Soil Dynamics and Earthquake Engineering 172 (2023) 108008

Contents lists available at ScienceDirect

Soil Dynamics and Earthquake Engineering

journal homepage: www.elsevier.com/locate/soildyn

The potential of region-specific machine-learning-based ground motion models: Application to Turkey

Amirhossein Mohammadi [a], Shaghayegh Karimzadeh [a,*], Seyed Amir Banimahd [b], Volkan Ozsarac [c], Paulo B. Lourenço [a]

[a] Department of Civil Engineering, ISISE, University of Minho, Campus de Azurém, 4800 - 058, Guimarães, Portugal
[b] Department of Civil Engineering, Faculty of Engineering, Ardakan University, P.O. Box 184, Ardakan, Iran
[c] Department of Science, Technology and Society, IUSS Pavia, University School for Advanced Studies, 27100, Pavia, Italy

ARTICLE INFO

Keywords:
Turkish ground motion dataset
Inter-event and intra-event residuals
Likelihood function
Ground motion model
Artificial neural network
Extreme gradient boosting

ABSTRACT

Conventional ground motion models have extensively been established worldwide based on classical regression analysis of records. Alternatively, advanced nonparametric machine-learning (ML) algorithms may capture the complex nonlinear behaviour of earthquake motions. This paper investigates the efficiency of artificial neural network (ANN) and extreme gradient boosting (XGBoost) in predicting peak ground acceleration (PGA), peak ground velocity (PGV) and pseudo-spectral acceleration (PSA) (period, T = 0.03–2.0 s) for the Turkish dataset. The dataset involves 1166 records of 383 events with a moment magnitude (M_w) of 4.0–7.6, Joyner and Boore distance (R_{JB}) of 0–200 km, focal depth (FD) less than 35 km, and site condition as the averaged shear wave velocity of the soil on the top 30 m (V_{S30}) of 131–1380 m/s. The performance of the models is compared against empirical models in terms of root-mean-square error (*RMSE*), coefficient of determination (R^2), Pearson correlation coefficient (r), and inter-event and intra-event residuals. To perform residual analysis, a likelihood function is developed. Findings reveal that the XGBoost approach gives an unbiased model with a higher correlation and lower residual than ANN. Finally, an online platform is provided for any interested users.

1. Introduction

Earthquakes have been the primary source of human losses throughout history, with large economic losses, particularly in seismically active regions. In earthquake engineering and engineering seismology applications, ground motion models (GMMs) are essential for estimating the intensity of ground shaking. They have been widely developed to predict ground motion intensity measures, IMs (e.g., peak ground acceleration, PGA, peak ground velocity, PGV, and pseudo-spectral acceleration, PSA, at different periods, T) along with the associated uncertainty in any site of interest. GMMs link ground motion IMs to variables involving fault mechanism (FM), event magnitude (mostly in terms of moment magnitude, M_w), focal depth (FD), source-to-site distance, and characteristics of the soil profile at the station. GMMs are commonly used in civil and earthquake engineering fields, ranging from performing deterministic or probabilistic seismic hazard analyses and developing seismic hazard maps for building codes. GMMs are also employed in assessing site-specific seismic hazard levels for designing infrastructures and seismic loss estimation studies. A literature survey reveals that the former studies have mainly developed global or region-specific empirical GMMs based on classical regression analysis [1–17]. In recent years, the functional forms of the empirical GMMs have been largely modified to account for the nonlinearity, in addition to soil amplification, source mechanism, geometric and anelastic attenuation, and uncertainties involved in real motions. Therefore, recently proposed models became very intricate. Moreover, a key challenge in developing empirical models is the priori definition of functional forms with an adequate level of accuracy.

Nonparametric models (e.g. Refs. [18–23]), which do not require fixed functional forms, have been proposed as alternatives to deal with the high nonlinearity (complexity) of the ground motions and the difficulties involved in the parametric (empirical) models [24–27]. Furthermore, Kong et al. [28] and Alimoradi and Beck [29] demonstrated the widespread applicability of machine learning (ML) algorithms (e.g., artificial neural network, ANN, random forest, RF, gradient boosting, GB, extreme gradient boosting, XGBoost, support vector

* Corresponding author.
E-mail addresses: id9251@civil.uminho.pt (A. Mohammadi), shaghkn@civil.uminho.pt (S. Karimzadeh), banimahd@ardakan.ac.ir (S.A. Banimahd), volkan.ozsarac@iusspavia.it (V. Ozsarac), pbl@civil.uminho.pt (P.B. Lourenço).

https://doi.org/10.1016/j.soildyn.2023.108008
Received 31 January 2023; Received in revised form 30 April 2023; Accepted 2 May 2023

machine, SVM) in seismology, highlighting their potential to enhance the understanding of seismic events and improve prediction accuracy. As such, with the advancements in artificial intelligence and soft computing techniques in recent years, a significant number of GMMs have been developed using various approaches. For example, Dhanya and Raghukanth [30] and Dhanya et al. [31] recently employed the ANN approach for developing a global GMM based on PEER NGA-West2 ground motion database and used a hybrid technique combining genetic algorithm and Levenberg–Marquardt technique to train the model. The developed model was able to capture the main ground motion characteristics of the existing GMMs from the NGA-West2 project [11,12,15,32] and the variability better than the previous ANN-based model developed by Derras et al. [33]. On the other hand, Dhanya and Raghukanth [34] proposed assigning regional flags to the records while using the same approach to develop GMMs for regions with sparse recorded data, such as North-Eastern India and the Western Himalayas. Moreover, to overcome the high-frequency (>1 Hz) limitation of physics-based simulations and to eventually enhance the ground motion predictions for future seismic events, especially those with high magnitudes [35], Paolucci et al. [36] proposed enriching the simulated time histories by iteratively scaling their Fourier spectrum to match the prediction of PSA at short periods by ANN-based GMMs. Ghalehjough and Mahinroosta [37] used the fuzzy logic model to predict the PGA of Iranian ground motions and showed that the proposed GMM is more efficient than empirical GMMs. Furthermore, Khosravikia et al. [38] employed three ML techniques, ANN, RF, and SVM, to develop GMMs for Oklahoma, Kansas and Texas and concluded that if the data is sufficient, all ML techniques tend to provide more accurate estimates compared to traditional GMMs, and specifically, RF outperforms other algorithms. Likewise, Seo et al. [39] evaluated the performance of classical regression-based models and the GMMs developed using ANN, RF, and GB algorithms for South Korea to predict PSA, while the GB-based GMM was recognised as the best performing model. In general, as demonstrated by the studies mentioned earlier and many others (see Refs. [40–63]), the main advantage of such sophisticated GMMs is that if the ground motion database used to train the models is sufficiently large, they have lower dispersion and more accurate predictions than traditional regression-based ones since they can capture complex nonlinear relationships between the input and variables. However, the main drawback of models based on fuzzy logic and ML algorithms is that they are "black box" models, meaning that providing a physical interpretation of them is difficult. Further, it is not usually permitted to extrapolate such models beyond the original data range due to the absence of a physical model. Nonetheless, this is not believed to be a drawback, as using empirical GMMs outside the original data range is also considered controversial [64].

In particular, to handle epistemic uncertainty in probabilistic seismic hazard analyses, Atkinson et al. [65] study demonstrates the need for different GMMs in a logic tree format [66,67]. It is worth noting that the applicability of different GMMs relies on their accuracy, model parameters, and, more importantly, the dataset used for the analyses. The study by Douglas [68] summarised the available parametric and nonparametric GMMs derived worldwide based on either real or simulated ground motion datasets between 1964 and early 2021. Overall, in addition to 87 empirical GMMs derived based on simulated ground motion datasets, that study summarises 485 and 316 empirical GMMs for predicting PGA and elastic PSA ordinates, respectively. Regarding nonparametric models, in addition to 18 backbone models, that study includes details of 39 models. The available models are mainly global, while implementing the NGA ground motion dataset or European records.

This work focuses on constructing a local GMM for Turkey, one of the distinguished seismic hazard zones, based on the most recent ground motion data. In the literature, some region-specific empirical GMMs are derived from the dataset of Turkey [1,2,16,58,69–74]. The model developed by Cabalar and Cevik [58] was proposed for predicting PGA, while the other studies were developed for predicting the full spectral ordinates. Among the available models, the studies of Özbey et al. [71] and Bindi et al. [1] have performed regression only using ground motions from events recorded in north-western Turkey, while the rest are developed based on all regions of Turkey. Regarding the nonparametric models, the study of Güllü and Erçelebi [41], where the datasets involve ground motions recorded up to 2004, predicted only PGA by employing the ANN approach. Later, Günaydın and Günaydın [57] proposed a nonparametric GMM to predict PGA using three different ANN methods: radial basis function, generalised regression neural networks, and feed-forward back-propagation. The proposed model was developed using the database of north-western Turkey between 1999 and 2000. The authors predicted the vertical and two horizontal components separately, employing M_w, FD, hypocentral distance, and site conditions. Yerlikaya-Özkurt et al. [63] recently derived a GMM for Turkey to predict PGA and PGV using the multivariate adaptive regression splines method. That model was developed based on three independent variables, including M_w, site condition as the averaged shear wave velocity of the soil on the top 30 m (V_{S30}), and Joyner and Boore distance (R_{JB}) as the source-to-site distance metric. The authors employed 726 strong ground motions of 156 events with strike-slip fault mechanisms. M_w range was within 3.8–7.6 while the R_{JB} range was within 0–200 km.

In this regard, this study introduces a novel nonparametric region-specific GMM capable of predicting the full PSA ordinates by investigating the effectiveness of alternative advanced ML algorithms. This study is novel in that it contributes to filling a gap in the literature by developing a GMM utilising advanced ML-based approaches, namely, ANN and XGBoost algorithms, to estimate the spectral ordinates of the Turkish dataset with minimal computational resources. This is significant because existing models rely on empirical methods and require numerous regression coefficients, resulting in complicated calculations. Moreover, this study introduces a new analytical maximum likelihood formula as an adjustment to the model developed by Abrahamson and Youngs [75], which rectifies their likelihood function. Compared to the other studies, the GMM developed in this study uses the ground shakings, including the most recent large-magnitude earthquakes in Turkey (e.g., the 2020 Elazig earthquake with $M_w = 6.7$, the 2020 Samos earthquake with $M_w = 6.6$). The database of this study is compiled from AFAD [76] and includes 383 different earthquake events with a total of 1166 ground motion time histories recorded in Turkey between 1976 and 2022. The records have M_w of 4.0–7.6, R_{JB} of 0.1–200.0 km, V_{S30} of 131–1380 m/s, and FD less than 35 km. To develop the GMMs based on alternative ML algorithms, the predictors are M_w, V_{S30}, R_{JB}, and FM. The present study investigates the efficiency of two alternative ML algorithms: ANN and XGBoost, for predicting peak ground motion parameters, including PGA, PGV, and the elastic PSA at 14 time periods for 5% damping within the range of 0.03–2.0 s in Turkey. To optimise the hyperparameters of the ML models and assess the most efficient values, the Bayesian optimisation algorithm (BOA) [77] for the XGBoost model, along with the trial and error [78,79] approach for the ANN model, are utilised. To investigate whether the model is unbiased with respect to any predictor and to reduce the aleatory uncertainty [80], the ML algorithms herein are adjusted by splitting the uncertainty in terms of inter-event (between-event) and intra-event (within-event) terms using the approach developed by Abrahamson and Youngs [75]. A correction is made to the likelihood function proposed by that study. Next, the performances of different ML algorithms are evaluated through root-mean-square error (*RMSE*), coefficient of determination (R^2), and Pearson correlation coefficient (r). The developed models are also compared against the empirical attenuation model by Akkar et al. [6] and Kale et al. [16] through training with the same database. Finally, the best nonparametric GMM developed by this study is determined and implemented in web-based application software for end-users.

2. Adopted strong ground motion database

Turkey is in a geologically active area, where most of the country lies on seismic faults dominated by mostly shallow active structures. The seismotectonic setting of Turkey can be explained by the interaction of the movement of the Arabian and African plates toward the relatively stable Eurasian plate in the north resulting in two major fault zones: North Anatolian Fault Zone and East Anatolian Fault Zone [81–83].

The instrumental dataset of the Turkish Disaster and Emergency Management Presidency (AFAD) [76] includes several earthquakes that occurred all over the country, several of which led to human losses and damage to the built environment [84]. This study takes the raw strong ground motions between 1967 and 2022 with M_w of 4.0–7.6, R_{JB} of 0–200 km, FD less than 35 km, and all recorded at stations with V_{S30} ranging from 131 to 1380 m/s from AFAD [76]. Additional raw recordings are retrieved from the dataset RESOURCE [85], while missing information related to specific events is gathered from additional sources [86–89] for the completeness of the dataset. The records are then filtered using baseline correction and a fourth-order band-pass Butterworth filter within the frequency range of 0.1–25 Hz. The dataset includes a variety of magnitude scales, such as M_s, M_b, M_L, M_w and M_d. The homogeneity of the magnitude is ensured by eliminating the records with magnitude scales other than M_w from the dataset. The collected database contains 1166 recordings from 383 distinctive earthquake events recorded at 269 seismic stations in Turkey since 1967. The spatial distribution of the stations and the earthquake events for this database is shown in Fig. 1. For statistical evaluation, Fig. 2 illustrates the histogram of seismic characteristics of the ground motion records in terms of M_w, V_{S30}, R_{JB}, and FD. The statistics reveal that large-magnitude events in the dataset are rare, while M_w between 4.5 and 5.0 has the highest probability. The plot corresponding to V_{S30} at the stations, which describes the local site conditions, reveals that most of the stations in Turkey have soil types C and D consistent with the soil classification system of the National Earthquake Hazards Reduction Program (NEHRP) [90]. The distribution plot of R_{JB}, which is the shortest distance from a station to the surface projection of the rupture plane selected for characterising the source-to-site distance, reveals that most records have R_{JB} less than 75 km. Finally, the distribution of FD reveals that most events are shallow earthquakes having a mean FD of approximately 12 km.

In this study, accelerograms from all events, including main-shocks and fore-/after-shocks, are included in the analysis, and this decision is based on the adequacy of their waveform characteristics for computing accurate ground-motion intensity metrics of interest, as noted by Kale et al. [16]. Previous research, including Douglas and Halldórsson [91], found no significant differences in spectral accelerations between main-shocks and after-shocks when using the same dataset as Ambraseys et al. [92]. This finding supports the decision to use all available strong-motion data to develop the GMM for this study. In addition, most of current GMMs for Turkey are constructed based on the entire dataset, further justifying the choice to retain all available data for the analysis. The information regarding the FM, including normal (N), reverse (R), and strike-slip (SS) for all earthquakes, is plotted in Fig. 3. The distribution of FM demonstrates that SS is the predominant fault mechanism in Turkey (almost 60%). In contrast, events with the R fault mechanism have the smallest occurrences.

Fig. 4 illustrates the distribution of M_w versus R_{JB} for different soil classes and fault mechanisms. The scatter plots reveal that the number of near-field records, particularly for $R_{JB} < 10$ km and large-magnitude events, is relatively small. In contrast, the dataset is abundant for the M_w range of 4.0–6.0 and R_{JB} larger than 10 km. As stated, most recorded motions for large-magnitude events have NEHRP-C and NEHRP-D soil types. Earthquakes with the R fault mechanisms within the M_w range of 5.0 and 6.0 are also rare. Finally, large-magnitude events with M_w more than 7.0 have happened primarily due to the rupture of faults with the SS fault mechanism. In contrast, no large-magnitude event ($M_w > 6.5$) struck due to the rupture of the N fault mechanism.

