DOI: 10.1111/mice.13015

RESEARCH ARTICLE

COMPUTER-AIDED CIVIL AND INFRASTRUCTURE ENGINEERING WILEY

Explainable probabilistic deep learning framework for seismic assessment of structures using distribution-free prediction intervals

Mohamed Noureldin[1] | **Tamer Abuhmed**[2] | **Melike Saygi**[1] | **Jinkoo Kim**[3]

[1]Department of Civil & Architectural Engineering, Sungkyunkwan University, Suwon, Republic of Korea

[2]Department of Computer Science and Engineering–College of Computing and Informatics, Sungkyunkwan University, Suwon, Republic of Korea

[3]Department of Global Smart City, Sungkyunkwan University, Suwon, Republic of Korea

Correspondence
Jinkoo Kim, Department of Civil & Architectural Engineering, Sungkyunkwan University, Suwon, Republic of Korea.
Email: jkim12@skku.edu

Tamer Abuhmed, College of Computing and Informatics, Sungkyunkwan University, Suwon, Republic of Korea.
Email: tamer@skku.edu

Funding information
National Research Foundation of Korea, Grant/Award Numbers: 2021R1A2C1011198, 2021R1A2C2006631

Abstract

A new probabilistic framework is proposed for providing a distribution-free prediction interval (PI) of seismic responses required for various earthquake engineering applications. The framework overcomes the limitation of point prediction models and avoids the complexity of traditional probabilistic methods. The framework utilizes a few assumptions of probability distributions and requires no prior assumed statistical distribution for the PI. Ensemble probabilistic deep learning models (DLMs) are used to provide quality-driven PIs of seismic responses for low- to mid-rise buildings with limited irregularity. Considering these systems and ground motions with the aid of Monte Carlo simulation and nonlinear time-history analysis (NLTHA), huge datasets are generated for training. To have an insight into the probabilistic DLM, explainable artificial intelligence techniques are used. The superiority of the proposed framework in quantifying uncertainties is validated by comparison with the conventional Bayesian method. In addition, its applicability is investigated by providing bounds of seismic fragility curves, life cycle cost, and resilience index obtained by NLTHA for a benchmark case study model. The results showed that the proposed framework is robust and outperforms the conventional Bayesian method in uncertainty quantification for the considered dataset.

1 | INTRODUCTION

Uncertainty propagation has become an important aspect of modern seismic design and assessment of buildings, especially for next-generation performance-based seismic design, where uncertainties and probable consequences are explicitly considered (FEMA-P58, 2018). Generally, the probabilistic relationship between seismic intensity and the corresponding engineering demand parameters (EDPs) is considered in modern seismic codes and guidelines by frameworks that require conducting numerous nonlinear time-history analyses (NLTHAs). To overcome the huge computational burden, some modern codes (e.g., ASCE 41, 2017) recommend nonlinear static procedures (NSPs) to obtain seismic demands. However, NSPs are based on a limited number of data (for both earthquakes and structural systems) to form simple regression equations that have little information about the overall uncertainty involved (FEMA 440, 2005). In addition, NSPs come with a number of assumptions and simplifications that may not accurately reflect the actual behavior of a structure during seismic excitation. They are generally

only applicable to simple structural systems and may not be suitable for systems with more complex configurations.

Machine learning and deep learning models (DLMs) are gaining momentum in civil engineering applications. A few to mention are the prediction of material properties (Alam et al., 2020; Rafiei & Adeli, 2017a, 2017b; Rafiei et al., 2017), assessment of civil infrastructure resilience (Franchin & Cavalieri, 2015), ground motion prediction equation (Mu & Yuen, 2016), seismic reliability analysis (Nabian & Meidani, 2018), and structural damage detection (Eltouny & Liang, 2021). Many previous studies focused on using machine learning techniques to predict dynamic responses and assess the performance of structures (e.g., Li et al., 2021; Sun et al., 2021; Xie et al., 2020). Other studies focused on the application of DLMs to predict the dynamic response of systems (e.g., Feng et al., 2021; Stoffel et al., 2020; Wu et al., 2019; Zhang et al., 2019). Previous studies generally focused on the deterministic prediction of responses (e.g., Ali et al., 2023; Noureldin et al., 2023; Payán et al., 2017). Moreover, predicting the uncertainty of responses is highly affected by the input feature variability, which means that the heteroscedastic assumption needs to be considered, not the homoscedastic one. Based on that, the variance of the input data is considered a variable that assumes different values across the input data space (Kendall & Gal, 2017). This assumption is more suitable for earthquake engineering applications where the ground motion intensity measure is naturally uncertain (aleatory uncertainty).

Recent studies (e.g., Kim et al., 2020) used the Bayesian DLM to predict seismic responses of single degree of freedom (SDOF) systems. These studies considered input randomness only and assumed that the model uncertainty is reduced by using a huge number of training data. This assumption may not be suitable in case of limited training data, which is the case in many earthquake engineering applications. In addition, in many cases, huge training data may not be enough for reducing model uncertainty. This is because model uncertainty stems from three sources: bias (model misspecification), variance (training data uncertainty), and parameter uncertainty. Bias measures the error in predictions, variance reflects the representativeness of the sample data, and parameter uncertainty reflects the effect of nonoptimal parameters on the model's predictions, particularly in sparse regions of the input space (Pearce et al., 2018). The Monte Carlo dropout technique may be used for handling model uncertainty, but it requires assuming a probabilistic distribution of the hyperparameters in advance (Kendall & Gal, 2017), which might not be very accurate when not enough information is available.

A few limitations can be observed, in general, in recent studies concerning uncertainty quantification using DLMs in the earthquake engineering field. For example, most studies use "readymade" DLMs such as convolutional neural network (CNN; e.g., Kim et al., 2020), which is not tailored for uncertainty quantification and requires large dataset and computation resources that make it more suitable for image processing, classification, and so forth. Another limitation is related to the type of structural system that is commonly used in uncertainty quantification. SDOF systems are commonly used for simplified analysis; however, quantifying the latent uncertainty in the transformation of the SDOF system responses to multi-degree of freedom (MDOF) system responses is commonly overlooked. Another important limitation is to assume a specific distribution (e.g., Gaussian) for model variables, which may not be the case for all earthquake engineering applications and structural damage results (Ghiasi et al., 2021).

