852, 2681, and 6256 records from 412, 336, and 377 earthquakes, respectively. The average shear wave velocity at 30 m depth (V_{S30}) is one of the major inputs in the present ML model. The V_{S30} value at each station has been computed from the assumption of constant velocity extrapolation given by Boore et al. (2011), which assumes that the velocity of the last layer continues up to 30 ms. The depth-wise distribution of shear wave velocity has been given by K-NET seismic network archives available at (http://www.kyoshin.bosai.go.jp/). The range of datasets used in the model's training is one of the crucial factors that control the

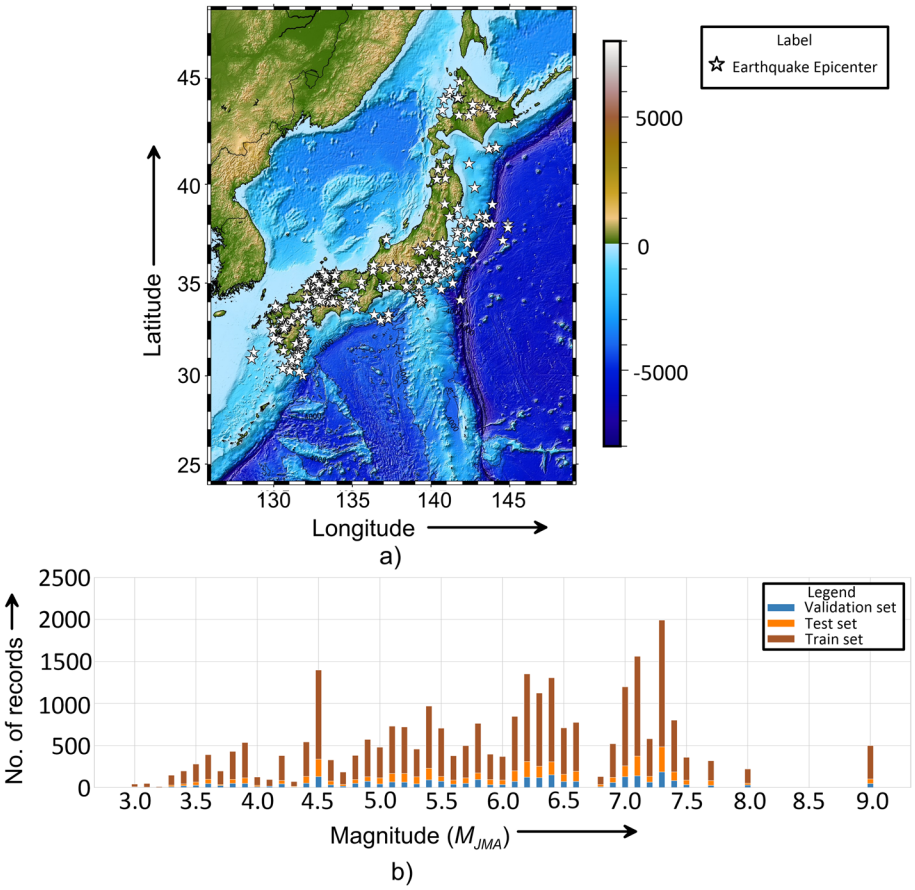

Fig. 3 (**a**) The epicenter of earthquakes from the Japan dataset taken for the test, train, and validation set are presented by the white color star. (**b**) Distribution of magnitudes of earthquakes used in our dataset based on validation, test, and train set given by blue, orange, and brown color

model's performance. The range of various parameters used in the model's training is given in Table 1. It is seen from this model that the testing and validation data set is within the range of data used for the training of the model. Other than this data, acceleration data from three earthquakes that have not been included in the training, testing, and validation data set have

Table 1 Range of various parameters used in the training, testing and validation of the model

Dataset	V_{S30}		Epicenter Distance(km)		Hypocenter Distance(km)		Magnitude(M_{JMA})		Focal Depth(km)	
	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max
Train	39	1579	0.02	985.23	5.57	985.53	3	9	0	256
Test	39	1579	1.08	955.20	7.62	955.50	3	9	0	256
Validation	39	1579	3.02	946.25	7.79	946.32	3	9	0	256

been utilized to prepare iso acceleration scenarios during these earthquakes. The parameters of these earthquakes are given in Table 2. A total of 428, 290, and 215 records have been used from these three earthquakes for the preparation of the iso acceleration scenario for comparison with the predictions from the ML model. The locations of epicenters of these three earthquakes are shown in Fig. 4.

The cross-region applicability of the trained model has been utilized for predicting the peak ground motion scenario of the Sarpol-e Zahab earthquake in Iran (Panahi 2017). Records from 37 stations shown in Fig. 4d that have recorded the Sarpol-e Zahab earthquake have been used in the present work. The parameters of this earthquake are given in Table 3. Once the data set is finalized, the next task in the training of the ML model is based on extracting various features from the record and available sources.

The regression models popularly used in various attenuation studies (Abrahamson and Litehiser 1989; Campbell 1997; Douglas 2003b) clearly indicate that the logarithm of peak ground acceleration (PGA_{ln}) shows linear dependency on various independent parameters like the magnitude of the earthquake, hypocentral distance, epicenter distance, focal depth, and site parameter. Therefore, PGA_{ln} has been used in this study for training the machine learning model. In this work, the model has been trained with several parameters that include hypocentral distance R_h (km), epicentral distance R_e (km), focal depth d_i, magnitude M, V_{S30} and site characteristics (s_t). The s_t is represented by the classifications based on the average shear wave velocity at 30 ms depth (V_{S30}) (Borcherdt 2012). The K-NET data consists of only up to 20 ms depth of V_{s30}. Therefore, the V_s values are extrapolated up to 30 ms from the obtained velocity at the last depth till the data is present (i.e., a constant-velocity extrapolation) (Boore et al. 2011). The following equation has been used to calculate the V_{S30} value by using the shear wave velocity profile available at a given site:

$$V_{S30} = \frac{30m}{\sum_i \Delta t_i} \tag{7}$$

In the above equation, the parameter t_i represents the wave travel time of a shear wave in a particular layer, which is calculated by using the following equation:

$$\sum_i \Delta t_i = \frac{\Delta d_i}{v_{si}} \tag{8}$$

In the above equation, d_i represents the thickness of the intervening i^{th} layer between the surface and 30 m depth, and v_{si} is the shear wave velocity in the respective layer. The value of V_{S30} obtained at a particular site serves as a basis for the classification of the site given by the NEHRP (BSSC 2004). Based on various ranges of V_{S30} value, the soil has been

Table 2 Parameters of earthquake case study dataset

Parameters	Case 1	Case 2	Case 3
Origin Date	2022/03/16	2022/01/22	2022/03/16
Origin Time	23:36:00	1:08:00	23:34:00
Epicenter	37.697, 141.622	32.715, 132.072	37.68, 141.605
Depth (km)	57	45	57
Magnitude (M_{JMA})	7.4	6.6	6.1
Total Records	428	290	215