To develop the GMMs of this study based on two alternative ML approaches, the predictive variables are considered as M_w, V_{S30}, R_{JB}, and FM. Given the input variables M_w, R_{JB}, V_{S30}, and FM, the vector of IMs,

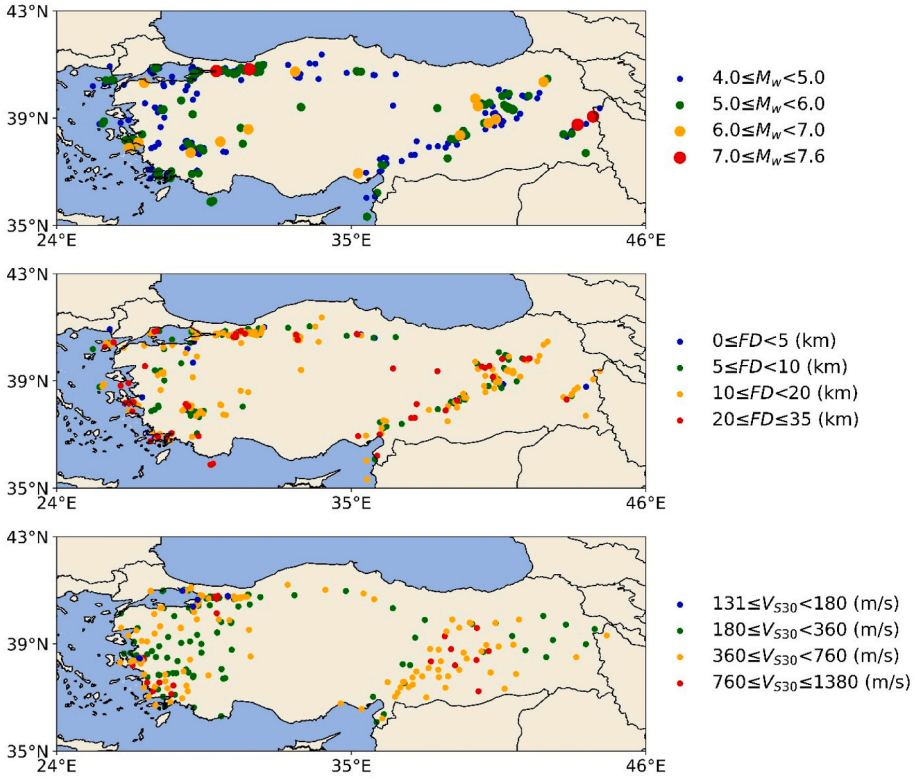

Fig. 1. Spatial distribution of the stations and earthquake events for the Turkish dataset between 1967 and 2022.

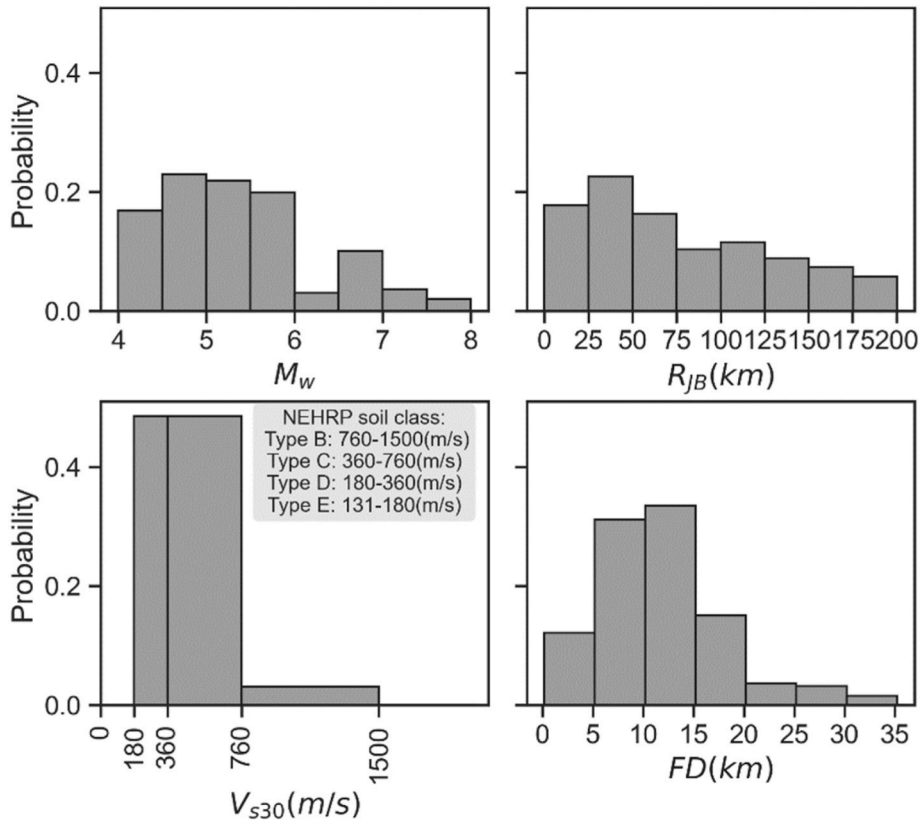

Fig. 2. Histograms of seismological features of the Turkish ground motion records.

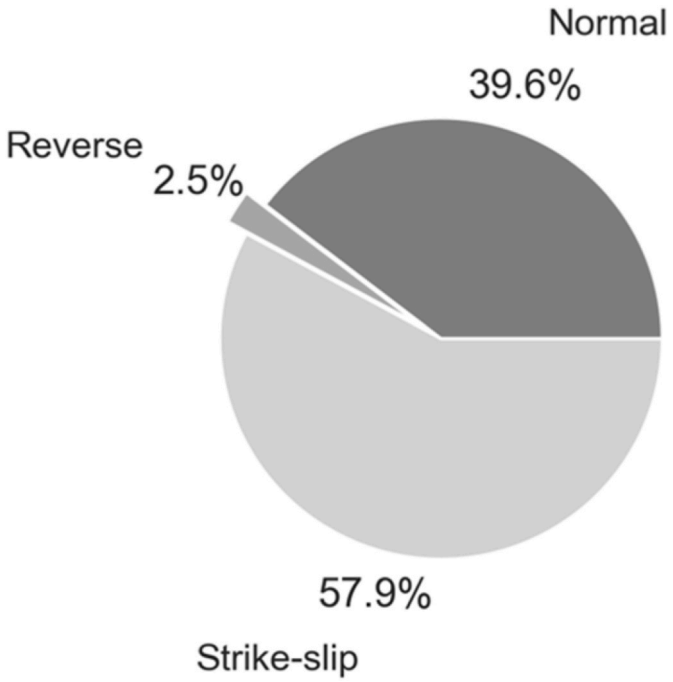

Fig. 3. Distribution of earthquakes with respect to the focal mechanism.

including PGA, PGV, and 5% damped elastic PSA at various periods (T = 0.03–2 s) are estimated. The IMs of each record are computed by means of the open-source toolbox introduced by Ozsarac et al. [93]. The normalisation approach enabled a fair comparison between ground motions of varying magnitudes and facilitated the identification of differences in ground motion amplification across spectral ordinates. The PSA of all databases under consideration normalised by their PGA values is shown in Fig. 5. The median PSA for these records and one standard deviation above the median are also shown to illustrate the range of possible PSA values.

Finally, the characterisation of intra-event spatial correlations is recognised as a valuable tool for evaluating the performance of GMMs, particularly for near-field records where the correlation is known to be prominent. Nevertheless, it should be noted that the spatial correlation model is deemed adequate and dependable for datasets obtained from dense seismic array networks with ample recordings of each event [94] or for developing simulated-based GMMs [95]. As the seismic network for past events was limited in Turkey, and most events had a small number of near-field stations, this phenomenon was not accounted for in the present study. Furthermore, it should be noted that even if spatial correlation affects estimated GMMs, its overall impact on predictions is relatively insignificant [96].

3. Methodology

This section summarises the techniques used to generate the GMMs for this study. Next, a discussion of conventional methodologies, ML approaches and optimisation algorithms for tuning the hyperparameters of the developed models will be presented. Following this, the mixed-effect algorithm will be reviewed. Indicators of model performance and an overview of the research methodology will be provided at the end.

3.1. Conventional GMMs

The familiar approach for predicting a ground motion IM, such as PSA, is to employ ground motion prediction equations, the most recently known as GMMs. Empirical GMMs are typically developed using a sta-

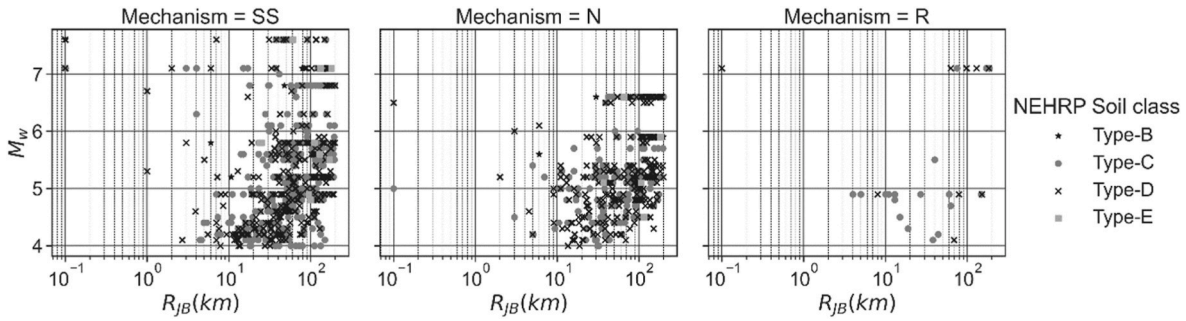

Fig. 4. Magnitude-distance (M_w-R_{JB}) distribution of the dataset with respect to the focal mechanism (strike-slip, SS; normal, N; and reverse, R) and soil class (Types B, C, D, and E) according to NEHRP guidelines [90].

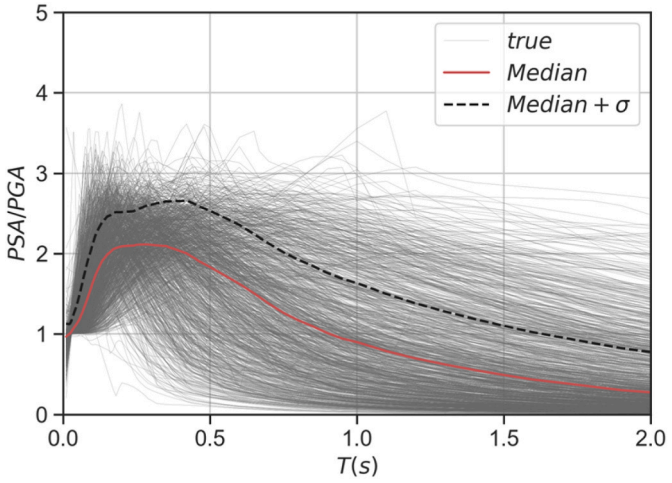

Fig. 5. Normalised 5% damped pseudo-spectral acceleration (PSA) for the Turkish dataset.

tistical regression [97] on the large sets of ground motion intensities observed in past earthquakes. Since significant scatter is present in the observed data for each IM, GMMs, in general, deliver a probability distribution instead of a single value as follows:

$$\begin{aligned} &\ln y_{ij} = \mu\left(X_i, X_{ij}, \theta\right) + \eta_i + \varepsilon_{ij} \\ &X_i : M_w, FM, FD, \ldots \\ &X_{ij} : V_{S30}, R_{JB}, \ldots \end{aligned} \tag{1}$$

where ln y_{ij} is the natural logarithm of the IM, i denotes the index of the earthquake event, and j represents the station's index. $\mu(X_i,X_{ij},\theta)$ indicates the median ground motion prediction function, with X_i representing event-related parameters, X_{ij} defining station-related parameters and θ being the vector of model parameters. η_i is the inter-event (between-event) and ε_{ij} is the intra-event (within-events) residual components in the natural logarithm scale. The term "between-event" refers to the average difference between the median estimates of the GMM and the observed ground motions for the i^{th} earthquake. The term "within-event" refers to the difference between the record of the i^{th} earthquake at the j^{th} station and the median prediction for the i^{th} earthquake. Both residuals, i.e., inter-event and intra-event components, are supposed to be normally distributed independent random variables with zero mean and standard deviations of τ and σ, respectively. Finally, the total standard deviation corresponding to the GMM is reported by:

$$\Phi = \sqrt{\sigma^2 + \tau^2} \tag{2}$$

The efforts toward seismic hazard characterisation of Turkey gained momentum after İzmit 1999 (M_w = 7.4) and Düzce 1999 (M_w = 7.1) earthquakes. Consequently, some empirical local GMMs have been proposed for Turkey [69,70,98] to estimate either PGA or PSA values. Moreover, some regional GMMs were proposed for the north-western Turkey [71,99]. Akkar and Çağnan [2] evaluated some of these pioneering GMMs, and found a bias, potentially due to the data used for regression analysis. The authors proposed an empirical model (AC10) that considers the faulting mechanism and the magnitude scaling, geometric decay and site effects, which were also considered in previous GMMs. The latest local GMM (KAAH15) for Turkey was proposed by Kale et al. [16], considering estimators to account for anelastic attenuation.

Moreover, it is common to use GMMs for seismic characterisation of hazards, developed based on databases containing ground motion recordings of worldwide events or events coming from a broader region, such as pan-European records. Among others [100], a set of new empirical GMMs (ASB14) is proposed for the Middle East and Europe [6]. These GMMs use the ergodic assumption, which leads to large aleatory variability in source effects, attenuation and path effects, and site seismic processes. Kotha et al. [101] suggested a new non-ergodic GMM (KO16) to reduce the estimated aleatory variability. Both ASB14 and KO16 utilised the dataset RESOURCE [85] assembled for the Middle East and Europe.

In the present study, the GMMs based on alternative ML approaches have the form given in Eq. (1) in which X_i includes M_w and FM, and X_{ij} contains R_{JB} and V_{s30}. Moreover, to compare the predictive capabilities of the considered ML algorithms, a comparison is provided with the GMM of ASB14 proposed based on R_{JB}, in addition to a recent local GMM for Turkey, KAAH15.

3.2. Machine-learning algorithms

In this study, two alternative ML algorithms are utilised and tested to derive GMMs for the ground motion dataset of Turkey. It is noted that ML algorithms are sensitive to the scale of data; thus, it is recommended to transform the input features into a similar scale. To accelerate the training speed and to minimise the possible errors, the input dataset of this study is normalised using the following expression:

$$\theta_{si} = 0.6\frac{\theta_i - \theta_{\min}}{\theta_{\max} - \theta_{\min}} + 0.2 \tag{3}$$

where θ_{si} is the scaled value of θ_i, which is the input parameter, and the terms θ_{max} and θ_{min} are, respectively, the maximum and minimum values of that parameter in the training dataset. The range between 0.2 and 0.8 is selected to avoid analysis failure at the value of zero [102]. Nonetheless, the analysis has shown that selecting any range within (0,1] does not significantly affect the results.

In the following sections, the two adopted techniques are described in detail.

3.2.1. Artificial neural network

ANN is a prevailing computational tool, particularly for solving

complex regression and classification problems, which can imitate the cognitive skills of thinking human minds to solve real-world problems that conventional approaches cannot follow [103]. Empirical equations may yield a complex and not applicable expression when a problem involves many explanatory parameters. In such cases, ANN models can predict the solution of highly nonlinear and complex problems better than statistical or empirical models. Fig. 6 gives a schematic representation of the structure of the ANN algorithm employed in this work. The model consists of an input layer, a hidden layer, and an output layer. The neurons of the input layer directly receive initial signals from the explanatory variables for further processing in the adjacent layers. A nonlinear transformation is applied through an activation function, f (.), on the summation of weighted input signals arriving in the network of Fig. 6. The linear summation of the weighted inputs with bias is the net input (u_k), which can be expressed as:

$$u_i = \sum_{i=1}^{m} w_{ki} x_k + b_i \tag{4}$$

where m is the number of hidden layers, x_k is the k^{th} input variable, w_{ki} is the weight of k^{th} input variable given for i^{th} neuron, and b_i is bias in i^{th} neuron. Subsequently, the output layer performs a linear transformation on the summation of weighted signals entering this layer. Finally, the output of the model is obtained in the output layer.

In this study, the error between the desired targets and outputs of the model is minimised using an error backpropagation algorithm. This algorithm adjusts the connection weights (w) and biases (b) of the ANN model. Among various backpropagation algorithms, the Marquardt-Levenberg algorithm [104,105], developed for solving least square problems, is used. Design of the architecture of neural networks consists of identifying the number of hidden layers and its neurons, and the type of activation function. These hyperparameters are decided based on the minimum calculated RMSE of the test dataset for each possible combination using a trial-and-error method. In this work, the optimal choices for the hyperparameters of the ANN model are obtained as one hidden layer, four neurons, and a log-sigmoid activation function (Table 1).

Table 1
Hyperparameters of the artificial neural network (ANN) approach and their values or types.

