Structural Monitoring and Maintenance, *Vol. 12, No. 1 (2025) 71-91*
https://doi.org/10.12989/smm.2025.12.1.071

Machine learning-based methodologies for probabilistic prediction of random seismic frame structural response

Zheng Wu, Meiling Xiao*, Yiwu Sun and Houming Wang

School of Architecture and Planning, Yunnan University, Kunming 650051, P.R. China

(*Received August 7, 2024, Revised November 10, 2024, Accepted November 14, 2024*)

Abstract. This paper proposes an innovative methodology that synergistically combines machine learning techniques with probabilistic learning on manifolds for generating samples to predict the response distribution of frame structures. Through a rigorous feature engineering process, 11 seismic feature parameters and one structural feature parameter were judiciously selected. Leveraging a small-scale dataset, an exhaustive model selection process was undertaken, evaluating the performance of Support Vector Regression, Random Forest, and Gradient Boosting Trees, ultimately identifying the optimal machine learning model. By concurrently accounting for the stochastic nature of seismic motions and structural characteristics, this methodology is employed to predict the distribution of structural responses of multi-story reinforced concrete frame structures subjected to stochastic seismic events. The results demonstrate that this methodology achieves a high degree of prediction accuracy on the test dataset and can reasonably predict the seismic damage to reinforced concrete frame structures, thereby providing valuable guidance for post-earthquake disaster assessment and emergency response efforts.

Keywords: K-means clustering; machine learning; probabilistic learning on manifolds; random seismic response prediction

1. Introduction

In recent years, the frequent occurrence of major earthquakes globally has posed a significant threat to human life and property safety. Traditional seismic response analysis methods require substantial computational resources and time, which limits their practical application (Yue *et al.* 2019). Especially with the increasing complexity of building structures and the significant randomness of seismic loads, how to rapidly and accurately predict the response of structural systems under seismic actions has become a focal point for researchers (Wang e*t al.* 2022).

Various traditional methods for predicting seismic structural responses, such as numerical simulation methods based on time history analysis and response spectrum analysis (De Domenico *et al*. 2018, Yang *et al.* 2022), can predict structural responses with reasonable accuracy through detailed analysis of structures and seismic waveforms. However, they incur extremely high computational costs and lack flexibility when faced with variable seismic waveforms and complex structural systems. In contrast, machine learning methods can quickly establish predictive models by learning from a large amount of seismic and structural response data, significantly reducing

*Corresponding author, Professor, E-mail: mlxiao@ynu.edu.cn

http://www.techno-press.org/?journal=smm&subpage=7 ISSN: 2288-6605 (Print), 2288-6613 (Online)

computation time. Particularly in applications focused on stochastic seismic response prediction, machine learning methods can offer more flexible solutions than traditional methods by identifying and learning complex data patterns. They have also been successfully applied in various fields such as healthcare, finance, and transportation, demonstrating their ability to solve complex problems and provide innovative solutions (Patil *et al*. 2022, Bowman *et al.* 2022, Wang *et al.* 2024, Mao *et al*. 2024, Byun *et al*. 2023).

Machine learning methods, especially deep learning and ensemble learning techniques, have been widely applied in seismic response prediction in recent years. Through nonlinear mapping and feature extraction capabilities, they can effectively address complex patterns in seismic responses, such as artificial neural networks (ANNs) used for predicting nonlinear seismic responses of multi-degree-of-freedom systems and building assemblies. These models can significantly reduce computational costs while maintaining high predictive accuracy (Kalakonas and Silva 2021, Guo *et al.* 2021). Mangalathu *et al.* (2020) utilized random forest and other machine learning algorithms to predict failure patterns of reinforced concrete columns and shear walls with high accuracy (84-86%). Additionally, machine learning models can predict the seismic responses of damped structures, with advanced models like the Seismic Wave Transformer (SWT) showing superior accuracy (Zhang *et al.* 2023). However, these methods often rely on a large amount of training data and have limited generalization ability when dealing with unseen data. With improvements in computational capabilities and the increase in data volume, deep learning methods are gradually becoming mainstream. Dail *et al*. (2021) employed convolutional neural networks (CNNs) for automatic feature extraction and prediction of seismic response time histories, achieving high computational efficiency. However, such methods perform poorly in small sample problems and struggle to adapt to the randomness of seismic motions and structural uncertainties.

To address these shortcomings, recent research has increasingly focused on ensemble learning methods and probabilistic learning approaches. Torky *et al*. (2021) combined CNN with other neural network architectures, such as long short-term memory (LSTM) networks, showing promise in predicting multi-component seismic responses. These hybrid models can map the relationship between base acceleration time series and superstructure responses, enhancing prediction stability and robustness through multi-model integration.

In the field of probabilistic learning, probabilistic learning on manifolds (PLoM) has gradually become a research hotspot. This method is particularly suited for handling high-dimensional and nonlinear data, as it establishes probabilistic models in manifold spaces to capture the underlying distribution of the data more effectively (Soize and Ghanem 2019). Consequently, PLoM has also begun to be introduced into the field of seismic response prediction (Zhong *et al.* 2023), achieving significant results. To address deterministic prediction methods, probabilistic prediction methods, such as natural gradient boosting (NGBoost), have been used to assess the conditional probability distribution of structural responses, yielding good results (Ding *et al*. 2023). Kim *et al.* (2020) proposed a Bayesian deep learning method to enhance prediction accuracy by quantifying uncertainties in structural responses.

However, despite the progress made in existing research, many challenges and shortcomings remain. First, most studies still rely on deterministic methods for seismic motion modeling, failing to adequately account for the randomness and diversity of seismic motions (Zhang *et al*. 2018, Hwang *et al.* 2021). Secondly, while PLoM has significant advantages in handling high-dimensional data, it still suffers from high computational complexity. Moreover, existing research mostly focuses on linear predictions of single structures, with little consideration given to

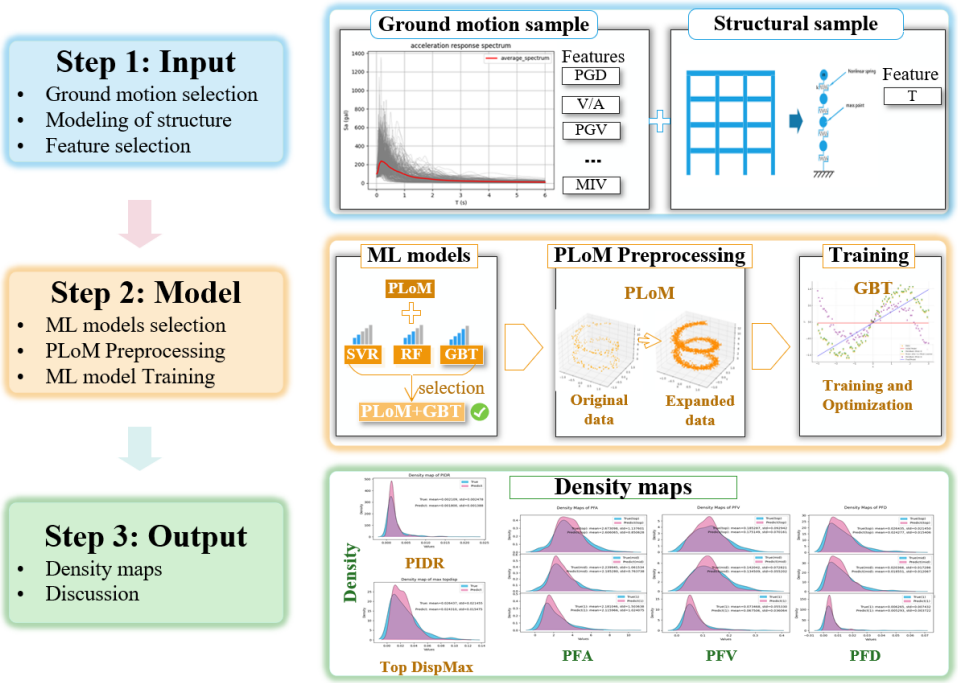


Fig. 1 The flowchart of prediction seismic random response of frame structure

structural characteristics and nonlinear behaviors (Khan *et al*. 2019). Additionally, achieving a balance between computational efficiency and predictive accuracy while better capturing the randomness of seismic motions remains a key challenge in this field.

Therefore, this paper proposes a seismic response probabilistic prediction method that combines probabilistic learning on manifolds (PLoM) with traditional machine learning models for predicting the response distribution of frame structures. Through meticulous feature engineering, we selected 11 seismic and one structural feature parameter. Using a small dataset, we conducted an exhaustive model selection process, assessing Support Vector Regression, Random Forest, and Gradient Boosting Trees to find the best machine learning model. This methodology, which considers both the randomness of seismic motions and structural properties, is used to forecast the structural response distribution of multi-story reinforced concrete frames under stochastic seismic events. This innovative method offers a dual benefit: it delivers new theoretical insights into stochastic seismic response prediction and provides the technical foundation for efficient and precise seismic response forecasting in real-world engineering scenarios.