Earthquake engineering applications such as seismic fragility, life cycle cost (LCC), and resilience index (RI) require the calculation of prediction interval (PI) rather than point prediction to estimate lower and upper bounds. For example, lower-bound seismic fragility is crucial for critically important structures such as nuclear power plants. Upper-bound LCC is highly required by building owners to prepare a suitable budget for their assets. Upper and lower bound values of RI are highly important for hospitals and power stations, which are crucial for overall community resilience after natural hazards, and for decision-makers to allocate available resources efficiently. In addition, probability-based information may not be convenient to decision-makers who prefer lower and upper bounds for better-informed decisions (FEMA-P58, 2018). Moreover, PI is considered a more robust uncertainty quantification alternative than confidence interval (CI) since PI incorporates more uncertainties and is more likely to capture the true value in estimation than CI (Kabir et al., 2018). Also, PI considers both the epistemic input uncertainty and the aleatory model uncertainty, whereas CI considers model uncertainty only (Pearce et al., 2018). Moreover, designers and stakeholders may be highly interested in how much influence an important input feature has on the output response; in other words, the degree of association between predictors (input) and output (response). This means that not only "accuracy" of the DLM is important, but also "explainability" plays an important role (Sun et al., 2021).

This study provides an innovative framework for researchers to account for seismic response uncertainty while maintaining the simplicity and accuracy required for various earthquake engineering applications (for low- to mid-rise buildings with limited irregularity). In the current study:

1. A new probabilistic explainable quality-driven (distribution-free) PI ensemble deep neural network (NN; for short QDN: quality-driven deep NN) model is proposed to predict structural responses considering different types of uncertainties. The main advantage of QDN over NLTHA is that it provides both upper and lower bounds for the EDPs and can capture the nonlinear relationship between input parameters and responses within the domain of the input space.
2. A total of 188,000 dataset points are used for training, validation, and testing (137,000 dataset points are obtained using NLTHAs, and 50,000 dataset points are obtained using Monte Carlo simulation [MCS]).
3. Explainable artificial intelligence (XAI) techniques such as partial dependence plots (PDPs), individual conditional expectation (ICE), and analysis of variance (ANOVA) are used to highlight the important features and to explain the degree of association between the input features of the earthquake and the output EDPs.
4. The framework accounts for heteroscedastic uncertainties stemming from the seismic response prediction as well as the mapping between SDOF and MDOF system responses.
5. NLTHAs are conducted for benchmark models to validate the framework.

2 | THE PROPOSED PROCEDURE

2.1 | The steps of the overall framework

Figure 1 shows the overall procedure of the proposed framework. The main input feature used for the DLM is related to the structural and earthquake features. The former includes structural stiffness and strength. The latter includes the frequency content (FC; represented by response spectrum), peak ground acceleration (PGA), peak ground velocity (PGV), peak ground displacement (PGD), magnitude, fault distance, Arias intensity (AI), fault type, 30-m average shear wave-velocity (Vs_{30}), lowest usable frequency, and duration. The main output prediction of the DLMs is three EDPs, which are maximum displacement (D), maximum acceleration (A), and maximum base shear (V) of the MDOF system. The framework can be summarized in the following steps:

1. In the first step, the main structural input features of different types of structural systems and earthquake events are prepared to cover a wide range of systems and ground motion excitations. Structural systems (represented as 900 SDOF systems) with varying stiffness and strength are utilized, with stiffness represented by the systems' natural period (ranging from 0.45 to 4 s in 30 increments) and strength represented by the system strength ratio (varying from 0.04 to 0.9 in 30 intervals). All input features are used at this stage for the earthquake event, such as magnitude, rupture distance, PGA, and so forth. Numerous NLTHAs are conducted at this stage (138,000 runs) using a wide range of SDOF systems. Three different EDPs are obtained from these analyses, which are D, A, and V.
2. The most important input features are selected by applying the ANOVA method. The selected input features will be used for both the probabilistic DLM (QDN1, Step 4) and to explain the relationship between the input features and the final output EDPs using XAI (Step 6). This is to ensure that the model is trained on the most relevant and informative features in the data. By selecting the most important features, the model's accuracy and efficiency will be improved.
3. Another dataset will be prepared for the second DLM (QDN2, Step 5). In this dataset, different types of multistory frame systems are selected to represent low- to mid-rise frame systems. NLTHAs will be conducted on MDOF systems and their equivalent SDOF systems to obtain D, A, and V. The relationship between the same EDP obtained from both SDOF and MDOF systems is plotted. The input and output of the dataset are the EDP obtained from the equivalent SDOF and MDOF systems, respectively. Then this dataset is augmented using MCS, considering different sources of uncertainties such as construction quality assurance uncertainty and nonlinear analytical structural modeling uncertainty.
4. In this step, the first probabilistic DLM (QDN1) is trained, validated, and tested to provide a PI for each output response variable for all EDPs (i.e., D, A, and V) of the SDOF systems obtained from Step 1 using the selected input features obtained from step 2. QDN1 quantifies the uncertainties stemming from data uncertainty, model optimization uncertainty, and misspecification bias. Details of the uncertainty sources and the architecture of the QDN1 will be provided in the next sections. The obtained PI takes into consideration the aforementioned uncertainties for each output response value of the dataset.
5. In this step, the second probabilistic DLM (QDN2) is trained, validated, and tested to provide the PIs of the responses of the MDOFs. These PIs will be obtained from QDN2 using the PIs obtained from the previous step (Step 4) and the augmented dataset prepared in Step 3.
6. In this step, the degree of association between the important input features obtained from Step 2 and the output response variable PIs is investigated using XAI techniques. This relationship between the important input features and the response variables is important