Hyperparameter	Short explanation	Value/Type
number of hidden layers	a layer in between input layers and output layers	1
number of neurons	small individual units as connection points	input layer: 4 hidden layer: 4 output layer: 16
activation function	controls if a neuron should be triggered or not	log-sigmoid

3.2.2. Extreme gradient boosting

Chen and Guestrin [106] proposed XGBoost as a practical implementation of the gradient boosting technique. To eliminate model complexity and prevent overfitting, XGBoost includes a regularisation function. Due to its capacity to handle large-scale problems with significant functioning and execution speed, XGBoost has recently gained popularity as an ML method. It is, however, more challenging than other boosting algorithms to understand and interpret [107].

Fig. 7 presents the structure of the XGBoost approach. In the first step, a tree is trained with randomly selected data to predict the given data. Then, the residuals of the predictions are used to train the next prediction tree. The residuals of trained trees are consecutively used for training another tree. This iterative approach updates the model parameters to optimise the objective function through division into two parts: one part represents the loss function (L). In contrast, another part penalises the model's complexity and prevents overfitting, as shown below:

$$\sum_{i=1}^{N} L\left(y_i, \hat{y}_i^t\right) + \sum_{i=1}^{t} \Omega(f_i) \tag{5}$$

$$\Omega(f) = \gamma K + \frac{1}{2}\lambda \sum_{j=1}^{K} w_j^2 \tag{6}$$

where, γ is the complexity parameter of each tree leaf, K is the number of leaves, λ is the regularisation parameters, and w_j is the score of j^{th} leave. To obtain the best XGBoost model, the hyperparameters of the algorithm, as listed in Table 2 should be optimised. The BOA is used for this purpose, as explained in the following section.

3.3. Optimisation algorithms for tuning the hyperparameters of the ML algorithms

In general, there is no easy way to define the best parameters of a neural network, which is an optimisation problem beyond the scope of this study. An efficient yet straightforward way to define reasonable values for the parameters of an ANN model is trial and error. This algorithm, which is generally used in the literature [78,79], is utilised here for tuning parameters of the ANN model.

On the other hand, to identify the optimal hyperparameters of the XGBoost approach, BOA, which is effective in contrast to different known optimisation approaches (e.g., manual, random search, grid, particle swarm optimisation), is used in this study. The term

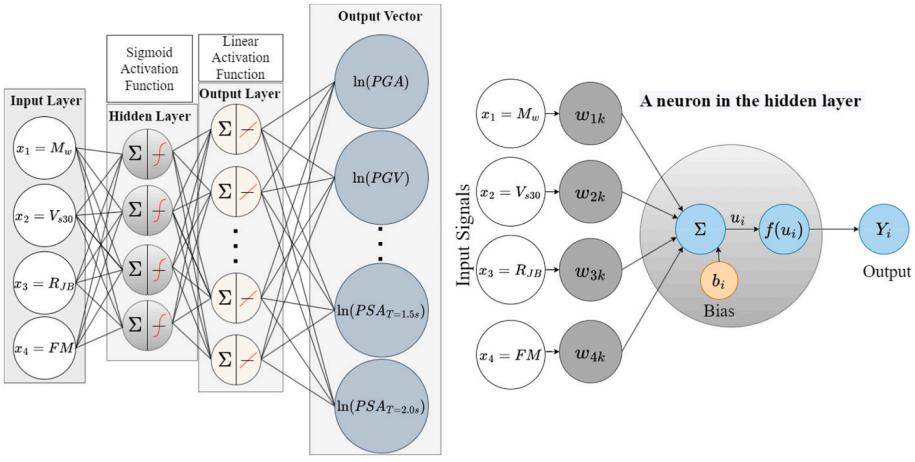

Fig. 6. Structure of the artificial neural network (ANN) model and illustration of artificial neurons of the hidden layer.

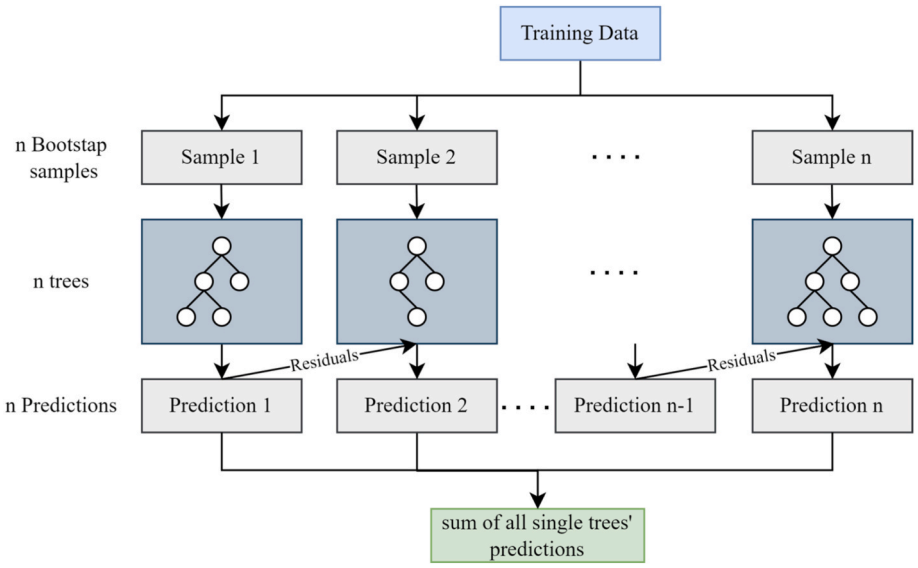

Fig. 7. The structure of the extreme gradient boosting (XGBoost) approach.

Table 2
Hyperparameters of the extreme gradient boosting (XGBoost) approach and their search spaces.

Hyperparameter	Short explanation	Space
n_estimators	number of estimators	20–1000
learning_rate	learning rate	0.1–0.5
max_depth	maximum depth of trees	3–8
reg_alpha	L1 regularisation term on weights	0–1
reg_lambda	L2 regularisation term on weights	1–2
subsample	subsample ratio of the training dataset	0.1–1
min_child_weight	minimum child weight	0–10
colsample_bytree	subsample ratio of column	0.5–1

"optimisation" in BOA refers to the global optimisation of a black-box function for which the formula and derivatives are unknown [108]. This optimisation stems from Bayes' theorem as below:

$$p(\omega|D) = \frac{p(D|\omega)p(\omega)}{p(D)} \tag{7}$$

where ω denotes an unseen value, $p(\omega)$ is the prior probability distribution, $p(D)$ is the evidence, $p(D|\omega)$ denotes the probability, and, finally, $p(\omega|D)$ represents the posterior probability distribution. Prior knowledge is employed by Bayes' rule in order to define the posterior possibility in which the outcomes of earlier iterations are considered for determining the values of the upcoming iteration. Two sub-models, the acquisition, and the substitute, can be used with the BOA. The substitute model assesses the objective function through the Gaussian process (GP), a common surrogate for objective function modelling. This is a Gaussian distribution generalisation. In general, GP describes a prior over function, which can be changed into a posterior over function after observation of specific values of the function. This method assumes that the function $\mathscr{F}(x)$ is a realisation of GP with the mean of μ and the covariance of K [109]:

$$\mathscr{F}(x) \sim GP(\mu, K) \tag{8}$$

The acquisition function of BOA is maximised over repetitions and depends on the prior observations. The acquisition model recommends iteration using the findings of the substitute model as the next step. The hyperparameter optimisation through BOA is expressed mathematically as:

$$x^* = \underset{x \in X}{\operatorname{argmin}} \mathscr{F}(x) \tag{9}$$

The best set of hyperparameters (x^*) for any space ($x \epsilon X$) can be assessed by finding the optimised value for the objective score (i.e., $\mathscr{F}(x)$).

An overview of BOA steps is summarised below.

Step 1: Defining the objective function by setting hyperparameters of the selected machine
Step 2: Constructing a surrogate probability model of the objective function
Step 3: For the surrogate probability model finding the best-performing set of hyperparameters
Step 4: Employing the hyperparameters of Step 3 in the real objective function
Step 5: Rebuilding the surrogate probability model by incorporating the new results
Step 6: Iterating steps 3 to 5 for the maximum iteration number
Step 7: Training the selected machine using the obtained hyperparameters

3.4. Mixed effect model

Mixed effect models are beneficial for cases where data are acquired through repeated measurements. In this case, both fixed and random effects components are included in the residuals. This study employs seismic events recorded at different stations with various site characteristics and distance information. It is well known that the variability between seismic events and even inside the records of each earthquake is high, which requires splitting the total residual into different components [75].

Here, the well-known procedure for the mixed effect model suggested by Abrahamson and Youngs [75] is used to perform residual analysis. This procedure is modified herein by proposing an algebraic maximum likelihood function for computing model parameters and variances using the expectation-maximisation algorithm. The solution to Eq. (3) of Abrahamson and Youngs [75], given in Eq. (7) of the same reference, is revised here. The new likelihood function is derived in Appendix A. It should be highlighted that computationally solving this function, Eq. (A- 10), is substantially more straightforward and faster than solving Eq. (A- 1).

Here, the artificial bee colony (ABC) [110] and genetic algorithm (GA) [111] are respectively used in ANN and XGBoost approaches to maximise the log-likelihood function (minimising the $-\ln L$) of Eq. (A-

10). ABC and GA are numerical approaches for finding the optimal configuration that minimises the objective function of interest. The intelligent foraging behaviour of honeybees and the mechanics of genetics besides natural selection form the foundation of ABC and GA algorithms, respectively. These metaheuristic algorithms utilise iterative search techniques for solving a function of complex nature (as they do not require knowledge of the derivatives). Despite being computationally basic, the algorithms are powerful tools for optimisation problems.

The ABC algorithm has three setting parameters, making it more flexible than other most known algorithms. The algorithm mimics the behaviour of three types of bees in a colony, namely employed bees, onlooker bees, and scout bees. In this method, the artificial bees in the hive are divided into two groups: employed and onlooker bees. Each employed bee flies into a specific food source, then randomly searches the neighbourhood of the food source and evaluates the nectar's quality and shares the information with onlooker bees in the hive. It is noteworthy that each food source is a candidate for the solution of the problem. In the first step, a random population of the artificial bee is generated as:

$$\Phi = [\Phi_1, \Phi_2, \ldots, \Phi_{SN}] \tag{10}$$

where *SN* is the number of food sources equal to that of employed or onlooker bees. Choosing the first setting parameter of the ABC algorithm, *SN*, depends on the complexity of the problem. Φ_i is a vector including the variances σ^2 and τ^2. Thus, Φ_i can be defined as follows:

$$\Phi_i = \left[\sigma_i^2, \tau_i^2\right] \tag{11}$$

Considering the random generation of the initial population of the artificial bee colony in the first phase of the optimisation process, σ_i^2 and τ_i^2 are considered as random numbers in the range $[0, 1]$. In this phase, each employed bee searches around the assigned food source by the following equation:

$$\Phi_i^{NEW} = \Phi_i + (\Phi_i - \Phi_k) \times X \tag{12}$$

where Φ_i^{NEW} is the new value found for Φ_i. k is defined randomly different from *i*. X is a random variable between [-1,1] ($X \sim U[-1, 1]$). When the employed bees return into the hive, the information related to each food source quality is evaluated by the fitness value as follows:

$$fit_i = \frac{1}{1 + F(\Phi_i)} \tag{13}$$

where $F(\Phi_i)$ denotes the value of objective function of i^{th} food source. When artificial employed bees share the information in the hive, onlooker bees select the food source based on the probability of i^{th} food source with the following equation:

$$P_i = 0.10 + 0.90 \frac{fit_i}{\max(fit_i)} \tag{14}$$

In the second phase, onlooker bees search the neighbourhood of the selected food source using Eq. (13). If the quality of the food source cannot be enhanced after a predetermined number of searches (limit value; the second setting parameter), the food source will be abandoned. Thus, in the third phase, the employed bees that could not find a better solution change into scout bees randomly search the solution space in the range $[0, 1]$. This process is terminated if iterations exceed a predefined maximum cycle number (MCN), the third setting parameter of the ABC algorithm. By minimising the objective function, the unknown parameters of the optimisation problem are obtained accordingly. The values considered for the setting parameters of the ABC algorithm were 10, 50 and 100 for *SN*, limit, and MCN, respectively. These values were found to be sufficient for determining the unknown parameters of the problem. It is noteworthy that the number of food sources (i.e., *SN*) can be increased. However, it would be with the cost of more computational effort. An in-depth study of the determination of the setting parameters is out of the scope of this study.

On the other hand, GA is a class of optimisation algorithm that is inspired by the process of natural selection in biological evolution. It is particularly useful in solving complex optimisation problems that involve a large search space and numerous constraints. The GA process involves modelling the desired solution as a set of parameters, which are then represented by a chromosome. The chromosomes are combined and mutated to generate new solutions, which are subsequently evaluated for fitness. The fittest solutions are selected and used to generate a new population of chromosomes, and the process is repeated until the desired solution is found.

Below is a step-by-step procedure to develop the mixed effect model, considering the explanation provided regarding GA and ABC.

Step 1: First, an initial model is trained by a fixed-effect training procedure, i.e., the random effect is assumed to be equal to zero as follows:

$$\ln y_{ij} = f\left(X_i, X_{ij}, \theta\right) + \varepsilon_{ij} \tag{15}$$

Step 2: Residual components are computed by maximising the log-likelihood function (Eq. (A- 10)) using the numerical algorithms (ABC and GA).
Step 3: Based on estimated values of (σ,τ) and vector of model parameters θ, the random-effect term is obtained through the following formula:

$$\eta_i = \frac{\tau^2 \sum_{j=1}^{n_i} \left(y_{ij} - \mu_{ij}\right)}{n_i \tau^2 + \sigma^2} \tag{16}$$

Step 4: A new model is trained using a fixed-effect training procedure for $\ln y_{ij} - \eta_i$.
Step 5: Steps 2, 3, and 4 are iterated until the termination criterium is fulfilled. The adopted termination criterion is 0.5% in terms of the difference between two successive likelihood values.

3.5. Evaluation of the model performance indicators

In this study, the results from the n-time repeated k folds are averaged to evaluate the performance of two alternative ML approaches. For this purpose, *RMSE*, R^2, and r are the three statistical performance indices used in this study. These indicators are calculated as follows:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \tag{17}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{18}$$

$$r = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2 \sum_{i=1}^{n} (\hat{y}_i - \bar{\hat{y}})^2}} \tag{19}$$

where n is the number of samples, y is the actual value, $\hat{y}$ is the predicted value, and $\bar{y}$ and $\bar{\hat{y}}$ are the arithmetic means of y and $\hat{y}$ values, respectively. In this study, to evaluate the accuracy of the proposed ML models, the estimated IMs in terms of PGA, PGV, and PSA at different periods within 0.03–2 s are compared with those of real values. Among the model performance parameters, R^2 quantifies the variance in the response variable that can be predicted using the predictor variables.

The error-related indicator (*RMSE*) indicates the average distance between the observed and predicted response values. Finally, the *r* measures the degree of linearity between the predicted and observed IMs. The performance of a given GMM increases by an increase in R^2 and *r* and a decrease in *RMSE*.

3.6. An overview of the research methodology

Fig. 8 shows an overview of the procedure used for developing the ML-based GMMs in this study. As can be seen, in the first step, 80% of records from different M_w ranges are randomly selected for training, and the rest of the data is stored for testing. Then, the training dataset is scaled and used to train the machine for the first time (with a repeated k-fold cross-validation approach). It is important to note that hyper-parameters of the machines are tuned using BOA (for XGBoost) or trial and error (for ANN). Through the likelihood function proposed in appendix A and given measured and predicted IMs, the intra-event and inter-event uncertainties (σ, τ) are obtained. Given the σ and τ values, the intra-event and inter-event residual terms are calculated (ε, η). The machine is retrained by $ln(IMs)$-η_i, and a new likelihood value is obtained through a similar approach. This procedure continues until the likelihood function converges to the maximum value based on the termination criteria (0.5% in the difference between two successive likelihood values). Finally, the model performance is validated using the test dataset, which is scaled by considering the scaling parameters from the training dataset.

4. Results and discussion

This section compares the suitability of alternative ML-based GMMs to conventional GMMs, including ASB14 and KAAH15. Results are initially evaluated schematically and then statistically by calculating model performance indicators and inter-/intra-event residuals. Finally, the superlative model is chosen by evaluating the findings, and the results for the proposed model are further analysed.

