



Developing machine learning-based ground motion models to predict peak ground velocity in Türkiye

Fahrettin Kuran · Gülüm Tanırcan ·
Elham Pashaei

Received: 29 February 2024 / Accepted: 19 August 2024 / Published online: 5 September 2024
© The Author(s), under exclusive licence to Springer Nature B.V. 2024

Abstract This paper introduces machine learning-based Türkiye-specific ground motion models for the geometric mean horizontal component of peak ground velocity (PGV). PGV is a significant intensity metric to measure and diagnose potential earthquake damage in structures. Reliable prediction of PGV is of essential importance in precise calculations of seismic hazard. The efficiencies, reliabilities, and

capabilities of various machine learning algorithms, including Random Forest, Support Vector Machine, Linear Regression, Artificial Neural Network, Gradient Boosting, and Bayesian Ridge Regression, are evaluated and compared. The most recently compiled Turkish strong motion database, which consists of over 950 earthquakes occurring from 1983 to 2023, is used for shaping the models' ability to learn and make accurate predictions. Three feature selection methods—Least Absolute Shrinkage and Selection Operator, Recursive Feature Elimination, and Pearson's Correlation—representing embedded, wrapper, and filter approaches, respectively, are applied to determine the most suitable estimator parameters to predict PGV. Residual analyses and statistical evaluation metrics are employed to measure the performance and effectiveness of the machine learning models. Among the algorithms applied, Gradient Boosting demonstrates exceptional success in predicting PGV, particularly when utilizing all estimator parameters (features) collectively. The Gradient Boosting model exhibits superior predictive capabilities compared to existing ground motion models. It is applicable to shallow crustal strike-slip and normal faulting earthquakes with moment magnitude ranging from 3.5 to 7.8 and Joyner and Boore distance up to 200 km.

Highlights

- A Türkiye-specific machine learning-based ground motion model for peak ground velocity (PGV) is developed using the most updated strong motion database of Türkiye, including data from the 2023 earthquake sequences.
- Six machine learning algorithms are applied to predict PGV, with Gradient Boosting demonstrating a superior ability to represent the strong motion characteristics of earthquakes based on the given features.
- The effectiveness and reliability of feature selection and outlier elimination methods are tested for the first time on Türkiye's strong motion dataset.

F. Kuran · G. Tanırcan (✉)
Earthquake Engineering Department, Kandilli Observatory
and Earthquake Research Institute, Bogazici University,
Istanbul, Türkiye
e-mail: birgore@bogazici.edu.tr

F. Kuran
e-mail: fahrettin.kuran@bogazici.edu.tr

E. Pashaei
Software Engineering Department, Istanbul Gelisim
University, Istanbul, Türkiye
e-mail: epashaei@gelisim.edu.tr

Keywords Ground motion model · Peak ground velocity (PGV) · Feature selection · Outlier elimination · Machine learning · Türkiye

1 Introduction

Among several ground motion intensity measures, peak ground velocity (PGV), which is the maximum absolute amplitude of ground velocity during an instantaneous time, is a significant and frequently used parameter in various areas of earthquake engineering studies (Fang et al. 2020; Karakas et al. 2021; Kostadinov and Towhata 2002; Omine et al. 2008; Orense 2005; Razin et al. 2021; Zhang et al. 2018) mainly due to its direct correlation with shaking and damage intensity (Atkinson and Kaka 2007; Michel et al. 2017). PGV is the essential design parameter of several infrastructure systems, such as tunnels and pipelines. In liquefying ground, PGV is the most accurate ground motion metric for estimating the dynamic performance of the soil-pile system (Zhang et al. 2018). In addition, PGV is one of the intensity measures used as a threshold quantity for various early warning algorithms (Carranza et al. 2013; Colombelli et al. 2015; Picozzi et al. 2015). Overall, the utility of PGV predictions extends far beyond academic interest, playing a vital role in informing emergency response efforts, urban planning decisions, and building code regulations. Accurate predictions of PGV enable stakeholders to proactively mitigate seismic risks, enhance structural resilience, and safeguard human lives and infrastructure in earthquake-prone regions.

For the prediction of PGV, Empirical Ground Motion Prediction Equations (GMPE) such as Akkar and Bommer (2007), Boore and Atkinson (2008), Chiou and Youngs (2014) (hereafter CY14), Boore et al. (2014) (hereafter BSSA14), Bindi et al. (2014), and Zhang et al. (2022) have been built up to using parametric regressions considering a fixed functional form.

BSSA14 (Boore et al. 2014) is one of the most comprehensive Ground Motion Models (GMM) for PGV derived from a subset of the NGA-West2 database consisting of $3.0 \leq \text{moment magnitude } (M_w) \leq 7.9$ earthquakes (Ancheta et al. 2014). Their model is appropriate for $3.0 \leq M_w \leq 8.5$ Strike-slip (SS) and $3.0 \leq M_w \leq 7.0$ Normal (N) earthquakes at sites with shear wave velocity of the top 30 m of the soil (V_{s30}) in the range of 150 m/s and 1500 m/s and distances between 0 and 400 km. M_w , Joyner-Boore distance (R_{JB}), V_{s30} , and style-of-faulting (SoF) are the estimator parameters utilized for PGV prediction.

The CY14 (Chiou and Youngs 2014) PGV ground motion model is an update of their previous equation (Chiou and Youngs 2008) with an enlarged database. The essential estimator variables used in their analyses are M_w , distances (R_{rup} , R_{JB} , R_x), shear wave velocity parameters (including V_{s30} , depth from the ground surface to the 1 km/s shear-wave velocity horizon ($Z_{1.0}$), depth to the top of the ruptured plane (Z_{TOR})), hanging wall parameters, faulting properties, such as dip angle (δ) as well as directivity effect properties. Their model is appropriate for $3.5 \leq M_w \leq 8.5$ SS, $3.5 \leq M_w \leq 8.0$ reverse (R) and N earthquakes in active tectonic regions at sites with V_{s30} in the range of 180 m/s and 1500 m/s and rupture distance (R_{rup}) between 0 and 300 km.

On one hand, conventional regression equations, which are easily applied with just a few parameters and have a high overall prediction success, are often used in various engineering applications. On the other hand, it is difficult to fully identify the complex and nonlinear features of the data in conventional GMPEs produced via parametric regression due to the presumption of a fixed functional form. It is particularly well known that, at near-fault regions, large PGVs formed by pulse-type low-frequency motions may require higher seismic demands from structures than those specified in design codes. Hence, the prediction of PGV is of utmost importance for the seismic design of long-period structures. However, PGVs in the near-field have yet to be predicted precisely due to unexpected properties of the source rupture process. For instance, the observed geometric mean of horizontal component PGVs at a 10–20 km distance from the fault trace is almost four times larger than the median predictions for the 2023 ($M_w 7.8$) Kahramanmaraş Earthquake of Türkiye.

Machine learning (ML) algorithms that have taken their place in a variety of earthquake engineering studies (Bhatta and Dang 2023; Demir and Sahin 2023; Farahani and Barari 2023; Hussaini et al. 2024; Impraïmakis 2023; Jeddi et al. 2022; Kim et al. 2023; Meenakshi et al. 2023; Mousavi and Beroza 2020; Thaler et al. 2021) enable us to get beyond these restrictions. ML-based models can predict PGVs more reliably and precisely because of their ability to handle nonlinear relationships, adapt to changing situations, potentially provide predictions in real-time, and be efficient. We have listed the ML-based GMMs that have been carried out so far in a clear

and understandable chronological order in our previous study (Kuran et al. 2023a). In addition to this literature review, seven ML-based GMMs for PGV are available (Karimzadeh et al. 2023a, 2023b; Kuran et al. 2023b; Mohammadi et al. 2023a, 2023b; Sedaghati and Pezeshk 2023a, 2023b).

Karimzadeh et al. (2023a) (hereafter, KMH-CAL23) studied artificial neural network (ANN) algorithms with 7359 simulated recordings for the prediction of peak ground acceleration (PGA), PGV, and pseudo-spectral accelerations (PSAs). The focal depth (FD), SoF, M_w (ranging from 5.0 to 7.5), R_{JB} (ranging from 0 to 272 km), and V_{s30} are used in their GMM for Türkiye. Karimzadeh et al. (2023b) also used the eXtreme Gradient Boosting (XGBoost) algorithm to obtain a local GMM for PGA, PGV, and PSAs. They utilized 247,710 simulated recordings with M_w (ranging from 5.0 to 6.8), R_{JB} (ranging from 0 to 150 km), and FD (ranging from 5 to 17 km) as estimator parameters for Portugal. Their XGBoost model indicated a persuasive performance; however, the attenuation rate of the model decreased rapidly. Mohammadi et al. (2023a) studied ANN and XGBoost algorithms to develop Türkiye-based GMMs for PGA, PGV, and PSAs. They utilized 1166 real recordings of 383 earthquakes from the Disaster and Emergency Management Presidency of Türkiye (AFAD) and the RESORCE database to Türkiye-specific GMMs. Their estimator parameters are M_w , R_{JB} , V_{s30} , and SoF. Their findings show that the XGBoost algorithm produces an unbiased model across the M_w , V_{s30} , and R_{JB} parameters with a more significant correlation and a lower residual than ANN.

Considering the overall advantages of ML-based prediction algorithms, we attempted to build an ML-based Türkiye-specific GMM for PGV using real earthquake recordings. While ML-based GMM studies are rapidly increasing in the literature, a detailed investigation of the dataset preprocessing steps and additional techniques to enhance prediction accuracy for PGV has yet to be conducted. In addition to creating a Türkiye-specific GMM, we explore the potential applications of the dataset, address the consideration of outliers in model development, and assess the effectiveness of feature selection methods. The objective of this paper is to develop a regional ML-based PGV prediction algorithm by meticulously selecting estimator parameters, ML algorithms, and datasets, with the aim of advancing the field and paving the

way for future studies. The training and evaluation processes involved residual analysis with a likelihood function. Six well-known ML techniques—Random Forest (RF), Support Vector Machine (SVM), Gradient Boosting (GB), Bayesian Ridge Regression (BRR), Linear Regression (LR), and Artificial Neural Network (ANN)—were adapted to predict the geomean PGV of two horizontal components. The primary components of the assessment include the acceptability of the models, reliability, precision in the output, and validity for future data.

2 Strong ground motion database

Two strong ground motion databases—The New Turkish Strong Motion Database (N-TSMD) (Tanırcan et al. 2022) and the Strong-Motion Database of Türkiye (SMD-TR) (Sandıkkaya et al. 2023)—are combined and utilized for both training and testing processes in this study.

In the N-TSMD, the acceleration time histories belong to 904 strong motion stations operated by the Disaster and Emergency Management Presidency of Türkiye (AFAD) (<https://tadas.afad.gov.tr/>). A total of 743 earthquakes in the time span between 1983 and 2020 constitutes the preliminary version of N-TSMD (Fig. 1). The N-TSMD consists of 15,110 recordings of $3.5 \leq M_w \leq 7.6$ earthquakes and hypocentral depths ≤ 30 km if at least five recordings exist within rupture distances 0 to 200 km. The earthquake's missing location, depth, M_w , and fault information were verified and set up from earthquake catalogues maintained by several national and international authorities (Tanırcan et al. 2022). Among the recordings in the N-TSMD, the stations with measured V_{s30} information were used. Forty-six percent of the stations have a V_{s30} range between 306 m/s and 760 m/s and thirty-eight percent have a V_{s30} range between 180 m/s and 360 m/s (Fig. 2b). Recordings were visually checked for any non-standard errors (Douglas 2003). A great majority of the N-TSMD is composed of Strike-slip (SS) (approx. 52%) and Normal (N) (approx. 45%) earthquakes (Fig. 2a). In the N-TSMD, the geometric mean of two horizontal components of PGV varies between 0.001 cm/s and 74.532 cm/s. Also, the distance parameters, including R_{JB} , R_{rup} , and epicentral distance (R_{epi}), are limited to 200 km (Fig. 3 and Table 1). The number of $M_w \geq 7$ events in