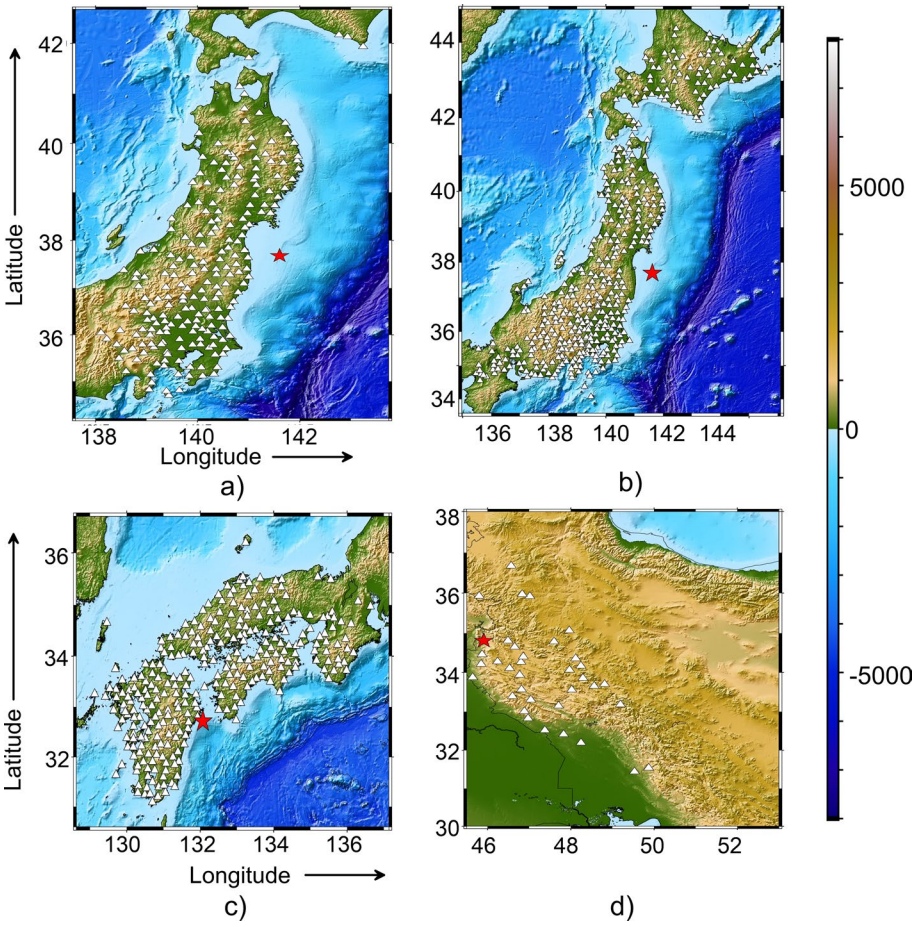

Fig. 4 (**a**) The Off Fukushima earthquake magnitude 7.4, (**b**) foreshock of the Off Fukushima earthquake of 6.1 M_{JMA} and (**c**) the Hyuganada earthquake (6.6 M_{JMA}). (**d**) Sarpol-e Zahab earthquake from the region of Iran of 7.3 M_w. The stations and epicenter of the earthquake are represented by a red-colored star and a white-colored triangle, respectively

Table 3 Parameters of earthquake used for cross-region prediction of Iran earthquake

Parameters	Sarpol-e Zahab (Iran)
Origin date	12/11/2017
Origin time	18:18:17 (UTC)
Epicenter	34.81, 45.91
Depth (h)	18 km
Magnitude (M_w)	7.3
Total records	37

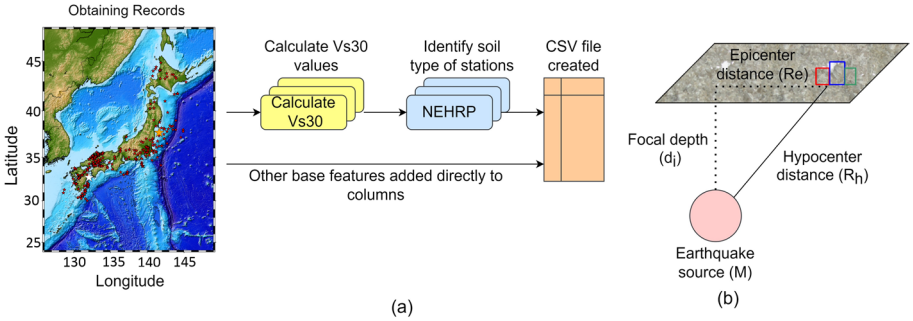

Fig. 5 (**a**) Feature extraction pipeline that shows an example of seismic records from which features are to be extracted. calculated V_{S30} value and the example which shows how to calculate V_{S30} values. Identify the soil type by calculating V_{S30} values based on NEHRP code and create a csv file of features using the python program. (**b**) Various parameters have been used in the training of the model

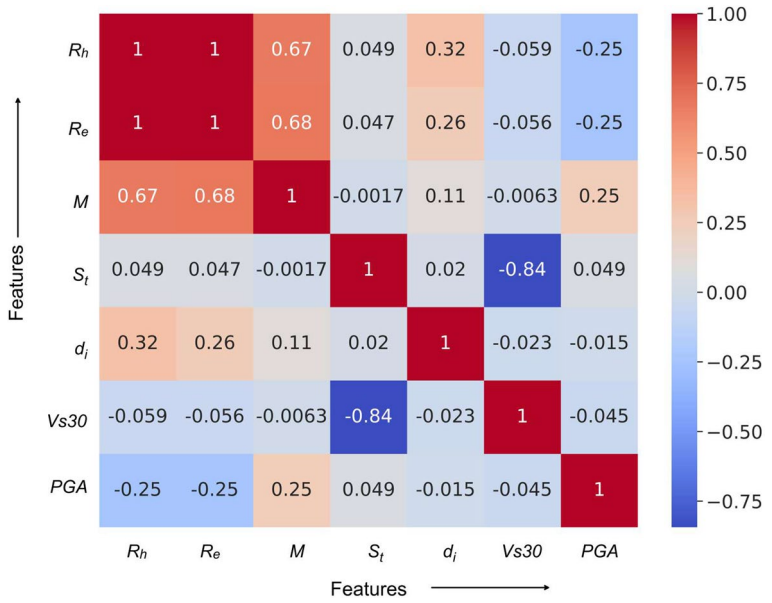

Fig. 6 Pearson correlation coefficient matrix of features from the Japan dataset

classified into A, B, C, D, and E types. According to NEHRP (BSSC 2004), the respective range of V_{S30} for A, B, C, D, and E is taken into account. One dimensional shear wave velocity profile is available at each station of the Kyoshin Network (K-NET) and has been used to calculate the value of V_{S30}, which has been further used for site classification.

The flow diagram for the computation of s_t based on the V_{S30} value at each station using the shear wave velocity profile is shown in Fig. 5a. Various parameters that have been used in the training of the model are shown in Fig 5b.

The Pearson correlation coefficient matrix, which depicts the correlation and direction of the relationship between two variables (Saidi et al. 2019), is shown in Fig. 6. The

Pearson correlation coefficient is represented by the symbol r and is calculated using the following formula (Saidi et al. 2019):

$$R = \frac{N\sum x'y' - (\sum x' \sum y')}{\sqrt{[N\sum x'^2 - (\sum x')^2][N\sum x'^2 - (\sum y')^2]}} \tag{9}$$

In the above equation, N represents the total number of instances. The x' and y' represent two different variables. The positive, negative and uncorrelated relation between two variables is represented by the +1, -1 and 0 values, respectively (Adler and Parmryd 2010). The multicollinearity between various parameters used in the training of the model has been studied by calculating the correlation coefficient of a variable with an individual or among a group of variables. Multicollinearity between various parameters used in the training of the model is visualised using the Pearson correlation coefficient shown in Fig. 6. The Pearson correlation coefficient between various parameters used in training of the model shows that the conventional parameters like R_h, R_e and M have a good negative or positive correlation with PGA, while the newly introduced parameters like V_{S30}, S_t and d_i shows a weak correlation with PGA. However, it can be observed that these parameters show a relatively good correlation with other parameters that have a direct influence on PGA.