4.1. Evaluation of the developed ML-based GMMs (ANN versus XGBoost)

This section compares the performances of the two ML algorithms against the empirical approaches of ASB14 and KAAH15. It is noted that for the sake of consistency, the empirical models are trained using the same training subset of this study. Fig. 9 presents the distribution of the observed versus predicted values for a sample IM, herein ln(PGA), evaluated through different algorithms for the entire dataset. In each figure, dashed lines reflect the ideal estimate (i.e., where the predicted and observed values are identical). The concentration of data along dashed lines demonstrates the correlation between estimated and observed values. The findings show that datasets are sufficiently close to the optimum fit line from all approaches. However, depending on the approach, the accuracy of the model changes. It is observed that XGBoost provides a better match than ANN and the other two empirical methods. The results of the ANN model are more consistent with those of the two empirical models. In addition to schematic comparisons, it is well-known that a model is only legitimate if it provides good model performance indicators. For this purpose, the effectiveness of the two ML algorithms for constructing GMMs against the aforementioned empirical GMMs is assessed by comparing the outcomes in terms of the model performance indicators as specified in section 3.5. Fig. 10 compares the results in terms of *RMSE*, R^2, and *r* for different IMs, including ln(PGA), ln(PGV), ln($PSA_{T=0.2\ s}$), ln($PSA_{T=0.5\ s}$), ln($PSA_{T=1.0\ s}$), and ln($PSA_{T=2.0\ s}$) from all GMMs. The *RMSE* from the ANN approach varies between 0.65 and 0.81. At the same time, the *RMSE* based on the XGBoost approach ranges between 0.55 and 0.68 while the other two empirical models provide *RMSE*, almost varying between 0.60 and 0.95.

According to the general hypothesis provided by (Smith 1986), the *r* values above 0.8 indicate a significant linear correlation concerning the estimated and observed values. As shown in Fig. 10 the *r* values for all models are above 0.8, meaning that the models capture the real values of the IMs in general. Nevertheless, the *r* values obtained from the XGBoost method are above 0.9, the highest compared to the others. Similarly, when the results are assessed in terms of R^2 for the investigated IMs, the mean value from the XGBoost approach is roughly obtained as 0.85, whereas the mean value from the ANN approach is around 0.80. For the other empirical models, including ASB14 and KAAH15, this value is estimated as 0.72 and 0.70, respectively. Therefore, the XGBoost model provides minimum error boundaries and maximum correlation coefficients compared to the other models.

In the final phase, it is essential to validate the model's bias to input variables such as M_w, R_{JB}, and V_{S30}. For this purpose, it is necessary to evaluate the inter-event and intra-event uncertainties, which indicate the variance of residuals concerning the seismic earthquakes and sites, respectively. To this end, the inter-event, intra-event and total uncertainties of the developed GMMs based on the ANN and XGBoost algorithms are calculated for PSA at all periods. For the sake of comparison, these plots are also developed for the two empirical models, including ASB14 and KAAH15. Results are illustrated in Fig. 11. It is observed that for all spectral values, the inter-event uncertainty is smaller than the intra-event uncertainty from all models in all period ranges. Compared to the empirical models, the ML-based GMMs have acceptable uncertainty ranges and can perform well. Among the two ML-based models, the trend in the total uncertainty for the ANN model is closer to the KAAH15 model at all period ranges. When the results of the

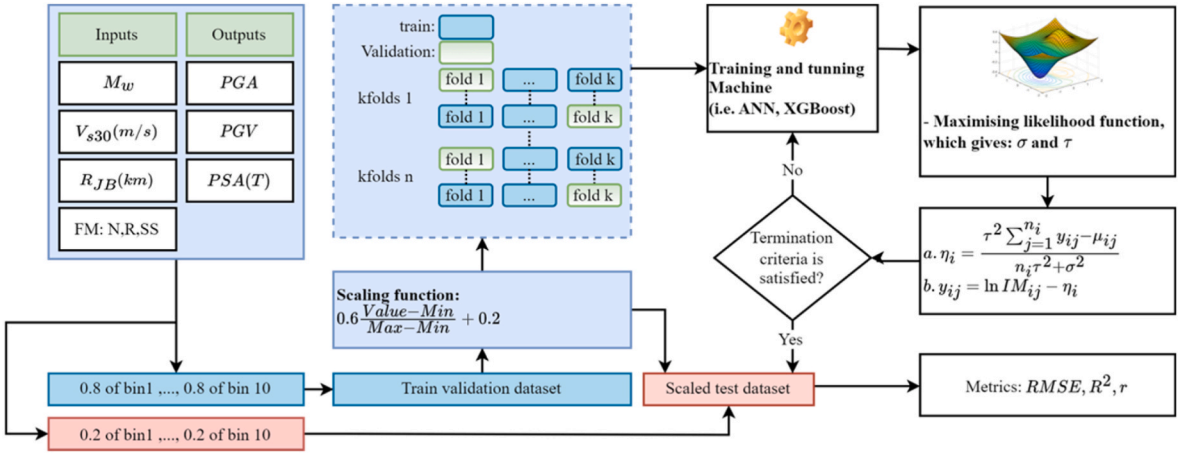

Fig. 8. Iterative process for finding random and fixed-effect residuals.

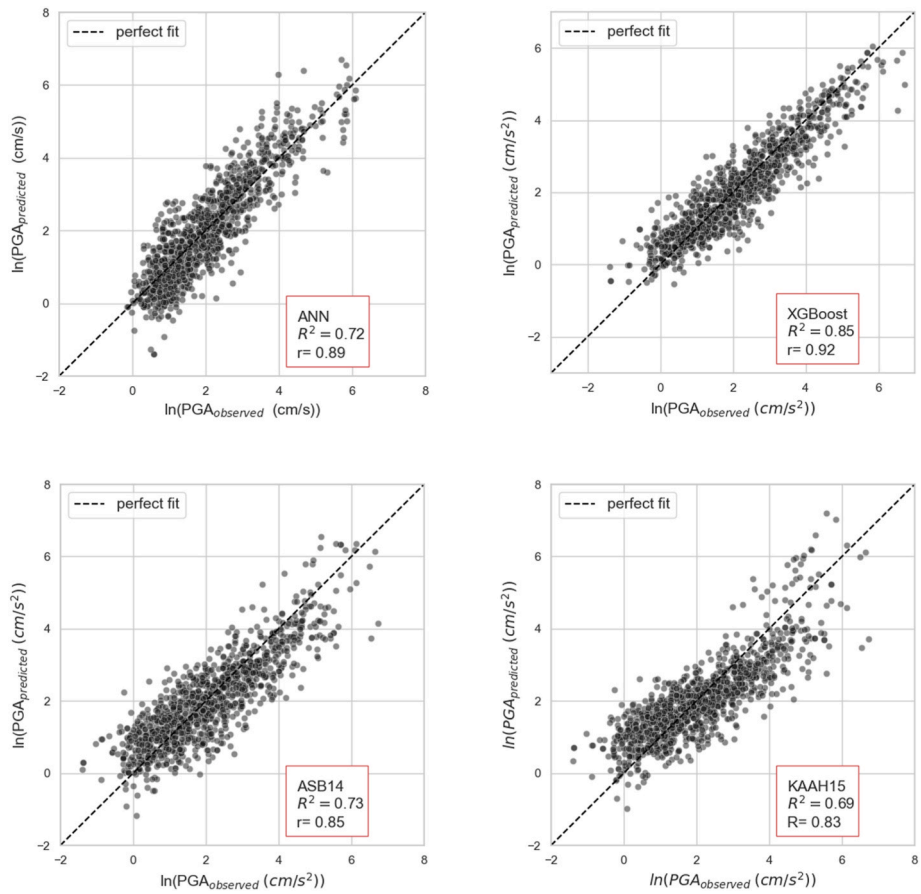

Fig. 9. Observed (targets) versus predicted (outputs) values in terms of a selected intensity measure, ln(PGA, from different machine-learning (ANN and XGBoost) and conventional algorithms (ASB14 [6] and KAAH15 [16]).

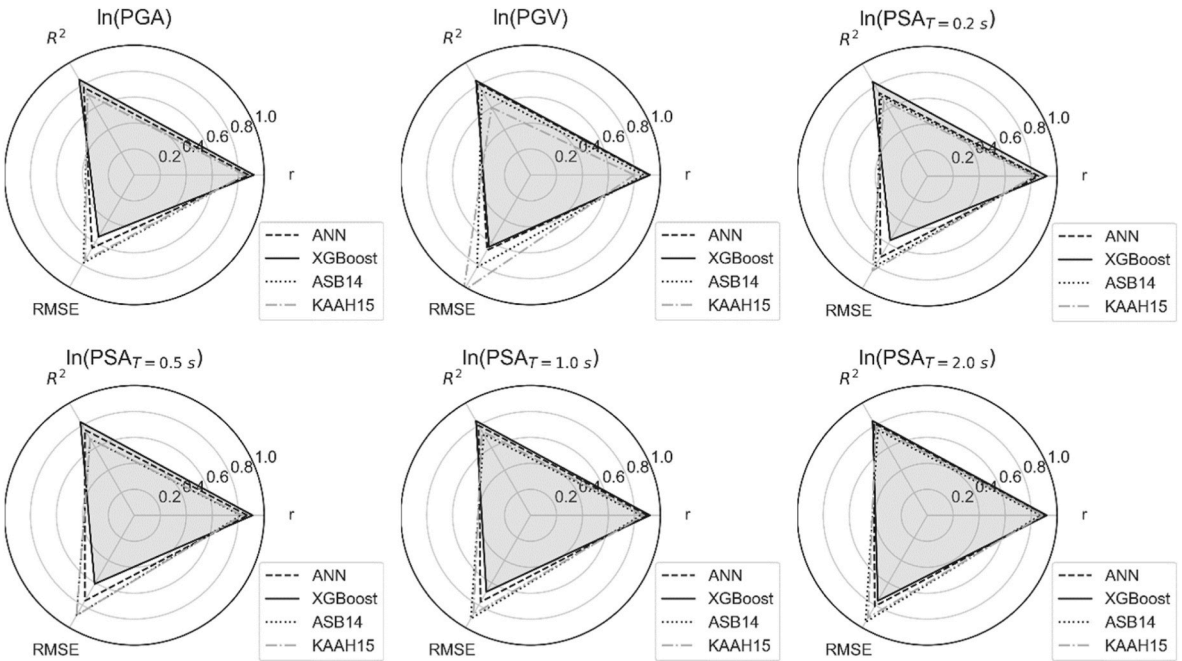

Fig. 10. Model performance indicators in terms of root-mean-square error (*RMSE*), coefficient of determination (R^2), and Pearson correlation coefficient (r) given for selected intensity measures, including ln(PGA), ln(PGV), ln($PSA_{T=0.2\ s}$), ln($PSA_{T=0.5\ s}$), ln($PSA_{T=1.0\ s}$), and ln($PSA_{T=2.0\ s}$) from different machine-learning (ANN and XGBoost) and conventional algorithms (ASB14 [6] and KAAH15 [16]). The smaller *RMSE* and higher R^2 and r indicate the better performance of each model.

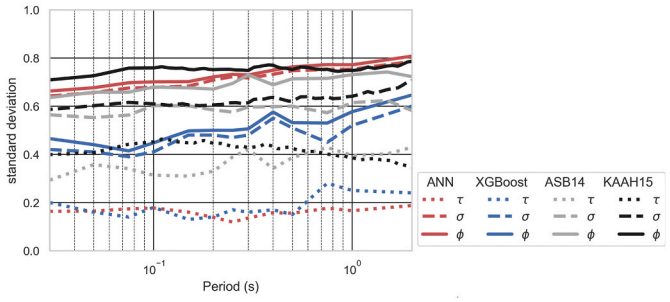

Fig. 11. Distribution of the inter-event (τ), intra-event (σ), and total (Ø) uncertainties for pseudo-spectral acceleration with respect to the period from different machine-learning (ANN and XGBoost) and conventional algorithms (ASB14 [6] and KAAH15 [16]).

two ML-based GMMs are compared, the inter-event residuals of the two models are approximately the same (except for the periods greater than 0.8 s). Nevertheless, the XGBoost model results in smaller intra-event residuals than the ANN model. This observation leads to smaller values of total uncertainty at all period ranges from the XGBoost model compared to the ANN model. It is also observed that the two empirical GMMs for the database of this study provide higher uncertainty than the XGBoost algorithm.

In conclusion, the ML-based GMM developed by the XGBoost algorithm is a robust predictive model because it has a lower *RMSE* and uncertainty and a higher R^2 than the models of ANN and ASB14 and KAAH15. This study, therefore, proposes the XGBoost-based GMM for the Turkish dataset. The outcomes of this model will be reviewed in depth from this point on.

4.2. The proposed GMM for the Turkish dataset (XGBoost)

XGBoost results in higher accuracy, and the results are further analysed with this model. Fig. 12 presents the distribution of the inter-event residuals from the XGBoost-based GMM regarding the source-related parameter (M_w) for different IMs. Similarly, Figs. 13 and 14 illustrate the distribution of the intra-event residuals, with respect to the site-related parameters, including R_{JB} and V_{S30}. The selected IMs are ln (PGA), ln(PGV), ln($PSA_{T=0.2\ s}$), ln($PSA_{T=0.5\ s}$), ln($PSA_{T=1.0\ s}$), and ln ($PSA_{T=2.0\ s}$). It is noted that the reason for considering different IMs herein is to investigate the performance of the developed model for estimating a bandwidth frequency, including low-, intermediate-, and high-frequencies. In these figures, the top box plots show frequency distribution of earthquakes with respect to a specific IM (i.e., M_w for inter-event, and R_{JB}, and V_{s30} for intra-event residuals). The boxplots on the right-hand side of these figures present the frequency distribution of the inter-/intra-event residuals. The diamonds represent data points beyond the third quartile of the data distribution. The fitted red lines to residuals versus explanatory variables indicate the means of residuals along those variables, and the shaded area around these lines represent the 95% confidence intervals for the true mean of the residuals. It is noted that the size of the confidence interval is proportional to the number of data points used in the analysis. The absence of any trend in the mean of residuals with tight confidence intervals suggests a high level of confidence in the unbiasedness of the model errors across the M_w, V_{s30}, and R_{JB} parameters. Nevertheless, this was tested using p-values, which are computed at a significance level of 0.05, and are presented in the subplots to facilitate the decision of accepting or rejecting the null hypothesis regarding the unbiasedness of the estimates. When the p-value of the IM is close to 1.0, it suggests that the resulting residual is less biased with respect to the input parameter. As

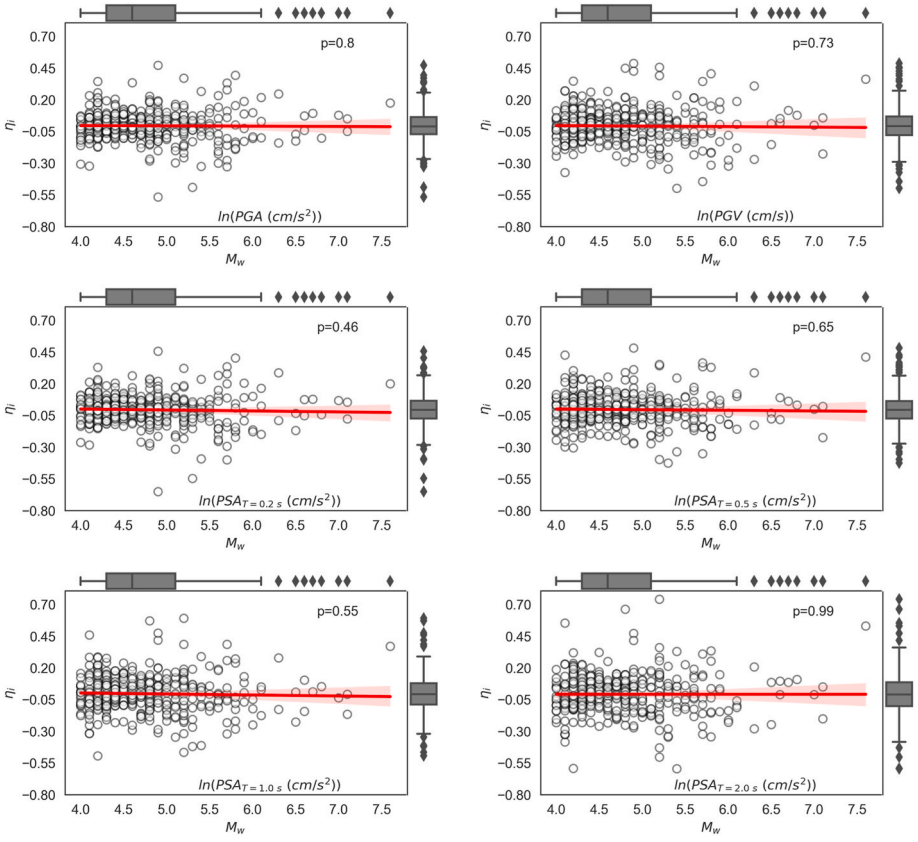

Fig. 12. Distribution of the inter-event residuals (η_i) with respect to magnitude (M_w) for selected intensity measures, including ln(PGA), ln(PGV), ln($PSA_{T=0.2\ s}$), ln ($PSA_{T=0.5\ s}$), ln($PSA_{T=1.0\ s}$), and ln($PSA_{T=2.0\ s}$) from the XGBoost-based ground motion model. The top boxplot shows distribution of M_w, while the boxplot on the right-hand side presents inter-event residual. The diamonds represent data points beyond the third quartile of the data distribution.