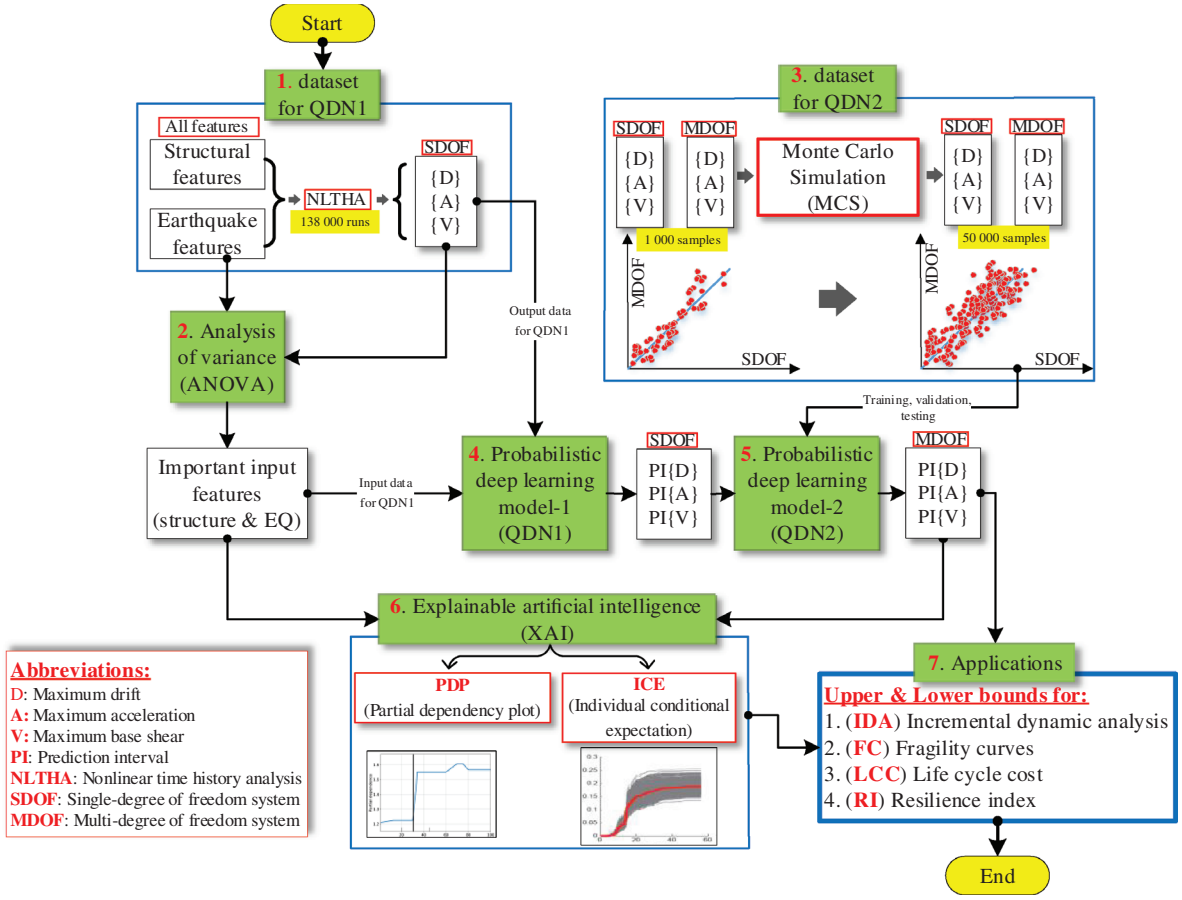


FIGURE 1 Overall procedure of the proposed framework.

in explaining the DLM's outcome and indicates the input thresholds that make a significant change in the output response, which will support decision-making for the application in the next step.

7. In this step, the PIs of the response variables obtained from Step 5 are used in any specific application to provide the upper and lower bounds of the assessment tool under consideration. The degree of association and thresholds obtained from the previous step will provide insights that will support the decision-making process related to the bounds of the assessment metric considered. In the current study, seismic fragility curves, LCC, and RI are used as assessment metrics.

2.2 | Uncertainty propagation and evaluation metrics

2.2.1 | Uncertainty sources

Commonly, two main types of uncertainties should be considered when a DLM is used for regression. The first is the model uncertainty (i.e., DLM estimated hyperparameter uncertainties), and the second is the input data uncertainty (i.e., inherent randomness in the input data parameters; Kendall & Gal, 2017). The former is considered epistemic uncertainty, whereas the latter is considered an aleatoric uncertainty (Pearce et al., 2018). The proposed QDN model estimates the total variance of predicted target responses σ^2_{target} (Equation 1; Pearce et al., 2018):

$$\sigma^2_{target} = \sigma^2_{model} + \sigma^2_{data} \quad (1)$$

where σ^2_{model} and σ^2_{data} are model variance (also termed model uncertainty or epistemic uncertainty) and data variance (also termed data noise variance or aleatoric uncertainty).

There are three main sources of model uncertainty; namely, model misspecification (or bias), training data uncertainty (or variance), and parameter uncertainty. Model misspecification refers to a situation in which the statistical model used to make predictions does not accurately represent the underlying relationship between the input variables and the response variable. This can occur when important variables are omitted from the model, the functional form of the relationships between variables

is incorrect, or other assumptions of the model are not met. Training data uncertainty reflects how the selected sample data are representative of the entire input dataset distribution and how much the model is sensitive to other samples in the input space. Parameter uncertainty reflects the effect of selecting nonoptimum parameters on the prediction potential of the model, especially in sparse regions of the input space. Many studies applying DLMs in earthquake engineering fields ignore model uncertainties (e.g., Kim et al., 2019, 2020; Zhang et al., 2019). The meaning of model uncertainty in the context refers to the uncertainty in the hyperparameters estimated by the DLM. It is considered an "epistemic uncertainty," meaning it arises from a lack of knowledge or incomplete understanding of the underlying model. In addition, when model uncertainty is considered, only data and parameter uncertainties are considered, and model misspecification is generally ignored (Pearce et al., 2018).

It is important to highlight that the conditional variance can be modeled following one of the two assumptions: homoscedasticity and heteroscedasticity. Homoscedasticity is a situation where the variance of the response variable is constant, leading to a simple linear regression model with a constant variance. In the current study, heteroscedasticity approach is used where the variance of the response variable changes, requiring a more complex regression model with a varying variance.

2.2.2 | Quality-driven loss function

The term "quality-driven" means that the proposed framework incorporates a gradient descent method, designed through qualitative assessment, and includes model uncertainty, which is different from the conventional lower-upper boundary estimation approach. The loss function used in the proposed framework is distribution-free; in other words, it requires no assumption with a specific distribution for the dataset. The loss function (i.e., the objective function) that needs to be minimized to obtain the optimal NN for a specific dataset can be defined as in Equation (2) (Pearce et al., 2018):

$$L = MPIW_{capt.} + \lambda \frac{n}{\alpha(1-\alpha)} \max\{0,\ (1-\alpha) - PICP\}^2 \tag{2}$$

where $MPIW_{capt.}$ is the captured mean PI width, $PICP$ is the PI coverage probability, λ is a Lagrangian multiplier that controls the relative importance of the width, compared to the coverage of the PI, n is the number of data points, and $(1-\alpha)$ is the desired proportion of coverage, where α is assumed 0.05.

The PI should be bounded by the predicted upper bound, $\hat{y}_U$, and lower bound, $\hat{y}_L$ (Equation 3):

$$\Pr(\hat{y}_{Li} < y_i < \hat{y}_{Ui}) \geq\ (1-\alpha) \tag{3}$$

where y_i is the target observation of an input covariate x_i, where $1 \leq i \leq n$.