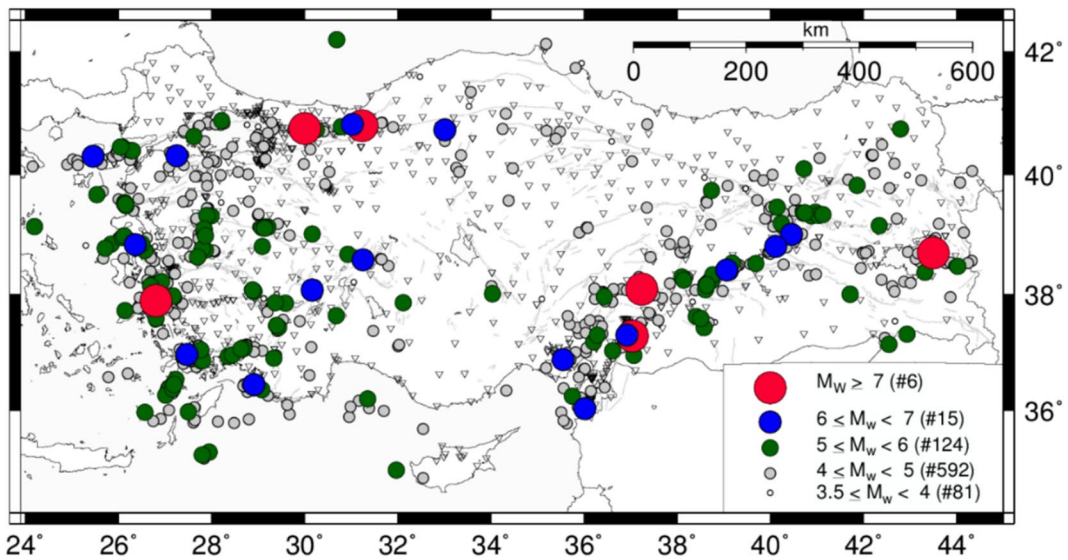


Fig. 1 Epicentral distribution of earthquakes in the expanded N-TSMD (Grey lines are the fault traces (Emre et al. 2018))

Fig. 2 Style-of-faulting distribution of earthquakes (a) and site class of recordings according to NEHRP site classification (b) in the N-TSMD and SMD-TR (A: $V_{s30} > 1500$ m/s, B: $760 \text{ m/s} < V_{s30} \leq 1500$ m/s, C: $360 \text{ m/s} < V_{s30} \leq 760$ m/s, D: $180 \text{ m/s} < V_{s30} \leq 360$ m/s, and E: $V_{s30} < 180$ m/s)

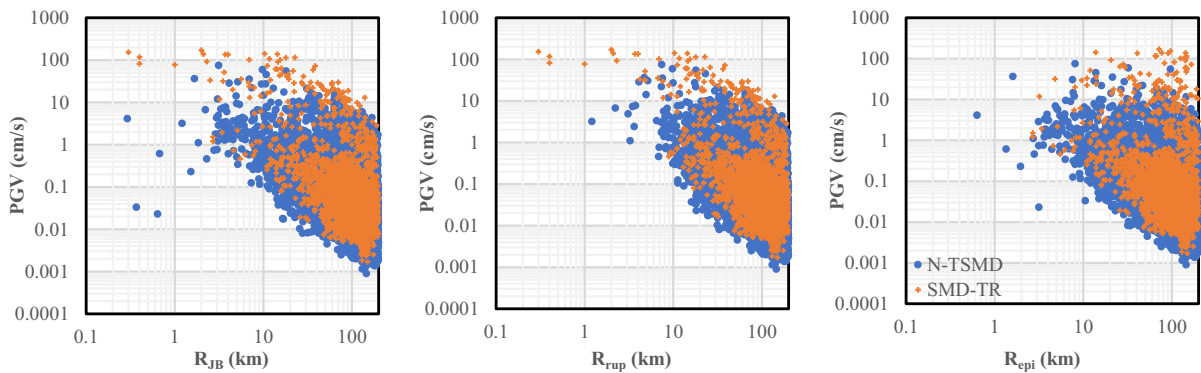
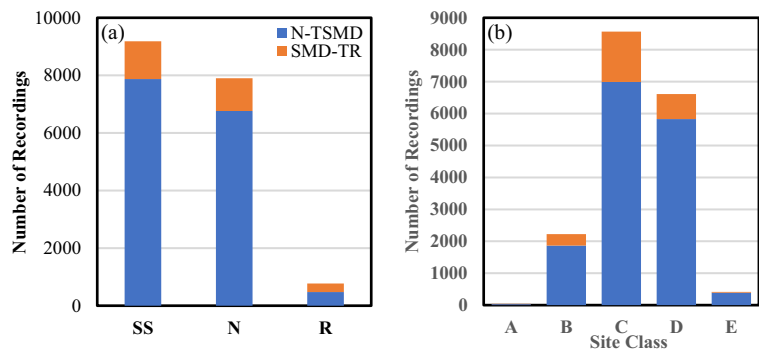


Fig. 3 Distribution of PGVs vs. distance parameters; R_{JB} , R_{rup} , and R_{epi}

Table 1 Minimum and maximum values of parameters in the N-TSMD (* the expanded N-TSMD with the SMD-TR)

Index	N-TSMD & SMD-TR		
	Min	Max	Max*
PGV (cm/s)	0.001	74.5	171.3
M_w	3.5	7.6	7.8
R_{JB} (km)	0.3	199.4	199.4
R_{rup} (km)	1.21	200	200
R_{epi} (km)	0.6	200	200
δ (°)	10.6	90	90
V_{s30} (m/s)	131	1862	1862

the database is only four. They constitute 0.71% of the total database (23 recordings from the 1999 Düzce (M_w 7.1), two recordings from the 2011 Van (M_w 7.1), 60 recordings from the 2020 Samos (M_w 7.0), and 22 recordings from the 1999 (M_w 7.6) Kocaeli earthquakes).

The 2023 Kahramanmaraş doublet earthquakes (M_w 7.8 and M_w 7.7) and subsequently hundreds of aftershocks with magnitudes as large as M_w 6.8 occurred at southeastern Türkiye. Preparation of the PGV GMM without these valuable data would narrow the applicability intervals in terms of magnitude and distance measures. Hence, the N-TSMD is complemented by a newly released updated strong-motion database of Türkiye (SMD-TR) (Sandikkaya et al. 2023) to further train and test the best-performing ML algorithm with the N-TSMD. The SMD-TR consists of 95,890 recordings of 9244 earthquakes with $2.0 \leq M_w \leq 7.8$ that occurred between 1976 and

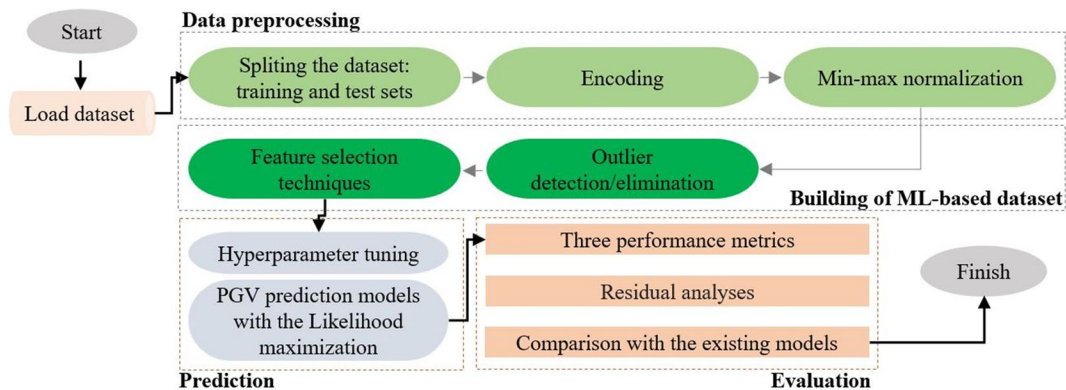
2023. A total of 148 earthquakes with $3.7 \leq M_w \leq 7.8$ from 2020 to February 2023 are used from the SMD-TR. It corresponds to 2746 recordings with $0.3 \text{ km} \leq R_{JB} \leq 199 \text{ km}$ and $131 \text{ m/s} \leq V_{s30} \leq 1822 \text{ m/s}$ (Figs. 2–3). This inclusion enables us to enrich the number of recordings, especially for $M_w \geq 5.5$ earthquakes with 231 recordings including the 2023 Kahramanmaraş doublet events. Finally, the expanded N-TSMD includes 17,856 recordings of 818 earthquakes ($3.5 \leq M_w \leq 7.8$), R_{JB} up to 200 km, and V_{s30} ranging from 131 m/s and 1862 m/s. For the development of Türkiye-specific GMMs for PGV, the expanded N-TSMD is utilized.

3 Methodology

This section presents the state-of-the-art methods and techniques used to derive an applicable non-parametric Türkiye-specific GMM for predicting PGV. Figure 4 provides an overview of the entire process. The following subsections explain the methodology, with results discussed in Section 4.

Six different ML algorithms (described in Section 3.1) are trained with each dataset variation, considering seven parameters as estimators, and the results are assessed using three different performance metrics (Section 3.2). First, data pre-processing techniques (Section 3.3) are applied to the expanded N-TSMD. The effects of the outlier elimination procedure are then investigated.

Next, feature selection methods (Section 3.4) are applied to the original dataset to determine the most

**Fig. 4** Flowchart followed in this study

effective set of estimator parameters and to evaluate the impact of feature selection methods on the earthquake dataset. Hyperparameter tuning (Section 3.5) is performed on six different ML algorithms using the most effective dataset determined in the previous section. PGV prediction models are then reconstructed using different input sets, with the best input set is determined based on performance metrics. Up to this point, three performance metrics were utilized in the assessments.

In the final phase, PGV prediction models are created using six different ML algorithms, adopting the mixed effect model (Section 3.6), and the results are examined in comparative detail. Additionally, the proposed model is compared with existing models in the literature, as presented in Section 5.

3.1 Machine learning algorithms

Machine learning, a subset of artificial intelligence, has gained widespread popularity and frequent use in recent years. Its adaptation across various disciplines is attributed to its robust learning capabilities, versatility, and efficiency. In this study, ground motion models for PGV were developed using supervised ML algorithms, including ANN, BRR, GB, LR, RF, and SVM. The predictive performance of these models was systematically examined and compared to identify the most suitable Türkiye-specific GMM for PGV.

Linear Regression establishes a relationship between the independent variables (M_w , V_{S30} , δ , R_{JB} , R_{rup} , R_{epi} , and SoF) and dependent variable (PGV) by fitting the best straight line (Han et al. 2001). The logic of the linear regression algorithm compatible with this study can be explained clearly in Eq. 1.

$$\ln(PGV) = f_{int} + f_m M_w + f_v V_{S30} + f_\delta \delta + f_{JB} R_{JB} + f_{rup} R_{rup} + f_{epi} R_{epi} + f_{sof} SoF \quad (1)$$

where M_w , V_{S30} , δ , R_{JB} , R_{rup} , R_{epi} , and SoF are the independent variables, $\ln(PGV)$ is the dependent variable, f_{int} is the intercept, and f_m , f_v , f_δ , f_{JB} , f_{rup} , f_{epi} , and f_{sof} are the slopes.

Artificial Neural Network stands out as a highly effective supervised learning techniques, particularly in the context of feed-forward neural networks. In this respect, a Multilayer Perceptron Neural Network (MLP) is a well-known type of ANN widely utilized

in prediction studies related to realistic regression problems. The structure of MLP consists of input, hidden, and output layers. Each layer contains a set number of neurons, all of which are connected to the neurons in the neighboring layer through fully-weighted connections (Tunay et al. 2022).

Bayesian Ridge Regression proves highly beneficial in scenarios where there is insufficient data or the data is poorly distributed. The output of the Bayesian regression technique is derived from a probability density function, indicating that BRR provides probabilistic models for regression problems. Our focus lies in evaluating the effectiveness of BRR algorithms in predicting PGV.

Random Forest stands out as one of the adaptable and simple-to-operate supervised ML algorithms. It leverages multiple decision trees to construct a more reliable and accurate "forest" composed of its individual components. In the initial step, RF selects random recordings from the dataset. Subsequently, it generates decision trees for each recording and gets initial PGVs. Each decision tree independently contributes its predicted PGVs through voting. Finally, RF aggregates the voting results from each decision tree, calculating the average of the PGVs (Pashaei et al. 2017).

Gradient Boosting stands out as a fundamental and powerful ensemble algorithm, renowned for its exceptional predictive capabilities. Similar to other boosting techniques, GB constructs the model step-by-step and extends it by permitting optimization of any differentiable loss function (Fafalios et al. 2020). Leo Breiman observed that boosting can be thought of as an optimization algorithm on an appropriate cost function, which brought about the structure of gradient boosting (Breiman 1997). Jerome H Friedman (Friedman 2001, 2002), proceeded on to create techniques for explicit regression gradient boosting.