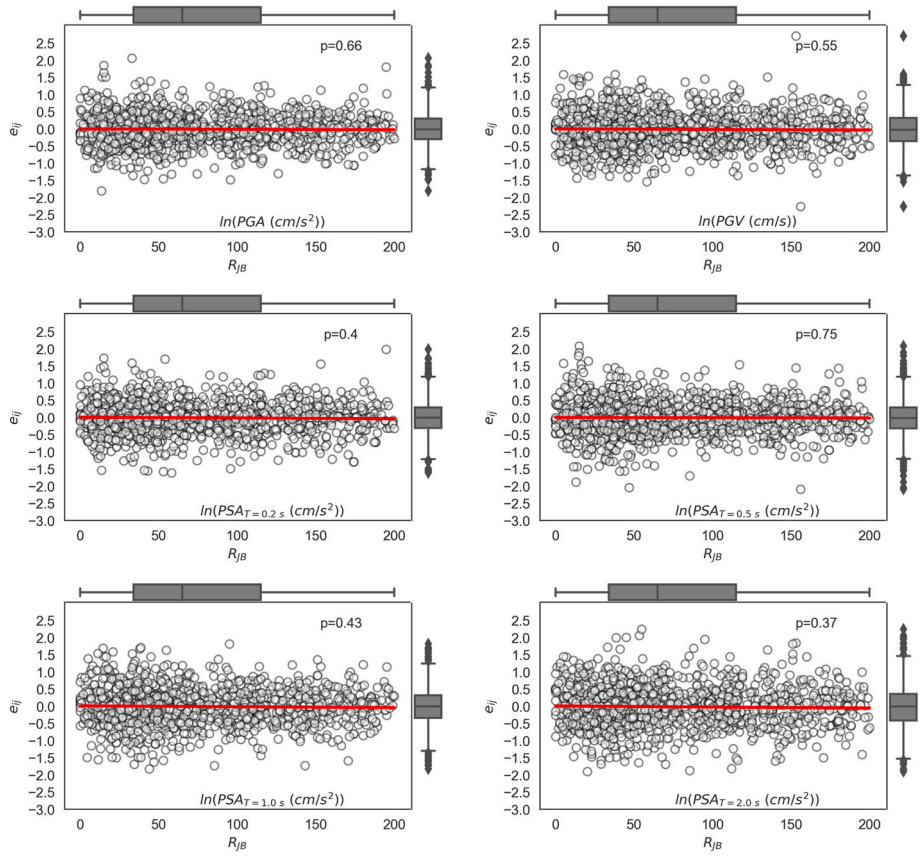

Fig. 13. Distribution of the intra-event residuals (ε_{ij}) with respect to distance (R_{JB}) for selected intensity measures, including ln(PGA), ln(PGV), ln($PSA_{T=0.2\ s}$), ln($PSA_{T=0.5\ s}$), ln($PSA_{T=1.0\ s}$), and ln($PSA_{T=2.0\ s}$) from the XGBoost-based ground motion model. The top boxplot shows distribution of R_{JB}, while the boxplot on the right-hand side presents intra-event residual. The diamonds represent data points beyond the third quartile of the data distribution.

shown Figs. 12–14 for all IMs, the inter-event residual varies between −0.75 and 0.75 and intra-event residual varies between −2.0 and 2.0, which is consistent with the observations of other studies [6,16]. Overall having a p-value over 0.05 for all considered IMs supports the assumption that the mean residual does not exhibit any discernible pattern, indicating that the model is free of source-related or site-related bias for all frequency bands. In Fig. 12, the confidence interval of the residuals is wider for earthquake events with Mw above 6, which may be attributed to insufficient datasets in the large-magnitude range. Finally, a comparison of the inter-event and intra-event residuals reveals that the intra-event residuals are greater than the inter-event residuals. This observation is consistent with the outcomes of other studies [6,16].

Although XGBoost is highly efficient, interpretation of its results is challenging compared to other predictive models such as ANN. There are techniques to address this problem, among which the one employed in this study and referred to as Shapley additive explanation (SHAP) [112]. SHAP is developed based on the game theory to interpret the outputs of any ML-based model, including XGBoost. In this technique, predictions are made with or without each of the input variables. Then, the importance of each input variable is measured by comparing these predictions. Fig. 15 presents the SHAP values for the entire database and the input features of the GMM, where the x-axis shows the SHAP value for each earthquake record. On the y-axis of these graphs, input variables are ordered from the most significant (on top) to the least effective (at the bottom). The value of input variables (feature value) is displayed on a scale from lowest to highest, with blue representing the most inferior and red representing the most superior. As seen in these plots, depending on the output of interest, M_w or R_{JB} has the highest effect on the model. For IMs, including ln(PGV), ln($PSA_{T=0.5s}$), ln($PSA_{T=1.0s}$), and ln($PSA_{T=2.0s}$), M_w provides the highest effect, while for ln(PGA) and ln($PSA_{T=0.2s}$) R_{JB} has the highest impact. The results of this study are consistent with a recent study by Withers et al. [113] which developed a ML-based GMM and showed that distance is the most critical factor, with decreasing importance as a function of period. Additionally, the influence of M_w increases at longer periods. Finally, for all outputs, FM has a minor effect on the predictions. These findings are consistent with the known physical behaviour of ground motion records and the results support and extend current knowledge of input parameter importance in GMMs.

The proposed GMM is evaluated further to determine if it can represent physics-based phenomena regarding the behaviour of real earthquakes. For this purpose, the results for various magnitude and distance combinations using $V_{S30} = 760$ m/s and FM of SS are compared. Fig. 16 displays the estimated PGA, PGV, $PSA_{T=0.2\ s}$, $PSA_{T=0.5\ s}$, $PSA_{T=1.0\ s}$, and $PSA_{T=2.0\ s}$ from the XGBoost-based GMM for various M_w ranges between 4.0 and 7.6. This effect is evaluated for four R_{JB} values, namely 15, 50, 75, and 150 km. An increase in M_w and a decrease in R_{JB} leads to a rise in the PGA, PGV, and PSA levels at all period ranges. In addition, the trend of the GMM is compared against the change in R_{JB} for different moment magnitudes (4.5, 5.5, 6.5, and 7.5) using various values of R_{JB} between 0 and 200 km. The outcomes are plotted in Fig. 17. Results show that an increase in R_{JB} leads to a decrease in the PGA, PGV and PSA levels at all period ranges, indicating that the suggested GMM effectively captures the distance-dependent attenuation. Consistent with the former observation, an increase in the magnitude results in an increase in the ground motion amplitudes. Upon analysing the ground motion data, we observed that as the distance between the source and site decreases, the difference between the magnitudes of 6.5 and 7.5 narrows. This phenomenon is a well-known characteristic of earthquake ground motions in shallow crustal earthquakes in interplate

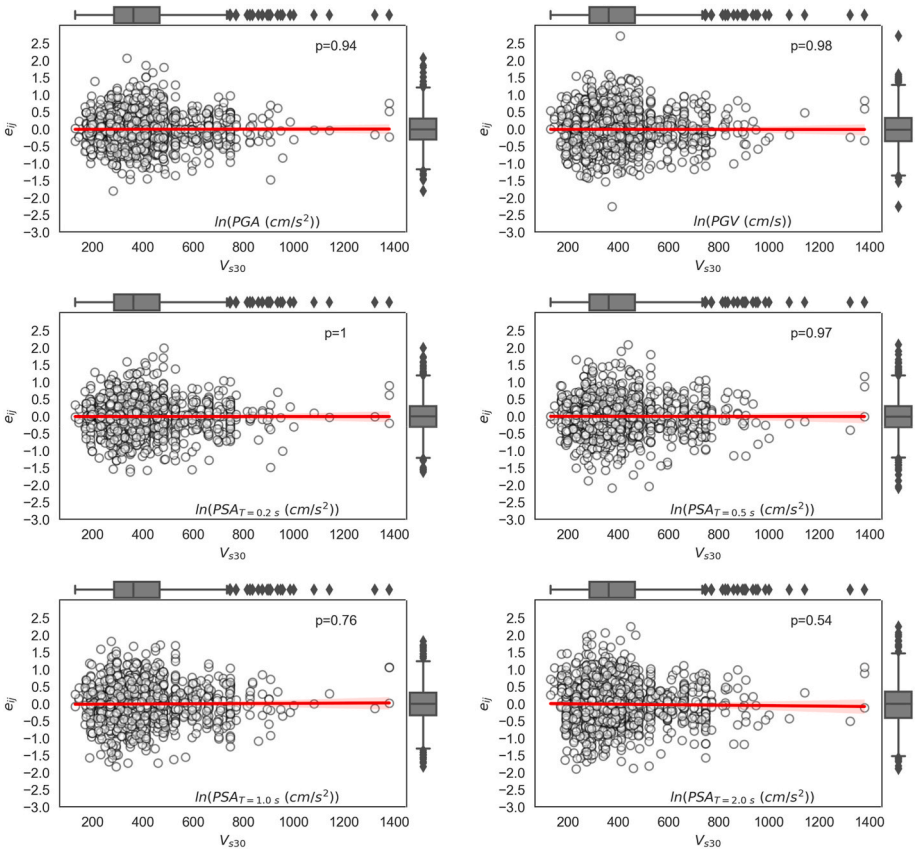

Fig. 14. Distribution of the intra-event residuals (ε_{ij}) with respect to shear wave velocity (V_{s30}) for selected intensity measures, including ln(PGA), ln(PGV), ln($PSA_{T\ =\ 0.2\ s}$), ln($PSA_{T\ =\ 0.5\ s}$), ln($PSA_{T\ =\ 1.0\ s}$), ln($PSA_{T\ =\ 2.0\ s}$) from the XGBoost-based ground motion model. The top boxplot shows distribution of V_{s30}, while the boxplot on the right-hand side presents intra-event residual. The diamonds represent data points beyond the third quartile of the data distribution.

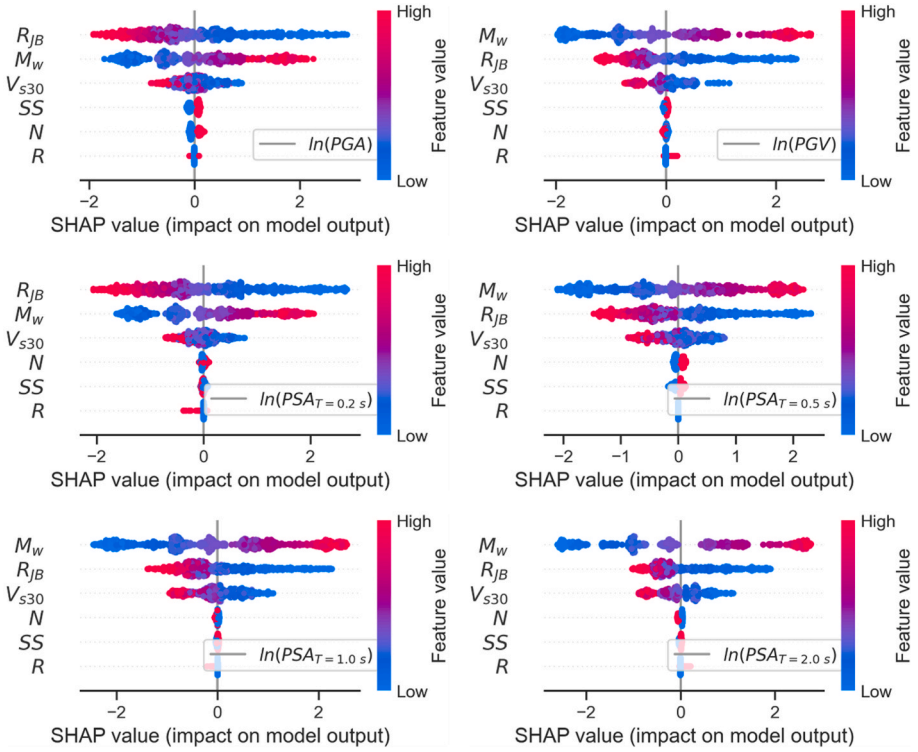

Fig. 15. Shapley additive explanation (SHAP) plots of the developed XGBoost-based ground motion model for selected intensity measures, including ln(PGA), ln (PGV), ln($PSA_{T\ =\ 0.2\ s}$), ln($PSA_{T\ =\ 0.5\ s}$), ln($PSA_{T\ =\ 1.0\ s}$), and ln($PSA_{T\ =\ 2.0\ s}$).

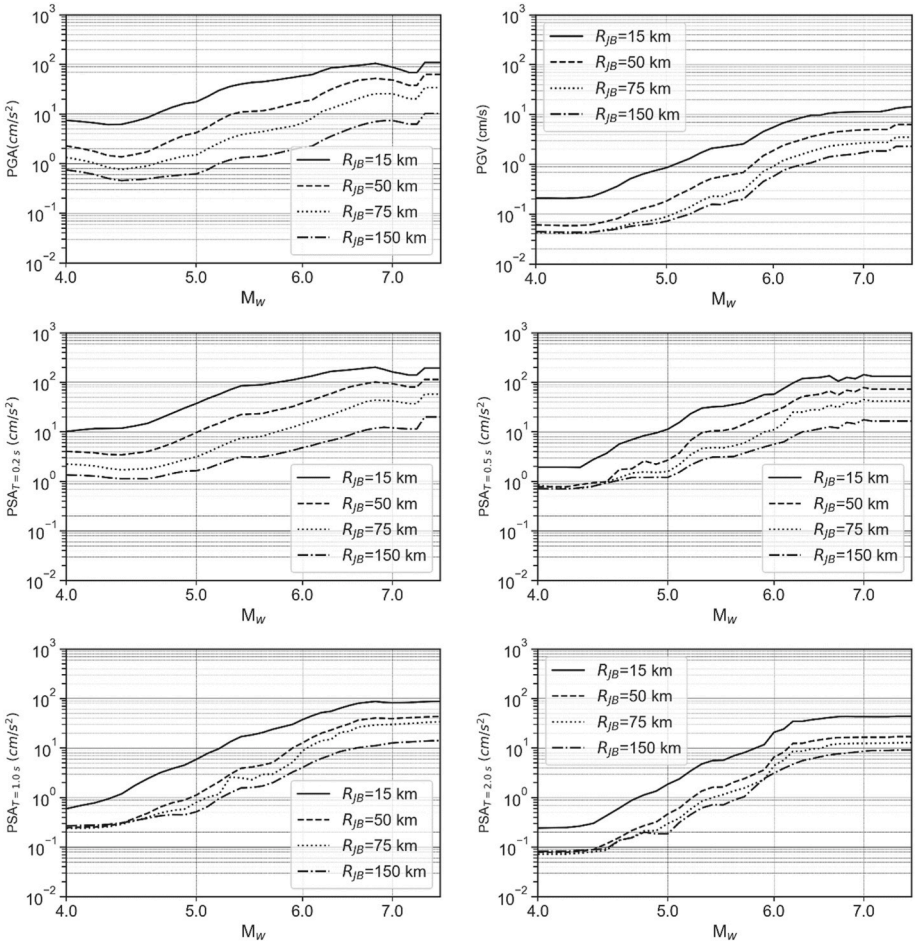

Fig. 16. Variation of PGA, PGV, $PSA_{T=0.2\,s}$, $PSA_{T=0.5\,s}$, $PSA_{T=1.0\,s}$, and $PSA_{T=2.0\,s}$ with respect to magnitude (M_w) for the focal mechanism (FD): strike-slip (SS), shear wave velocity (V_{S30}): 760 m/s, and distance (R_{JB}): 15, 50, 75 and 150 km using the developed XGBoost-based ground motion model.

tectonic regions. At a given distance from the source, the amplitude of ground motion depends on both the earthquake's magnitude and the distance from the source, leading to a narrowing effect between the ground motion amplitudes of earthquakes of different magnitudes at shorter distances. This narrowing effect is due to non-self-similar ground motion scaling and a magnitude-distance dependent saturation of earthquake ground motion amplitudes at larger magnitudes [114]. The good agreement between the ground motion model developed in this study and the observed behaviour of real earthquakes confirms the validity of the proposed approach, which incorporates this physical phenomenon, and supports its use for seismic hazard assessment in the study region.