3.1 Training, testing and validation of model

The SeisEML model developed in this paper includes various base learners that belong to hybridized models, kernel methods, tree regression, and ridge regression, respectively. The current study partitions the dataset into training, validation, and testing sets. The validation set serves the purpose of selecting the optimal model by comparing their performances. The entire dataset is divided into 20852 records from 421 earthquakes, 6256 records from 383 earthquakes, and 2681 records from 340 earthquakes randomly selected to prepare training, testing, and validation data sets for the developed machine learning model. The performance of the developed model has been checked based on R^2, MAE, and ME, respectively.

The model ablation study for SeisEML is presented in Table 4. The outcome is expressed as the logarithm of the predicted peak ground acceleration, which represents the higher value between the horizontal components denoted by PGA_{lg}. Table 4 indicates that the SeisEML model gives better results in terms of statistical parameters such as R^2 and MAE than all other individual models. The quantitative analysis,

Table 4 The model ablation study is based on R^2 and MAE for the test and validation dataset

Models	Test		Validation	
	R^2	MAE	R^2	MAE
GPR	0.68	0.145	0.68	0.147
ExTR	0.69	0.1407	0.69	0.1419
RSubCatB	0.69	0.1477	0.70	0.147
AdaBoost	0.41	0.203	0.37	0.207
DTree	0.66	0.150	0.66	0.151
RR	0.41	0.204	0.41	0.204
SeisEML	**0.77**	**0.123**	**0.76**	**0.130**

Table 5 An analysis of feature ablation for optimal input feature set selection in the SeisEML model

SNo.	Feature set	R^2	*MAE*	*RMSE*
1	[R_h,R_e,M,S_t,d_i,V_{S30}]	0.74	0.49	0.62
2	[R_e, M, S_t, d_i, V_{S30}]	0.72	0.51	0.64
3	[M, S_t, d_i, V_{S30}]	0.16	0.89	1.12
4	[S_t, d_i, V_{S30}]	0.09	0.93	1.16
5	[d_i, V_{S30}]	0.09	0.93	1.16
6	[V_{S30}]	0.04	0.96	1.2

Table 6 A model ablation study for the determination of the optimal degree (d) in kernel ridge regression

Degree (d)	*MAE*	*RMSE*
1	0.13	0.16
2	0.13	0.16
3	0.13	0.16
4	0.12	0.15
5	0.12	0.16

as depicted in Table 4 shows that the individual models, such as GPR, Extra Tree Regressor (ExTR), Random Subspace Catboost (RSubCatB), AdaBoost, Decision Tree (Dtree), and SeisEML also give a better result after SeisEML. The quantitative analysis of diverse statistical indicators unambiguously demonstrates that the SeisEML model, which incorporates hybridized models, kernel-based algorithms, tree regression algorithms, and regression algorithms, respectively, exhibits superior performance compared to the standalone models. The results obtained by SeisEML model have been displayed by bold numerical values.

The feature ablation study has been performed to select the best feature set that gives minimum *MAE* and *RMSE* error when trained and tested with the XGBoost model, as shown in Table 5. The best performance is obtained in the first set. Table 6 presents the ablation analysis conducted to determine the optimal degree (d) for kernel ridge regression within our algorithm, revealing the selection of $d = 4$. The model receives input data consisting of predictions generated by the LightGBM, XGB1, XGB2, XGB3, CatB, and BO-CatB models. The evaluation metrics values such as *MAE* and R^2 score obtained from the transformed test dataset using various individual models and stacking layers of the SeisEML model are shown in Fig. 7a and b for PGA_{ln} prediction. Figure 7a suggests that among all developed models, the minimum *MAE* of value 0.123 has been obtained for the Stack4 model, which leads to the five-layer SeisEML model. The R^2 score for this model is obtained as 0.77. The model ablation study reveals that utilizing the five-layer model in the trained SeisEML model leads to an increase in the R^2 score and a reduction in *MAE* values. The *MAE* of 0.12, 0.13, 0.089 and R^2 scores of 0.77, 0.76, and 0.88 are obtained in the predictions of PGA_{ln} from SeisEML model using testing, validation, and training data sets, respectively. Figure 8a, b, and c compare the actual and the predicted PGA values from the SeisEML model using training, validation and testing data sets. This suggests that the performance of the trained model is satisfactory in predicting the PGA_{ln} values for both the testing and validation data sets.

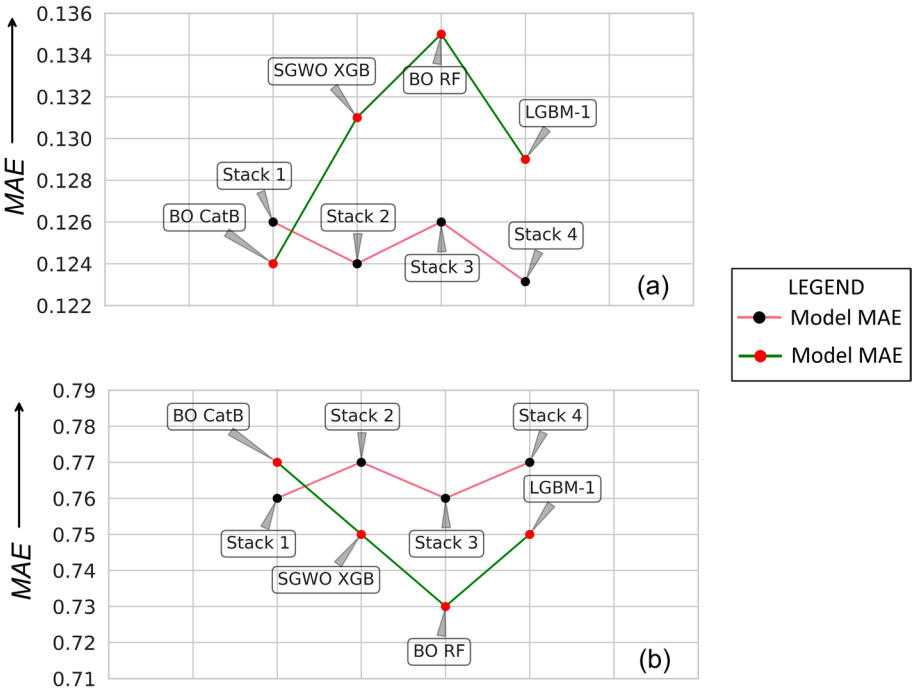

Fig. 7 (**a**) *MAE* and (**b**) R^2 score of the individual models and stacking layer within the SeisEML model are presented. The errors are calculated based on the predictions of the PGA_{ln} obtained from different machine learning models and the stacking layer of the SeisEML model. The green and red lines in the plot indicate the results obtained from the various ML models and the stacking layer of the SeisEML model, respectively

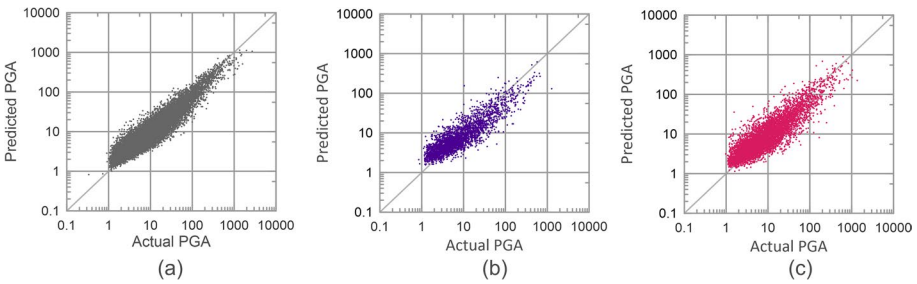

Fig. 8 The gray, purple, and red solid circles show the predicted *PGA* using the (**a**) train, (**b**) validation, and (c) test set with respect to the actual *PGA*