The PI of each point should be calculated such that $MPIW_{capt.}$ is the minimum while maintaining $PICP \geq (1-\alpha)$. The $MPIW_{capt.}$ and $PICP$ can be mathematically quantified as in Equations (4) and (5).

$$MPIW_{capt.} = \frac{1}{c} \sum_{i=1}^{n} (\hat{y}_{Ui} - \hat{y}_{Li}) . k_i \tag{4}$$

$$PICP = \frac{c}{n} \tag{5}$$

where c is the total number of data points captured by the PI, $c = \sum_{i=1}^{n} k_i$, $\hat{y}_{Ui}$ and $\hat{y}_{Li}$ are the estimated upper and lower bounds of the point under consideration, and k_i is a binary variable $k_i \in \{0, 1\}$ that represents the occurrence of the data point within the estimated PI (Equation 6):

$$k_i = \begin{cases} 1, & if\ y_{Li} \leq y_i \leq y_{Li} \\ 0, & else \end{cases} \tag{6}$$

It is assumed that k_i can be represented as a Bernoulli random variable [i.e., $k_i \sim\ Bernoulli\ (1-\alpha)$. In addition, k_i is assumed independent and identically distributed variable. The former assumption can be used to justify that c can be represented by a Binomial distribution [i.e., $c \sim Binomial\ (n, (1-\alpha))$].

Utilizing the likelihood-based approach, the optimum NN parameters, θ, are optimized to maximize, $\mathcal{L}_\theta = \mathcal{L}(\theta|\mathbf{K}, \alpha)$, where $\mathbf{K}$ is the vector of length n with each element in the vector represented by k_i. Based on that, the probability mass function can be calculated as

$$\mathcal{L}_\theta = \binom{n}{c} (1-\alpha)^c \alpha^{n-c} \tag{7}$$

Using the central limit theorem and the negative log-likelihood, the following Equation (8) can be obtained:

$$-\log \mathcal{L}_\theta \propto \frac{(n(1-\alpha)-c)^2}{n\alpha(1-\alpha)} = \frac{n}{\alpha(1-\alpha)} ((1-\alpha) - PICP)^2 \tag{8}$$

Equation (8) is used to formulate the second term (on the right-hand side) in the main loss equation (Equation 2), considering that there should be a penalty applied if $PICP < (1-\alpha)$.

2.2.3 | Evaluation accuracy metrics

Two main characteristics of the PI are commonly used for assessing its quality, which are sharpness and calibration. The former is represented by the small width of the PI; whereas, the latter is represented by the coverage probability of the PI (Kabir et al., 2018). In the current study, more than one accuracy metric is used to demonstrate the uncertainty quantification potential of the proposed framework. $MPIW_{capt.}$ (for simplicity MPIW) and PICP will be used as accuracy metrics. Generally, the smaller the width, the lower the coverage probability. This means that the optimal PI is the one that maintains $PICP < (1-\alpha)$ with the smallest width (MPIW) possible. Another accuracy metric called coverage width-based criterion (CWC) is used, which combines MPIW and PICP (Equation 9):

$$CWC = MPIW\{1 + \gamma\,(PICP)\, e^{\eta(\mu - PICP)} \tag{9}$$

where $\gamma(PICP)$ is defined as in Equation (10):

$$\gamma\,(PICP) = \begin{cases} 1, & PICP < \mu \\ 0, & PICP \geq \mu \end{cases} \tag{10}$$

where $\eta = 50$ and $\mu = 1-\alpha$ are two hyperparameters. In addition, the point prediction potential of the proposed QDNs will be evaluated by the mean absolute error (MAE) as follows:

$$MAE = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n} \tag{11}$$

where y_i and $\hat{y}_i$ are the target and predicted observations of an input covariate x_i, where $1 \leq i \leq n$.

2.3 | First probabilistic DLM for SDOF system response PI

To estimate seismic responses of a structure subjected to earthquake excitation, nonlinear static or dynamic analyses are generally conducted. The former has a few limitations since it uses coefficients based on simple linear regression equations (using limited experiments and earthquake events) that add uncertain errors to the estimated response. The latter may suffer from convergence problems besides being sensitive to modeling uncertainty. The proposed framework offers advantages over NLTHA, such as providing both upper and lower bounds for the EDP, capturing nonlinear relationships between input parameters and response within the domain of the input space, and offering XAI-based explainability for greater transparency and understanding of the predictions. This means that not only will the framework provide accurate results, but it also offers greater transparency and understanding of the predictions made, making it easier to identify and address potential issues. In this section, the first probabilistic DLM QDN1 is developed to overcome the limitations of both nonlinear static and nonlinear dynamic analysis approaches. In addition, this model takes into consideration uncertainties stemming from different sources. The main components of QDN1, such as dataset, input features, and model architecture, will be discussed in the following sections.

2.3.1 | Dataset and nonlinear dynamic analysis (Step 1)

In the current study, the input dataset comprises a wide range of SDOF systems. Equation (12) describes mathematically the motion of an elastoplastic SDOF system subjected to a ground motion excitation:

$$\ddot{u} + 2\xi\varpi\dot{u} + \varpi^2 u_y\, \tilde{f}_S\,(u) = -\ddot{u}_g\,(t) \tag{12}$$

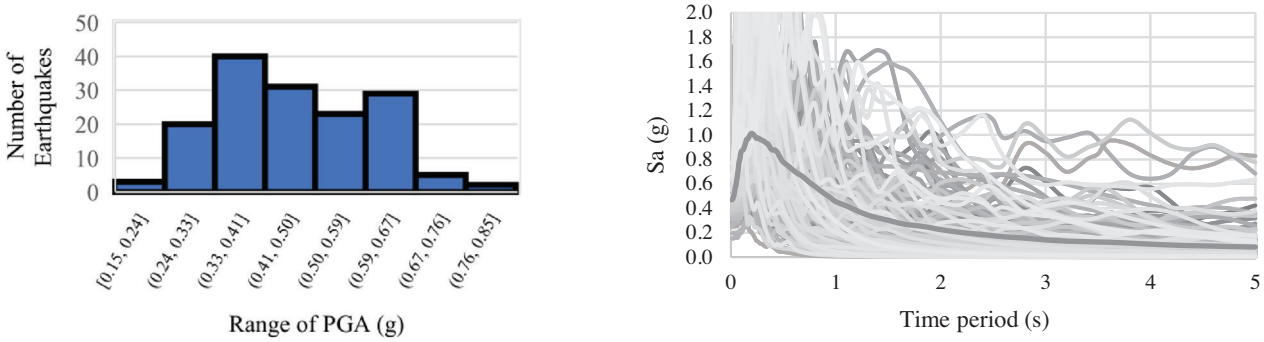
where ζ is the damping ratio; ω is the circular natural frequency; $\ddot{\mu}_g(t)$ is the ground excitation, and $u(t)$, $\dot{u}(t)$, and $\ddot{u}(t)$ denote the displacement, velocity, and acceleration, respectively; $\tilde{f}_S\,(u)$ represents the force–deformation relation in a partially dimensionless form of the elastoplastic SDOF system and can be represented as $\tilde{f}_S\,(u) = (f_S(u)/f_y)$, and $f_S(u)$, f_y, and u_y are the resisting force, yield strength, and yield displacement of an elastoplastic SDOF system, respectively.