This study also delved into radiation damping in ground motion records and its implications in the developed GMM. To investigate this phenomenon, the variation of the PSA concerning R_{JB} is investigated for the SS fault mechanism, V_{S30} = 760 m/s and two different moment magnitudes (M_w = 4.5 and M_w = 7.5). Results are plotted in Fig. 18a. The analysis revealed that the peak value of PSA decreases and shifts towards longer periods as the distance increases, which agrees with physical properties regarding the distance-dependent damping of ground motions as observed in previous studies [30,115]. This behaviour is attributed to the attenuation of seismic energy as it propagates away from the source due to the dissipative properties of the earth's crust. Additionally, it is verified that the developed GMM captures the radiation-damping characteristics of ground motion records, indicating the model's efficacy. Consistent with earthquake physics, the event magnitude affects how far the peak shifts. Finally, the efficiency of the developed GMM is investigated for soil classes C and D, which are the predominant soil types in the region, according to the NEHRP soil classification [90]. For this purpose, a representative V_{S30} value of 300 m/s and 560 m/s are used for soil types C and D, respectively. All estimations are carried out for the SS fault mechanism using R_{JB} = 30 km and for two different moment magnitudes (M_w = 4.5 and M_w = 7.5). Results are plotted in Fig. 18b. It is evident that as the soil type shifts from stiffer soil to softer soil (i.e., type C to type D), the ground motion amplitudes increase (particularly for longer periods), and the peak of the spectra moves near longer periods. Results also demonstrate that the earthquake magnitude affects how far the peak shifts, which is consistent with the physics of earthquakes.

Overall, the interpretation of the results reveals that the proposed XGBoost-based GMM can capture the behaviour of empirical GMMs with a need for minimal seismological data and without the necessity for nonlinear regression with multiple coefficients. The proposed model features a predefined closed-form function to estimate PGA, PGV and PSA for the Turkish dataset and is accessible to users without requiring many computations (Appendix B). Finally, the proposed XGBoost-based GMM could be implemented in future studies for large and more homogeneous datasets to improve its accuracy and minimise limitations, particularly for large-magnitude events and closer distances, by either using real worldwide datasets or combining real with region-specific simulated ground motions.

5. Summary and conclusions

This study investigates the efficiency of two alternative ML algorithms for predicting peak ground motion parameters and spectral

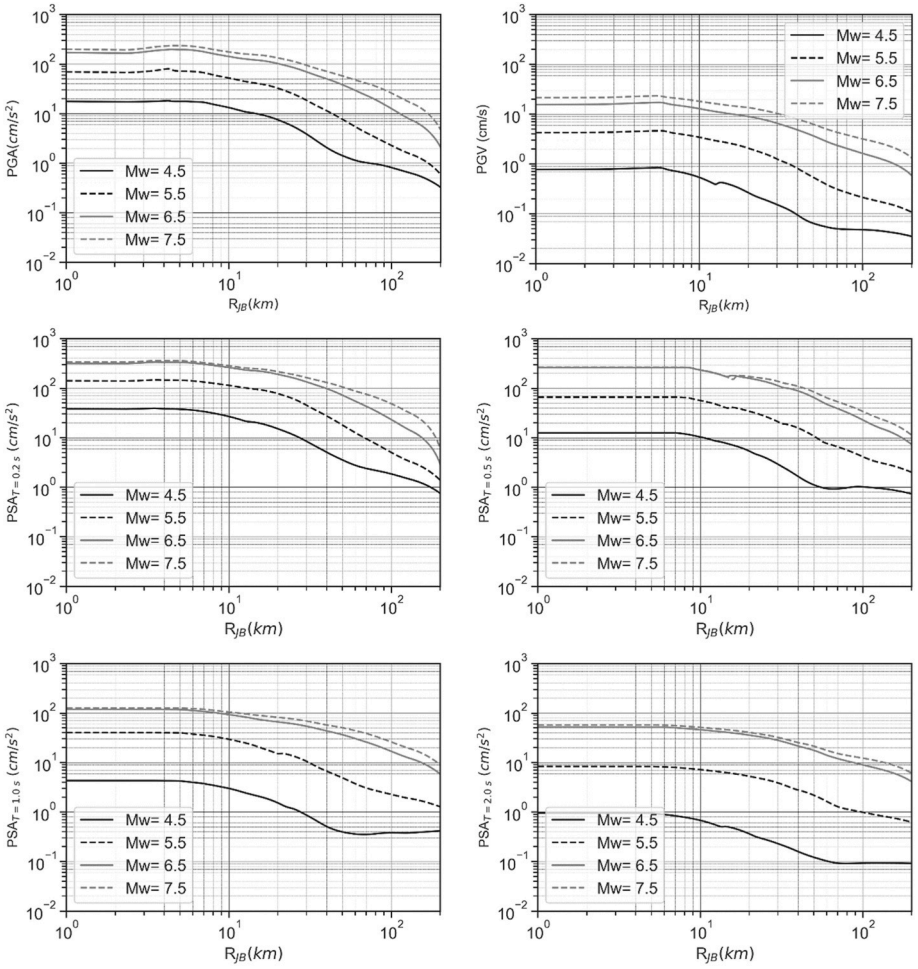

Fig. 17. Variation of selected intensity measures including PGA, PGV, $PSA_{T=0.2\,s}$, $PSA_{T=0.5\,s}$, $PSA_{T=1.0\,s}$, and $PSA_{T=2.0\,s}$ with respect to distance (R_{JB}) for the focal mechanism (FD): strike-slip (SS), shear wave velocity (V_{S30}): 760 m/s, and magnitude (M_w): 4.5, 5.5, 6.5, and 7.5 using the developed XGBoost-based ground motion model.

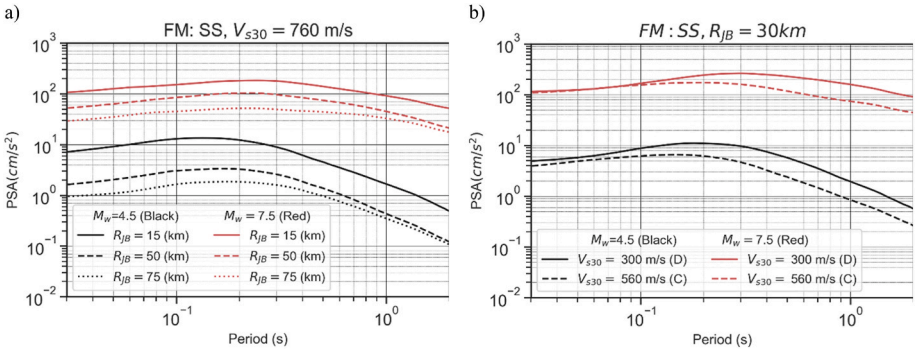

Fig. 18. Variation of pseudo-spectral acceleration (PSA) with respect to different (a) distance (R_{JB}): 15, 50, 75 km using shear wave velocity (V_{S30}): 760 m/s and (b) soil classes (type C and type D) [90] based on R_{JB}: 30 km both for the focal mechanism (FM): strike-slip (SS) and magnitude (M_w): 4.5 and 7.5 using the developed XGBoost-based ground motion model.

ordinates: ANN and XGBoost. The comparison includes PGA, PGV, and the PSA for 5% damping at 14 time periods within the range of 0.03–2.0 s. Turkey is used as a case study, and the dataset consists of 1166 ground motions with an M_w range of 4.0–7.6 and R_{JB} of 0–200 km observed during 383 seismic events since 1976. The stations feature V_{S30} ranging from 131 to 1380 m/s. To optimise the hyperparameters of the ML models, the Bayesian optimisation and trial-and-error procedures are used, respectively, for the XGBoost and ANN approaches, where the most effective hyperparameters are determined. To determine if the model is biased toward any predictor and to reduce the aleatory uncertainty [80], the ML algorithms of this study are modified by dividing the uncertainty into inter-event (between-event) and intra-event (within-event) terms. For this purpose, the method proposed by Abrahamson and Youngs [75] is implemented using a modified version of the likelihood function originally proposed. Next, the performance of the ML algorithms is determined using a set of model performance

indicators, including R^2, *RMSE*, and *r*. The developed models are compared to alternative empirical attenuation models existing in literature utilising the same database.

Interpretation of the results of this study reveals that developing nonparametric GMMs with modern ML techniques yields results better than those of conventional GMMs. Among the two ML algorithms, the best approach is chosen to be the XGBoost model since it provides the minimum error and maximum correlation for peak ground motion parameters and all spectral coordinates. Consistent with conventional GMMs, residual analysis generates acceptable uncertainty for all spectral values. The residuals are further evaluated regarding the inter-event and intra-event uncertainties with respect to explanatory factors. For this purpose, the inter-event residual is examined relative to the magnitude, whilst the intra-event residual is investigated against the soil and distance information of the dataset. Overall, the inter-event uncertainty for all spectral values is less than the intra-event uncertainty. It is also demonstrated that inter-event and intra-event residuals contain no substantial bias. with respect to the input variables, indicating that the constructed GMMs, in general, adequately describe the overall behaviour of the ground motion dataset.

The proposed XGBoost-based GMM accurately captures the physical properties of ground motion records, including distance-, magnitude-, soil-, and radiation-damping effects. The results also reveal a narrowing effect between ground motion amplitudes of earthquakes of large magnitudes at shorter distances, which is consistent with earthquake physics of shallow interplate tectonic regions. The good agreement between the developed ground motion model and the observed behaviour of real earthquakes in the region confirms the validity and effectiveness of the proposed approach for seismic hazard assessment in the study region.

Finally, the results of this study demonstrate that the proposed XGBoost-based GMM can capture the behaviour of empirical GMMs using minimum seismological data and without a need for nonlinear regression with numerous coefficients. This research introduces a novel nonparametric local GMM for the Turkish dataset by designing and implementing a web-based application platform for end users (Appendix B). The proposed model might be employed for other regions. Still, it is recommended to consider the range of the seismological parameters of the original dataset by accounting for the uncertainties involved. Last but not least, to increase accuracy and reduce limitations of the proposed model, particularly for large-magnitude events and closer distances, the suggested ML-based GMM could be further studied in future research for other tectonic zones with vast and more homogeneous real datasets. Other limitations of this study are the lack of consideration of spatial correlation and additional input parameters to capture near-field effects in the ground motion records. To improve the model, future studies could address these behaviours. This could also be fulfilled by combining real catalogues with region-specific simulated ground motions for regions with limited datasets of large-magnitude near-field records.

Funding

This work has received funding from multiple sources. This work was partly financed by FCT / MCTES through national funds (PIDDAC) under the R&D Unit Institute for Sustainability and Innovation in Structural Engineering (ISISE), under reference UIDB / 04029/2020, and under the Associate Laboratory Advanced Production and Intelligent Systems ARISE under reference LA/P/0112/2020. Additionally, the research was partly funded by the STAND4HERITAGE project, which received financial support from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program, Grant agreement No. 833123, as an Advanced Grant. The first author also acknowledges the support of national funds through FCT, under grant agreement 2020.08876.BD.

Author contributions

The XGBoost model and all figures are developed in Python by Amirhossein Mohammadi. Shaghayegh Karimzadeh developed the earthquake catalogue, prepared the first draft of the manuscript, and collaborated with the first and third authors to establish the ground motion models and codes regarding residual analyses. The ANN model was developed in MATLAB with Seyed Amir Banimahd. Volkan Ozsarac worked on the post-processing of the earthquake catalogue. Paulo B Lourenço mainly contributed to funding acquisition, supervision, manuscript reviewing and editing. All authors reviewed the results and approved the final version of the manuscript.

Data statement

The data of this study are available on request from the authors.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

This work has received funding from multiple sources. The national funds from FCT/MCTES through national funds (PIDDAC) under the R&D Unit Institute for Sustainability and Innovation in Structural Engineering (ISISE), reference UIDB/04029/2020, and the Associate Laboratory Advanced Production and Intelligent Systems ARISE, reference LA/P/0112/2020, provided partial financial support for this study. Additionally, the research was partly funded by the STAND4HERITAGE project, which received financial support from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program, Grant agreement No. 833123, as an Advanced Grant. The first author also acknowledges the support of national funds through FCT, under grant agreement 2020.08876.BD.

Appendix A. Derivation of the likelihood function

To derive the terms of the likelihood function, the primary form of this equation is considered as follows:

$$\ln L = -\frac{N}{2}\ln 2\,\pi - \frac{1}{2}\ln|\boldsymbol{C}| - \frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^T\boldsymbol{C}^{-1}(\mathbf{y}-\boldsymbol{\mu}) \tag{A-1}$$

where N is the total number of records. $\mathbf{y}$ and $\boldsymbol{\mu}$ are, respectively, the observed and estimated vectors of IMs. The term $\mathbf{C}$ is the covariance matrix of total residuals. The matrix $\mathbf{C}$, its determinant $|\mathbf{C}|$, and its inverse $\mathbf{C}^{-1}$ are expressed as follows:

$$\boldsymbol{C}=\sigma^2\boldsymbol{I}_N+\tau^2\sum_{i=1}^{M+}\boldsymbol{I}_{n_i} \tag{A-2}$$

$$|\boldsymbol{C}|=\sigma^{2(N-M)}\prod_{i=1}^{M}\left(\sigma^2+n_i\tau^2\right) \tag{A-3}$$

$$\boldsymbol{C}^{-1}=\frac{1}{\sigma^2}\boldsymbol{I}_N-\frac{1}{\sigma^2}\sum_{i=1}^{M+}\frac{\tau^2}{\sigma^2+n_i\tau^2}\boldsymbol{I}_{n_i}=\frac{1}{\sigma^2}\left(\boldsymbol{I}_N-\sum_{i=1}^{M+}\frac{\tau^2}{\sigma^2+n_i\tau^2}\boldsymbol{I}_{n_i}\right) \tag{A-4}$$

where σ and τ are the intra-event and inter-event standard deviations, M is the total number of events, n_i is the number of records for the i^{th} event, $\mathbf{I}_N$ is the identity matrix of size N, and $\sum_{i=1}^{M+}\mathbf{1}_{n_i}$ is the direct sum of $n_i\times n_i$ matrixes of ones for M events. The terms y_{ij} and μ_{ij} are, respectively, the observed and predicted values of IM for the i^{th} event at the j^{th} station. The likelihood function is obtained by substituting Eqs. (A- 2),(A- 3), and (A- 4) into Eq. (A-1), as follows:

$$\ln L=-\underbrace{\frac{N}{2}\ln 2\,\pi}_{1^{th}\text{ term}}\underbrace{-\frac{1}{2}\ln\left(\sigma^{2(N-M)}\prod_{i=1}^{M}\left(\sigma^2+n_i\tau^2\right)\right)}_{2^{nd}\text{ term}}\underbrace{-\frac{1}{2\sigma^2}(\mathbf{y}-\boldsymbol{\mu})^T\mathbf{I}_N(\mathbf{y}-\boldsymbol{\mu})}_{3^{rd}\text{ term}}\underbrace{+\frac{1}{2\sigma^2}\left((\mathbf{y}-\boldsymbol{\mu})^T\left(\sum_{i=1}^{M+}\frac{\tau^2}{\sigma^2+n_i\tau^2}\mathbf{1}_{n_i}\right)(\mathbf{y}-\boldsymbol{\mu})\right)}_{4^{th}\text{ term}} \tag{A-5}$$