2. Methodology

The main objective of this study is to propose a new method for predicting the response of frame structures under stochastic seismic actions. The approach leverages newly generated samples from probabilistic learning on manifolds (PLoM) to train traditional simple machine learning algorithms, thereby providing a probabilistic distribution prediction of the results.

Compared to other existing methods, the proposed approach has the following features:

- By utilizing PLoM to efficiently generate new samples from the learning distribution of high-dimensional problems, the original elastoplastic time history analysis dataset is expanded, avoiding extensive elastoplastic time history analysis processes and reducing the consumption of computational resources. After generating new samples using PLoM, the dataset is reasonably expanded, and the predictive performance of traditional machine learning algorithms is enhanced.
- The method accounts for the uncertainty of seismic motions by analyzing historical earthquake records and statistically distributing seismic parameters across different locations and periods. It also considers structural diversity by establishing frame structures with different periods.
- The proposed method provides a prediction of the structural seismic response distribution that fully considers the complexity of structural behavior under seismic conditions, aiding decision-makers in making balanced decisions rather than relying solely on single deterministic predictions.

The implementation of the method involves three steps, as illustrated in Fig. 1.

Step 1: Dataset Preparation

- Ground motions: Using the k-means clustering algorithm, seismic samples are unsupervisedly classified to select a representative subset. For each seismic record in the subset, seismic parameters are calculated to identify key parameters that are relatively independent and strongly correlated with structural responses.
- Frame structures: A multi-degree-of-freedom (MDOF) shear model is used as the theoretical calculation model for reinforced concrete (RC) frame structures. The software OpenSees is employed to establish shear models with different periods and perform nonlinear time history analysis (NLTHA). Through NLTHA, the seismic responses of structures in terms of engineering demand parameters (EDP) are obtained, including peak inter-story drift ratio (PIDR), maximum roof displacement, peak floor acceleration (PFA), peak floor velocity (PFV), and peak floor displacement (PFD). This generates a database with seismic parameters and structural periods as input variables and seismic responses as output variables.

Step 2: PLoM Data Preprocessing

Probabilistic learning on manifolds (PLoM) is used to expand the database from Step 1, generating multiple samples statistically faithful to the original data distribution. This step is significantly less time-consuming than generating an equivalent amount of data through NLTHA.

Step 3: Development of the Machine Learning Model

The maximum inter-story drift ratio of a 7-story RC frame under 284 sets of seismic motions is used as training data to evaluate the performance of traditional machine learning models (SVR, RF, GBT) combined with PLoM. The Gradient Boosting Trees combined with PLoM performed the best, thus it was selected as the base model. The PLoM-expanded database was split into training and testing sets in an 8:2 ratio. The optimal model was obtained based on the training set and random grid search optimization. Finally, the model's performance was validated using the test set.

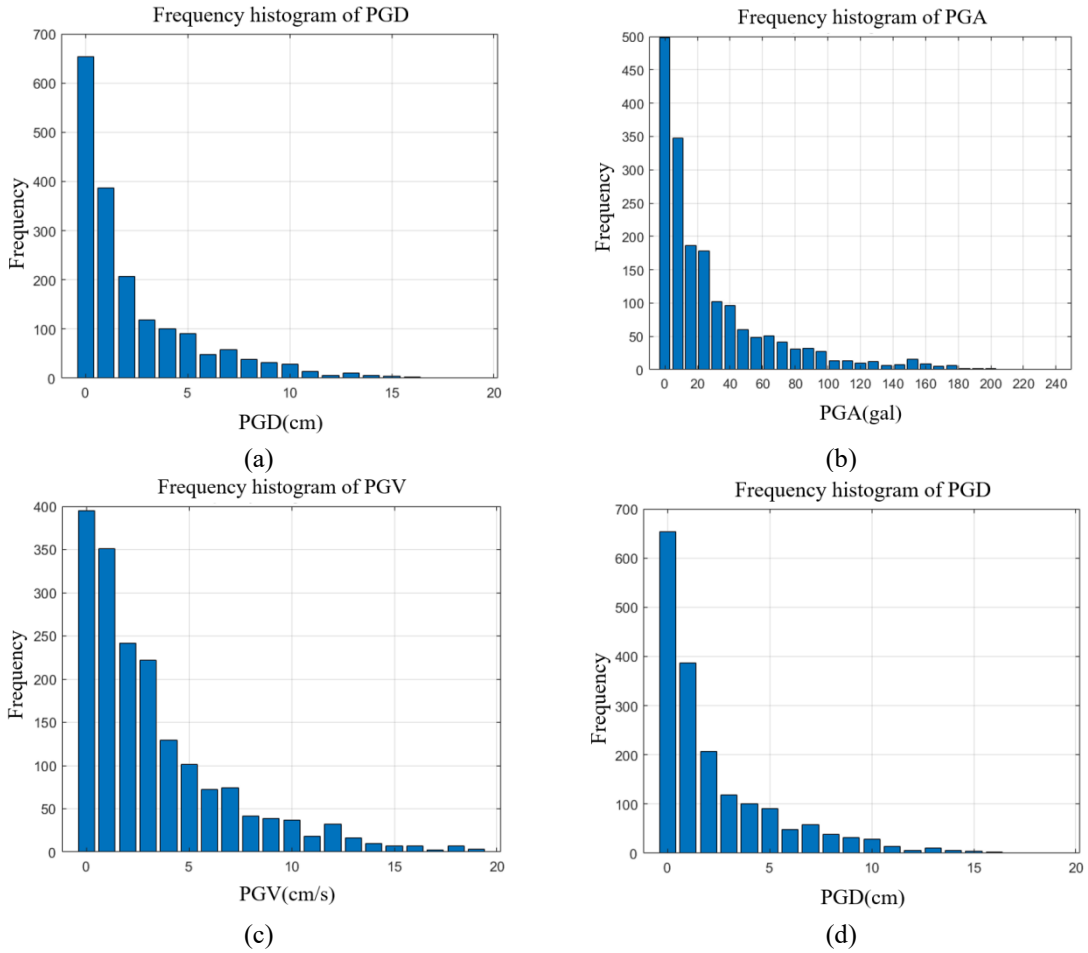


Fig. 2 Ground motions distribution after outlier removal. (a) 3D distribution map of ground motions after outlier removal, (b) Frequency histogram of PGA, (c) Frequency histogram of PGV and (d) Frequency histogram of PGD

3. Examples

This study considers the uncertainty of ground motions and frame structural characteristics.

3.1 Selection of representative seismic records

The Pacific Earthquake Engineering Research Center (PEER) database (https://ngawest2.berkeley.edu/) provides numerous historical seismic records available for developing a machine learning predictive model. Ground motions were selected according to the following criteria, resulting in 1953 seismic records that meet the research objectives: According to statistics (Wair *et al.* 2012), Shear wave velocity 265 m/s≤ Vs_{30} ≤550 m/s, and the earthquake intensity magnitude≥6.5, and 30km≤Epicentral Distance≤1000 km.

3.1.1 Ground motions samples choosing based on K-means clustering

From the PEER database, 1953 seismic records were obtained. Statistical analysis of these records quantifies the distribution characteristics of seismic parameters peak ground acceleration (PGA), peak ground velocity (PGV) and peak ground displacement (PGD). The distribution of PGA, PGV, and PGD is relatively concentrated: most PGA values range from 0 to 200 gal, with fewer values between 200 and 520 gal; most PGV values are between 0 and 20 cm/s, with fewer values between 20 and 53 cm/s; most PGD values are between 0 and 20 cm, with fewer values between 20 and 53 cm. Based on these distributions, K-means clustering was used for sampling selection to simulate seismic events with different parameter combinations.

3.1.2 Outlier removal based on KNN

The seismic samples contain a few outliers that need to be removed. The k-nearest neighbors (KNN) algorithm is a simple and widely used unsupervised learning algorithm suitable for outlier detection. With k set to 500 and using PGA, PGV, and PGD as features, 147 outliers were removed, resulting in 1806 seismic records. Their three-dimensional distribution and frequency histograms of each parameter are shown in Fig. 2.

3.1.3 Seismic clustering and selection based on K-means

After outlier removal, the remaining seismic samples still number too many, and their distribution in PGA, PGV, and PGD is concentrated on the left. K-means clustering was used to further filter seismic motions by unsupervised classification of seismic samples. Two data points were randomly chosen as initial cluster centers from the dataset. The clustered groups are shown in Fig. 3(a). To ensure a more uniform and widespread distribution of seismic records, the cluster with larger amplitude characteristics (yellow part) was chosen, resulting in 284 seismic samples. Their three-dimensional distribution and frequency histograms of each parameter are shown in Fig. 3, with seismic response spectra in Fig. 3(b).