Equation (12) shows that the controlling parameters that govern the dynamic response of an elastoplastic SDOF system subjected to a ground-excitation $\ddot{\mu}_g(t)$ are ξ, ω, and f_y. In the current study, the natural frequency (ω) will be represented by the natural period ($T = 2\pi/\omega$) of the elastoplastic SDOF system. The yield strength (f_y) of the elastoplastic SDOF system will be represented by the strength ratio ($SR = f_y/(m.g)$, which represents the ratio between the system strength (f_y) and its weight ($m.g$). The system's natural period and strength ratio are considered reliable parameters to represent building structure characteristics as indicated in several studies (FEMA 440, 2005; Gharehbaghi et al., 2020). A total of 900 SDOF systems are generated to cover practical ranges of T (0.45 to 4 s with 30 steps) and SR (0.04 to 0.9 with 30 steps). These ranges are considered a reasonable and practical representation of a wide range of real structures considering previous studies (FEMA 440, 2005). It is worth mentioning that structural engineers often employ elastoplastic models when characterizing structural behavior due to their efficiency and dependability. This stability may assist in enhancing the

TABLE 1 Ranges of the input earthquake parameters used for the input dataset.

		ESDOF		PGA	Magnitude	Source to site	Vs_{30}	Lowest useable	Source-fault
	Limit	*T* (s)	*SR*	(g)	(M_w)	distance (km)	(m/s)	frequency (Hz)	mechanism
Limits of parameters	Upper	4.0	0.9	1.8	7.62	218.13	1428.14	3.75	Normal; reverse; reverse oblique; strike-slip
	Lower	0.4	0.04	0.017	4.2	0.56	169.84	0.025	

Abbreviations: ESDOF, equivalent single degree of freedom; PGA, peak ground acceleration; *SR*, strength ratio; Vs_{30}, 30-m average shear wave-velocity.



(a) Histogram of the earthquake PGAs

(b) Response spectra of the earthquakes

FIGURE 2 Diversity and distribution of earthquake input dataset for the machine learning model. (a) Histogram of the earthquake peak ground accelerations (PGAs). (b) Response spectra of the earthquakes.

predictive capabilities of the modeling approach. If more sophisticated models that enable strength deterioration were incorporated for the nonlinear response of the structural components, the prediction of structural response could become more challenging. Despite this, the choice of elastoplastic models is justifiable as a basic step.

Different natural earthquake events (153 events), which are obtained from the PEER database (2021) and used in a previous study (Noureldin et al., 2022), are used to represent various earthquake characteristics (Table 1). Figure 2a,b, respectively, show the PGA histogram and response spectra of the ground motions used for QDN1. As shown in these figures, the selected earthquake events cover a wide range of PGA, with more events selected to represent the most commonly used range of design (approximately from 0.3 to 0.6 g). Moreover, a wide range of spectral accelerations is covered by the selected earthquakes as shown in Figure 2b. Numerous NLTHAs are conducted (900 systems x 153 earthquakes = 137,700 runs) to obtain maximum drift (*D*), *A*, and *V* of each system. OpenSees (2011) and Matlab (2020) programs are used to conduct NLTHAs with the aid of a parallel computing scheme utilizing multiple cores; more details can be found in Noureldin et al. (2022).

Since the dataset variables have different scales, a minimum–maximum normalization technique is used to have a uniform scale of all variables. This technique is a linear transformation mapping that preserves the original data characteristics. Equation (13) can be used for this normalization technique (Noureldin, Ali, et al., 2021):

$$y_n = \left(\frac{y_{un} - y_{omin}}{y_{omax} - y_{omin}}\right)(y_{nmax} - y_{nmin}) + y_{nmin} \quad (13)$$

where y_n and y_{un} are the normalized and un-normalized values, respectively; y_{omin} and y_{omax} are the minimum and maximum of the original (un-normalized) data range, respectively; and y_{nmin} and y_{nmax} are the minimum and the maximum of the normalized data range, respectively.

2.3.2 | Important feature selection using ANOVA (Step 2)

In the current study, several features are used to represent the ground motion excitations in the dataset such as the FC (represented by response spectrum), PGA, PGV, PGD, magnitude, fault distance, AI, fault type, Vs_{30}, lowest usable frequency, and duration. All of these input features can be used as input for the DLM, but it is a better practice to reduce the dimensionality of the input space by removing irrelevant or less important features. This will lead to improved permeance, faster processing, and a more easily understandable model. Important input features can be selected by investigating the relative importance of all

input features on the output response variable. This can be done by using the ANOVA method to quantify the statistical significance of the input feature on the response variable. In this method, the null hypothesis that a particular input feature has an insignificant effect on the response variable is statistically tested by decomposing the different sources of variance in the model. This statistical test considers how much influence one feature can have on another feature's effect on the response variable. In other words, the effect of interaction among the input features on the response variable is considered. The test statistic for the hypothesis of no differences in feature group means can be expressed as in Equation (14) (Montgomery, 2013):

$$F_0 = \frac{SS_{FG}/(a-1)}{SS_E/(N-a)} \tag{14}$$

where $SS_{FG} = \frac{1}{n}\sum_{i=1}^{a}(y_i^2 - \frac{y^2}{N})$; $SS_E = SS_T - SS_{FG}$; $SS_T = \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij}^2 - \frac{y^2}{N})$ where SS_{FG} is the sum of squares between input feature groups; SS_T is the total corrected sum of squares that measures the overall variability in the data; SS_E is the sum of squares due to error within input feature groups; a is the number of input feature groups (a feature group represents different levels of a single feature), y_{ij} is the jth observation obtained from a particular input feature level in the ith feature group; n is the number of observations (response outputs) in the ith group; N is the total number of observations; y_i is the total of observations under the ith feature group; and y is the total of all observations. The hypothesis test result can be represented as a p-value that provides the probability of the null hypothesis being true. Input features that significantly affect the output response variable should have a p-value of less than an arbitrary threshold (commonly 0.05 is used; Montgomery, 2013). The input features that have a significant effect on the response variable will be used in training the probabilistic DLMs (Steps 4 and 5) and in explaining the relationship between the important features and the output response variable under consideration using XAI techniques (Step 6).