4 Discussion

The ground motion prediction equations (GMPE) are treated as an effective and simple means for predicting peak ground accelerations by various workers. This paper compares the PGA predicted by the ML model with the GMPE given by Shoushtari et al. (2018) and Abrahamson and Litehiser (1989). The GMPE given by Shoushtari et al. (2018) and Abrahamson and Litehiser (1989) is based on moment magnitude (M_w) and surface wave

magnitude (M_s), while the test data set consists of the magnitude in the JMA scale (M_{JMA}). The conversion formula given by Scordilis (2005) has been used in the present work to convert M_{JMA} into M_w and M_s for the calculation of PGA using GMPEs. The M_{JMA} to M_w conversion has been shown in the equation below:

$$M_w = 0.58M_{JMA} + 2.25, 3.0 \leq M_{JMA} \leq 5.5 \tag{10}$$

$$M_w = 0.97M_{JMA} + 0.04, 5.6 \leq M_{JMA} \leq 8.2 \tag{11}$$

The GMPE given by Shoushtari et al. (2018) is one of the relations which has been prepared for subduction zone earthquakes using earthquake data from the Malay Peninsula and Japan. The following equation represents the GMPE given by Shoushtari et al. (2018):

$$\log_{10} PGA_{pred} = aM_w + bR - \log_{10}(R + c10^{aM_w}) + d + s_T \tag{12}$$

In the above equation, PGA_{pred} is the value of the predicted PGA in cm/s^2, R is hypocentral distance (km), and M_w is the moment magnitude of the earthquake. Coefficients a, b, c, and d are 0.4683, −0.002159, 0, and 0.6524, respectively. The parameter s_t is dependent on the soil classification given by NEHRP and its values as given by Shoushtari et al. (2018). The range of magnitude used in the data set for generating the regression model given by Shoushtari et al. (2018) is from 5.0 to 9.1 M_w.

The ground motion prediction equation given by Abrahamson and Litehiser (1989) has been utilized by different workers for calculating peak ground accelerations from several worldwide earthquakes. The M_s value is computed by utilizing M_w obtained through the conversion equation provided by Scordilis (2005). In the present work, this relation has been used to compare predicted peak ground acceleration using the SeisEML model for cross-region learning of the developed model. The following attenuation relation has been given by Abrahamson and Litehiser (1989):

$$\log_{10} PGA_{pred} = \alpha + \beta M_s - \bar{C}\log_{10}[r + e^{h_2 M_s}] + F\phi + Ebr \tag{13}$$

$$PGA_{pred} = 10^{\alpha + \beta M_s - \bar{C}\log_{10}[r + e^{h_2 M_s}] + F\phi + Ebr} * 980 \tag{14}$$

In this expression, parameters r and M_s represent the hypocentre distance and the surface wave magnitude of the earthquake. The variable $\bar{C}$, α, β, h_2, ϕ, and b have the value as 0.982, −0.62, 0.177, 0.284, 0.132, and −0.0008 for horizontal PGA prediction. The local conditions for Japanese earthquakes favor setting $F = 1$ and $E = 1$. Therefore, the same values are used for PGA prediction. The range of magnitude used in the dataset by Abrahamson and Litehiser (1989) is between 5.0 to 8.1. Consequently, adhering to the prescribed upper and lower limits defined by Abrahamson et al. (2014) and Shoushtari et al. (2018), the range for comparison between the SeisEML model predictions and the PGA estimates obtained from GMPEs has been established. Specifically, out of 6256 records in the test set, a subset of 3911 records from 336 earthquakes recorded at 841 stations has been selected for this comparative analysis.

The Fig. 9a, b, and c represent the density plot of the actual and predicted PGA values using the SeisEML model, the GMPEs given by Shoushtari et al. (2018) and Abrahamson and Litehiser (1989), respectively. The comparison in Fig. 9 shows that the linear trend between the actual and predicted model is maintained clearly in the predictions from the SeisEML model, while deviations from linearity are seen for predictions from GMPEs. The

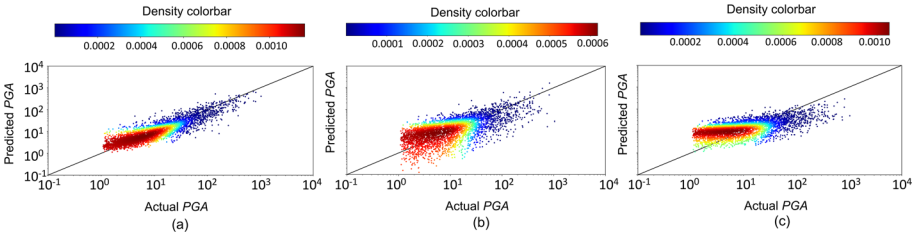

Fig. 9 The solid circle shows the scatter plot colored by the density of PGA predicted using (**a**) SeisEML model, (**b**) GMPE by Shoushtari et al. (2018), and (**c**) Abrahamson and Litehiser (1989) attenuation relation, respectively. The scale of the respective plot is given below every plot

R^2 score between actual and predicted values obtained from the developed model, GMPE relations by Shoushtari et al. (2018) and Abrahamson et al. (2014) is obtained as.80,.34 and.32, respectively. The R^2 score indicates the predictions using SeisEML are strong to very strong, while they are weak to moderate for GMPEs (Akoğlu 2018; Dancey and Reidy 1999; Chan 2003). This clearly validates the efficacy of the developed ML model in this work.

The performance of the prediction methodology can be evaluated by observing the difference between predicted and observed PGA. The following formula calculates this difference (ERR):

$$ERR = PGA_{pred} - PGA_{actual} \tag{15}$$

Where PGA_{pred} and PGA_{actual} are the predicted and actual PGA. The *MAE* for PGA predictions on the entire Japan region's test dataset, as detailed in Table 1 and Fig. 3, using the SeisEML model, Shoushtari et al. (2018), and Abrahamson and Litehiser (1989) GMPEs, are 10.90, 17.82, and 18.63 gal, respectively. The *RMSE* obtained for predictions using SeisEML, Shoushtari et al. (2018) and Abrahamson and Litehiser (1989) is 31.66, 50.99, and 48.08 gal, respectively. This clearly shows that the developed machine learning model performs better than GMPEs.

The ERR value for individual records obtained in the PGA predictions with respect to hypocenter distance by the SeisEML model and GMPEs given by Shoushtari et al. (2018) and Abrahamson and Litehiser (1989) is shown in Fig. 10a, b and c, respectively. There are 3911 records selected in the test set after filtering the magnitude range using Shoushtari et al. (2018) and Abrahamson et al. (2014) methods. This clearly indicates that the performance of the developed SeisEML model is better in the near-source regions than that of GMPEs. This is equally important as maximum damage during an earthquake is in the near-source region close to the hypocenter of an earthquake. The comparison of prediction error (ERR) with respect to the actual PGA for the SeisEML model and GMPE relations by Shoushtari et al. (2018), and Abrahamson and Litehiser (1989) is shown in Fig. 10d, e, and f. This represents that the performance of GMPEs is relatively poor as compared to that of the SeisEML model for high values of PGAs.