The 2nd term of Eq. (A- 5) can be simplified as follows:

$$-\frac{1}{2}\ln\left(\sigma^{2(N-M)}\prod_{i=1}^{M}\left(\sigma^2+n_i\tau^2\right)\right)=-\frac{1}{2}(N-M)\ln\sigma^2-\frac{1}{2}\sum_{i=1}^{M}\ln\left(\sigma^2+n_i\tau^2\right) \tag{A-6}$$

The 3rd term of Eq. (A- 5) can be rewritten as:

$$-\frac{1}{2\sigma^2}(\mathbf{y}-\boldsymbol{\mu})^T\mathbf{I}_N(\mathbf{y}-\boldsymbol{\mu})=-\frac{1}{2\sigma^2}\sum_{i=1}^{M}\sum_{j=1}^{n_i}\left(y_{ij}-\mu_{ij}\right)^2 \tag{A-7}$$

The 4th term of Eq. (A- 5) can be converted from a matrix form to an algebraic expression through the expansion of the formula as follows:

$$\begin{aligned}&\frac{1}{2\sigma^2}\left((\mathbf{y}-\boldsymbol{\mu})^T\left(\sum_{i=1}^{M+}\frac{\tau^2}{\sigma^2+n_i\tau^2}\mathbf{1}_{n_i}\right)(\mathbf{y}-\boldsymbol{\mu})\right)=\frac{1}{2\sigma^2}\sum_{i=1}^{M}\left(\frac{\tau^2}{\sigma^2+n_i\tau^2}(\mathbf{y}_i-\boldsymbol{\mu}_i)^T\mathbf{1}_{n_i}(\mathbf{y}_i-\boldsymbol{\mu}_i)\right)=\\&\frac{\tau^2}{2\sigma^2}\sum_{i=1}^{M}\left(\frac{1}{\sigma^2+n_i\tau^2}\left(\sum_{j=1}^{n_i}(y_{ij}-\mu_{ij})\right)^2\right)=\frac{\tau^2}{2\sigma^2}\sum_{i=1}^{M}\left(\frac{1}{\sigma^2+n_i\tau^2}(n_i(\overline{Y}_i-\overline{\mu}_i))^2\right)=\\&\frac{\tau^2}{2\sigma^2}\sum_{i=1}^{M}\left(\frac{n_i^2}{\sigma^2+n_i\tau^2}(\overline{Y}_i-\overline{\mu}_i)^2\right)\end{aligned} \tag{A-8}$$

where $\mathbf{y}_i$ and $\boldsymbol{\mu}_i$ are the vector of observed and estimated values of IM for i^{th} event. The terms $\overline{Y}_i$ and $\overline{\mu}_i$ are, respectively, the mean values of observed and predicted IM for i^{th} event:

$$\overline{Y}_i=\frac{1}{n_i}\sum_{j=1}^{n_i}y_{ij},\overline{\mu}_i=\frac{1}{n_i}\sum_{j=1}^{n_i}\mu_{ij} \tag{A-9}$$

Finally, by replacing the 2nd, 3rd, and 4th terms of Eq. (A- 5) by extended formulas expressed in Eqs. (A- 6),(A- 7), and (A- 8) the Likelihood function can be given as follows:

$$\begin{aligned}&\ln L=-\frac{N}{2}\ln 2\,\pi-\frac{N-M}{2}\ln\sigma^2-\frac{1}{2}\sum_{i=1}^{M}\ln\left(\sigma^2+n_i\tau^2\right)-\frac{1}{2\sigma^2}\sum_{i=1}^{M}\sum_{j=1}^{n_i}\left(y_{ij}-\mu_{ij}\right)^2+\\&\frac{\tau^2}{2\sigma^2}\sum_{i=1}^{M}\left(\frac{n_i^2}{\sigma^2+n_i\tau^2}(\overline{Y}_i-\overline{\mu}_i)^2\right)\end{aligned} \tag{A-10}$$

Appendix B. Development of web-based application software

In this study, Streamlit is used to build a graphical user interface (GUI) tool that provides easy access to the GMM developed by XGBoost (the code is available at https://github.com/amirxdbx/GMM). Figure B1 illustrates the interface of the tool, which is available at https://amirxdbx-gmm-deploy-zc0z7k.streamlit.app/. As shown in this figure, the user defines the characteristics of a scenario earthquake in terms of M_w, R_{JB}, V_{s30}, and FM. The outcome of the software is given in terms of PGA, PGV, and PSA.

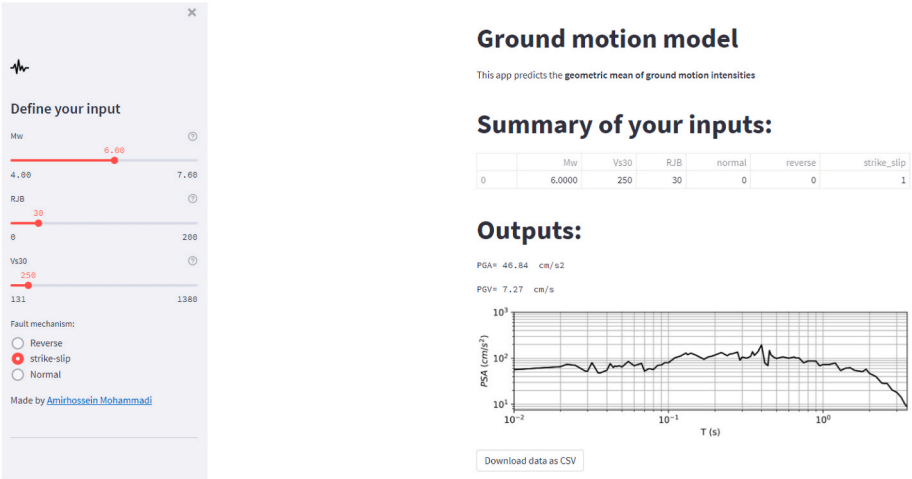

Fig. B1. GUI of the XGBoost-based GMM .

References

[1] Bindi D, Parolai S, Grosser H, Milkereit C, Durukal E. Empirical ground-motion prediction equations for northwestern Turkey using the aftershocks of the 1999 Kocaeli earthquake. Geophys Res Lett 2007;34. https://doi.org/10.1029/2007GL029222.

[2] Akkar S, Çağnan Z. A local ground-motion predictive model for Turkey, and its comparison with other regional and global ground-motion models. Bull Seismol Soc Am 2010;100:2978–95. https://doi.org/10.1785/0120090367.

[3] Gülerce Z, Kargioğlu B, Abrahamson NA. Turkey-adjusted NGA-W1 horizontal ground motion prediction models. Earthq Spectra 2016;32:75–100. https://doi.org/10.1193/022714EQS034M.

[4] Montalva GA, Bastías N, Rodriguez-Marek A. Ground-motion prediction equation for the Chilean subduction zone. Bull Seismol Soc Am 2017;107:901–11. https://doi.org/10.1785/0120160221.

[5] Boore DM, Stewart JP, Skarlatoudis AA, Seyhan E, Margaris B, Theodoulidis N, et al. A ground-motion prediction model for shallow crustal earthquakes in Greece. Bull Seismol Soc Am 2021;111:857–74. https://doi.org/10.1785/0120200270.

[6] Akkar S, Sandıkkaya MA, Bommer JJ. Empirical ground-motion models for point- and extended-source crustal earthquake scenarios in Europe and the Middle East. Bull Earthq Eng 2014;12:359–87. https://doi.org/10.1007/s10518-013-9461-4.

[7] Phung VB, Nguyen CN, Huang BS. On the development of region and site-specific ground motion prediction model for the region of I-lan, Taiwan. Eng Geol 2023;312:106933. https://doi.org/10.1016/j.enggeo.2022.106933.

[8] Yaghmaei-Sabegh S, Ebrahimi-Aghabagher M. Non-isotropic and isotropic ground motion prediction models. Pure Appl Geophys 2020;177:801–19. https://doi.org/10.1007/s00024-019-02303-3.

[9] Yaghmaei-Sabegh S, Karimzadeh S, Ebrahimi M, Ozsarac V, Du W. A new region-specific empirical model for prediction of ground motion significant duration in Turkey. Bull Earthq Eng 2022;20:4919–36. https://doi.org/10.1007/s10518-022-01417-9.

[10] Campbell KW, Bozorgnia Y. A ground motion prediction equation for JMA instrumental seismic intensity for shallow crustal earthquakes in active tectonic regimes. Earthq Eng Struct Dynam 2011;40:413–27. https://doi.org/10.1002/eqe.1027.

[11] Campbell KW, Bozorgnia Y. NGA-West2 ground motion model for the average horizontal components of PGA, PGV, and 5% damped linear acceleration response spectra. Earthq Spectra 2014;30:1087–114. https://doi.org/10.1193/062913EQS175M.

[12] Chiou BSJ, Youngs RR. Update of the Chiou and Youngs NGA model for the average horizontal component of peak ground motion and response spectra. Earthq Spectra 2014;30:1117–53. https://doi.org/10.1193/072813EQS219M.

[13] Bindi D, Massa M, Luzi L, Ameri G, Pacor F, Puglia R, et al. Pan-European ground-motion prediction equations for the average horizontal component of PGA, PGV, and 5%-damped PSA at spectral periods up to 3.0 s using the RESORCE dataset. Bull Earthq Eng 2014;12:391–430. https://doi.org/10.1007/s10518-013-9525-5.

[14] Idriss IM. An NGA-West2 empirical model for estimating the horizontal spectral values generated by shallow crustal earthquakes. Earthq Spectra 2014;30:1155–77. https://doi.org/10.1193/070613EQS195M.

[15] Abrahamson NA, Silva WJ, Kamai R. Summary of the ASK14 ground motion relation for active crustal regions. Earthq Spectra 2014;30:1025–55. https://doi.org/10.1193/070913EQS198M.

[16] Kale Ö, Akkar S, Ansari A, Hamzehloo H. A ground-motion predictive model for Iran and Turkey for horizontal PGA, PGV, and 5% damped response spectrum: investigation of possible regional effects. Bull Seismol Soc Am 2015;105:963–80. https://doi.org/10.1785/0120140134.

[17] Mu HQ, Yuen KV. Ground motion prediction equation development by heterogeneous bayesian learning. Comput Civ Infrastruct Eng 2016;31:761–76. https://doi.org/10.1111/mice.12215.

[18] Schnabel PB, Seed HB. Accelerations in rock for earthquakes in the western United States. Bull Seismol Soc Am 1973;63:501–16.

[19] Anderson JG, Lei Yutian. Nonparametric description of peak acceleration as a function of magnitude, distance, and site in Guerrero, Mexico. Bull Seismol Soc Am 1994;84:1003–17.

[20] Katayama T. An engineering prediction model of acceleration response spectra and its application to seismic hazard mapping. Earthq Eng Struct Dynam 1982;10:149–63. https://doi.org/10.1002/eqe.4290100111.

[21] Fajfar P, Peruš I. A non-parametric approach to attenuation relations. J Earthq Eng 1997;1:319–40. https://doi.org/10.1080/13632469708962371.

[22] Peruš I, Fajfar P. How reliable are the ground motion prediction equations? 20th. Int Conf Struct Mech React Technol 2009;1–9.

[23] Peruš I, Fajfar P. Ground-motion prediction by a non-parametric approach. Earthq Eng Struct Dynam 2010;39:1395–416. https://doi.org/10.1002/eqe.1007.

[24] Podili B, Raghukanth STG. Ground motion prediction equations for higher order parameters. Soil Dynam Earthq Eng 2019;118:98–110. https://doi.org/10.1016/j.soildyn.2018.11.027.

[25] Kubo H, Kunugi T, Suzuki W, Suzuki S, Aoi S. Hybrid predictor for ground-motion intensity with machine learning and conventional ground motion prediction equation. Sci Rep 2020;10. https://doi.org/10.1038/s41598-020-68630-x.

[26] Lekshmy PR, Raghukanth STG. A hybrid genetic algorithm-neural network model for power spectral density compatible ground motion prediction. Soil Dynam Earthq Eng 2021;142. https://doi.org/10.1016/j.soildyn.2020.106528.

[27] Li C, Ji D, Zhai C, Ma Y, Xie L. Vertical ground motion model for the NGA-West2 database using deep learning method. Soil Dynam Earthq Eng 2023;165:107713. https://doi.org/10.1016/j.soildyn.2022.107713.

[28] Kong Q, Trugman DT, Ross ZE, Bianco MJ, Meade BJ, Gerstoft P. Machine learning in seismology: turning data into insights. Seismol Res Lett 2019;90:3–14. https://doi.org/10.1785/0220180259.

[29] Alimoradi A, Beck JL. Machine-learning methods for earthquake ground motion analysis and simulation. J Eng Mech 2015;141. https://doi.org/10.1061/(asce)em.1943-7889.0000869.

[30] Dhanya J, Raghukanth STG. Ground motion prediction model using artificial neural network. Pure Appl Geophys 2018;175:1035–64. https://doi.org/10.1007/s00024-017-1751-3.

[31] Dhanya J, Sagar D, Raghukanth STG. Predictive models for ground motion parameters using artificial neural network. Lect. Notes Civ. Eng. 2019;12:93–105. https://doi.org/10.1007/978-981-13-0365-4_8.

[32] Boore DM, Stewart JP, Seyhan E, Atkinson GM. NGA-West2 equations for predicting PGA, PGV, and 5% damped PSA for shallow crustal earthquakes. Earthq Spectra 2014;30:1057–85. https://doi.org/10.1193/070113EQS184M.

[33] Derras B, Bard PY, Cotton F. Site-condition proxies, ground motion variability, and data-driven GMPEs: insights from the NGA-West2 and RESORCE data sets. Earthq Spectra 2016;32:2027–56. https://doi.org/10.1193/060215EQS082M.

[34] Dhanya J, Raghukanth STG. Neural network-based hybrid ground motion prediction equations for Western Himalayas and North-Eastern India. Acta Geophys 2020;68:303–24. https://doi.org/10.1007/s11600-019-00395-y.

[35] Sharma V, Dhanya J, Gade M, Sivasubramonian J. New generalized ANN-based hybrid broadband response spectra generator using physics-based simulations. Nat Hazards 2022;116:1879–901. https://doi.org/10.1007/s11069-022-05746-5.

[36] Paolucci R, Gatti F, Infantino M, Smerzini C, Özcebe AG, Stupazzini M. Broadband ground motions from 3D physics-based numerical simulations using artificial neural networks. Bull Seismol Soc Am 2018;108:1272–86. https://doi.org/10.1785/0120170293.

[37] Karimi Ghalehjough B, Mahinroosta R. Peak ground acceleration prediction by fuzzy logic modeling for Iranian plateau. Acta Geophys 2020;68:75–89. https://doi.org/10.1007/s11600-019-00394-z.

[38] Khosravikia F, Clayton P. Machine learning in ground motion prediction. Comput Geosci 2021;148. https://doi.org/10.1016/j.cageo.2021.104700.

[39] Seo H, Kim J, Kim B. Machine-learning-based surface ground-motion prediction models for South Korea with low-to-moderate seismicity. Bull Seismol Soc Am 2022;112:1549–64. https://doi.org/10.1785/0120210244.

[40] Pathak J, Paul DK, Godbole PN. ANN based attenuation relationship for estimation of PGA using Indian strong-motion data. In: First eur. Conf. Earthq. Eng. Seismol. (a Jt. event 13th ECEE 30th gen. Assem. ESC); 2006. Geneva, Switzerland.

[41] Güllü H, Erçelebi E. A neural network approach for attenuation relationships: an application using strong ground motion data from Turkey. Eng Geol 2007;93:65–81. https://doi.org/10.1016/j.enggeo.2007.05.004.

[42] Gandomi M, Soltanpour M, Zolfaghari MR, Gandomi AH. Prediction of peak ground acceleration of Iran's tectonic regions using a hybrid soft computing technique. Geosci Front 2016;7:75–82. https://doi.org/10.1016/j.gsf.2014.10.004.

[43] Tezcan J, Hazirbaba YD, Cheng Q. A kernel-based mixed effect regression model for earthquake ground motions. Adv Eng Software 2016;120:26–35. https://doi.org/10.1016/j.advengsoft.2016.06.002.