2.3.3 | Architecture of the first probabilistic DLM (QDN1)

The application of DLMs in the earthquake engineering domain is gaining momentum; however, most previous studies are using available conventional DLMs such as CNN (Kim et al., 2020; Sun et al., 2021) and long-short term memory (e.g., Mangalathu & Burton, 2019; Xu et al., 2021; Zhang et al., 2019). These models are most suitable for image processing, classification, segmentation, computer vision, and so forth; however, they are not highly customized for uncertainty quantification, especially for seismic assessment applications. In addition, these conventional DLMs suffer from several limitations. For example, CNN requires a large dataset to process and train the NN, which might not be the case for earthquake engineering applications where limited datasets are available. In addition, CNN operates a down-sampling strategy called "max pooling," which significantly slows down the overall process and also does not account for important spatial hierarchies among input features (Xi et al., 2017).

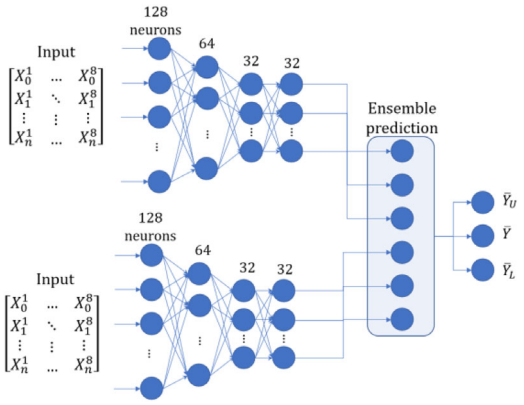


FIGURE 3 Architecture of probalistic deep learning model-1 (QDN1).

The important input features selected based on ANOVA (Step 2) will be used as input for QDN1, and the output will be a PI of each EDP, considering all sources of uncertainties discussed previously.

To assess the predictive uncertainty of the proposed probabilistic ensemble deep NN (QDN), we employ ensemble NNs with three outputs to construct the network (Lakshminarayanan et al., 2017; Pearce et al., 2018). Ensembles of models improve predictive performance and can capture more complex patterns in the data and reduce the chances of overfitting. Also, they can be more robust to changes in the data distribution as well as to variations in the NN models themselves. As shown in Figure 3, two NNs were designed and trained without resampling the data. Optimizing the number of ensembled NNs yields a faster and more accurate prediction. Each network receives earthquake input data (i.e., earthquake characteristics) and predicts a response variable for all EDPs (i.e., D, A, and V) as well as the lower-upper boundary of the response variable. The NN hyperparameters were optimized using Bayesian optimization with Gaussian processes. As shown in Figure 3, the final NN has four hidden layers of sizes 128, 64, 32, and 32 neurons. The model has batch size = 300, epochs = 140, learning rate (lr = 0.01), weight decay = 0.97, and Adamptive moment estimation (Adam) optimizer. The model uses the rectified linear unit

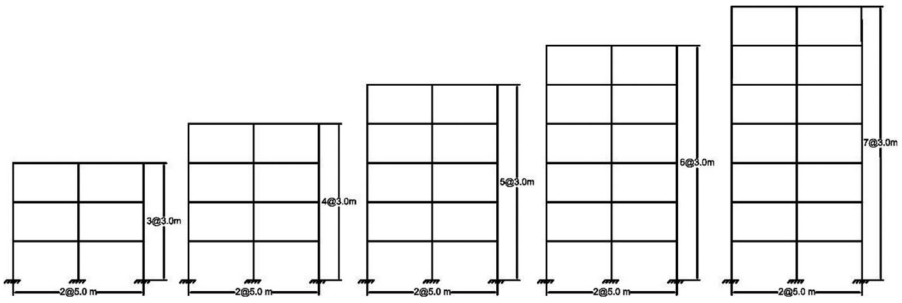


FIGURE 4 Structural frame models used for augmentation of the dataset.

(ReLU) activation function and optimizes the model using L loss function of Equation (2), where $\lambda = 15$ and $\alpha = 0.1$.

2.4 | Second probabilistic DLM for MDOF system response PI

To transform the maximum displacement of an MDOF system to its equivalent SDOF systems and vice versa, most current seismic codes and guidelines are using coefficients that are primarily based on the mass participation factor of the fundamental mode shape of the structure (ASCE 41-16). However, under a real scenario earthquake, the actual shape vector will change during the time-varying profile of the structure, especially when the structure is responding inelastically. Using coefficients based on linear regression functions for transformation between SDOF and MDOF systems may result in overly conservative seismic demands of actual structures (FEMA 440, 2005; Kim et al., 2019). In this section, the second probabilistic DLM (QDN2) is used to construct a nonlinear regression model that maps between SDOF and MDOF system responses taking into consideration different uncertainty sources. QDN2 is constructed using the output response PIs of the SDOF system obtained in Step 4 and the augmented dataset from Step 3 to obtain the MDOF system response PIs for the EDPs under consideration.

2.4.1 | Dataset of SDOF–MDOF response pairs using NLTHA

Since the main scope of the current study is low- to mid-rise buildings with limited irregularity, five structural framed 2D models are used to represent common prototype framed buildings as shown in Figure 4. The structural details of the model structures can be found in Noureldin et al. (2022). Acknowledging that it is practically impossible to include all building configurations in the study, the results obtained based on these models will be augmented in the following section and used for training QDN2.

Hundred earthquakes are selected randomly from the earthquake set used in Section 2.3.1. After that, NLTHAs are conducted for two versions of each model, which are the full structural model (i.e., MDOF system) and its corresponding equivalent SDOF system. OpenSees (McKenna et al., 2011) software is used for nonlinear modeling required for NLTHA. Beams and columns are modeled as elastic beam–column elements connected with zero-length elements at the ends. These elements represent the nonlinearity concentration during plastic hinge formation at the ends of beams and columns. For the zero-length element, two nodes at the same location are defined and connected by a uniaxial material element to represent the force–deformation relationship for the element. A uniaxial material element with bilinear behavior for flexure (without strength deterioration) is utilized for dynamic analysis. It is worth mentioning that elastoplastic force–deformation model provides reasonably accurate drift values when compared to other force–deformation models for structural system (FEMA 440 and ASCE 41-17). The lumped mass technique is used to distribute the modal mass of the structure to the nodes of the beams and columns (D'Angela et al., 2021). Newmark integration technique is used to implement the direct integration method for conducting NLTHA. Modal damping of 5.0% of the critical is used for nonlinear dynamic analyses considering a fixed base at the bottom of columns as the boundary condition. Rayleigh damping approach is used for NLTHAs, and the P-delta effect is not considered.