The iso acceleration contour map prepared from the PGA values recorded at different stations gives valuable information about the destruction pattern during an earthquake. The performance of the trained model to predict the PGA distribution due to an individual earthquake has been checked by comparing the observed and predicted iso acceleration contours of three strong to major earthquakes that occurred in Japan in 2022. The

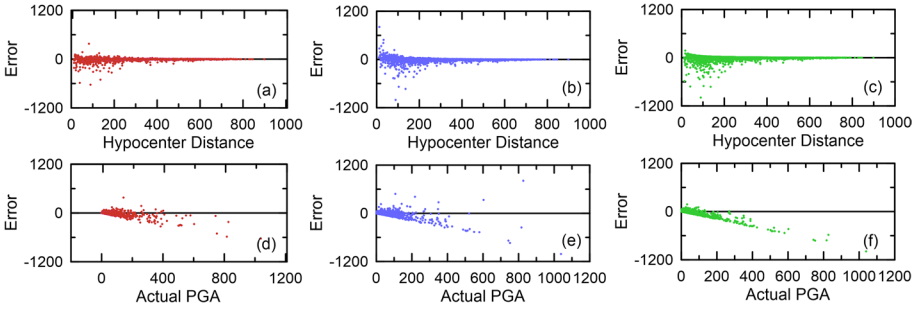

Fig. 10 (**a**) The error obtained from PGA predicted using the SeisEML model and (**b**) Shoushtari et al. (2018), and (**c**) Abrahamson and Litehiser (1989) relation in comparison to hypocenter distance. The error obtained from the predicted PGA by (**d**) SeisEML model and (**e**) Shoushtari et al. (2018), and (**f**) Abrahamson and Litehiser (1989) with relation to actual PGA

parameters for these earthquakes are provided in Table 2. The iso acceleration contours of these three earthquakes have been prepared by using PGA values obtained from observed data, SeisEML model, and regression relation given by Shoushtari et al. (2018). The comparison of observed and predicted PGA using different methods in Fig. 11a, b, and c shows that the PGA predicted using SEisEML model matches closely with the observed data as compared to that obtained from the regression relation given by Shoushtari et al. (2018) in all three earthquakes. The trend of iso acceleration contours obtained by the observed data shown in Fig. 11d, e and f matches closely with that prepared from predicted PGA using SeisEML model shown in Fig. 11g, h, and i in the near-source region. However, the trend of iso acceleration contours prepared using the Shoushtari et al. (2018) relation and shown in Fig. 11j, k, and l shows clear deviation from observed trends in the near-source region. This shows that the developed SeisEML model has a high potential in modeling a realistic earthquake scenario for Japanese earthquakes, especially in the near-source region where damage is expectantly high.

In order to validate the applicability of the developed ML model for another region, the cross-region prediction has been made to an earthquake that occurred in a different region having similar tectonic settings. The Sarpol-e Zahab earthquake (M 7.3) that occurred in the subduction zone boundary has been utilized for this purpose. Figure 12a and b illustrate the geospatial map, station locations, and the epicenter of the earthquake. Furthermore, Table 3 provides the parameters associated with these earthquakes. The comparison of PGA obtained from the SeisEML model with that from the observed data is shown in Fig. 12c. The trend of iso acceleration contours prepared from observed data and that obtained from SeiSEML shows that over-prediction in some stations is visible in Fig. 12d and e, respectively. The comparison in Fig. 12c shows that higher PGA values match at many stations. The mismatch in PGA values can be attributed to the use of Japanese data in which the epicentre of a large number of earthquakes falls in the oceanic region that contributes to higher hypocentral and epicentral distances, while the epicenter of the Iran earthquake is within the continent that amount to low hypocentral and epicentral distances. The distribution of error between observed and predicted PGA clearly shows that the given error is high in some of the near-source stations shown in Fig. 12f. It matches substantially well with the observed values at most of the stations, as shown in Fig. 12c.

The reliability of the developed SeisEML model can be checked by comparing the *RMSE* and *MAE* score obtained from the comparison of the logarithm of the PGA values

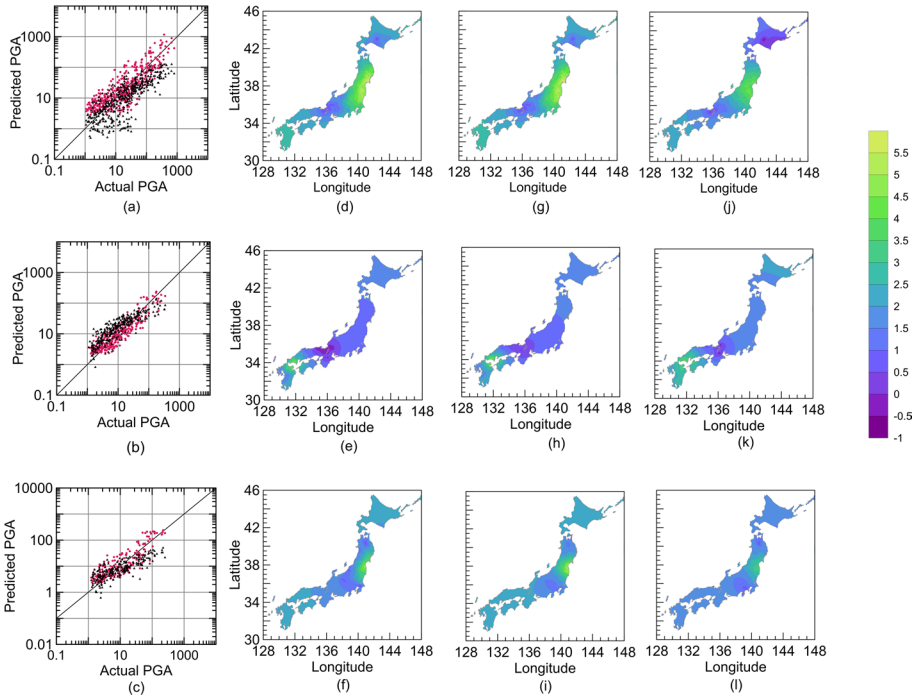

Fig. 11 Comparison of actual and predicted PGA for (**a**) 7.4 M_{JMA} Off Fukushima earthquake that occurred in Japan on 16-March-2022 (**b**) 6.6 M_{JMA} earthquake that occurred as foreshock of the Off Fukushima earthquake that occurred in Japan on 16-March-2022 (**c**) 6.6 M_{JMA} earthquake. Iso acceleration contour maps were prepared using actual acceleration data from (**d**) 7.4, (**e**) 6.6, and (**f**) 6.1 M_JMA earthquake, respectively. Iso acceleration contour map prepared using predicted acceleration data from the SeisEML model for (**g**) 7.4, (**h**) 6.6, and (**i**) 6.1 M_{JMA} earthquake. Iso acceleration contour maps (**j**) 7.4,(k) 6.6, and (l) 6.1 M_{JMA} earthquake, respectively, using Shoushtari et al. (2018) relation

obtained from predictions using the SeisEML model and GMPEs given by Shoushtari et al. (2018) and Abrahamson and Litehiser (1989), respectively, in Tables 7 and 8. It is seen clearly that among all methods of predictions of PGA, the SeisEML model gives the lowest score of *RMSE* and *MAE* for overall test data as well as for all individual earthquakes used in the present study. The comparison of *RMSE* and *MAE* scores in Tables 7 and 8 shows that the SeisEML model performed better for the Japanese earthquake when compared with the Iran earthquake. However, predictions of the PGA from the SeisEML model are relatively good for a cross-region earthquake compared to the popular regression relations (GMPE1 and GMPE2). This clearly establishes the efficacy of the improved GMPE developed using the ML model for the prediction of PGA for cross-region predictions of earthquakes in a similar tectonic environment.