3.1.4 Selection of Seismic Parameters Based on Correlation Analysis

Similar to Xing et al. (2024), this study uses data-driven method to deal with the uncertainty of ground motion data. Seismic ground motions can be quantified using various parameters (Fidarova *et al*. 2023). To fully reflect seismic characteristics, 24 seismic parameters covering amplitude, spectral, and duration were initially selected and calculated. Table 1 presents some of the main seismic parameters used for machine learning.

To reduce input variables and enhance the efficiency of the sample set, it is crucial to select key seismic parameters that are relatively independent and strongly correlated with structural seismic responses. The Pearson correlation coefficient measures the strength and direction of the linear relationship between two continuous variables (Lee Rodgers *et al*. 1988), defined as the covariance of the variables divided by the product of their standard deviations. To quantitatively analyze the importance of key seismic parameters, the following steps were taken:

- Calculate the Pearson correlation coefficient R [Eq. (1)] between each parameter and others to select relatively independent seismic parameters.
- Evaluate the correlation between seismic parameters and structural seismic responses by calculating the Pearson correlation coefficient r [Eq. (1)] between each parameter and the maximum structural response. The maximum structural response is determined for single-degree-of-freedom structures with periods of 0.2s, 2.0s, 4.0s, and 6.0s, using OpenSees modeling and analyzing the 284 selected seismic motions.

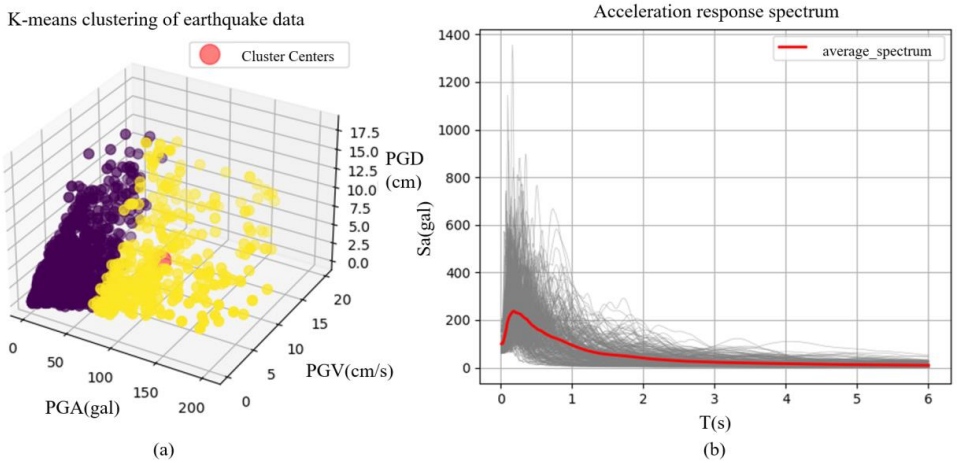


Fig. 3 (a) K-means clustering and (b)Response spectrum of 284 sets of ground motions

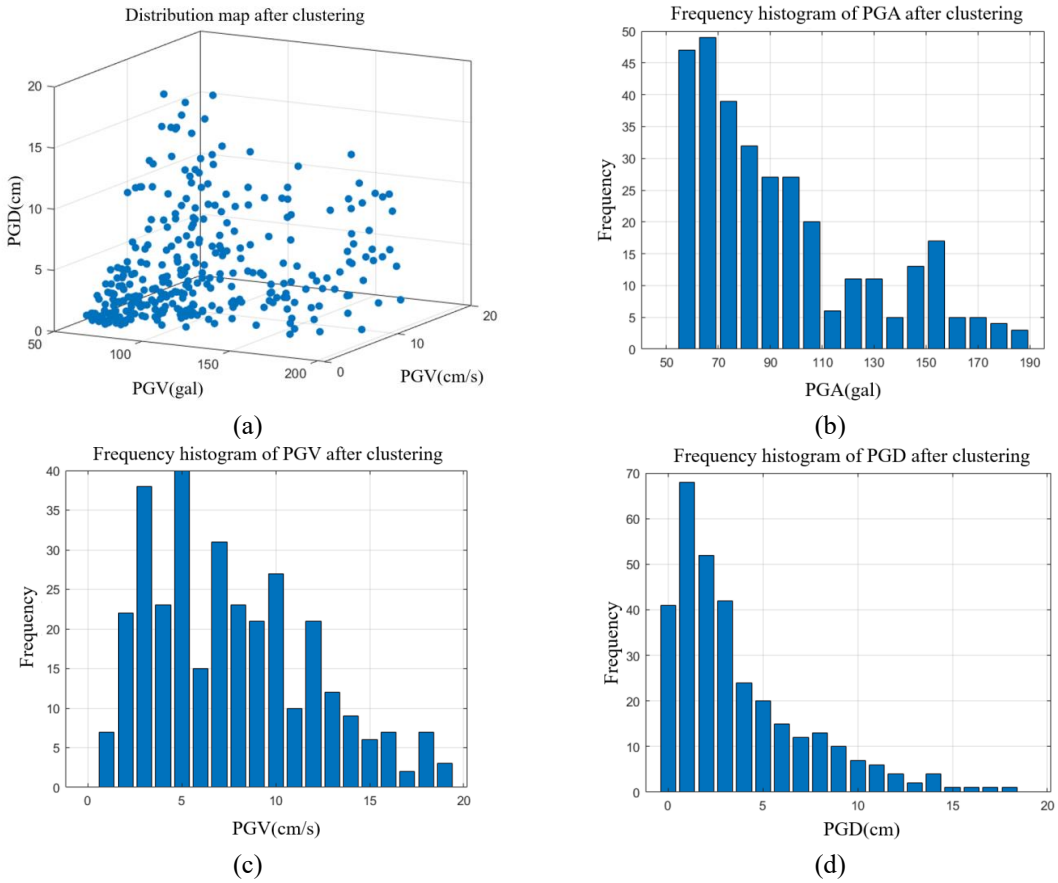


Fig. 4 Ground motion records distribution after K-means clustering. (a) 3D distribution map of seismic recordings after outlier removal, (b) Frequency histogram of PGA, (c) Frequency histogram of PGV and (d) Frequency histogram of PGD

Table 1 Ground motion parameters for machine learning

Ground Motion Characteristics	Parameters	Description
Amplitude	PGA; PGV; PGD	PGA=max|a(t)| / PGV=max|v(t)|/ PGD=max|d(t)|
	PGV / PGA	$v_{max}/a_{max}=\frac{\max\|v(t)\|}{\max\|a(t)\|}$
	SMA	Sustained Maximum Acceleration
	SMV	Sustained Maximum Velocity
	EDA	Effective Design Acceleration
	V_{rms}	Root Mean Square Velocity $V_{rms}=\sqrt{\frac{1}{t}\int_0^t \dot{u}_g^2(t)dt}$
	D_{rms}	Root Mean Square Displacement $D_{rms}=\sqrt{\frac{1}{t}\int_0^t u_g^2(t)dt}$
	MIV	Maximum Incremental Velocity $MIV=\max\left\|\int_{t_1}^{t_2}\ddot{u}_g dt\right\|$
	HI	Housner Intensity $HI=\int_{0.1}^{2.5} PSV(\xi=0.05,T)dT$
Spectrum	PSA	Peak Spectral Acceleration
	PSV	Peak Spectral Velocity
	Sa_{avg}	Average Spectral Acceleration $Sa_{avg}=\frac{1}{N}\sum_{T_1}^{T_N} Sa(T_i,\zeta)$
Duration	T_D	Effective Duration
Energy	SED	Specific Energy Density $SED=\int_0^{t_{mr}}[v(t)]^2dt$

The smaller the average absolute value of R between a parameter and others, the more independent the parameter. Conversely, the larger the average absolute value of r between a parameter and the maximum structural response, the stronger the correlation with the structural response. Thus, the importance of a seismic parameter is negatively correlated with R and positively correlated with r. To emphasize the parameters strongly correlated with structural responses, the weight of r was increased. The weighted average correlation coefficient p representing the importance of a seismic parameter is calculated as follows

$$p=-[(|R_1|+|R_2|+\ldots+|R_m|)/m]\times 0.2+[(|r_1|+|r_2|+\ldots+|r_n|)/n]\times 0.8 \quad (1)$$

Where R_m is the Pearson correlation coefficient between a parameter and the m-th ground motion parameter, and r_n is the Pearson correlation coefficient between a parameter and the n-th structural response. Finally, based on the value of p, the top 11 seismic parameters were selected as key input parameters for seismic features, as shown in Table 2.