The proposed framework aims to offer a useful alternative for predicting the response of low- to medium-rise regular buildings rather than replacing NLTHA completely. The framework has the advantage of providing both upper and lower bounds for the results and offering an explanatory aspect through the XAI technique. Meanwhile, NLTHA results are limited to specific earthquake characteristics and do not show the uncertainty involved, requiring multiple nonlinear dynamic analyses to obtain reliable results. Importantly, the training and computational cost of the framework is done once, making future predictions quick and efficient.

2.4.2 | Augmented dataset using MCS

Several techniques are used to overcome the challenge of limited existing data for training DLMs by generating synthetic data. Among these techniques are generative adversarial networks (Goodfellow et al., 2014) and variational autoencoders (Kingma & Welling, 2014) and MCS (Sun et al., 2021). In the current study, augmentation of the dataset using MCS is used. It is not feasible to consider all building types, so the most relevant models are chosen to represent low- to mid-rise frame models.

Structural designers prefer simplified structural models such as bare frame models, which may create uncertainties in the structural behavior. Based on that, in the current study, three different sources of uncertainty are considered (Celarec & Dolšek, 2013). Structural modeling uncertainty originates from inaccuracies in component modeling, damping, and mass assumptions. Modeling dispersion (β_m) can be associated with the building definition and construction quality assurance dispersion (β_c) and the quality of the analytical model (β_q). β_c accounts for the difference between the actual characteristics of structural components (e.g., rebar location, material strength, section dimensions, etc.) and the assumed ones in design. β_q accounts for the difference between the actual component force–deformation relationship and the assumed relationships in the design stage, considering deterioration and failure mechanisms of the structural component. Recommended values for β_c and β_q are provided by recent seismic guidelines (e.g., FEMA-P58, 2018) considering three different levels, which are superior quality, average quality, and limited quality corresponding to 0.10, 0.25, and 0.40, respectively, for both parameters. Modeling dispersion (β_m) can be estimated as $\beta_m = \sqrt{\beta_c^2 + \beta_q^2}$. Record-to-record variability accounts for the difference in peak response quantities due to the difference in ground motions in the same set of earthquakes. Generally, using a suitable number of ground motions (more than 30) is sufficient to account for record-to-record variability and EDP correlation coefficients for seismic assessment applications (FEMA P-58, 2018). In the current study, 100 different ground motions are used to account for record-to-record variability. Ground motion variability, β_{gm}, accounts for the uncertainty in the attenuation relationship used to obtain the target response spectrum. In the current study, the attenuation relationship for Western North America is used (β_{gm} ranges between 0.6 and 0.7; FEMA P-58, 2018).

The augmented dataset is based on generating a large number of synthetic NLTHA (simulated) realizations based on a limited suite of actual ones using mathematical transformation. In this process, a demand matrix is formed such that each column represents the required EDP (i.e., *D, A*, and *V*). This matrix is assumed to represent a jointly lognormal distribution. The median value of each EDP and the covariance matrix (which represents the relationship of each EDP to others in the same set) are calculated. Using a random selection mathematical process, the covariance matrix, and median values, a large number of synthetic seismic demand vectors are mathematically generated. In the current study, 200 NLTHAs are conducted for each model (100 for the MDOF system and another 100 for the equivalent SDOF system of the same model) with a total of 1000 NLTHAs for all models. These analyses are considered the original dataset that will be augmented using the MCS method considering the uncertainties mentioned before. Mathematically, the process can be formulated as follows: If m NLTHAs are conducted for n EDPs, a matrix of demand parameters $\mathbf{X}$ can be formed with m rows and n columns and is assumed to be a jointly lognormal matrix (FEMA P-58, 2018). Another matrix, $\mathbf{Y}$, is formed by taking the logarithm of each entry of $\mathbf{X}$. $\mathbf{Y}$ is assumed to be jointly normal and can be represented by a vector of the natural logarithms of each EDP, $\mathrm{M_Y}$, and an $n \times n$ covariance matrix, $\Sigma_{\mathbf{YY}}$. A vector of the natural logarithm of EDPs, $\mathbf{Z}$, is mathematically generated assuming the same statistical characteristics of $\mathbf{Y}$ (Equation 15) (Yang et al., 2009):

$$\mathbf{Z} = \mathbf{AU} + \mathbf{B} = \mathbf{L}_{\mathrm{np}}\,\mathbf{D}_{\mathrm{pp}}\,\mathbf{U} + \mathbf{M}_{\mathrm{y}} \tag{15}$$

where $\mathbf{U}$ is the random variable vector with a mean of $\mathbf{0}$ (i.e., a vector of zeros) and unit covariance matrix, $\mathbf{A}$ and $\mathbf{B}$ are matrices of constant coefficients needed for the linear transformation of $\mathbf{U}$ to $\mathbf{Z}$, p is the rank of the covariance matrix $\Sigma_{\mathbf{YY}}$, $\mathbf{L}_{\mathrm{np}}$ is obtained by partitioning the $n \times n$ eigenvector matrix of $\Sigma_{\mathbf{YY}}$, $\mathbf{D}_{\mathrm{pp}}$ is the matrix obtained by partitioning the square root of eigenvalues of the matrix $\Sigma_{\mathbf{YY}}$. The covariance matrix $\Sigma_{\mathbf{YY}}$ can be decomposed into a correlation coefficient matrix and a variance matrix (which contains the variances of the EDPs). The latter matrix (i.e., variance matrix) is inflated with β_m and β_{gm} to account for the uncertainty due to structural modeling and ground motion variability, respectively. Figure 5 shows an example augmented dataset for the third model from 100 original samples to 10,000 samples.

2.4.3 | Architecture of the second probabilistic DLM (QDN2)

QDN2 model has similar architecture as QDN1 with fewer layers and inputs. Ensemble NNs with three outputs are employed to construct the network (Lakshminarayanan et al., 2017; Pearce et al., 2018). As shown in Figure 6, each network receives a response variable for all EDPs with its upper and lower boundaries generated from QDN1 and

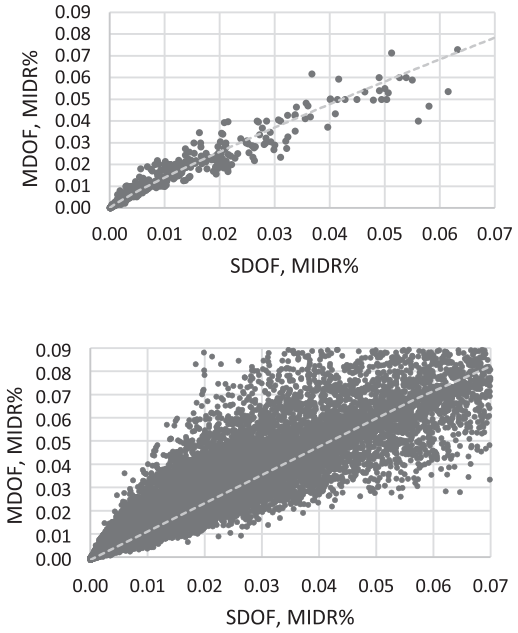


FIGURE 5 Augmented samples using Monte Carlo simulation (MCS) for the third model, (a) D-MDOF-SDOF relation for 100 samples and (b) D-MDOF-SDOF relation for 10,000 samples.