[44] Thomas S, Pillai GN, Pal K, Jagtap P. Prediction of ground motion parameters using randomized ANFIS (RANFIS). Appl Soft Comput J 2016;40:624–34. https://doi.org/10.1016/j.asoc.2015.12.013.

[45] Thomas S, Pillai GN, Pal K. Prediction of peak ground acceleration using ε-SVR, ν-SVR and Ls-SVR algorithm. Geomatics, Nat Hazards Risk 2017;8:177–93. https://doi.org/10.1080/19475705.2016.1176604.

[46] Hamze-Ziabari SM, Bakhshpoori T. Improving the prediction of ground motion parameters based on an efficient bagging ensemble model of M5′ and CART algorithms. Appl Soft Comput J 2018;68:147–61. https://doi.org/10.1016/j.asoc.2018.03.052.

[47] Kaveh A, Hamze-Ziabari SM, Bakhshpoori T. Feasibility of pso-anfis-pso and ga-anfis-ga models in prediction of peak ground acceleration. Int J Optim Civ Eng 2018;8:1–14.

[48] Khosravikia F, Zeinali Y, Nagy Z, Clayton P, Rathje EM. Neural network-based equations for predicting PGA and PGV in Texas. Oklahoma, and Kansas; 2018. p. 538–49. https://doi.org/10.1061/9780784481462.052.

[49] Derakhshani A, Foruzan AH. Predicting the principal strong ground motion parameters: a deep learning approach. Appl Soft Comput J 2019;80:192–201. https://doi.org/10.1016/j.asoc.2019.03.029.

[50] Wiszniowski J. Estimation of a ground motion model for induced events by Fahlman's Cascade Correlation Neural Network. Comput Geosci 2019;131:23–31. https://doi.org/10.1016/j.cageo.2019.06.006.

[51] Raghucharan MC, Somala SN, Erteleva O, Rogozhi E. Seismic attenuation model for data gap regions using recorded and simulated ground motions. Nat Hazards 2021;107:423–46. https://doi.org/10.1007/s11069-021-04589-w.

[52] Ahmad I, El Naggar MH, Khan AN. Neural network based attenuation of strong motion peaks in Europe. J Earthq Eng 2008;12:663–80. https://doi.org/10.1080/13632460701758570.

[53] Huang H, Ramkrishnan R, Kolathayar S, Garg A, Yadav JS. Development of region-specific new generation attenuation relations for north India using artificial neural networks. Lect. Notes Civ. Eng 2021;123:85–101. https://doi.org/10.1007/978-981-33-4324-5_6. Springer Science and Business Media Deutschland GmbH.

[54] Ji D, Li C, Zhai C, Dong Y, Katsanos EI, Wang W. Prediction of ground-motion parameters for the NGA-west2 database using refined second-order deep neural networks. Bull Seismol Soc Am 2021;111:3278–96. https://doi.org/10.1785/0120200388.

[55] Kashani AR, Akhani M, Camp CV, Gandomi AH. A neural network to predict spectral acceleration. Basics Comput. Geophys. 2020:335–49. https://doi.org/10.1016/B978-0-12-820513-6.00006-0.

[56] Hu J, Zhang H. Support vector machine method for developing ground motion models for earthquakes in western part of China. J Earthq Eng 2022;26:5679–94. https://doi.org/10.1080/13632469.2021.1884146.

[57] Günaydın K, Günaydın A. Peak ground acceleration prediction by artificial neural networks for northwestern Turkey. Math Probl Eng 2008. https://doi.org/10.1155/2008/919420.

[58] Cabalar AF, Cevik A. Genetic programming-based attenuation relationship: an application of recent earthquakes in Turkey. Comput Geosci 2009;35:1884–96. https://doi.org/10.1016/j.cageo.2008.10.015.

[59] Alavi AH, Gandomi AH. Prediction of principal ground-motion parameters using a hybrid method coupling artificial neural networks and simulated annealing. Comput Struct 2011;89:2176–94. https://doi.org/10.1016/j.compstruc.2011.08.019.

[60] Kuehn NM, Riggelsen C, Scherbaum F. Modeling the joint probability of earthquake, site, and ground-motion parameters using Bayesian networks. Bull Seismol Soc Am 2011;101:235–49. https://doi.org/10.1785/0120100080.

[61] Derras B, Bard PY, Cotton F. Towards fully data driven ground-motion prediction models for Europe. Bull Earthq Eng 2014;12:495–516. https://doi.org/10.1007/s10518-013-9481-0.

[62] Hermkes M, Kuehn NM, Riggelsen C. Simultaneous quantification of epistemic and aleatory uncertainty in GMPEs using Gaussian process regression. Bull Earthq Eng 2014;12:449–66. https://doi.org/10.1007/s10518-013-9507-7.

[63] Yerlikaya-Özkurt F, Askan A, Weber GW. An alternative approach to the ground motion prediction problem by a non-parametric adaptive regression method. Eng Optim 2014;46:1651–68. https://doi.org/10.1080/0305215X.2013.858141.

[64] Tezcan J, Cheng Q. Support vector regression for estimating earthquake response spectra. Bull Earthq Eng 2012;10:1205–19. https://doi.org/10.1007/s10518-012-9350-2.

[65] Atkinson GM, Bommer JJ, Abrahamson NA. Alternative approaches to modeling epistemic uncertainty in ground motions in probabilistic seismic-hazard analysis. Seismol Res Lett 2014;85:1141–4. https://doi.org/10.1785/0220140120.

[66] Bommer JJ, Scherbaum F. The use and misuse of logic trees in probabilistic seismic hazard analysis. Earthq Spectra 2008;24:997–1009. https://doi.org/10.1193/1.2977755.

[67] Douglas J. Capturing geographically-varying uncertainty in earthquake ground motion models or what we think we know may change. Geotech. Geol. Earthq. Eng. 2018;46:153–81. https://doi.org/10.1007/978-3-319-75741-4_6.

[68] Douglas John. Ground motion prediction equations. 2022. Glasgow.

[69] Gülkan P, Kalkan E. Attenuation modeling of recent earthquakes in Turkey. J Seismol 2002;6:397–409. https://doi.org/10.1023/A:1020087426440.

[70] Kalkan E, Gülkan P. Site-dependent spectra derived from ground motion records in Turkey. Earthq Spectra 2004;20:1111–38. https://doi.org/10.1193/1.1812555.

[71] Özbey C, Sari A, Manuel L, Erdik M, Fahjan Y. An empirical attenuation relationship for Northwestern Turkey ground motion using a random effects approach. Soil Dynam Earthq Eng 2004;24:115–25. https://doi.org/10.1016/j.soildyn.2003.10.005.

[72] Akinci A, Malagnini L, Herrmann RB, Gok R, Sørensen MB. Ground motion scaling in the Marmara region, Turkey. Geophys J Int 2006;166:635–51. https://doi.org/10.1111/j.1365-246X.2006.02971.x.

[73] Akyol N, Karagöz Ö. Empirical attenuation relationships for western Anatolia, Turkey. Turk J Earth Sci 2009;18:351–82. https://doi.org/10.3906/yer-0705-2.

[74] Kayabali K, Beyaz T. Strong motion attenuation relationship for Turkey-a different perspective. Bull Eng Geol Environ 2011;70:467–81. https://doi.org/10.1007/s10064-010-0335-6.

[75] Abrahamson NA, Youngs RR. A stable algorithm for regression analyses using the random effects model. Bull Seismol Soc Am 1992;82:505–10. https://doi.org/10.1785/bssa0820010505.

[76] AFAD. Turkish accelerometric database and analysis system. Disaster Emerg Manag Pres 2020. https://tadas.afad.gov.tr.

[77] Wu J, Chen XY, Zhang H, Xiong LD, Lei H, Deng SH. Hyperparameter optimization for machine learning models based on bayesian optimization. J Electron Sci Technol 2019;17:26–40. https://doi.org/10.11989/JEST.1674-862X.80904120.

[78] Naderpour H, Kheyroddin A, Amiri GG. Prediction of FRP-confined compressive strength of concrete using artificial neural networks. Compos Struct 2010;92:2817–29. https://doi.org/10.1016/j.compstruct.2010.04.008.

[79] Tran VL, Thai DK, Kim SE. Application of ANN in predicting ACC of SCFST column. Compos Struct 2019;228. https://doi.org/10.1016/j.compstruct.2019.111332.

[80] Bommer JJ, Abrahamson NA. Why do modern probabilistic seismic-hazard analyses often lead to increased hazard estimates? Bull Seismol Soc Am 2006;96:1967. https://doi.org/10.1785/0120060043. –77.

[81] McKenzie D. The East Anatolian Fault: a major structure in eastern Turkey. Earth Planet Sci Lett 1976;29:189–93. https://doi.org/10.1016/0012-821X(76)90038-8.

[82] Barka AA. The North Anatolian Fault Zone. Ann Tect 1992;6:164–95.

[83] Reilinger R, McClusky S, Vernant P, Lawrence S, Ergintav S, Cakmak R, et al. GPS constraints on continental deformation in the Africa-Arabia-Eurasia continental collision zone and implications for the dynamics of plate interactions. J Geophys Res Solid Earth 2006;111. https://doi.org/10.1029/2005JB004051.

[84] Duman TY, Çan T, Emre Ö, Kadirioğlu FT, Başarır Baştürk N, Kılıç T, et al. Seismotectonic database of Turkey. Bull Earthq Eng 2018;16:3277–316. https://doi.org/10.1007/s10518-016-9965-9.

[85] Akkar S, Sandıkkaya MA, Şenyurt M, Sisi AA, Ay B, Traversa P, et al. Reference database for seismic ground-motion in Europe (RESORCE). Bull Earthq Eng 2014;12:311–39. https://doi.org/10.1007/s10518-013-9506-8.

[86] Sandikkaya MA, Yilmaz MT, Bakir BS, Yilmaz Ö. Site classification of Turkish national strong-motion stations. J Seismol 2010;14:543–63. https://doi.org/10.1007/s10950-009-9182-y.

[87] Tan O, Tapirdamaz MC, Yörük A. The earthquake catalogues for Turkey. Turk J Earth Sci 2008;17:405–18.

[88] Örgülü G. Seismicity and source parameters for small-scale earthquakes along the splays of the north Anatolian fault (NAF) in the marmara sea. Geophys J Int 2011;184:385–404. https://doi.org/10.1111/j.1365-246X.2010.04844.x.

[89] Wollin C, Bohnhoff M, Martínez-Garzón P, Küperkoch L, Raub C. A unified earthquake catalogue for the Sea of Marmara Region, Turkey, based on automatized phase picking and travel-time inversion: seismotectonic implications. Tectonophysics 2018;747–748:416–44. https://doi.org/10.1016/j.tecto.2018.05.020.

[90] Federal Emergency Management Agency. NEHRP guidelines for the seismic rehabilitation of buildings. 1997.

[91] Douglas J, Halldórsson B. On the use of aftershocks when deriving ground-motion prediction equations. In: 9th US natl. 10th can. Conf. Earthq. Eng. 2010, incl. Pap. From 4th int. Tsunami symp. vol. 9; 2010. p. 7456–65.

[92] Ambraseys NN, Douglas J, Sarma SK, Smit PM. Equations for the estimation of strong ground motions from shallow crustal earthquakes using data from Europe and the middle east: horizontal peak ground acceleration and spectral acceleration. Bull Earthq Eng 2005;3:1–53. https://doi.org/10.1007/s10518-005-0183-0.
[93] Ozsarac V, Monteiro R, Calvi GM. Probabilistic seismic assessment of reinforced concrete bridges using simulated records. Struct Infrastruct Eng 2023;19:554–74. https://doi.org/10.1080/15732479.2021.1956551.
[94] Goda K, Atkinson GM. Intraevent spatial correlation of ground-motion parameters using SK-net data. Bull Seismol Soc Am 2010;100:3055–67. https://doi.org/10.1785/0120100031.
[95] Schiappapietra E, Smerzini C. Spatial correlation of broadband earthquake ground motion in Norcia (Central Italy) from physics-based simulations. Bull Earthq Eng 2021;19:4693–717. https://doi.org/10.1007/s10518-021-01160-7.
[96] Hong HP, Zhang Y, Goda K. Effect of spatial correlation on estimated ground-motion prediction equations. Bull Seismol Soc Am 2009;99:928–34. https://doi.org/10.1785/0120080172.
[97] Joyner WB, Boore DM. Methods for regression analysis of strong-motion data. Bull Seismol Soc Am 1993;83:469–87. https://doi.org/10.1785/bssa0830020469.
[98] Ulusay R, Tuncay E, Sonmez H, Gokceoglu C. An attenuation relationship based on Turkish strong motion data and iso-acceleration map of Turkey. Eng Geol 2004;74:265–91. https://doi.org/10.1016/j.enggeo.2004.04.002.
[99] Schwarz J, Ende C, Habenberger J, Lang DH, Baumbach M, Grosser H, et al. Horizontal and vertical response spectra on the basis of strong-motion recordings from the 1999 Turkey earthquakes. Proc XXVIII Gen Assem Eur Seismol Comm (ESC) Sep 2002 2002.
[100] Douglas J, Akkar S, Ameri G, Bard PY, Bindi D, Bommer JJ, et al. Comparisons among the five ground-motion models developed using RESORCE for the prediction of response spectral accelerations due to earthquakes in Europe and the Middle East. Bull Earthq Eng 2014;12:341–58. https://doi.org/10.1007/s10518-013-9522-8.
[101] Kotha SR, Bindi D, Cotton F. Partially non-ergodic region specific GMPE for Europe and Middle-East. Bull Earthq Eng 2016;14:1245–63. https://doi.org/10.1007/s10518-016-9875-x.
[102] Naderpour H, Alavi SA. A proposed model to estimate shear contribution of FRP in strengthened RC beams in terms of Adaptive Neuro-Fuzzy Inference System. Compos Struct 2017;170:215–27. https://doi.org/10.1016/j.compstruct.2017.03.028.
[103] Haykin S. Neural networks and learning machines, vol. 3; 2008. 978-0131471399.
[104] Marquardt DW. An algorithm for least-squares estimation of nonlinear parameters. J Soc Ind Appl Math 1963;11:431–41. https://doi.org/10.1137/0111030.
[105] Levenberg K. A method for the solution of certain non-linear problems in least squares. Q Appl Math 1944;2:164–8. https://doi.org/10.1090/qam/10666.
[106] Chen T, Guestrin C. XGBoost: a scalable tree boosting system. Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. 2016;13–17:785–94. https://doi.org/10.1145/2939672.2939785. Augu, Association for Computing Machinery.
[107] Bakouregui AS, Mohamed HM, Yahia A, Benmokrane B. Explainable extreme gradient boosting tree-based prediction of load-carrying capacity of FRP-RC columns. Eng Struct 2021;245:112836. https://doi.org/10.1016/j.engstruct.2021.112836.
[108] Shahriari B, Swersky K, Wang Z, Adams RP, De Freitas N. Taking the human out of the loop: a review of Bayesian optimization. Proc IEEE 2016;104:148–75. https://doi.org/10.1109/JPROC.2015.2494218.
[109] Alam MS, Sultana N, Hossain SMZ. Bayesian optimization algorithm based support vector regression analysis for estimation of shear capacity of FRP reinforced concrete members. Appl Soft Comput 2021;105:107281. https://doi.org/10.1016/J.ASOC.2021.107281.
[110] Karaboga D, Akay B. A comparative study of Artificial Bee Colony algorithm. Appl Math Comput 2009;214:108–32. https://doi.org/10.1016/j.amc.2009.03.090.
[111] Golberg DE. Genetic algorithms in search, optimization, and machine learning, vol. 1989. Addion Wesley; 1989. p. 36.
[112] Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst 2017;30:4766–75.
[113] Withers KB, Moschetti MP, Thompson EM. A machine learning approach to developing ground motion models from simulated ground motions. Geophys Res Lett 2020;47. https://doi.org/10.1029/2019GL086690.
[114] Yenier E, Atkinson GM. Equivalent point-source modeling of moderate-to-large magnitude earthquakes and associated ground-motion saturation effects. Bull Seismol Soc Am 2014;104:1458–78. https://doi.org/10.1785/0120130147.
[115] Huang SK, Chen CT, Loh CH, Chang LM. Extracting ground motion characteristics of distant earthquakes for mitigating displacement-sensitive equipment. J Low Freq Noise Vib Act Control 2018;37:859–80. https://doi.org/10.1177/1461348418781984.