3.2 Parameters of structural characteristics

3.2.1 MDOF shear model and feature extraction

Simplified models are useful in many cases (Zhu *et al.* 2021). Multistory RC frame structures typically exhibit a clear shear-type lateral displacement pattern under seismic action. Hence, the MDOF (Multiple Degrees of Freedom) shear model is commonly used to simulate the seismic

Table 2 Ranking of the importance of seismic parameters

Seismic Parameters	D_{rms}	PGD	V_{rms}	SMV	V/A	PGV	SED	PSV	HI	$Sa_{\_avg}$	MIV
p	0.49	0.47	0.39	0.39	0.38	0.36	0.35	0.27	0.25	0.24	0.24

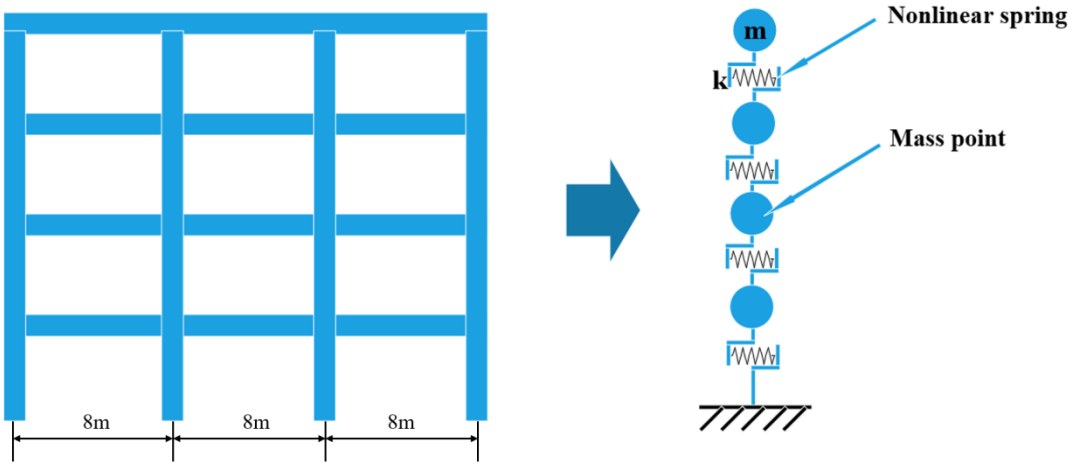


Fig. 5 MDOF (Multiple Degrees of Freedom) shear model

response of multistory RC frame structures (Xiong 2016), as shown in Fig. 5. The assumption of the MDOF shear model for multistory RC frame structures is that the mass of each floor is concentrated at the floor level, and the floors are rigid in-plane while neglecting rotational displacements. Therefore, each floor can be simplified to a mass point that generates corresponding forces when seismic acceleration is input.

Several researchers, including Lu *et al*. (2017), Xiong (2016), and Wu *et al*. (2017) have studied the simplification of the MDOF model for frame structures. Xiong Chen's model, in particular, included a large number of structural samples. Based on the parameters calibrated from Xiong C's model, this study extracts four different structural dynamic characteristics: natural period, damping ratio, story stiffness, and story mass. The characteristic parameters are shown in Table 3, with the number of stories ranging from 3 to 15. Each structure is assumed to have uniform properties across stories, resulting in 33 basic MDOF shear models.

The MDOF shear models are established using the OpenSees software. The shear spring elements are modeled using TwoNodeLink elements, and the uniaxialMaterial Hysteretic material is used to simulate the hysteretic behavior of the frame structure. The 284 seismic records are input into the model, and the outputs include the peak inter-story drift ratio (PIDR), peak roof displacement, and peak floor accelerations (PFA), peak floor velocities (PFV), and peak floor displacements (PFD) for the first floor, an intermediate floor (higher one in the case of an even number of floors), and the top floor. The response values for each structure are averaged over the 284 seismic analyses.

Table 3 Structural characteristic parameters

	Structural Parameters			
	Floor Weight (t)	Floor Stiffness ($N \cdot m^{-1}$)	Damping ratio	Period (s)
Value Range	[500,700]	[0.85×10^9, 1.25×10^9]	[0.04,0.06]	[0.33, 1.50]

Table 4 Ranking of the importance of structural characteristic parameters

Parameters	Period	Floor Weight	Floor Stiffness	Damping ratio
p	0.383	-0.04	-0.04	-0.04

3.2.2 Correlation analysis and feature selection

Following the procedure for selecting key seismic parameters described in Section 3.3, the four dynamic characteristics of the structure (natural period, damping ratio, story stiffness, and story mass) are analyzed for their correlation with structural responses (PIDR, peak roof displacement, and PFAs, PFVs, PFDs of the first, intermediate, and top floors). The parameters are ranked by their weighted average correlation coefficients p calculated according to Eq. (1), as shown in Table 4, with the natural period ranked first and thus selected as the key structural feature.

3.2.3 Selection and development of machine learning models

The dataset provided by the aforementioned method includes 11 seismic parameters and 1 structural period as input features, totaling 12 features or dimensions, which is considered low-dimensional data. The output sample size consists of the maximum inter-story drift ratio, peak roof displacement, and peak floor accelerations (PFA), peak floor velocities (PFV), and peak floor displacements (PFD) for 33 different story frame structures under 284 sets of seismic excitations. This results in a total of 284 × 33 = 9372 data sets, which is considered a medium data volume. To accurately and efficiently predict the seismic response of frame structures under medium sample sizes, it is necessary to expand the data set using generative methods and select a simple and efficient machine learning model to handle the generative samples.

Probabilistic Learning on Manifolds (PLoM) and generative samples

The PLoM code package from Zhong (Zhong, Gual, Govindjee, PLoM python package v1.0. https://github.com/sanjayg0/PLoM. 2021.) implements the probabilistic learning on manifolds algorithm (Soize and Ghanem 2016, Soize and Ghanem, 2019) for generating random vectors statistically consistent with the given dataset within a finite Euclidean space. The algorithm involves four main steps: data normalization, principal component analysis (PCA), kernel density estimation, and the generation of random variables using ISDE.

The implementation steps of the method are briefly introduced as follows. The initial input for the PLoM model is a random matrix [X], which is an n×N matrix where n is the dimension of the random data to be modeled and N is the number of available samples. The data matrix $[X]_{n\times N}$ is projected onto a set of orthogonal bases (PCA basis) to obtain the matrix $[H]_{v\times N}$ ($v\leq N$). A matrix transformation $[Z]=[H][g]([g]^T[g])^{-1}$ is then performed, where $[g]_{Nxm}$ ($m\leq N$) is the diffusion map basis, representing the dimensionality reduction process and generation of the manifold. The final step is to generate new samples $[Z]_{vxm}$ and transform them back to the original coordinate system to obtain new samples $[X_{new}]$. This process estimates the probability model on the manifold, learns

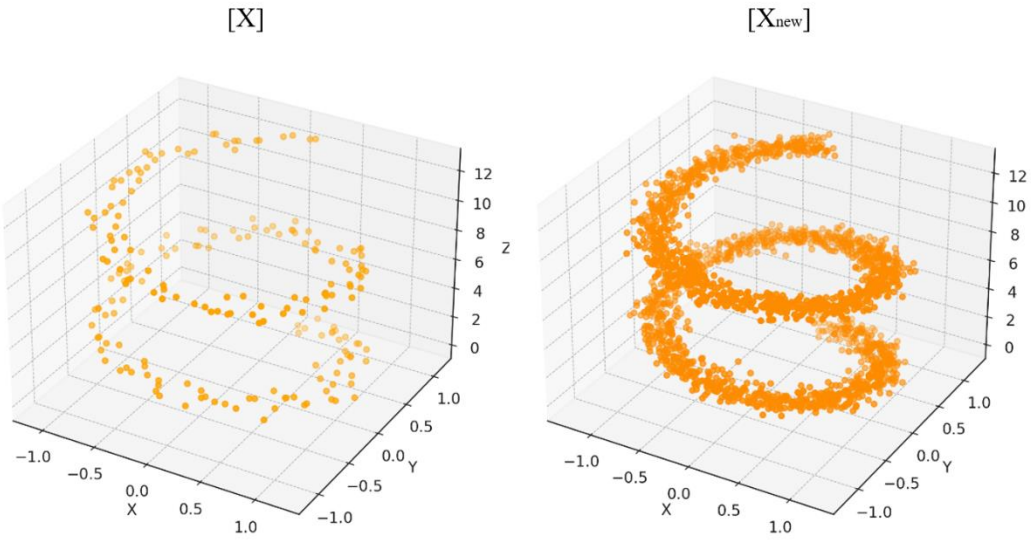


Fig. 6 The generation of new samples on the original manifold by PLoM

the probability distribution, solves the Itô stochastic differential equation using Galerkin projection, and generates samples [X_{new}] faithful to the original data. More details can be found in references (Zhong *et al.* 2023, Soize *et al.* 2016, Soize *et al.* 2017, Soize *et al.* 2020, Soize *et al*. 2020). Fig. 7 illustrates the generation of new samples on the original manifold by PLoM.

Selection of Machine Learning Models

Through the processing of the PLoM model, the original sample is expanded into an n×N dataset consistent with the manifold distribution, n is the generation multiple of the samples. For low-dimensional, medium-sized datasets with nonlinear relationships, simpler algorithms often have the advantages of strong interpretability, high computational efficiency, and strong generalization ability compared to complex neural networks. Such algorithms include support vector regression (SVR), random forests (RF), and gradient boosting trees (GBT). Selecting a suitable basic model for predicting the seismic response of frame structures under random seismic events is crucial.