Support Vector Machine presents a distinctive method for achieving balanced predictive performance in classification tasks. Based on statistical learning theory, SVM provides a mathematically grounded approach to challenges. A portion of the training estimator parameters is used by SVM to create its solution (Awad and Khanna 2015). Support Vector Regression (SVR) is an SVM version for function approximation and regression with continuous values instead of classification tasks. Hence, the SVR

is used and evaluated to obtain ML-based GMMs for the prediction of PGV.

3.2 Statistical evaluation metrics

Statistical evaluation metrics serve as tools for measuring error and/or accuracy and are commonly used in the fields of ML-based studies to assess and compare the effectiveness of the models. For accurate comparisons, this study employs three frequently used evaluation metrics: mean absolute percentage error (MAPE), root mean square error (RMSE), and Pearson correlation coefficient (R). The details of the equations of these performance metrics can be found in (Kuran et al. 2023a).

3.3 Data preprocessing

Three well-known data processing techniques, including encoding, min–max normalization, and outlier detection/elimination, are adopted in this study.

The training and testing split of the expanded N-TSMD, consisting of 818 earthquakes, is performed based on events. First, five magnitude bins (3.5, 4.5, 5.5, 6.5, 7.5, 7.8) are created. Then, 80% of the recordings in each magnitude bin are reserved for training, while the remaining 20% are set aside for testing (see Fig. 5). The recordings used in the testing process are distinct from those used in the training process, ensuring that the test data remains unseen during training. Figure 5 illustrates the distribution of the number of recordings in the training and test datasets, categorized by magnitude bins.

Style-of-faulting (SoF), a significant estimator parameter representing the fault mechanism of earthquakes, is classified as Reverse (R), Normal (N), and

Strike-slip (SS). To use the SoF categorical variable as an estimator parameter in the ML-based GMM, it needs to be transformed into a numerical or binary vector. One-hot encoding is chosen for the SoF estimator parameter in the expanded N-TSMD. This approach converts each category in the SoF variable into a binary value, allowing the machine learning model to process it effectively.

The parameters utilized in constructing ML-based models may exhibit varying ranges and units. These ranges can differ significantly, as illustrated by V_{s30} and δ in Table 1. If such estimator parameters are incorporated into the model development process without adjustment, the V_{s30} value may disproportionately influence PGV prediction compared to δ . This imbalance in parameter contribution could compromise the model's accuracy and interpretability. To address this issue, normalization techniques are frequently employed to bring all parameters to a consistent scale. One widely utilized method is min–max normalization (Han et al. 2001; Wang et al. 2020), which rescales each parameter to a predefined range, typically between 0 and 1. This method is applied to estimator parameters in the expanded N-TSMD for data transformation. The process adjusts the minimum and maximum values of estimator parameters and converts all other values to a decimal between 0.2 and 0.8.

The Local Outlier Factor (LOF) (Breunig et al. 2000) is a high-precision and widely used unsupervised ML technique for identifying density-based local outliers. This method has applications across various domains (Cheng et al. 2024; Qi et al. 2024; Zou et al. 2023). To identify outliers, the LOF evaluates the local density of each parameter in relation to its neighbors. Breunig et al. (2000) indicated that the LOF can detect meaningful outliers in real-world

Fig. 5 Distribution of the number of recordings in the training and test dataset considering M_w bins

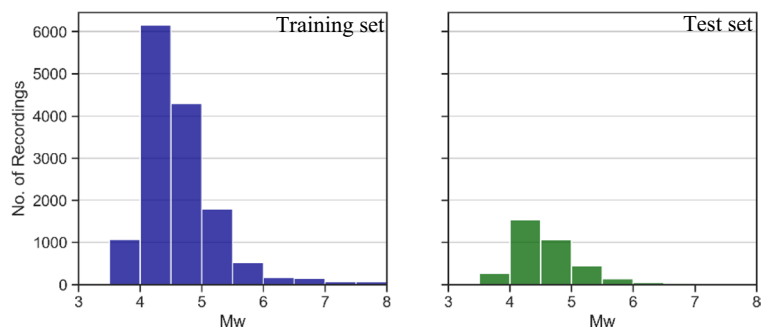


Table 2 Hyperparameter tuning of six ML algorithms

Algorithms	Hyperparameter	Definition	Values
BRR	lambda_1	Shape parameter	1e-4, 1e-6, 1e-12
	lambda_2	Inverse scale parameter	1e-4, 1e-6, 1e-12
	alpha_1	Shape parameter	1e-4, 1e-6, 1e-12
	alpha_2	Inverse scale parameter	1e-4, 1e-6, 1e-12
	lambda_init	Initial value for lambda	None, 1e-1, 1e-5, 1e-9
	alpha_init	Initial value for alpha	None, 1, 1.5, 1.9
	tol	Stop the algorithm if w has converged	1e-3, 1e-4, 1e-6
	n_iter	Maximum number of iterations	50, 100, 300
SVM	kernel	Kernel type	linear, RBF
	C	Penalty parameter	0.1, 1, 10
	nu	An upper bound on the fraction of training errors	0.1, 0.5, 0.9
	gamma	Kernel coefficient	0.1, 0.3, 0.4, 0.6, 0.9
GB	learning_rate	Learning rate	0.01, 0.1, 1.0
	n_estimators	Number of boosting stages	50, 100, 200, 500, 2000
	max_leaf_nodes	the maximum number of leaf nodes per tree	None, 5, 20, 100
RF	max_depth	Maximum depth of the individual regression estimators	3, 5, 10, 18
	max_depth	Maximum depth of the tree	None, 80, 200
	max_features	Number of features	1, 3
	n_estimators	Number of trees in the forest	50, 100, 200, 500, 2000
	min_samples_split	Minimum number of samples	1, 2, 3, 8
ANN	bootstrap	Bootstrap samples	True, False
	learning_rate	Learning rate	constant, adaptive
	hidden_layer_sizes	Number of neurons	(50,50,50), (50,100,50), (100,)
	solver	Solver for weight optimization	sgd, adam
	alpha	Strength of the L2 regularization term	0.0001, 0.05
	activation	Activation function for the hidden layer	relu, tanh, logistic

datasets that cannot be identified using existing approaches. Therefore, to determine whether the outlier removal method is suitable for PGV prediction, the LOF is adapted to the estimator parameters in the expanded N-TSMD, which consists of recordings of real earthquakes.

3.4 Feature selection

Feature selection methods in the machine learning domain are crucial due to the presence of many irrelevant, redundant, and noisy features in the data. These factors can negatively impact predictive accuracy and increase computational burden (Pashaei and Pashaei 2022). However, real earthquake recordings are not homogeneous, as moderate-to-large magnitude earthquakes occur less frequently than small earthquakes, leading to differences in feature representation.

Therefore, it is also important to understand if feature selection is a feasible application for the prediction of GMIMs. The Least Absolute Shrinkage and Selection Operator (LASSO) (Tibshirani 1996), Recursive Feature Elimination (RFE) (Guyon et al. 2002), and Pearson's Correlation were chosen and applied to the dataset due to their widespread popularity and ease of use.

As a filter method, Pearson correlation evaluates the linear relationship between each feature (in this case, seven estimator parameters) and the target variable (PGV) by calculating the Pearson correlation coefficient. Features with high absolute correlation coefficients (close to 1 or -1) are considered highly correlated with the target variable and are likely to contain valuable predictive information, while features with low absolute correlation coefficients (close to 0) are deemed weakly correlated and may be less

informative for prediction. This method aids in selecting or ranking features based on their correlation with the target variable, facilitating the identification of relevant features for predictive modeling in machine learning tasks.

RFE, a wrapper feature selection method, plays a crucial role in identifying critical features within a dataset by iteratively constructing models. It selectively retains features that contribute the most to predictive performance while discarding less informative or redundant ones. Typically, RFE requires specifying the desired number of features to retain, a decision influenced by various factors such as algorithm constraints, cross-validation, performance metrics, domain knowledge, and techniques like the elbow method. In this study, we evaluated model performance across different feature counts to pinpoint the optimal number that maximizes performance while mitigating overfitting risks. The top-ranked features identified by RFE are referred to as the selected features.

LASSO, an embedded feature selection method, is instrumental in identifying key features within a dataset by imposing a penalty on the absolute size of regression coefficients. This technique encourages sparsity in the model, effectively shrinking the coefficients of less important features to zero and

retaining only the most relevant ones. In the context of our study, LASSO aids in selecting the most predictive features for modeling PGV. By leveraging the regularization parameter, LASSO effectively balances model complexity and predictive accuracy. The features selected by LASSO are those with non-zero coefficients, highlighting their importance in predicting PGV.

3.5 Hyperparameter tuning

Hyperparameters, set before training ML algorithms, significantly influence the performance of ML-based GMMs. Hyperparameter tuning, the process of selecting appropriate hyperparameter values, is essential for optimizing model performance. In this study, we carefully consider domain knowledge, previous research, computational resources, and tuning methodology when defining the search space for ML hyperparameters. Our objective is to establish a range of hyperparameter values that balance computational efficiency and model performance. To achieve this, we employ the Grid-search method to obtain the highest accuracy from five algorithms used in developing Türkiye-specific GMMs, as detailed in Table 2, while consulting

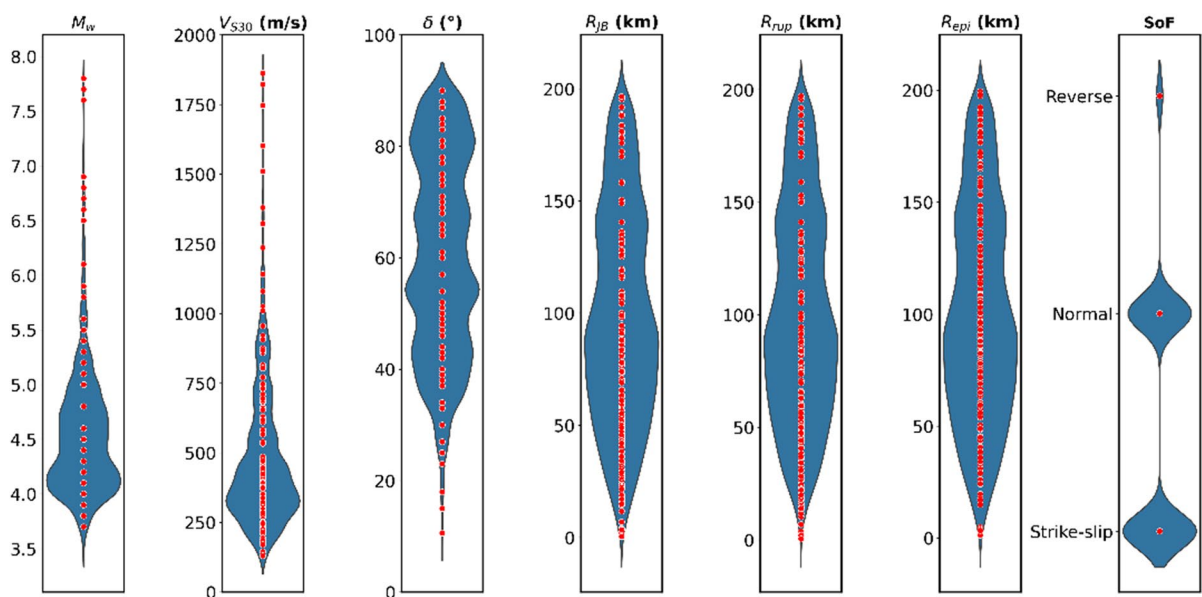


Fig. 6 Outlier detection in the expanded N-TSMD (Red dots represent outliers)

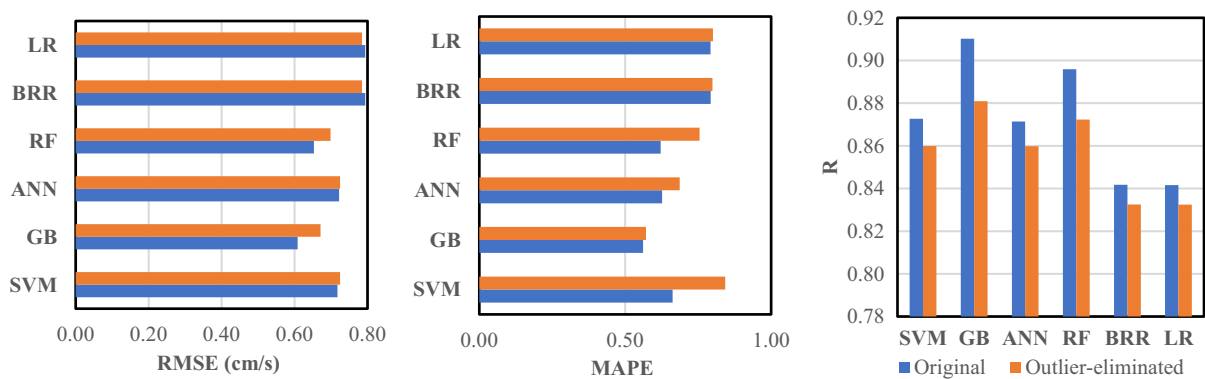


Fig. 7 Results of ML-based PGV models obtained with the original and outlier-eliminated expanded N-TSMD according to performance metrics