In a recent study conducted by Somala et al. (2021), a range of machine learning algorithms, including linear regression, support vector regression, k-nearest neighbors, random forest, and XGBoost, were employed to predict peak ground acceleration. It is found by Somala et al. (2021) that the random forest algorithm gives a minimum *RMSE* of about 0.036 g in the test data set. The work by Thomas et al. (2017) utilizes a support vector machine to predict peak ground acceleration. It is seen that the *MAE* obtained in the prediction of PGA using the least square support vector machine is.0316 g (Thomas et al.

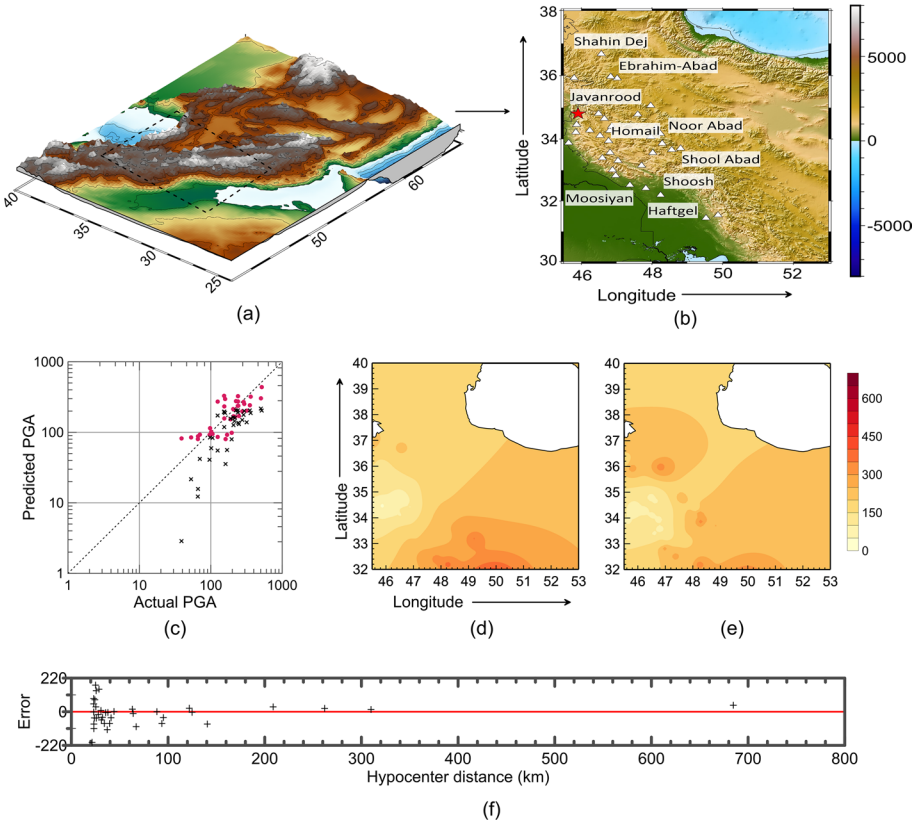

Fig. 12 (**a**) Three-dimensional geospatial map of Iran and (**b**) a region targeted for cross-regional prediction. (**c**) Relationship between the model's predicted *PGA* and the *PGA* predicted using Abrahamson and Litehiser (1989). (**d**, **e**) The iso acceleration contours produced by the actual and the predicted *PGA* by the SeisEML model. (**f**) *PGA* error that is anticipated concerning the hypocenter distance

Table 7 Generalized result analysis based on *RMSE* score

Parameters	Number of records	SeisEML	GMPE 1	GMPE 2
Japan test set	3911	0.58	1.06	1.05
Case 1 (7.4 M Japan)	428	0.83	1.22	1.16
Case 2 (6.6 M Japan)	290	0.46	0.83	0.66
Case 3 (6.1 M Japan)	215	0.70	0.93	0.85
Iran (7.3 M)	37	0.38	0.81	–

GMPE 1 is the GMPE method given by Abrahamson and Litehiser (1989)

GMPE 2 is the GMPE method given by Shoushtari et al. (2018)

2017) using test data sets. Although the training and testing datasets are different in these machine learning models, the efficacy of the SeisEML model can be established by comparing the prediction errors in the test data. Observations reveal that the SeisEML model

Table 8 Generalized result analysis based on *MAE* score

Parameters	Number of records	SeisEML	GMPE 1	GMPE 2
Japan test set	3911	0.45	0.88	0.86
Case 1 (7.4 M Japan)	428	0.69	1.00	0.91
Case 2 (6.6 M Japan)	290	0.46	0.83	0.66
Case 3 (6.1 M Japan)	215	0.57	0.80	0.69
Iran (7.3 M)	37	0.38	0.62	–

GMPE 1 is the GMPE method given by Abrahamson and Litehiser (1989)

GMPE 2 is the GMPE method given by Shoushtari et al. (2018)

has achieved a *RMSE* of 0.035g and *MAE* of 0.010g. These values are relatively less than those observed in recent studies by Somala et al. (2021) and Thomas et al. (2017), respectively. The comparison of R^2 score and *MAE* with different ML models in Table 9 clearly shows that the developed model has the potential for realistic prediction of PGA for earthquakes in a similar tectonic environment.

5 Conclusion

In this work, a machine learning model named SeisEML has been trained to predict PGA at a given site using various earthquake and site-dependent parameters. This machine learning model uses hybridized models (SGWO-XGB, BO-XGB, BO-CatB, SGWO-RF, and BO-RF), kernel-based algorithms (K-RR), tree regression algorithm (XGB-1, XGB-2, XGB-3, CatB, LGBM-1, LGBM-2, RF-1, and RF-2), and regression algorithm (RR). The developed model utilizes various input parameters like R_e, R_h, d_i, soil type (s_t), V_{S30}, and M for predicting PGA at different sites. The statistical performance indicators obtained by using the validation set on several individual ML models allowed us to select the top seventeen base models for the SeisEML model. The developed model underwent training, testing, and validation using the K-NET data from Japan. The testing, validation, and training datasets comprise 6256, 2681, and 20852 strong motion records recorded from various earthquakes at different stations. The predictions obtained from the SeisEML model have been compared with those obtained from popular GMPEs developed by Shoushtari et al. (2018), and Abrahamson and Litehiser (1989) in terms of *RMSE* and *MAE*. The

Table 9 SeisEML prediction model compared with different prediction models

Metric	SeisEML	Ls-SVR RBF kernel	ANN/SA	GP/OLS	MEP
R^2	0.770	0.76	0.73	0.66	0.70
MAE	0.01	0.0316	0.13	0.488	0.697

Ls-SVR RBF kernel method given by Thomas et al. (2017)

ANN/SA method given by Alavi and Gandomi (2011)

GP/OLS method given by Gandomi et al. (2011)

MEP method given by Alavi et al. (2011)

comparison shows that the SeisEML model performs better than the regression models both in terms of error and values.

The performance of the developed model has been evaluated on its capability to predict PGA scenarios of individual earthquakes. For this purpose, additional strong motion data from three distinct strong to major earthquakes, which were not included in the training and testing datasets, have been utilized. The results from the SeisEML model have been compared with the popular GMPE relations given by Shoushtari et al. (2018) and Abrahamson and Litehiser (1989). The statistical performance indicators such as *MAE* and *RMSE* indicate that the trained SeisEML model performs better than these regression models. The developed model has been tested for cross-region prediction of an earthquake in a similar tectonic environment. The Iran earthquake (M 7.3) recorded at 37 stations has been utilized for this purpose. It is seen that the predictions from the SeisEML model for this earthquake are comparatively better than those obtained using the regression model. The comparison of PGA predicted using the SeisEML model for Japanese earthquakes and an earthquake in a similar tectonic setup clearly establishes the efficacy of the trained model for predicting PGA values for earthquakes in similar tectonic scenarios. The predictions from the ML model show that it can considerably enhance the reliability of PGA prediction. Therefore, an improved GMPE using the ML model has been proposed in this study.