To improve efficiency and determine the best-performing machine learning model combined with PLoM-generated samples, we selected a dataset consisting of the maximum inter-story drift ratios of a 7-story frame structure under 284 sets of seismic excitations, totaling 284 data sets. These data were divided into a training set (227 data sets) and a test set (57 data sets) in an 8:2 ratio. The SVR, RF, and GBT models were trained on both the original and PLoM-generated samples, resulting in six models: SVR, SVR+PLoM, RF, RF+PLoM, GBT, and GBT+PLoM, where +PLoM indicates the use of PLoM-generated samples. The performance of these models on the test set was compared to select the best combination.

The average prediction performance of each model on the test set is shown in the fit plot in Fig. 7, which includes performance metrics such as R^2, MAE, MARE, and RMSE.

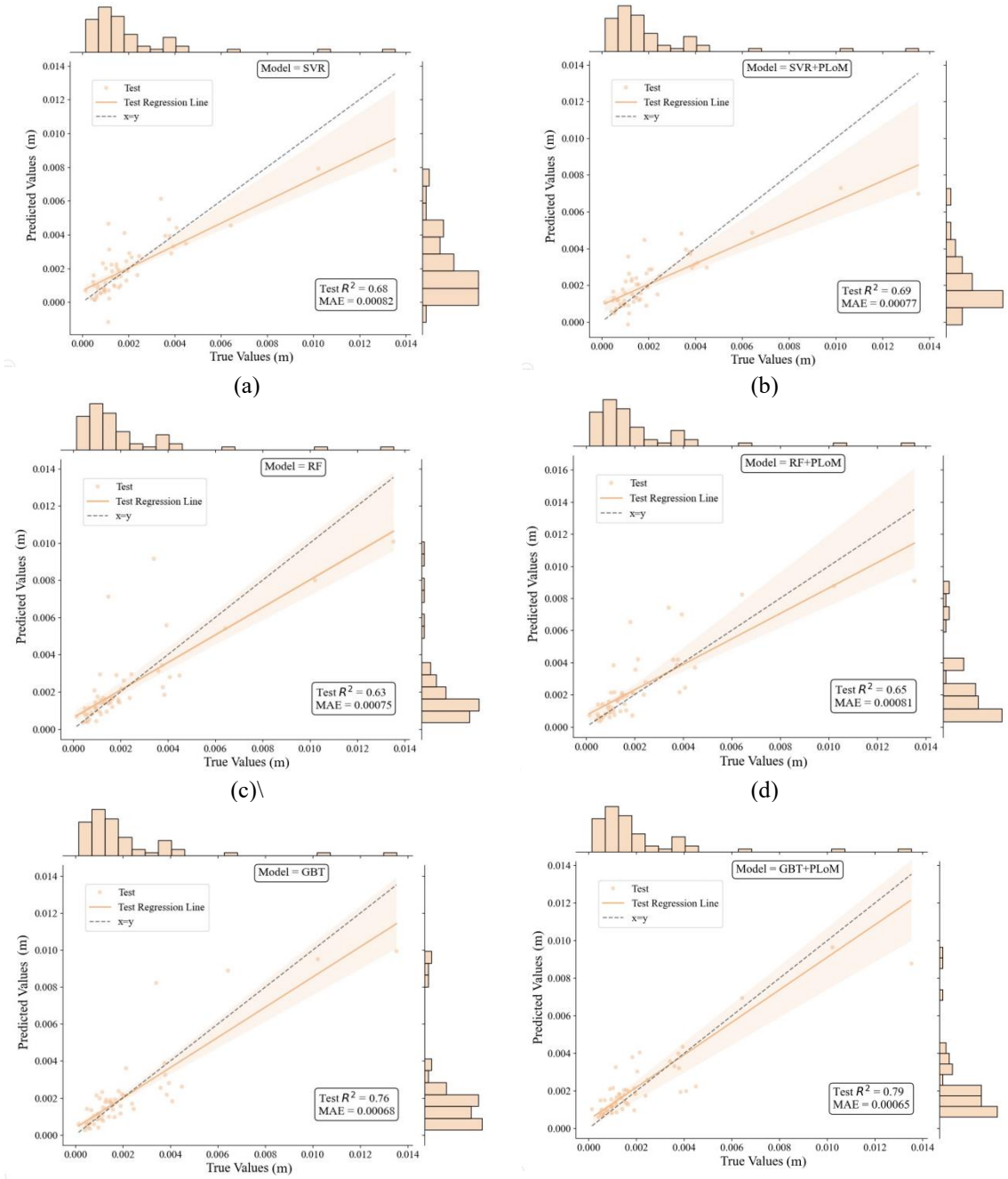


Fig. 7 The average prediction of SVR(a), SVR+PLoM(b), RF(c), RF+PLoM(d), GBT(e), and GBT+PLoM(f) on the test set

A successful prediction is defined as having a prediction error within 10%, and the number of successful predictions for each model is shown in Fig. 8.

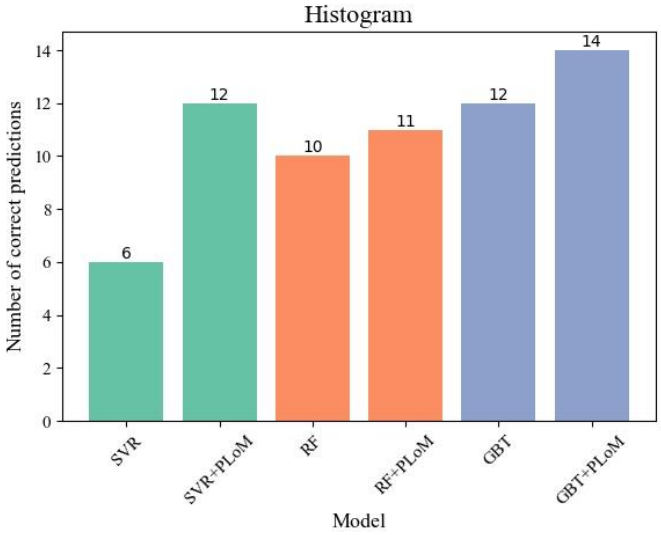


Fig. 8 Histogram of successful predictions for each model

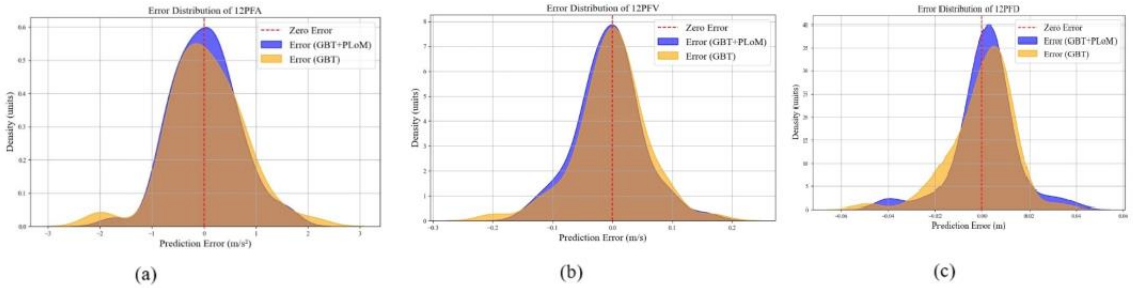


Fig. 9 Error distributions of (a)PFA, (b)PFV and (c)PFD at layer 12 of the two models on the test data

Comparing the performance of the six models on the test set, we found that regardless of the machine learning model, the use of PLoM-generated samples improved all metrics, with the GBT+PLoM combination performing the best. The GBT+PLoM model performance indicator R^2 reached 0.79, and the number of successful predictions was 21, indicating that the combination of gradient boosting trees (GBT) with PLoM-generated samples might be an effective method for predicting structural responses under random seismic events. This was further validated by using the GBT+PLoM and GBT models to train on the PFA, PFV, and PFD data sets of a 12-story RC frame structure under 284 sets of seismic excitations. The training and test sets were divided in the same 8:2 ratio, and the prediction distributions of the two models on the test set were compared. The box plots of predicted vs. actual values for both models are shown in Fig. 10. The results indicated that, compared to the GBT model, the GBT+PLoM model had a wider distribution of response predictions, with most medians closer to the actual values. In addition, the error distribution for the 12th layer of these two models on the test dataset was plotted to analyze and