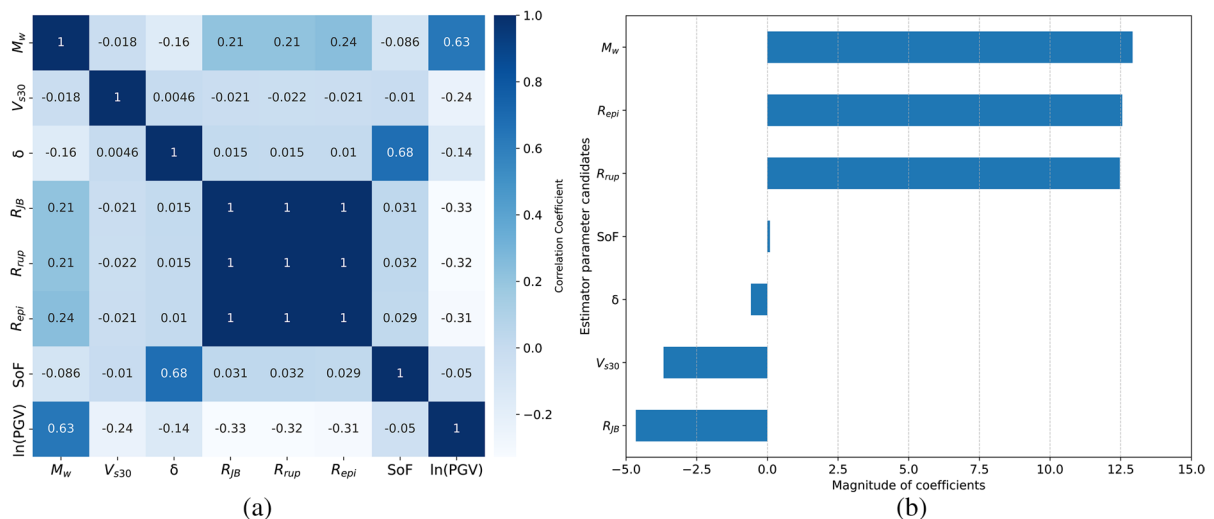


Fig. 8 Results of Pearson's correlation (a) and LASSO (b) methods on feature selection process

each algorithm's documentation for guidance on hyperparameter selection and their value ranges.

3.6 Mixed effect model (log-likelihood function)

Conventional ground motion models are often expressed in logarithmic functional forms. This form includes a prediction equation $f(X, \theta)$, inter-event variations (η_i), and intra-event variations (ε_{ij}), as shown in Eq. 2 (Abrahamson and Youngs 1992):

$$\ln(PGV) = f(X, \theta) + \eta_i + \varepsilon_{ij} \quad (2)$$

Mohammadi et al. (2023a) enhanced this approach by incorporating an algebraic maximum likelihood function to calculate model parameters and variances. Their modified mixed-effect model is adopted in this study, following Mohammadi et al. (2023a)'s modified likelihood function and iterative process outlined in their paper. In the iterative process, the genetic algorithm (GA) (Golberg 1989) is used to maximize the log-likelihood function. The total standard deviations of ML-based GMMs are calculated using Eq. 3:

$$\Phi = \sqrt{\sigma^2 + \tau^2} \quad (3)$$

Table 3 Feature sets for the prediction of PGV

Feature Sets	M_w	V_{s30}	δ	R_{JB}	R_{rup}	R_{epi}	SoF
All	✓	✓	✓	✓	✓	✓	✓
LASSO (Embedded)	✓	✓	✓	✓	✓	✓	✓
Pearson (Filter)	✓	✓	✓	✓	✓	✓	
RFE (Wrapper)	✓	✓		✓			
User-selected	✓	✓		✓			✓

where Φ , σ , and τ represent the standard deviations of total, inter-event (η_i), and intra-event (ε_{ij}) residuals, respectively.

4 Result and discussion

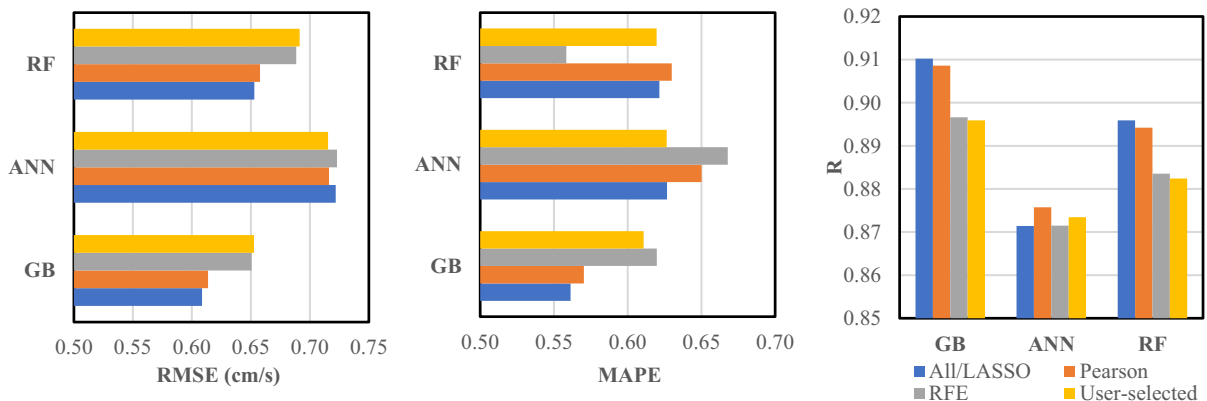
This section presents the results of developing the best ML-based Türkiye-specific ground motion model for PGV using the expanded N-TSMD. As briefly mentioned in Section 3, our key question during the dataset evaluation process is whether outliers should be eliminated from the dataset. This question is addressed in Section 4.1. The estimator parameter sets determined by feature selection methods and their comparisons are presented and evaluated in Section 4.2. The performance of Türkiye-specific GMMs developed using six ML algorithms and the likelihood function is reported in Section 4.3.

4.1 Effect of outlier elimination

The expanded N-TSMD consists of 17,856 recordings of earthquakes with magnitudes ranging from

$3.5 \leq M_w \leq 7.8$. Using the LOF method, 179 recordings were detected as outliers and subsequently removed (Fig. 6). GMMs for PGV prediction were created with six different ML algorithms (SVM, LR, GB, ANN, BRR, and RF) using all estimator parameters (M_w , R_{JB} , R_{rup} , R_{epi} , δ , SoF, and V_{s30}) to evaluate the effect of outlier elimination. Although the source and faulting parameters of the recordings are similar for the same event, the path and site parameters differ for each recording. The predictive ability of the models was evaluated using three different performance metrics. The test results of performance metrics form the evaluation criteria to determine the effectiveness and accuracy of the models.

Figure 7 presents the average performance results of the PGV prediction based on statistical metrics. In all ML-based models, the lowest MAPE and the highest R term are observed in the models developed with the original dataset. Additionally, the lowest RMSE for all ML models (except for LR and BRR) is achieved with the original dataset. The model obtained using the GB algorithm and original dataset demonstrates remarkable success, with the minimum RMSE, MAPE, as well as the highest R-value.

**Fig. 9** Results of ML-based PGV models obtained with different estimator sets

The original dataset showed superior success in four of the six algorithms based on performance metrics. Therefore, for the sake of simplicity and a better understanding of the evaluations performed, subsequent analyses were continued using only the original dataset consisting of 17,856 recordings.

4.2 Application and evaluation of feature selection

To assess the effectiveness of feature selection methods in PGV prediction, LASSO, RFE, and Pearson were applied to the expanded N-TSMD. Variables with an absolute weight value of 0.1 or higher in Pearson's correlation were selected as the model's estimator parameters. Pearson's correlation method accepts all variables except for SoF, which was not deemed sufficient to serve as estimator parameters in the dataset (Fig. 8(a)).

Figure 8(b) shows the magnitude of coefficients determined by the LASSO method versus estimator parameter candidates. The length of each bar in

this plot represents the magnitude of the coefficient. Features with longer bars (whether positive or negative) have larger absolute coefficients and are more influential in the development of GMMs. Features with positive coefficients contribute positively to the prediction, while features with negative coefficients contribute inversely to the prediction. Features with coefficients of zero are less important in the model development. LASSO identifies M_w , R_{epi} , and R_{rup} as maximum positive coefficients, while R_{JB} and V_{s30} are picked as estimator parameters with more negative significance. The δ and SoF have the least magnitude of coefficient among the seven variables according to LASSO, although, it still has a negative and positive effects, respectively. Ultimately, LASSO selected all variables as estimator parameters in the dataset (Table 3).

RFE adopts a more stringent approach compared to other methods in this study, selecting only M_w , V_{s30} , and R_{JB} as the most important estimator parameters (Table 3). Both RFE and Pearson methods do not consider SoF as a significant parameter. However,

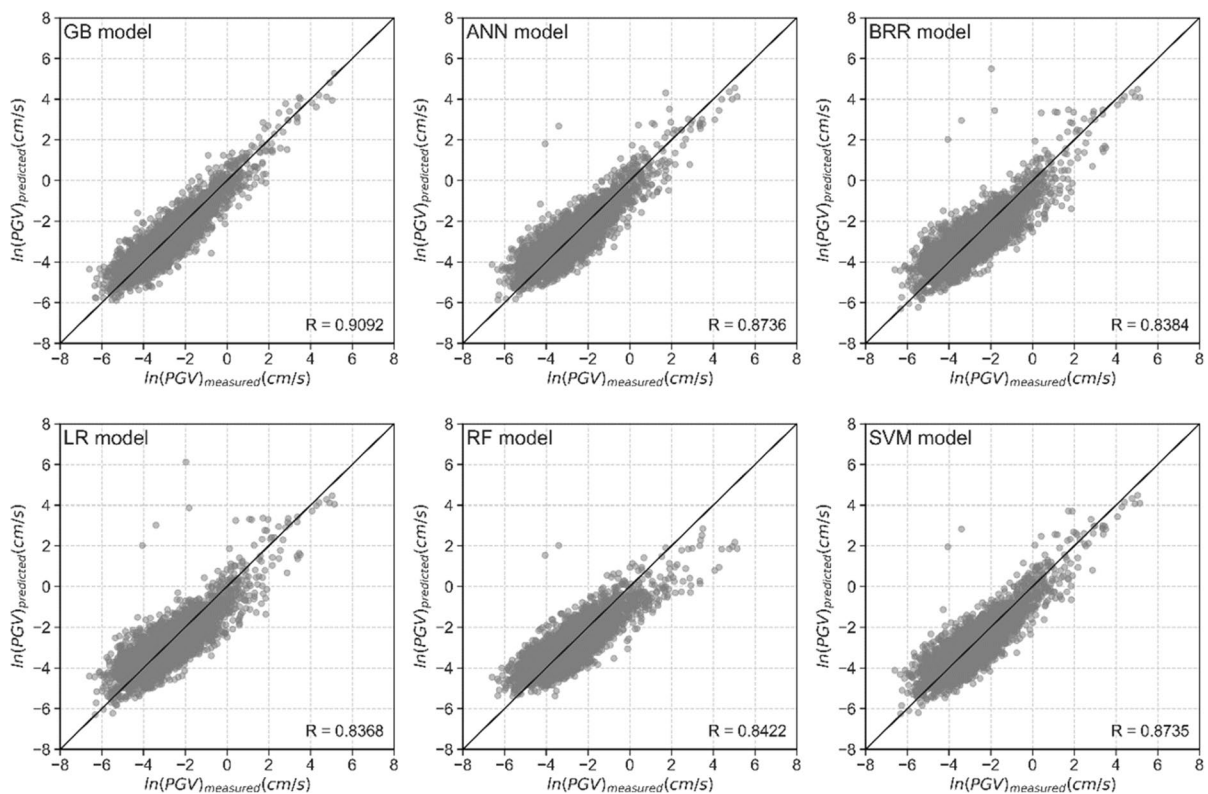


Fig. 10 Comparison of measured $\ln(\text{PGV})$ s versus predicted $\ln(\text{PGV})$ s