Acknowledgements We would like to thank Prof. John Clague and reviewers for manuscript suggestions. We thank Prof. Nagendra Kumar and Ms. Swati Singh for proof reading the manuscript and correcting grammatical errors. We would also like to thank National Research Institute for Earth Science and Disaster Resilience (NIED) for providing earthquake data.

Funding This project is supported from grant for PhD research by Prime Minister Research Fellowship (Grant code: PM-31-22-626-414).

Data availability The data used in this research work is taken from National Research Institute for Earth Science and Disaster Resilience (NIED).

Code availability (software application or custom code) Code used in this paper has been developed in Python.

Declarations

Conflict of interest We declare that there are no competing interests related to any financial concern with regard to the publication of a study. It is being declared that there are no personal relationships with people or organization that may influence or may be perceived to influence the research work described in this paper.

Ethical approval The work presented in this paper does not involve research using humans and/or animals.

References

Abrahamson NA, Litehiser JJ (1989) Attenuation of vertical peak acceleration. Bull Seismol Soc Am 79:549–580

Abrahamson NA, Silva WJ, Kamai R (2014) Summary of the ask14 ground motion relation for active crustal regions. Earthq Spectra 30:1025–1055

Adler J, Parmryd I (2010) Quantifying colocalization by correlation: The pearson correlation coefficient is superior to the mander's overlap coefficient. Cytometry A 77A

Akoğlu H (2018) User's guide to correlation coefficients. Turk J Emer Med 18:91–93

Alavi AH, Gandomi AH (2011) Prediction of principal ground-motion parameters using a hybrid method coupling artificial neural networks and simulated annealing. Comput Struct 89:2176–2194

Alavi AH, Gandomi AH, Modaresnezhad M et al (2011) New ground-motion prediction equations using multi expression programing. J Earthq Eng 15:511–536
Ambraseys N, Simpson K, Bommer JJ (1996) Prediction of horizontal response spectra in europe. Earthq Eng Struct Dyn 25:371–400
Anbazhagan P, Kumar A, Sitharam TG (2013) Ground motion prediction equation considering combined dataset of recorded and simulated ground motions. Soil Dyn Earthq Eng 53:92–108
Aoi S, Kunugi T, Nakamura H, et al. (2011) Deployment of new strong motion seismographs of k-net and kik-net. In: In: Akkar S, Gülkan P, van Eck T (eds) Earthquake Data in Engineering Seismology. Geotechnical, Geological, and Earthquake Engineering, Springer, Dordrecht
Boore DM (1983) Stochastic simulation of high-frequency ground motions based on seismological models of the radiated spectra. Bull Seismol Soc Am 73:1865–1894
Boore DM, Atkinson GM (1987) Stochastic prediction of ground motion and spectral response parameters at hard-rock sites in eastern north america. Bull Seismol Soc Am 77:440–467
Boore DM, Atkinson GM (2008) Ground-motion prediction equations for the average horizontal component of pga, pgv, and 5%-damped psa at spectral periods between 0.01 s and 10.0 s. Earthq Spectra 24:138–99
Boore DM, Thompson EM, Cadet H (2011) Regional correlations of vs30 and velocities averaged over depths less than and greater than 30 meters. Bull Seismol Soc Am 101:3046–3059
Boore DM, Stewart JP, Seyhan E et al (2014) Nga-west2 equations for predicting pga, pgv, and 5% damped psa for shallow crustal earthquakes. Earthq Spectra 30:1057–1085
Borcherdt R (2012) Vs30 - a site-characterization parameter for use in building codes, simplified earthquake resistant design, gmpes, and shakemaps. In: 15th World Conf. on Earthquake Engineering
Breiman L (1996) Bagging predictors. Mach Learn 24(2):123–140
Breiman L (2004) Random forests. Mach Learn 45:5–32
Breiman L, Friedman JH, Olshen RA, et al (1984) Classification and regression trees. In: Belmont, Calif.: Wadsworth
BSSC (2004) Nehrp recommended provisions for seismic regulations for new buildings and other structures (fema 450). 2003 edition. part 1: Provisions, buildin seismic safety council, national institute of building sciences, washington, d.c
Campbell KW (1997) Empirical near-source attenuation relationships for horizontal and vertical components of peak ground acceleration, peak ground velocity, and pseudo-absolute acceleration response spectra. Seismol Res Lett 68:154–179
Campbell KW, Bozorgnia Y (2014) Nga-west2 ground motion model for the average horizontal components of pga, pgv, and 5% damped linear acceleration response spectra. Earthq Spectra 30:1087–1115
Chan YH (2003) Biostatistics 104: correlational analysis. Singap Med J 44:614–619
Chen T, Guestrin C (2016) Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining
Dancey CP, Reidy J (1999) Statistics Without Maths For Psychology. Pearson Education Limited, https://api.semanticscholar.org/CorpusID:117870366
Derras B, Bard PY, Cotton F et al (2012) Adapting the neural network approach to pga prediction: an example based on the kik-net data. Bull Seismol Soc Am 102:1446–1461
Derras B, Bard PY, Cotton F (2013) Towards fully data driven ground-motion prediction models for europe. Bull Earthq Eng 12:495–516
Derras B, Bard PY, Cotton F (2016) Site-condition proxies, ground motion variability, and data-driven gmpes: insights from the nga-west2 and resorce data sets. Earthq Spectra 32:2027–2056
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 Sci Rev 61(1):43–104. https://doi.org/10.1016/S0012-8252(02)00112-5
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 Sci Rev 61:43–104
Douglas J, Edwards B (2016) Recent and future developments in earthquake ground motion estimation. Earth Sci Rev 160:203–219
Efron B, Tibshirani R (1994) An Introduction to the Bootstrap. Chapman and Hall/CRC
Frazier P (2018) A tutorial on bayesian optimization. ArXiv abs/1807.02811
Freund Y, Schapire RE (1995) A decision-theoretic generalization of on-line learning and an application to boosting. In: EuroCOLT
Freund Y, Schapire RE (1996) Experiments with a new boosting algorithm. In: ICML
Gandomi AH, Alavi AH, Mousavi M et al (2011) A hybrid computational approach to derive new ground-motion prediction equations. Eng Appl Artif Intell 24:717–732