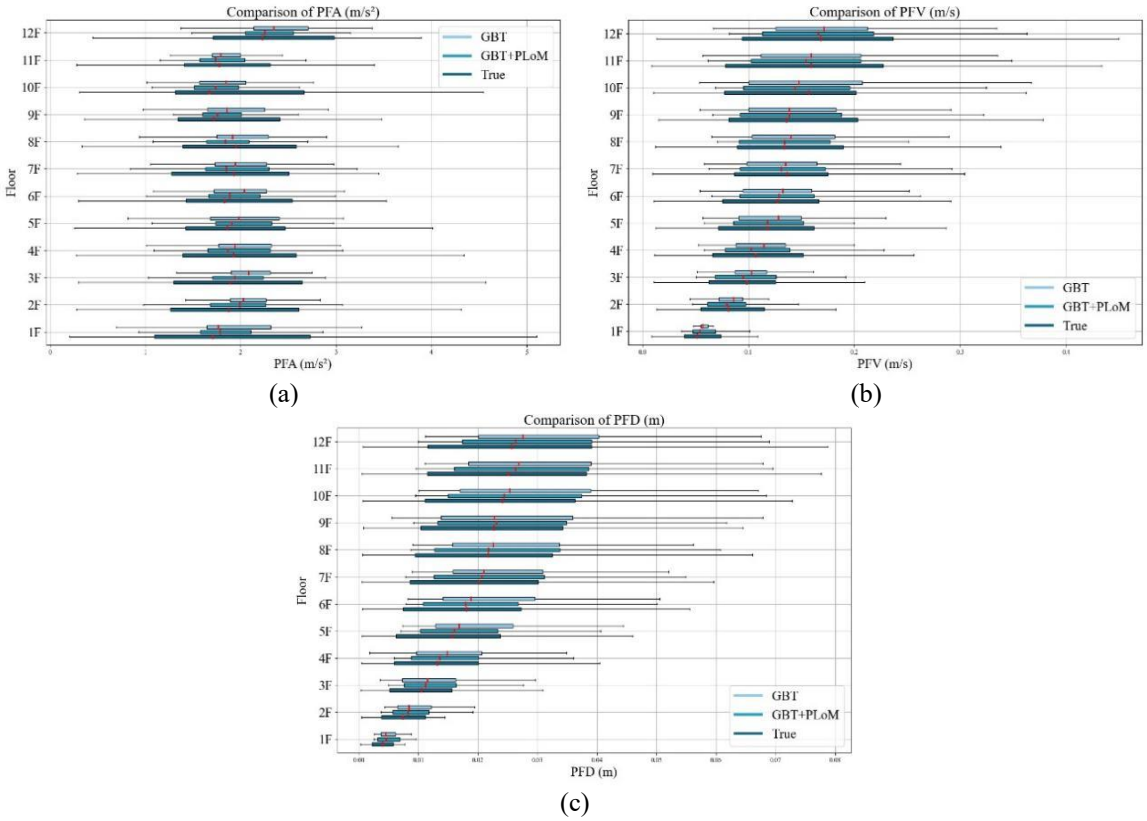


Fig. 10 Comparison of GBT predictions, GBT+PLoM predictions, and ground true values for each layer response PGA(a), PGV(b), PGD(c) of a 12-story frame structure

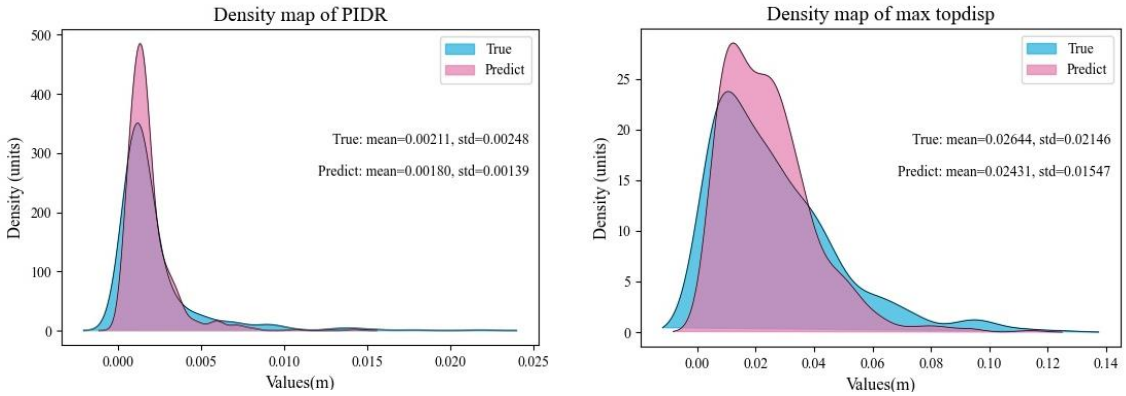
compare their performance in detail, as shown in Fig. 9. A close examination of the error distribution revealed that the GBT+PLoM model performed better across various metrics. Not only did it exhibit a higher density of errors around zero, but it also had a smaller error range and a more concentrated distribution. This result further validates the superiority of the GBT+PLoM model in practical applications and provides important references for the follow-up research.

3.3. Response prediction using gradient boosting tree model

3.3.1 PLoM preprocessing

The original sample dataset contains 9372 groups. To improve efficiency, we randomly selected 30% of the original dataset, which is 2812 groups, for all subsequent model training and testing. These 2812 data groups were divided into a training set (2250 groups) and a test set (562 groups) in an 8:2 ratio.

The PLoM package was used to generate samples from the 2250-group training set. After PLoM preprocessing, the sample size was increased by a factor of 20, resulting in 45,000 groups for subsequent Gradient Boosting Tree (GBT) model training.



(a) (b)

Fig. 11 (a) Density plot of PIDR and (b) peak roof displacement

Table 5 Hyperparameter settings

Hyperparameter	Meaning	Initial range	Value interval
n_estimators	Number of decision trees	[20,200]	20
max_depth	Maximum depth of decision trees	[1,100]	5
min_samples_split	Minimum samples per node	[2,20]	1
min_samples_leaf	Minimum samples per leaf	[2,20]	1
max_features	Maximum number of features to select	[1,20]	1

3.3.2 Training and testing the gradient boosting tree model

The GBT model was optimized using random grid search. The hyperparameter settings are detailed in Table 5.

To evaluate the model's performance, the 562-group test set was used for prediction. The prediction distributions were analyzed, and the results are shown in the following figures.

Fig. 11 show the density plots of the predicted and actual values for the PIDR and peak roof displacement, respectively. Fig. 12 display the density plots of PFA, PFV, and PFD for the first floor, middle floor, and top floor, respectively. Each plot provides the mean and standard deviation of both the predicted and actual values.

From these figures, it is evident that the prediction model successfully captures the distribution characteristics of the dataset. The maximum error in the mean values between the predicted and actual responses is 14.6%, while the minimum error is only 2.4%. The dispersion between the predicted data and the actual values also shows good consistency.

4. Discussions

- SHAP value feature importance is an effective method for interpreting machine learning mode ls (Lundberg *et al.* 2017). Fig. 13 shows the SHAP feature importance for the GBT model on the test dataset, where each bar of different colors represents the absolute SHAP

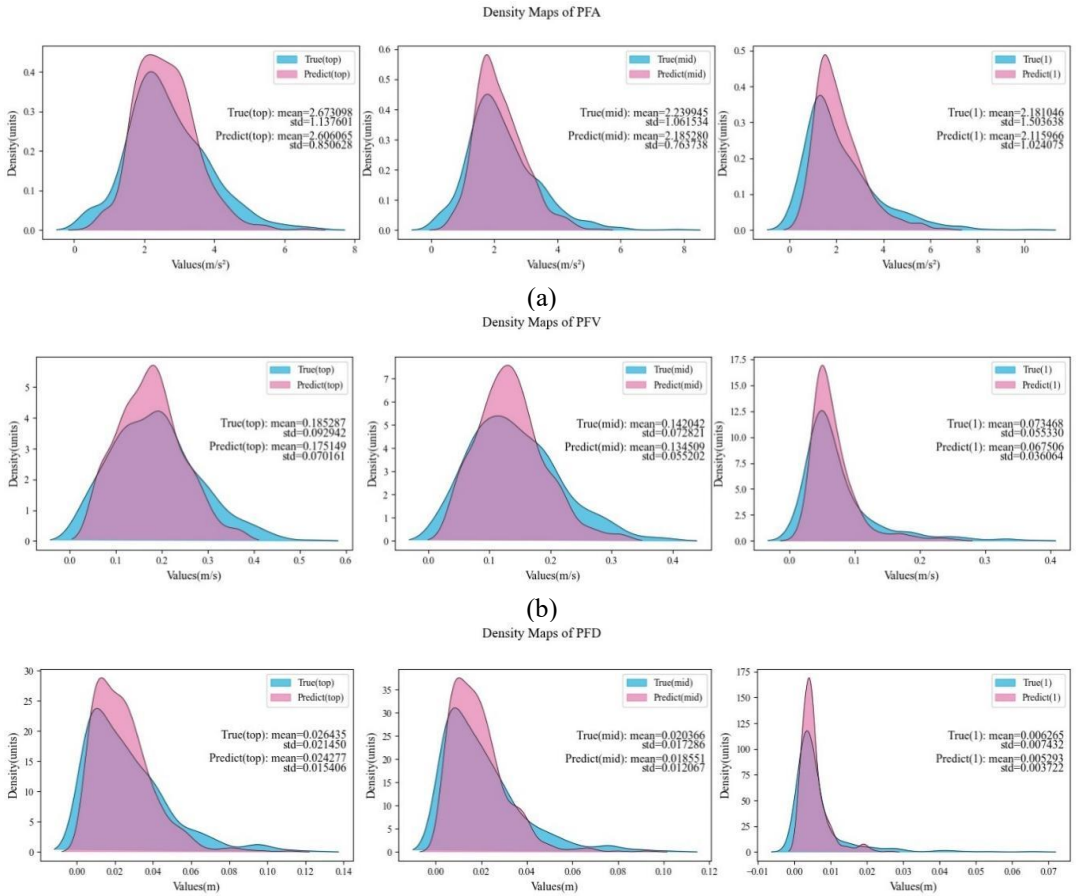


Fig. 12 Density plot of (a) PFA, (b) PFV, (c) PFD of the first floor, middle floor, and top floor