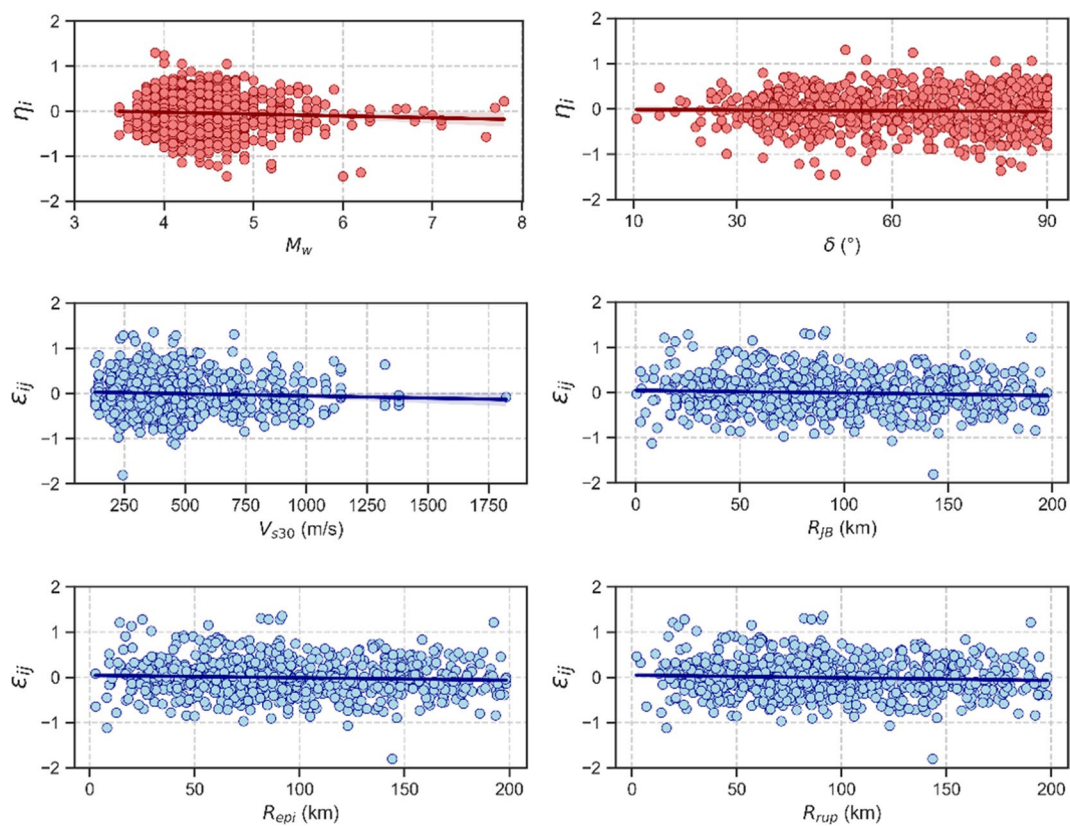


Fig. 11 Inter-event (η_i) and intra-event (ϵ_{ij}) residuals of $\ln(\text{PGV})$ s predicted with the GB model (unit in cm/s)

Table 4 Performance results of ML-based ground motion models

Model	RMSE (cm/s)	MAPE	R	σ	τ	Φ
SVM	0.717	0.648	0.874	0.530	0.444	0.691
GB	0.613	0.546	0.909	0.194	0.373	0.420
ANN	0.716	0.604	0.874	0.528	0.446	0.691
RF	0.798	0.705	0.842	0.214	0.586	0.624
BRR	0.802	0.784	0.838	0.601	0.467	0.761
LR	0.806	0.777	0.837	0.601	0.466	0.760

it is known that a definite correlation exists between SoF and PGV. In light of this, we proposed an additional feature set using the M_w , R_{JB} , SoF, and V_{s30} parameters, as shown in Table 3. This selection process is based on insights gathered from examining pertinent scholarly works and investigating the influence of various seismological parameters on ground motion.

In evaluating the feature selection methods, our aim is to assess models trained with different feature sets presented in Table 3 for PGV prediction. PGV is predicted using three cutting-edge ML algorithms

(RF, ANN, GB), which demonstrate superior performance in Section 4.1. The results are evaluated using the performance metrics and are presented in Fig. 9.

The RF algorithm trained with all/LASSO input set delivers better accuracy in predicting PGV, as indicated by the RMSE and R metrics. The ANN algorithm, when trained with either the user-selected or all/LASSO estimator sets, demonstrates comparable performance across all three performance metrics. However, the GB algorithm trained with all/LASSO input set achieves the highest accuracy in PGV prediction compared to the other models. Overall, models trained with the original

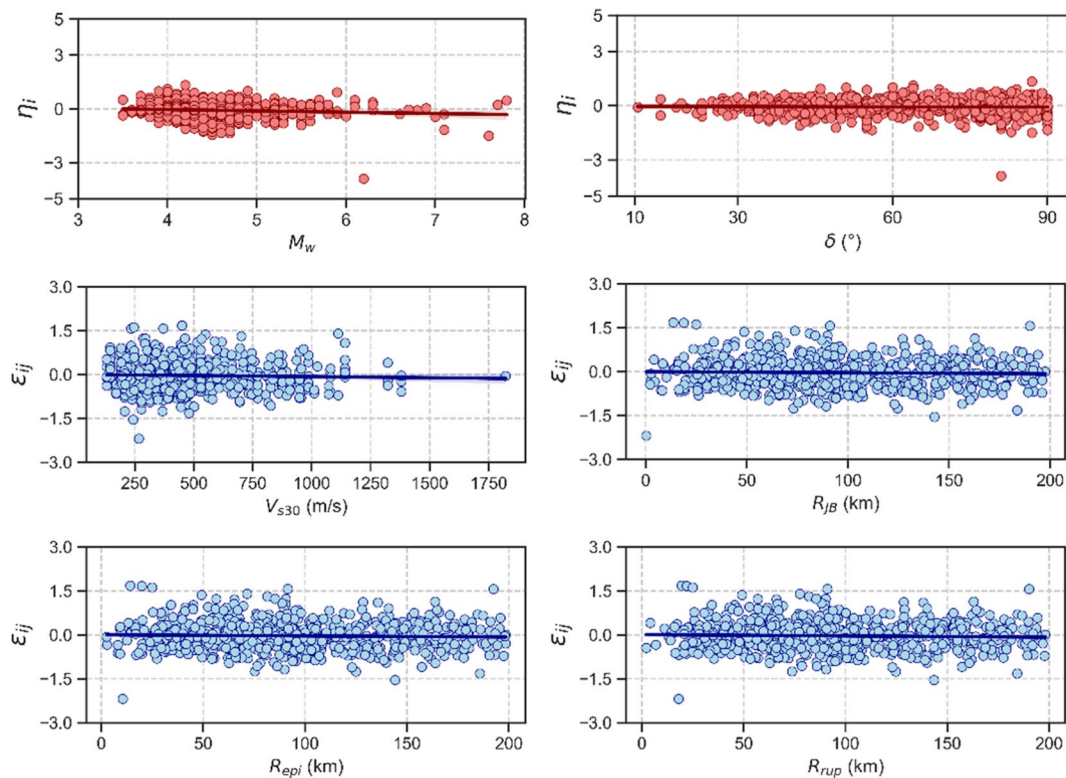


Fig. 12 Inter-event (η_i) and intra-event (ϵ_{ij}) residuals of $\ln(\text{PGV})$ s predicted with the ANN model (unit in cm/s)

dataset using the all/LASSO input set exhibit the best prediction capabilities for PGV.

In our previous study (Kuran et al. 2023a), we developed CAV GMMs using the N-TSMD. We discovered that for CAV prediction, models trained with the GB algorithms required only four estimator parameters (M_w , R_{JB} , SoF, and V_{s30}), whereas seven estimator parameters were necessary for PGV prediction. This highlights that the dataset and estimator parameters used to train algorithms can vary depending on the intensity measures. In particular, ML models might need a larger number of recordings and estimator parameters to establish a relationship for predicting distinct amplitude in time series, while modeling cumulative energy, such as CAV, may be achieved with fewer recordings and features.

4.3 Development of Türkiye-specific GMM for PGV using the mixed effect model

In the final phase of this study, the mixed effect model was incorporated into the training process of ML-based GMMs for PGV. Six different ML algorithms

(ANN, GB, RF, SVM, BRR, and LR) were integrated with the mixed effect model to develop PGV GMMs using the original dataset, which included all estimator parameters (M_w , R_{JB} , R_{rup} , R_{epi} , δ , SoF, and V_{s30}). The performances of these obtained models were assessed and compared using performance metrics and both inter- and intra-event residuals to identify the best algorithms for PGV prediction. Based on these evaluations, the most accurate PGV prediction model for Türkiye was determined.

Figure 10 displays the measured $\ln(\text{PGV})$ values versus $\ln(\text{PGV})$ values predicted by the six different ML-based models. The LR and BRR models exhibit similar trends and demonstrate relatively low performance compared to the other models. In contrast, the prediction results from the GB, ANN, and SVM models are closer to the ideal fit line. Notably, the GB model shows very low dispersion and delivers the best performance results for PGV prediction.

Inter-event and intra-event residuals of $\ln(\text{PGV})$ predicted by the six models are illustrated in Figs. 11–16 and supported by Table 4. As shown in

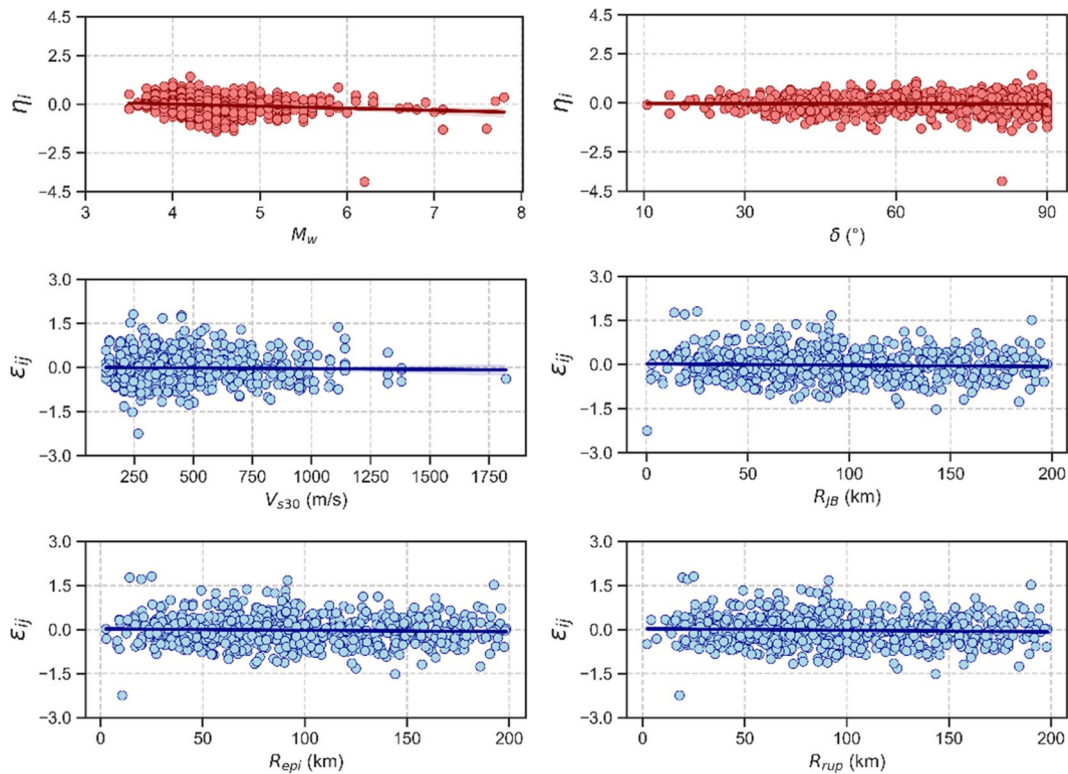


Fig. 13 Inter-event (η_i) and intra-event (ε_{ij}) residuals of $\ln(\text{PGV})$ s predicted with the SVM model (unit in cm/s)

Fig. 11, the inter-event (η_i) and intra-event (ε_{ij}) residuals for the GB model range from -1.5 to 1.5 and -2.0 to 2.0, respectively, which are the lowest ranges among all models. A general trend of slight overestimation is observed as the M_w increases; however, no systematic trends of η_i are detected with δ . Similarly, ε_{ij} distribution almost follows the zero line at all distances. A slight overestimation in the ε_{ij} of stiff/rock sites ($V_{s30} > 760$ m/s) is mainly related to the fact that there are a few recordings on these sites in the dataset.