Geurts P, Ernst D, Wehenkel L (2006) Extremely randomized trees. Mach Learn 63:3–42
Hadley DM, Helmberger DV (1980) Simulation of strong ground motions. Bull Seismol Soc Am 70:617–630
Hartzell S (1982) Simulation of ground accelerations for the May 1980 Mammoth Lakes, California, earthquakes. Bull Seismol Soc Am 72(6A):2381–2387. https://doi.org/10.1785/BSSA07206A2381
Hartzell SH (1978) Earthquake aftershocks as green's functions. Geophys Res Lett 5:1–4
Hoerl AE, Kennard RW (2000) Ridge regression: biased estimation for nonorthogonal problems. Technometrics 42:80–86
Hoshiba M, Kamigaichi O, Saito M et al (2008) Earthquake early warning starts nationwide in Japan. Eos, Trans Am Geophys Union 89:73–74
Hu P, Pan JS, Chu SC (2020) Improved binary grey wolf optimizer and its application for feature selection. Knowl Based Syst 195(105):746
Hutchings L (1985) Modelling earthquakes with empirical green's functions (abs). Earthq Notes 56:14
Irikura K (1983) Semi-empirical estimation of strong ground motions during large earthquakes. Bull Disaster Prev Res Inst 33:63–104
Irikura K, Muramatu I (1982) Synthesis of strong ground motions from large earthquakes using observed seismograms of small events. In: Proc 3rd Int Microzonation Conprence, Seattle, pp 447-458
Irikura KK (1986) Prediction of strong acceleration motions using empirical green's function. In: Proc 7th Japan Earthq Eng Symp, pp 151–156
Joshi AP, Midorikawa S (2005) Attenuation characteristics of ground motion intensity from earthquakes with intermediate depth. J Seismol 9:23–37
Kamae K, Irikura K (1992) Prediction of site specific strong ground motion using semi empirical methods. In: Earthq Eng 10th World Conf, pp 801-806
Kanamori H (1979) A semi-empirical approach to prediction of long-period ground motions from great earthquakes. Bull Seismol Soc Am
Ke G, Meng Q, Finley T et al. (2017) Lightgbm: a highly efficient gradient boosting decision tree. In: NIPS
Kinoshita S (1998) Kyoshin net (k-net). Seismol Res Lett 69:309–332
Kubo H, Kunugi T, Suzuki W, et al (2020) Hybrid predictor for ground-motion intensity with machine learning and conventional ground motion prediction equation. Sci Rep 10
Maclin R, Opitz DW (1999) Popular ensemble methods: an empirical study. J Artif Intell Res 11:169–198
Mandal P, Mandal P (2022) Peak ground acceleration prediction using supervised machine learning algorithm for earthquakes of mw5.6-7.9 occurring in India and Nepal. https://doi.org/10.21203/rs.3.rs-1806354/v1, preprint from Research Square
Midorikawa S (1993) Semi-empirical estimation of peak ground acceleration from large earthquakes. Tectonophysics 218(1):287–295. https://doi.org/10.1016/0040-1951(93)90275-O
Mikumo T, Irikura K, Imagawa K (1981) Near field strong motion synthesis front foreshock and aftershock records and rupture process of the main shock fault. IASPEI 21st General Assembly,London
Mirjalili SM, Mirjalili SM, Lewis A (2014) Grey wolf optimizer. Adv Eng Softw 69:46–61
Morikawa N, Fujiwara H (2013) A new ground motion prediction equation for japan applicable up to m9 mega-earthquake. J Disaster Res 8:878–888
Motazedian D, Atkinson GM (2005) Stochastic finite-fault modeling based on a dynamic corner frequency. Bull Seismol Soc Am 95:995–1010
Munguía L, Brune JN (1984) Simulations of strong ground motion for earthquakes in the mexicali-imperial valley region. Geophys J Int 79:747–771
Nath SK, Raj A, Thingbaijam KKS et al (2009) Ground motion synthesis and seismic scenario in Guwahati city-a stochastic approach. Seismol Res Lett 80:233–242
Okada Y (2004) Recent progress of seismic observation networks in Japan. J Phys: Conf Seri 433(012):039
Ostroumova L, Gusev G, Vorobev A, et al (2018) Catboost: unbiased boosting with categorical features. In: NeurIPS
Panahi H (2017) Sarpol zahab earthquake. https://doi.org/10.13140/RG.2.2.31717.50408
Sheng Paul Lai S (1982) Statistical characterization of strong ground motions using power spectral density function. Bull Seismol Soc Am 72:259–274
Prasanth S, Singh U, Kumar A et al (2020) Forecasting spread of covid-19 using google trends: a hybrid gwo-deep learning approach. Chaos, Solitons, and Fractals 142:110–336
Rasmussen CE, Williams CKI (2009) Gaussian processes for machine learning. In: Adaptive computation and machine learning
Saidi R, Waad B, Essoussi N (2019) Hybrid feature selection method based on the genetic algorithm and pearson correlation coefficient. In: Machine Learning Paradigms
Saikia CK (1993) Ground motion studies in great los angles due to m_w=7.0 earthquake on the elysian thrust fault. Bull Seismol Soc Am

Saikia CK, Herrmann RB (1985) Application of waveform modelling to determine focal mechanisms of four 1982 miramichi aftershocks. Bull Seismol Soc Am
Schapire RE (1990) The strength of weak learnability. Mach Learn 5(2):197–227. https://doi.org/10.1023/A:1022648800760
Scordilis E (2005) Globally valid relations converting ms, mb and mjma to mw. NATO advanced research workshop on earthquake monitoring and seismic hazard mitigation in Balkan Countries, 11–17 September 2005, the Rila Mountains-Resort Village Borovetz. Bulgaria, Abstracts book, pp 158–161
Shoushtari AV, bin Adnan A, Zare M (2018) Ground motion prediction equations for distant subduction interface earthquakes based on empirical data in the malay peninsula and Japan. Soil Dyn Earthq Eng
Si H, Midorikawa S (1999) New attenuation relationships for peak ground acceleration and velocity considering effects of fault type and site condition. J Struct Constr (transactions of Aij) 64:63–70
Somala SN, Chanda S, Karthikeyan K, et al (2021) Explainable machine learning on new zealand strong motion for pgv and pga. Structures
Thiagarajan B, Srinivasan L, Sharma A, et al. (2017) A machine learning approach for prediction of on-time performance of flights. 2017 IEEE/AIAA 36th Digital Avionics Systems Conference (DASC), pp 1–6
Thomas S, Pillai GN, Pal K (2017) Prediction of peak ground acceleration using ϵ-svr, ν-svr and ls-svr algorithm. Geomat Nat Hazards Risk 8:177–193
Trugman DT, Shearer PM (2018) Strong correlation between stress drop and peak ground acceleration for recent m 1–4 earthquakes in the san francisco bay area. Bull Seismol Soc Am 108:929–945
Victoria AH, Maragatham G (2021) Automatic tuning of hyperparameters using bayesian optimization. Evol Syst 12:217–223
Wang M, Chen H, Li H et al (2017) Grey wolf optimization evolving kernel extreme learning machine: application to bankruptcy prediction. Eng Appl Artif Intell 63:54–68
Yu GY (1994) Some aspects of earthquake seismology: slip partitioning along major convergent plate boundaries; composite source model for estimation of strong motion; and nonlinear soil response modeling. In: NIPS
Yu GY, Khattri KN, Anderson JG, et al. (1995) Strong ground motion from the uttarkashi, himalaya, india, earthquake: Comparison of observations with synthetics using the composite source model. Bull Seismol Soc Am
Zeng Y, Anderson JG, Yu GY (1994) A composite source model for computing realistic synthetic strong ground motions. Geophys Res Lett 21:725–728
Zhou T, Jiao H (2022) Exploration of the stacking ensemble machine learning algorithm for cheating detection in large-scale assessment. Educ Psychol Meas

Authors and Affiliations

Anushka Joshi[1] · Balasubramanian Raman[1] · C. Krishna Mohan[2] · Linga Reddy Cenkeramaddi[3]

✉ Anushka Joshi
anushka_j@cs.iitr.ac.in

Balasubramanian Raman
bala@cs.iitr.ac.in

C. Krishna Mohan
ckm@cse.iith.ac.in

Linga Reddy Cenkeramaddi
linga.cenkeramaddi@uia.no

[1] Department of Computer Science and Engineering, Indian Institute of Technology Roorkee, Roorkee, Uttarakhand 247667, India

[2] Department of Computer Science and Engineering, Indian Institute of Technology Hyderabad, Sangareddy, Telangana 502285, India

[3] Department of Information and Communication Technology, University of Agder, Grimstad 4879, Norway