value for the corresponding target variable. The vertical axis in the figure lists the feature abbreviations, with detailed information for each feature provided in Table 1. The features are arranged in descending order of importance. The results in Fig. 13 indicate that V/A, MIV, D_{rms}, SMV, PGV, and PSV are the six most critical features for the response predictions in this study, which closely aligns with the feature importance rankings discussed in Section 4.3. Specifically, PSV, HI, and MIV are the most important features for predicting the maximum inter-story drift ratio and the maximum roof displacement; V/A and MIV are crucial for predicting PFA; MIV and PSV are key for predicting PFV; and PSV and HI are essential for predicting PFD.

- In Section 4, different ML models were trained on a small-scale dataset and validated on the test set. Table 6 summarizes the performance of the six models on the test set, where MAPE is the average of the percentage of prediction error relative to the true value, representing the proportion of error to the true value. It is evident that the SVR and SVR+PLoM models had fewer than 20 correct predictions. Without the assistance of PLoM, the SVR model could only correctly predict 13 instances, significantly lagging behind the RF and GBT models.

Table 6 Performance of six ML models in Section 4

ML Model	Performance Metrics					
	MAE	MSE	RMSE	MAPE(%)*	R^2	Num of Correct Prediction
SVR	8.2×10^{-4}	1.6×10^{-6}	1.28×10^{-3}	60.44	0.68	6
SVR+PLoM	7.7×10^{-4}	1.6×10^{-6}	1.27×10^{-3}	57.82	0.69	12
RF	7.5×10^{-4}	1.9×10^{-6}	1.37×10^{-3}	47.22	0.63	10
RF+PLoM	8.1×10^{-4}	1.8×10^{-6}	1.34×10^{-3}	50.33	0.65	11
GBT	6.8×10^{-4}	1.2×10^{-6}	1.11×10^{-3}	50.08	0.76	12
GBT+PLoM	6.5×10^{-4}	1.1×10^{-6}	1.01×10^{-3}	51.98	0.79	14

*$\text{MAPE}(\%)=1/n\sum_{i=1}^{n}|(y_i-\hat{y}_i)/y_i|$

This is because SVR aims to find an optimal hyperplane in a high-dimensional space for regression, which works well for linear and simple nonlinear problems. However, for complex nonlinear problems like structural seismic response, SVR may require appropriate kernel functions and hyperparameters, and its performance is highly dependent on the data distribution, leading to suboptimal performance compared to RF and GBT. This is consistent with the conclusions from several related papers. In the study by Bhatta (2023), the prediction accuracy of Random Forest (RF) on the test dataset was 6% higher than that of Support Vector Machine (SVM). Ding (2023) used various machine learning models to predict the response of a 3x6 layer reinforced concrete (RC) framework structure, and the results of the gradient boosting model outperformed the SVM model by more than 10% in terms of the R^2 index. The comparison reveals that the GBT and GBT+PLoM models perform better, with R^2 values exceeding 0.75 and stable correct prediction counts. Both RF and GBT models capture different features and patterns in the data by constructing decision trees and training on various subsets. GBT, in particular, builds decision trees incrementally, correcting errors from previous iterations, which optimizes model performance and effectively captures complex patterns and nonlinear relationships.

In this study, even the best model among the six, GBT+PLoM, only achieved an R^2 of 0.79 and could not make more than 15 correct predictions within the 10% allowable error margin from the test set. Although these models were trained on a small-scale dataset, this highlights the randomness and complexity of structural behavior under seismic action. Despite the advanced machine learning models providing a certain degree of predictive capability, challenges remain in addressing the highly complex and random nature of seismic responses. Further research and data accumulation are needed to improve the models' predictive performance and reliability. Additionally, the development and enhancement of probabilistic prediction methods remain crucial. By incorporating uncertainty quantification and probabilistic models, we can better describe the complexity and uncertainty of seismic responses, providing more comprehensive and reliable prediction results. This will aid in making better-informed decisions in post-earthquake response and mitigation.

5. Conclusions

This study proposes a probabilistic prediction method for the response of frame structures

under random seismic events based on machine learning. The approach combines manifold-based probability learning (PLoM) for generating samples with a Gradient Boosting Tree (GBT) model. The structural response database considers the uncertainties and variability of seismic motions and the complexity of the structures, selecting 284 ground motion records and establishing 33 multi-degree-of-freedom (MDOF) models for frames with different periods. The feature engineering process used the maximum displacement of four single-degree-of-freedom (SDOF) structures under the 284 ground motion records and the mean responses of 33 frame structures under the same records. From this, 11 ground motion feature parameters and one structural feature parameter were selected through correlation analysis. In the machine learning model section, three basic models—Support Vector Regression (SVR), Random Forest (RF), and Gradient Boosting Tree (GBT)—and their combinations with PLoM-generated samples (SVR+PLoM, RF+PLoM, GBT+PLoM) were considered. These models were trained using the response data of a 7-story frame under 284 ground motion records. The best-performing model was found to be GBT+PLoM. This model was then retrained and used to predict structural responses under random seismic events, showing good performance in accurately reflecting the distribution characteristics of the structural responses. The predicted data's concentration and dispersion were consistent with the test set. The key findings include:

- Randomness of Structural Response under Seismic Events: The study highlights the significant randomness in structural responses under seismic motions. Traditional deterministic methods for estimating structural seismic responses can be overly conservative or lack safety margins. By quantitatively analyzing the uncertainties of seismic motions and structural responses, the proposed method predicts the probabilistic distribution of structural seismic responses, avoiding the shortcomings of traditional methods in accounting for uncertainties and complex scenarios. This provides more reliable support for decision-makers.
- Innovative Use of PLoM: One of the study's innovations is the use of PLoM, which improves the predictive capability of GBT regarding data distribution. This improvement was demonstrated using the PFA, PFV, and PFD of a 12-story frame structure as the dataset. Additionally, the study's dataset considered 284 ground motions and 33 frame structures, suggesting that the method can achieve higher accuracy with more ground motions and structural considerations in the future. The approach can also be applied to predict seismic damage for buildings made of various materials and structures.
- The ML model developed in this study can be used to derive vulnerability curves, a process that is crucial for assessing the likelihood of structural damage under different earthquake intensities. These curves not only show the probability of building damage at a specific earthquake intensity but also provide important insights for designing and strengthening strategies. Furthermore, the analysis of vulnerability curves can help engineers and decision-makers better understand the performance of buildings during earthquakes, offering a scientific basis for developing appropriate disaster mitigation measures, thereby enhancing the seismic resilience and safety of cities.
- The predictive method in this study provides an effective reference for the rapid assessment of structural safety after earthquakes. By incorporating advanced machine learning algorithms, this method can quickly process post-earthquake information to evaluate the extent of building damage and safety based on predicted results. Relying on these rapid assessments, decision-makers can promptly take necessary actions, such as initiating evacuations, rescue operations, or reinforcement measures, thereby effectively reducing potential risks. Moreover, this predictive method offers a basis for policymakers to develop effective risk management

policies and emergency response plans, aimed at minimizing the economic losses and casualties caused by earthquakes. Overall, this innovative assessment approach not only enhances the efficiency of earthquake risk management but also provides valuable insights for future urban construction and planning.

The combination of PLoM and GBT provides a robust framework for probabilistic prediction of structural response under random earthquake, and provides a new idea in seismic response analysis and prediction.

References

Bhatta, S., Liu, J., Zhang, H. and Wang, J. (2023), "Seismic damage prediction of RC buildings using machine learning", *Earthq. Eng. Struct. D.*, **52**(10), 2242-2262. https://doi.org/10.1002/eqe.3907.

Bowman, A.D., Prabhakar, S.P. and Jololian, L. (2022), "A framework for an automated development environment to support the data-driven machine learning paradigm", *Proceedings of Southeast Conference 2022*, 9764094. https://doi.org/10.1109/SoutheastCon48659.2022.9764094.