The η_i and ε_{ij} of ANN and SVM models (Figs. 12 and 13), show a similar shape of distributions with the GB model. However, the overestimation scatter ranges increase beyond $M_w 6$ in both models. The η_i and ε_{ij} ranges are approximately between -4.5 to 2.0 and -2.0 to 2.0, respectively. As observed in Fig. 10, the ANN and SVM models demonstrate nearly same residual results. BRR and LR models' residuals reflect almost similar distributions and ranges as in Figs. 14 and 15. The ε_{ij} and η_i values vary between -4.0 to 3.0 and -4.0 to 2.0, respectively. The prediction harmony between

the BRR and LR models is also evident in Fig. 10. The prediction accuracies of both models remain low compared to the other four models. The RF model shows a clear η_i trend against M_w compared to the other models, especially for $M_w > 5.5$ (Fig. 16). However, no trend is detected in the η_i against δ . The η_i range is between -5.0 to 3.0. The ε_{ij} range of RF model, on the other hand, is similar (between -2.0 to 2.0) to those of GB and SVM models.

Table 4 presents the performance results of six models with standard deviations of residuals (τ , σ , and Φ). The GB model exhibits the minimum errors and the highest R value among the models. Additionally, it has the lowest τ , σ , and Φ values out of six models.

Overall, the residual dispersions for the other models are broader compared to the GB model. According to statistical performance metrics, the GB model stands out as the best for PGV prediction. The standard deviations further indicate that the GB model shows lower variability compared to the other five

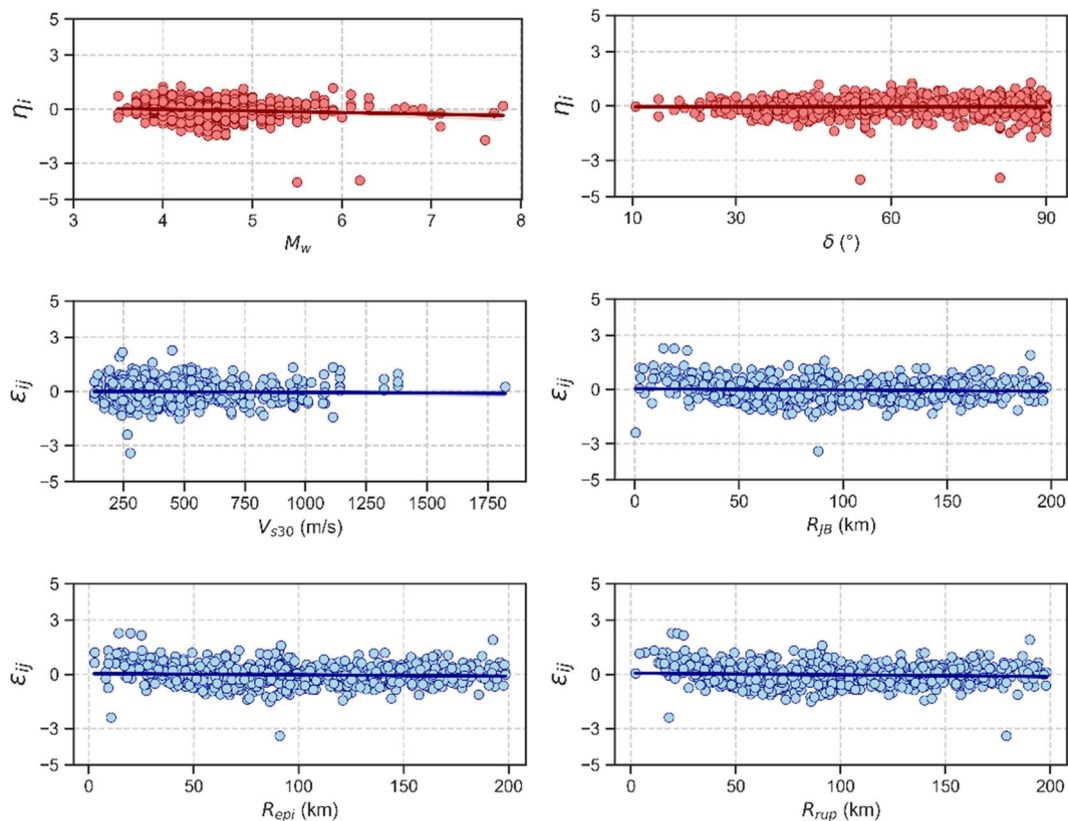


Fig. 14 Inter-event (η_i) and intra-event (ε_{ij}) residuals of $\ln(\text{PGV})$ s predicted with the BRR model (unit in cm/s)

models. Based on these results, the GB model is proposed as the final model in this study.

5 Comparison

In this section, two comprehensive GMMs (CY14 and BSSA14), and one ANN-based GMM (KMH-CAL23) were compared with the proposed GB-based Turkiye-specific model for PGV prediction.

PGV predictions were calculated for six earthquake scenarios with the CY14, BSSA14, KMH-CAL23, and proposed model. Those scenarios are the combination of M_w 5.0, 6.0, 7.0, N, and SS earthquakes predicted at $R_{JB}=1:200$ distances at a site with $V_{s30}=550$ m/s. Estimator parameters used in each GMM for PGV are presented in Table 5.

The Turkiye-specific proposed model, obtained by training seven estimators, is compared with CY14, BSSA14, and KMH-CAL23 models, as seen in Fig. 17. Nearly all PGVs predicted with the proposed

model are within the ± 1 standard deviation boundary of the CY14 and BSSA14 models for M_w 5–6 earthquakes, except at very far-field for M_w 5 earthquakes. Such a fast attenuation of PGV is also observable at the KMH-CAL23 for the M_w 5 SS events in the far-field. It might be attributed to the regional attenuation characteristics. The only difference is seen at the very near field region for M_w 7 SS and N earthquakes. PGV prediction for the first 3 km distance is lower than other models. In the same graph, PGV values (M_w 7 SS and N) predicted by KMH-CAL23 for the first 3 km fall between the $+1$ standard deviations of BSSA14 and CY14. Such a result also depicts a complementary effect of the simulated dataset on the near-field PGV prediction for M_w 7 events.

6 Conclusion

In this work, various ML models were systematically tested to identify the optimal model for

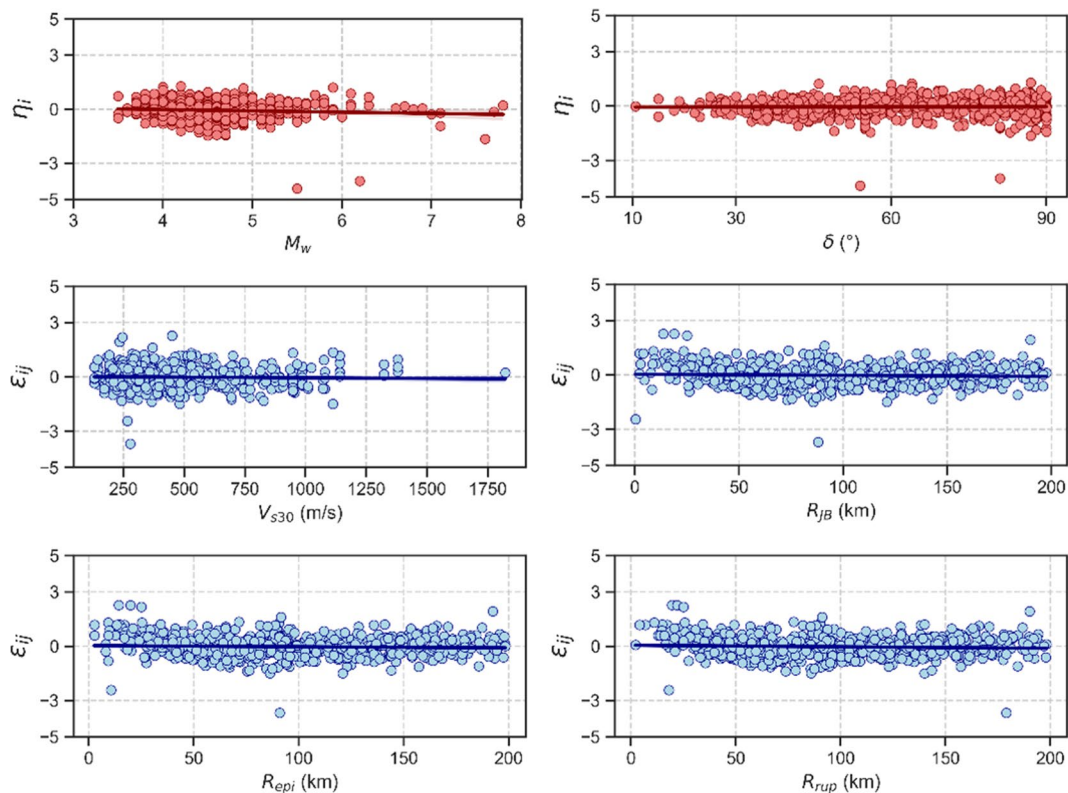


Fig. 15 Inter-event (η_i) and intra-event (ϵ_{ij}) residuals of $\ln(\text{PGV})$ s predicted with the LR model (unit in cm/s)

predicting PGV in Türkiye. Well-known and widely used ML techniques, including SVM, GB, BRR, ANN, LR, and RF, were employed with an extensive dataset comprising 17,856 recordings from 818 real earthquakes that occurred in Türkiye up to 2023. A comparative investigation of the earthquake database and estimator parameters was conducted to identify the most adaptable model for predicting PGVs. During the development of data-driven GMMs, seven estimator parameters— M_w , R_{JB} , R_{rup} , R_{epi} , δ , SoF , and V_{s30} —were evaluated. Three techniques were used to examine the effect of feature selection on the development of reliable prediction models. Five input sets—user-selected, All, LASSO, RFE, and Pearson—were used during the training and test phases of each model to assess the impact of estimator parameters on the model performance. In addition, an outlier elimination

method was applied to the dataset to determine the optimal approach for achieving the most accurate PGVs. MAPE, RMSE, and R evaluation metrics were employed to assess the models. PGV prediction models were then developed with the modified mixed effect model and residual analyses were conducted to validate the effectiveness of the proposed model. The best PGV model developed in this study was compared with the conventional model (BSSA14, CY14) and ML-based model (KMH-CAL23) to evaluate its efficiency. The key contributions of this work are as follows:

The outlier elimination process led to the exclusion of 179 recordings, which, although few, have great importance in the dataset. PGV models were constructed by training each algorithm with both the outlier-eliminated and original datasets, and their performance was tested. The findings indicate that

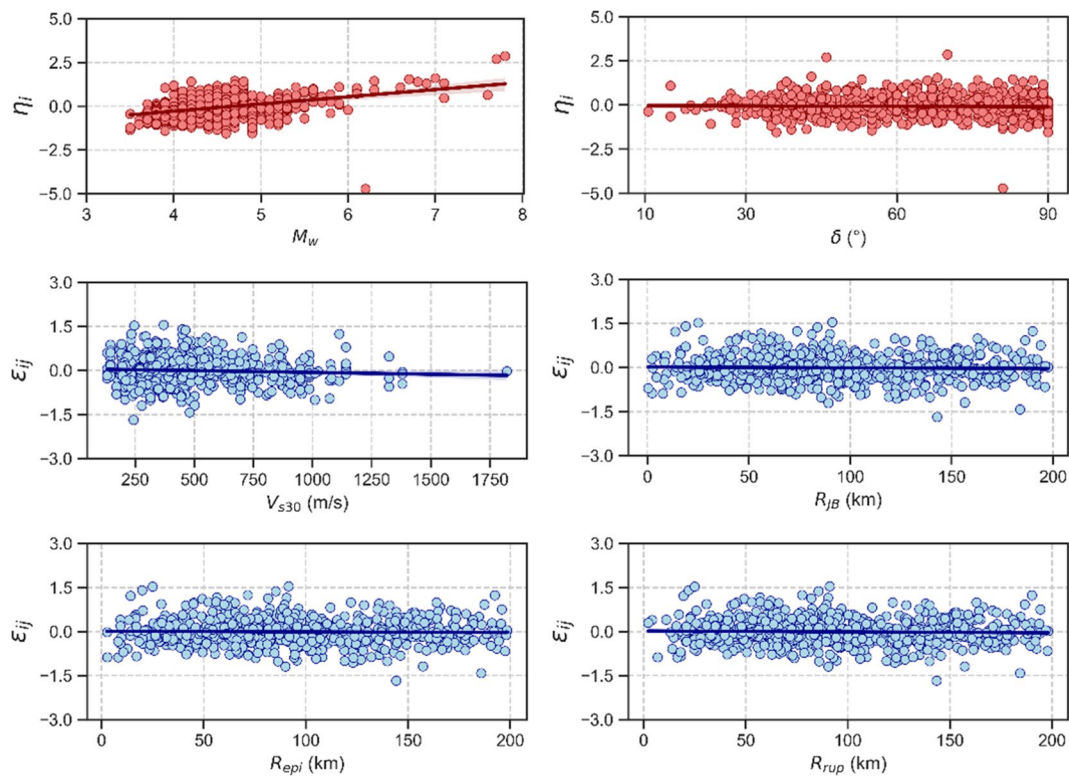


Fig. 16 Inter-event (η_i) and intra-event (ϵ_{ij}) residuals of $\ln(\text{PGV})$ s predicted with the RF model (unit in cm/s)

Table 5 Summary of estimator parameters in GMMs for PGV

Estimators	The proposed model	BSSA14	CY14	KMHCAL23
M_w	✓	✓	✓	✓
V_{s30}	✓	✓	✓	✓
$Z_{1.0}$		✓	✓	
Z_{TOR}			✓	
R_{JB}	✓	✓	✓	✓
R_{rup}	✓		✓	
R_{epi}	✓			
R_x			✓	
Hanging wall (HW)			✓	
SoF	✓	✓		✓
δ	✓		✓	
λ			✓	
FD				✓

the ML algorithms trained with the original dataset, without outlier elimination, achieve more accurate PGV predictions according to statistical performance metrics. Therefore, the outlier-elimination method should not be applied to a dataset consisting of real earthquakes.