Byun, N., Lee, J., Lee, K. and Kang, Y.J. (2023), "Extended artificial neural network for estimating the global response of a cable-stayed bridge based on limited multi-response data", *Smart Struct. Syst.*, **32**(4), 235-251. https://doi.org/10.12989/sss.2023.32.4.235.

Dail, H. and Wang, C. (2021), "Convolutional neural network estimation of bridge linear elastic seismic response", *Eng. Earthq. Eng.*, **6**(1), 188-199. https://doi.org/10.13197/j.eeev.2021.04.188.dail.019.

De Domenico, D., Falsone, G. and Ricciardi, G. (2018), "Improved response-spectrum analysis of base-isolated buildings: A substructure-based response spectrum method", *Eng. Struct.*, **162**, 198-212. https://doi.org/10.1016/j.engstruct.2018.02.037.

Ding, J.Y., Feng, D.C., Luo, Z. and Ghosh, J. (2023), "Efficient seismic fragility analysis method utilizing ground motion clustering and probabilistic machine learning", *Eng. Struct.*, **294**, 116739. https://doi.org/10.1016/j.engstruct.2023.116739.

Fidarova, M.I., Zaalishvili, V.B. and Melkov, D.A. (2023), "Correlation between the magnitude of macroseismic intensity and various indicators of instrumental records of fluctuations in the soil stratum", *Geol. Geophys. Russian South*, **13**(1), 59-75.

Guo, J., Enokida, R., Li, D. and Ikago, K. (2021), "Combination of physics-based and data-driven modeling for nonlinear structural seismic response prediction through deep residual learning", *Earthq. Eng. Struct. D.*, **50**(13), 3601-3624. https://doi.org/10.1002/eqe.3863.

Hwang, S.H., Eom, T. and Lee, D.G. (2021), "Machine learning-based approaches for seismic demand and collapse of ductile reinforced concrete building frames", *J. Build. Eng.*, **34**, 101905. https://doi.org/10.1016/j.jobe.2020.101905.

Kalakonas, P. and Silva, V. (2021), "Seismic vulnerability modelling of building portfolios using artificial neural networks", *Earthq. Eng. Struct. D.*, **50**(2), 310-327. https://doi.org/10.1002/eqe.3567.

Khan, T., Rabbani, M., Siddiquee, S.M.T. and Majumder, A. (2019), "An Innovative Data Mining Approach for Determine Earthquake Probability Based on Linear Regression Algorithm", *Proceedings of the 2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, 1-6. https://doi.org/10.1109/ICECCT.2019.8869286.

Kim, T., Song, J. and Kwon, O.S. (2020), "Probabilistic evaluation of seismic responses using deep learning method", *Struct. Saf.*, **84**, 101913. https://doi.org/10.1016/j.strusafe.2019.101913.

Lu, X., Xu, Z., Xiong, C. and Zeng, X. (2017), "High performance computing for regional building seismic damage simulation", *Procedia Eng.*, **198**, 836-844. https://doi.org/10.1016/j.proeng.2017.07.134.

Lundberg, S.M. and Lee, S.I. (2017), "A unified approach to interpreting model predictions", *Adv. Neural Inform. Process. Syst.*, 30. Available at: NeurIPS 2017 Proceedings.

Mangalathu, S., Hwang, S.H. and Jeon, J.S. (2020), "Failure mode and effects analysis of RC members

based on machine-learning-based SHapley Additive exPlanations (SHAP) approach", *Eng. Struct*., **219**, 110927. https://doi.org/10.1016/j.engstruct.2020.110927.

Mao, J., Su, X., Gui, G., Wang, H., Fu, Y. and Li, D. (2024), "ResNet transfer learning for accurate and efficient anomaly detection of bridge vibration data", *Smart Struct. Syst.*, **34**(6), 415-429. https://doi.org/10.12989/sss.2024.34.6.415.

Patil, P.S., Kappuram, K. and Bari, P. (2022), "Development of AMES: Automated ML expert system", *Proceedings of the International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON)*, **1**, 208-213. https://doi.org/10.1109/COM-IT-CON54601.2022.9850737.

Rodgers, J.L. and Nicewander, W.A. (1988), "Thirteen ways to look at the correlation coefficient", *The American Statistician*, **42**(1), 59-66. https://doi.org/10.1080/00031305.1988.10475524.

Soize, C. and Ghanem, R. (2017), "Polynomial chaos representation of databases on manifolds", *J. Comput. Phys.*, **343**, 197-221. http://dx.doi.org/10.1016/j.jcp.2017.01.031.

Soize, C. and Ghanem, R. (2019), "Physics-constrained non-Gaussian probabilistic learning on manifolds", *Int. J. Numer. Method. Eng.*, **119**(7), 604-632. https://doi.org/10.1002/nme.6202.

Soize, C. and Ghanem, R. (2020), "Sampling of Bayesian posteriors with a non-Gaussian probabilistic learning on manifolds from a small dataset", *Stat. Risk Model.*, **37**(1), 117-140. https://doi.org/10.1007/s11222-020-09954-6.

Soize, C. and Ghanem, R. (2016), "Data-driven probability concentration and sampling on manifold", *J. Comput. Phys.*, **321**, 242-258. https://doi.org/10.1016/j.jcp.2016.05.001.

Soize, C. and Ghanem, R. (2019), "Physics-constrained non-Gaussian probabilistic learning on manifolds", *Int. J. Numer. Method. Eng.*, **121**(1), 110-145. https://doi.org/10.1002/nme.6202.

Torky, A.A. and Ohno, S. (2021), "Deep learning techniques for predicting nonlinear multi-component seismic responses of structural buildings", *Comput. Struct.*, **252**, 106570. https://doi.org/10.1016/j.compstruc.2021.106570.

Wang, F. and Zhao, J. (2024), "Predicting the mechanical properties of high-performance concrete implementing boosting models integrated with metaheuristic algorithms", *Smart Struct. Syst*., **34**(6), 377-406. https://doi.org/10.12989/sss.2024.34.6.377.

Wang, T., Li, H., Noori, M., Ghiasi, R. and Altabey, W. A. (2022), "Probabilistic seismic response prediction of three-dimensional structures based on Bayesian convolutional neural network", *Sensors*, **22**(10), 3775. https://doi.org/10.3390/s22103775.

Wu, K., Liu, Y., Zhang, H. and Zhang, L. (2017), "Study on backbone curves of RC frame model for simulation of city-scale seismic responses of buildings", *Proceedings of the 26th National Structural Engineering Conference*, Changsha, China.

Xing, F., Du, W.L., Li, G., Dong, Z.Q. and Li, H.N. (2024), "A data-driven method for the reliability analysis of a transmission line under wind loads", *Steel Compos. Struct.*, **52**(4), 461-473. https://doi.org/10.12989/scs.2024.52.4.461.

Xiong, C. (2016), "Study on the regional building seismic damage simulation based on time-history analysis and 3D scene visualization", Ph.D. Dissertation; Tsinghua University, Beijing, China.

Yang, L., Fu, Z. and Wang, D. (2022), "Ground motion time history simulation for seismic response history analysis", *Front. Earth Sci.*, **10**, 908498. https://doi.org/10.3389/feart.2022.908498.

Yue, H., Zhang, S. and Li, X. (2019), "The source rupture model of large earthquakes: Progress and prospects in fast response and joint inversion techniques", *Science China Technol. Sci.*, **49**(6), 509-518. https://doi.org/10.1360/SSTe-2019-0078.

Zhang, T., Xu, W., Chen, G., Yang, S. and Zhang, Z. (2023), "Seismic response prediction of a damped structure based on data-driven machine learning methods", *Eng. Struct.*, **202**, 117264. https://doi.org/10.1016/j.engstruct.2023.117264.

Zhang, Y., Wang, Y., Wang, W. and Zhang, X. (2018), "A machine learning framework for assessing post-earthquake structural safety", *Struct. Saf.*, **72**, 1-16. https://doi.org/10.1016/j.strusafe.2018.01.002.

Zhong, K., Navarro, J.G., Govindjee, S. and Deierlein, G.G. (2023), "Surrogate modeling of structural seismic response using probabilistic learning on manifolds", *Earthq. Eng. Struct. D.*, **52**(8), 1923-1944. https://doi.org/10.1002/eqe.3839.

Zhong, K., Navarro, J.G., Govindjee, S. and Deierlein, G.G. (2023), "Surrogate modeling of structural seismic response using probabilistic learning on manifolds", *Earthq. Eng. Struct. D.*, **52**(8), 2407-2428. https://doi.org/10.1002/eqe.3839.

Zhu, L.H., Li, G. and Dong, Z.Q. (2021), "Dynamic test and numerical simulation on avoiding the weak-story failure mechanism in structures using LSFDs", *Steel Compos. Struct.*, **40**(2), 175-191. https://doi.org/10.12989/scs.2021.40.2.175.