-Overall, the algorithms trained with the all/ LASSO input set (M_w , R_{JB} , R_{rup} , R_{epi} , δ , SoF, and V_{s30}) exhibit better PGV prediction skills compared to other models.

-The GB model, trained with the original dataset including all estimator parameters, demonstrates superiority over other ML-based GMMs for predicting PGV. The ANN, SVM, and RF models follow the GB model in PGV prediction success. Conversely, the models developed with BRR and LR algorithms show comparatively lower success in predicting PGV.

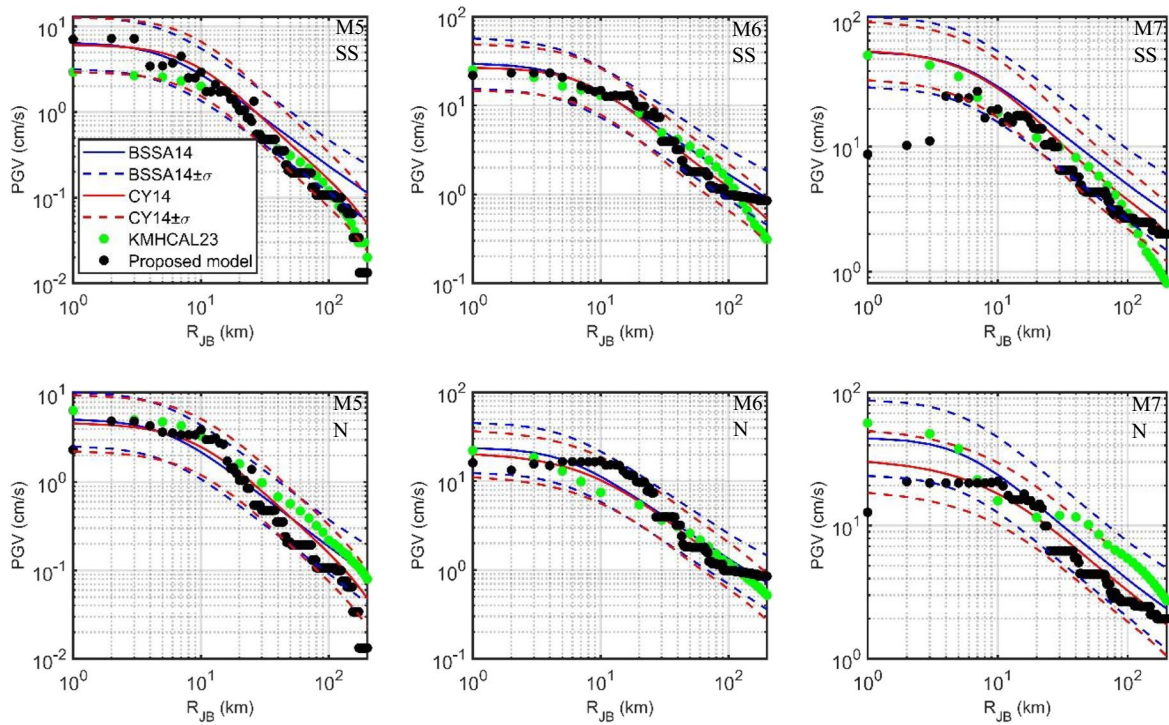


Fig. 17 PGVs predicted with BSSA14, CY14, KMHCAL23, and the proposed model for SoF=SS ($\delta=90^\circ$ and $\lambda=0^\circ$) and N ($\delta=60^\circ$ and $\lambda=-90^\circ$) with $R_{JB}=[1:200]$, $M_w=[5.0, 6.0, 7.0]$, $V_{s30}=[550 \text{ m/s}]$, and $FD=[7 \text{ km}]$

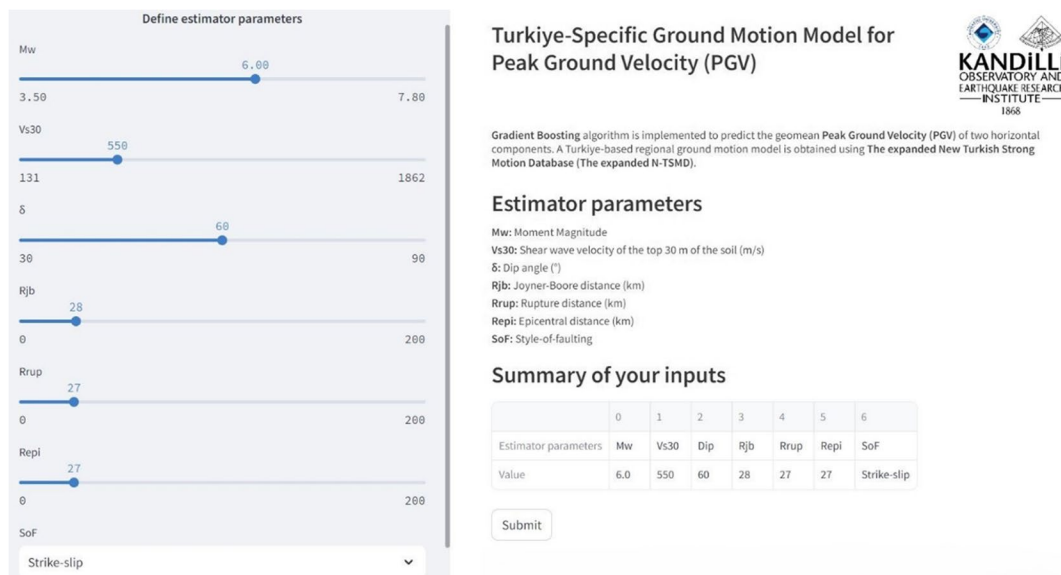


Fig. 18 The software interface of the Türkiye-specific ground motion model for peak ground velocity

-The proposed model, developed using the GB algorithm and a modified mixed effect approach, shows strong prediction skills for PGVs compared to its peer ground motion models. This model is made available to the user on the Internet platform, as shown in the Appendix.

-The success of the GB algorithms has been proven in the developed Türkiye-specific GMM with the N-TSMD, both in this study and our previous study on cumulative absolute velocity (Kuran et al. 2023a).

In future work, GMMs can be further improved for both PGV and CAV predictions by developing advanced boosting algorithms. Various state-of-the-art optimization techniques can be employed to enhance prediction accuracy during the hyper-parameter tuning process. Additionally, alternative approaches for detecting outliers in multivariate space, such as the Mahalanobis distance method, can be adopted (Mohammadi et al. 2023c).

Acknowledgements We would like to extend our gratitude to A. Mohammadi, S. Karimzadeh, and P.B. Lourenço for their guidance and generously sharing their code.

Author contributions F.K.: Conceptualization, methodology, software, writing—original draft preparation, investigation, visualization, developing the website of the model. G.T.: Data curation, supervision, writing—reviewing, editing. E.P.: Methodology, validation, reviewing, editing, developing the website of the model.

Funding No funding was received for conducting this study.

Data availability The data of this study will be available with its paper in 2024.

Declarations

Competing interests The authors declare no competing interests.

Financial interests The authors declare they have no financial interests.

Appendix The development of the website for the application of the proposed model

The best prediction model obtained as a result of this study, the Türkiye-specific GB ground motion model for peak ground velocity, is presented to the

user as seen in Fig. 18. The software tool can be accessed by clicking on the link: <https://ktp24-turkiyespecificgmmforgpv.streamlit.app/>.

The prediction of the PGV with the proposed model can be easily performed by defining the M_w , R_{JB} , R_{rup} , R_{epi} , δ , SoF, and V_{S30} estimator parameters within applicable ranges.

References

- Abrahamson NA, Youngs RR (1992) A stable algorithm for regression Analyses using the random effects model. *Bull Seismol Soc Am* 82(1):505–510. <https://doi.org/10.1785/BSSA0820010505>
- Akkar S, Bommer JJ (2007) Empirical prediction equations for peak ground velocity derived from strong-motion records from Europe and the Middle East. *Bull Seismol Soc Am* 97(2):511–530. <https://doi.org/10.1785/0120060141>
- Ancheta TD, Darragh RB, Stewart JP, Seyhan E, Silva WJ, Chiou BSJ, Wooddell KE, Graves RW, Kottke AR, Boore DM, Kishida T, Donahue JL (2014) NGA-West2 database. *Earthq Spectra* 30(3):989–1005. <https://doi.org/10.1193/070913EQS197M>
- Atkinson GM, Kaka SLI (2007) Relationships between felt intensity and instrumental ground motion in the Central United States and California. *Bull Seismol Soc Am* 97(2):497–510. <https://doi.org/10.1785/0120060154>
- Awad M, Khanna R (2015) Support vector machines for classification. In *Efficient Learning Machines* (pp. 39–66). https://doi.org/10.1007/978-1-4302-5990-9_3
- Bhatta S, Dang J (2023) Seismic damage prediction of RC buildings using machine learning. *Earthquake Eng Struct Dynam* 52(11):3504–3527. <https://doi.org/10.1002/eqe.3907>
- Bindi D, Massa M, Luzi L, Ameri G, Pacor F, Puglia R, Augliera P (2014) 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* 12(1):391–430. <https://doi.org/10.1007/s10518-013-9525-5>
- Boore DM, Atkinson GM (2008) Ground-motion prediction equations for the average horizontal component of PGA, PGV, and 5%-damped PSA at spectral periods between 0.01 s and 10.0 s. *Earthquake Spectra* 24(1):99–138. <https://doi.org/10.1193/1.2830434>
- Boore DM, Stewart JP, Seyhan E, Atkinson GM (2014) NGA-West 2 equations for predicting PGA, PGV, and 5%-damped PSA for shallow crustal earthquakes. *Earthq Spectra* 30(3):1057–1085. <https://doi.org/10.1193/070113EQS184M>
- Breiman L (1997) Arcing the edge. Technical Report 486, Statistics Department, University of California, Berkeley, pp 1–14
- Breunig MM, Kriegl HP, Ng RT, Sander J (2000) LOF: Identifying density-based local outliers. <https://doi.org/10.1145/335191.335388>

- Carranza M, Bufo E, Colombelli S, Zollo A (2013) Earthquake early warning for southern Iberia: A P wave threshold-based approach. *Geophys Res Lett* 40(17):4588–4593. <https://doi.org/10.1002/grl.50903>
- Cheng K, Wang Y, Yang X, Zhang K, Liu F (2024) An intelligent online fault diagnosis system for gas turbine sensors based on unsupervised learning method LOF and KELM. *Sens Actuators, A* 365:114872. <https://doi.org/10.1016/j.sna.2023.114872>
- Chiou BSJ, Youngs RR (2008) An NGA model for the average horizontal component of peak ground motion and response spectra. *Earthq Spectra* 24(1):173–215. <https://doi.org/10.1193/1.2894832>
- Chiou BSJ, Youngs RR (2014) Update of the Chiou and Youngs NGA model for the average horizontal component of peak ground motion and response spectra. *Earthq Spectra* 30(3):1117–1153. <https://doi.org/10.1193/072813EQS219M>
- Colombelli S, Caruso A, Zollo A, Festa G, Kanamori H (2015) A P wave-based, on-site method for earthquake early warning. *Geophys Res Lett* 42(5):1390–1398. <https://doi.org/10.1002/2014GL063002>
- Demir S, Sahin EK (2023) Application of state-of-the-art machine learning algorithms for slope stability prediction by handling outliers of the dataset. *Earth Sci Inf* 16(3):2497–2509. <https://doi.org/10.1007/s12145-023-01059-8>
- Douglas J (2003) What is a poor quality strong-motion record? *Bull Earthq Eng* 1:141–156. <https://doi.org/10.1023/A:1024861528201>
- Emre Ö, Duman TY, Özalp S, Şaroğlu F, Olgun Ş, Elmacı H, Çan T (2018) Active fault database of Turkey. *Bull Earthq Eng* 16(8):3229–3275. <https://doi.org/10.1007/s10518-016-0041-2>
- Fafalios S, Charonyktakis P, Tsamardinos I (2020) Gradient boosting trees. *Computer Science, Mathematics*. <https://api.semanticscholar.org/CorpusID:222288735>. Accessed 11 Dec 2023
- Fang R, Zheng J, Geng J, Shu Y, Shi C, Liu J (2020) Earthquake magnitude scaling using peak ground velocity derived from high-rate gnss observations. *Seismol Res Lett* 92(1):227–237. <https://doi.org/10.1785/0220190347>
- Farahani S, Barari A (2023) A simplified procedure for the prediction of liquefaction-induced settlement of offshore wind turbines supported by suction caisson foundation based on effective stress analyses and an ML-based group method of data handling. *Earthquake Eng Struct Dynam* 52(15):5072–5098. <https://doi.org/10.1002/eqe.4000>
- Friedman JH (2002) Stochastic gradient boosting. *Comput Stat Data Anal* 38:367–378. [https://doi.org/10.1016/S0167-9473\(01\)00065-2](https://doi.org/10.1016/S0167-9473(01)00065-2)
- Friedman JH (2001) Greedy function approximation: A gradient boosting machine. In *The Annals of Statistics* (Vol. 29, Issue 5).
- Golberg DE (1989) Genetic algorithms in search, optimization, and machine learning. Addison Wesley;1989. p. 36
- Guyon I, Weston J, Barnhill S (2002) Gene Selection for Cancer Classification using Support Vector Machines 46. <https://doi.org/10.1109/ICMLA.2007.35>
- Han J, Kamber M, Berzal F, Marín N (2001) Data mining: concepts and techniques 31. Morgan Kaufmann Publishers. <https://doi.org/10.1145/565117.565130>
- Hussaini SMS, Caicedo D, Mohammadi A, Karimzadeh S, Lourenço PB (2024) Nonparametric ground motion models of arias intensity and significant duration for the Italian dataset. *J Phys Conf Ser*, 2647(6). <https://doi.org/10.1088/1742-6596/2647/6/062001>
- Impraimakis M (2023) A convolutional neural network deep learning method for model class selection. *Earthquake Engineering and Structural Dynamics* <https://doi.org/10.1002/eqe.4045>
- Jeddi AB, Shafieezadeh A, Hur J, Ha JG, Hahm D, Kim MK (2022) Multi-hazard typhoon and earthquake collapse fragility models for transmission towers: An active learning reliability approach using gradient boosting classifiers. *Earthquake Eng Struct Dynam* 51(15):3552–3573. <https://doi.org/10.1002/eqe.3735>
- Karakas CC, Kuran F, Guvensoy G, Palanci M, Senel SM (2021) Investigation of parameters on the fragility response of existing buildings according to different seismic codes. *Proceedings of the 14th International Congress on Advances in Civil Engineering*. Istanbul, Turkey, pp 1219–1225
- Karimzadeh S, Mohammadi A, Hussaini SMS, Caicedo D, Askan A, Lourenço PB (2023a) ANN-based ground motion model for Turkey using stochastic simulation of earthquakes. *Geophys J Int* 236(1):413–429. <https://doi.org/10.1093/gji/ggad432>
- Karimzadeh S, Mohammadi A, Salahuddin U, Carvalho A, Lourenço PB (2023b) Backbone ground motion model through simulated records and XGBoost machine learning algorithm: An application for the Azores plateau (Portugal). *Earthquake Eng Struct Dynam*. <https://doi.org/10.1002/eqe.4040>
- Kim T, Kwon OS, Song J (2023) Deep learning based seismic response prediction of hysteretic systems having degradation and pinching. *Earthquake Eng Struct Dynam* 52(8):2384–2406. <https://doi.org/10.1002/eqe.3796>
- Kostadinov MV, Towhata I (2002) Assessment of liquefaction-inducing peak ground velocity and frequency of horizontal ground shaking at onset of phenomenon. *Soil Dyn Earthq Eng* 22:309–322
- Kuran F, Tanırcan G, Pashaei E (2023a) Performance evaluation of machine learning techniques in predicting cumulative absolute velocity. *Soil Dyn Earthq Eng* 174:108175. <https://doi.org/10.1016/j.soildyn.2023.108175>
- Kuran F, Tanırcan G, Pashaei E. (2023b) Prediction of peak ground velocity (PGV) and cumulative absolute velocity (CAV) of earthquakes using machine learning techniques. In Erberik, M.A., Askan, A., Kockar, M.K. (eds) *Proceedings of the 7th International Conference on Earthquake Engineering and Seismology*. ICEES 2023 401. https://doi.org/10.1007/978-3-031-57357-6_3
- Meenakshi Y, Sreenath V, Stg R (2023) Ground motion models for Fourier amplitude spectra and response spectra using Machine learning techniques. *Earthquake Engineering and Structural Dynamics* <https://doi.org/10.1002/eqe.4036>
- Michel C, Kelevitz K, Houlié N, Edwards B, Psimoulis P, Su Z, Clinton J, Giardini D (2017) The potential of high-rate GPS for strong ground motion assessment. *Bull Seismol Soc Am* 107(4):1849–1859. <https://doi.org/10.1785/0120160296>

- Mohammadi A, Barros JAO, Sena-Cruz J (2023c) A new model for predicting the shear strength of RC beams strengthened with externally bonded FRP sheets. *Compos Struct*, 319. <https://doi.org/10.1016/j.compstruct.2023.117081>
- Mohammadi AH, Hussaini SMS, Caicedo D, Karimzadeh S, Lourenço PB (2023b) Nonparametric ground motion models of cumulative absolute velocity and peak ground velocity for the Italian dataset. In Erberik, M.A., Askan, A., Kockar, M.K. (eds) *Proceedings of the 7th International Conference on Earthquake Engineering and Seismology*. ICEES 2023 401. Springer, Cham. https://doi.org/10.1007/978-3-031-57357-6_4
- Mohammadi A, Karimzadeh S, Banimahd SA, Ozsarac V, Lourenço PB (2023a) The potential of region-specific machine-learning-based ground motion models: Application to Turkey. *Soil Dyn Earthq Eng*, 172. <https://doi.org/10.1016/j.soildyn.2023.108008>
- Mousavi SM, Beroza GC (2020) A machine-learning approach for earthquake magnitude estimation. *Geophys Res Lett*, 47(1). <https://doi.org/10.1029/2019GL085976>
- Omine H, Hayashi T, Yashiro H, Fukushima I (2008) Seismic risk analysis method using both PGA and PGV. In *Proceeding of the 14th World Conference on Earthquake Engineering*, Beijing, China. https://www.iitk.ac.in/nicee/wcee/fourteenth_conf_china/. Accessed 11 Dec 2023
- Orense RP (2005) Assessment of liquefaction potential based on peak ground motion parameters. *Soil Dyn Earthq Eng* 25(3):225–240. <https://doi.org/10.1016/j.soildyn.2004.10.013>
- Pashaei E, Pashaei E (2022) Hybrid binary arithmetic optimization algorithm with simulated annealing for feature selection in high-dimensional biomedical data. *J Supercomput* 78(13):15598–15637. <https://doi.org/10.1007/s11227-022-04507-2>
- Pashaei E, Ozen M, Aydin N (2017) Splice site identification in human genome using random forest. *Heal Technol* 7:141–142. <https://doi.org/10.1007/s12553-016-0157-z>
- Picozzi M, Emolo A, Martino C, Zollo A, Miranda N, Verderame G, Boxberger T, Elia L, Del Gaudio S, Colombelli S, Amoroso O, Brondi P, De Risi MT, Parolai S, Bindi D, Boxberger T, Miranda N, Buonaiuto L, Amelia A (2015) Earthquake early warning system for schools: A feasibility study in southern Italy. *Seismol Res Lett* 86(2A):398–412. <https://doi.org/10.1785/0220140194>
- Qi S, Chen J, Chen P, Wen P, Niu X, Xu L (2024) An efficient GAN-based predictive framework for multivariate time series anomaly prediction in cloud data centers. *J Supercomput* 80:1268–1293. <https://doi.org/10.1007/s11227-023-05534-3>
- Razin T, Khatimah K, Annisa Y, Hamzah A, Massinai MFI (2021) Peak ground acceleration (PGA) and peak ground velocity (PGV) analyze for microzonation of earthquake hazard area: Case study in West Nusa Tenggara. *IOP Conference Series: Earth Environ Sci*, 873(1). <https://doi.org/10.1088/1755-1315/873/1/012046>
- Sandikkaya MA, Güryuva B, Kale Ö, Okçu O, İçen A, Yenier E, Akkar S (2023) An updated strong-motion database of Türkiye (SMD-TR). *Earthquake Spectra* <https://doi.org/10.1177/87552930231208158>
- Sedaghati F, Pezeshk S (2023b) Ensemble region-specific GMMs for subduction earthquakes. *Seismol Res Lett* 95(3):1735–1758. <https://doi.org/10.1785/0220230070>
- Sedaghati F, Pezeshk S (2023a) Machine learning-based ground motion models for shallow crustal earthquakes in active tectonic regions. *Earthq Spectra* 39(4):2406–2435. <https://doi.org/10.1177/87552930231191759>
- Tanırca G, Gülerce Z, Tetik T, Akbaş B (2022) Adapting a global fourier amplitude spectrum estimation model to the earthquake database of Türkiye. In *TUBİTAK 1002 Project Report No: 120M799*, [in Turkish]. (Accessible after 05/2025)
- Thaler D, Stoffel M, Markert B, Bamer F (2021) Machine-learning-enhanced tail end prediction of structural response statistics in earthquake engineering. *Earthquake Eng Struct Dynam* 50(8):2098–2114. <https://doi.org/10.1002/eqe.3432>
- Tibshirani R (1996) Regression Shrinkage and Selection Via the Lasso. *J Roy Stat Soc: Ser B (methodol)* 58(1):267–288. <https://doi.org/10.1111/j.2517-6161.1996.tb02080.x>
- Tunay M, Pashaei E, Pashaei E (2022) Hybrid hypercube optimization search algorithm and multilayer perceptron neural network for medical data classification. *Comput Intel Neurosc* 2022(1):1612468. <https://doi.org/10.1155/2022/1612468>
- Wang P, Zhao Q, Huang Y, Nie J, Man X, Mingxue Z (2020) Comprehensive evaluation of road risk in earthquake disaster area based on normalization method. 2020 International Conference on Computer, Information and Telecommunication Systems (CITS), 1–5. <https://doi.org/10.1109/CITS49457.2020.9232511>
- Zhang X, Tang L, Ling X, Chan AHC, Lu J (2018) Using peak ground velocity to characterize the response of soil-pile system in liquefying ground. *Eng Geol* 240:62–73. <https://doi.org/10.1016/j.enggeo.2018.04.011>
- Zhang B, Yu Y, Li X, Wang Y (2022) Ground motion prediction equation for the average horizontal component of PGA, PGV, and 5% damped acceleration response spectra at periods ranging from 0.033 to 8.0s in southwest China. *Soil Dyn Earthq Eng*, 159. <https://doi.org/10.1016/j.soildyn.2022.107297>
- Zou D, Xiang Y, Zhou T, Peng Q, Dai W, Hong Z, Shi Y, Wang S, Yin J, Quan H (2023) Outlier detection and data filling based on KNN and LOF for power transformer operation data classification. *Energy Rep* 9:698–711. <https://doi.org/10.1016/j.egy.2023.04.094>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.