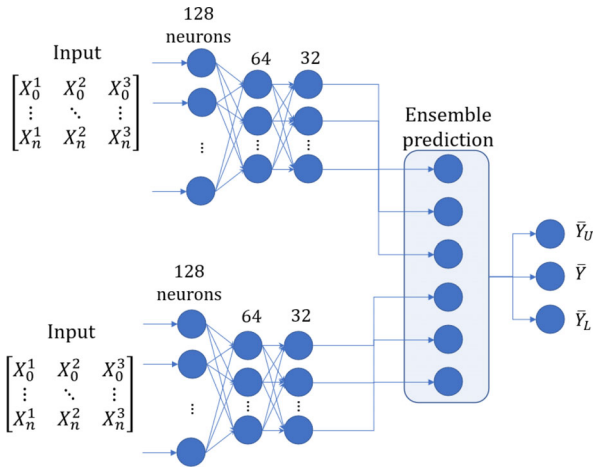


FIGURE 6 Architecture of proposed QDN2 model.

predicts the corresponding MDOF response variable for all EDPs (i.e., *D*, *A*, and *V*) as well as the MDOF response variable's lower-upper boundary. We have optimized the NN hyperparameters using Bayesian optimization with Gaussian processes. As shown in Figure 6, the final NN has three hidden layers with sizes of 64, 32, and 32 neurons. The model has batch size = 300, epochs = 100, lr = 0.01, weight decay = 0.95, and Adam optimizer. The model uses the ReLU activation function and optimizes the model using the L loss function of Equation (2), where $\lambda = 10$ and $\alpha = 0.1$.

QDN1 and QDN2 are stacked in series and the output from QDN1 is used as input for QDN2, which is called a stacked ensemble. The optimizations for each of the NNs in the stack would be focused on minimizing the MAE between the predicted outputs and the true values. The optimization of the entire system would involve tuning the parameters of each NN in the stack so that the overall prediction error of the system is minimized. This optimization is achieved using Bayesian optimization with Gaussian processes.

2.5 | Explainability of the DLM using XAI

DLMs; (Zou et al., 2022) are considered "black boxes" that cannot provide an understandable justification regarding the predictions obtained. In other words, the physical relationship between the input features and the output response is not clear (Mangalathu et al., 2022). XAI tries to enhance the transparency of the DLMs by explaining the relationship between the input features (predictors or input variables) and the response variable (Mangalathu et al., 2022). This explainability approach is highly appreciated in the earthquake engineering field where designers, stakeholders, and decision-makers are highly interested in having insights into the input–output variables relationship and the interpretability of the predicted response from the DLM. In this section, the PDP and ICE techniques are used to explain a model's predictions. PDP provides an insight regarding the effect of changes in the input feature on the mean value of predictions (or responses) of the DLM while maintaining other input features fixed. PDP provides visual interpretation regarding the type of the input–output relationship, which might be nonlinear, step-wise, linear, and so forth.

Assume that there is an input variable, *x*, which belongs to the input space, *X*, and is being mapped into an output function, *f(x)*, using a DLM. Then the partial dependence, $f_{s,}$ of an input variable, x_s, that belongs to the subset of input variables of interest, X_S, can be given as in Equation (16) (Friedman, 2001; Molnar, 2018)

$$f_s\ (x_s) = E_{X_C}\ [f_s\,(x_s, X_C)] = \int f_s\,(x_s, x_C)\,.P\,(x_C)\,dx_C \tag{16}$$

where X_C is the complement subset of X_S such that $X_C \cup X_S = X$, $f_s(x_s, x_C)$ is the response function for a given sample whose input variables are x_s and x_C, and $P(x_C)$ is the marginal probability density of x_C. The above integral can be approximated by computing the average over the dataset *X* (Equation 17).

$$\overline{f_s}\,(x_s) \approx \frac{1}{n} \sum_{i=1}^{n} f_s\left(x_s, x_C^{(i)}\right) \tag{17}$$

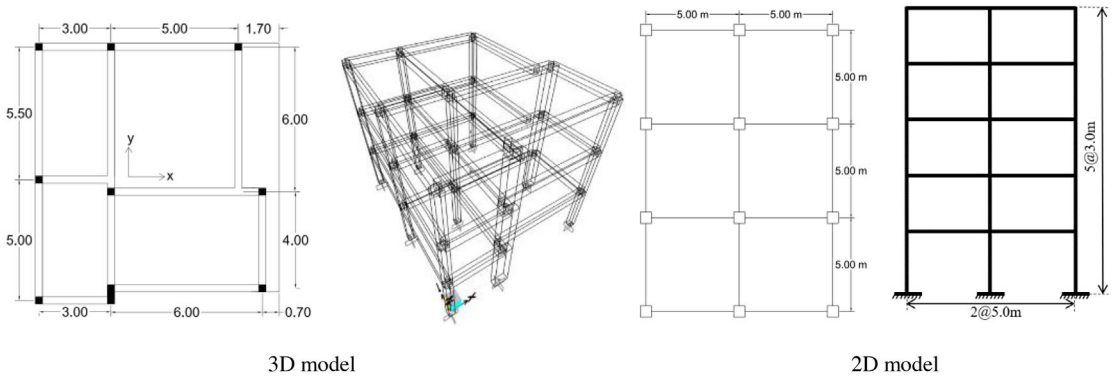


FIGURE 7 Configuration of the case study models (3D model details can be found in Fajfar et al., 2006).

where $\overline{f}_s(x_s)$ is the average partial dependence function at a point x_s, $x_C^{(i)}$is the ith sample value in X_C, and n is the number of instances in the dataset.

Based on that, the PDP provides a global relationship between the input and the output variables considering all instances in the dataset. On the other hand, ICE provides the dependence plot of the output variable on a particular input variable (i.e., local relationship) for each instance (Goldstein et, al., 2015) such that $f_S^{(i)}$ is plotted against $x_S^{(i)}$ for each instance in $\{(x_s, x_C^{(i)})\}_{i=1}^{n}$, while $x_C^{(i)}$ remains fixed. Based on that, the ICE plot provides a group of n curves for a particular x_s, these curves have the same shape pattern and are separated by certain levels according to the values of x_C. ICE plot has an advantage over PDP in that it provides more insights into the output response dependence on a particular input variable, especially in the case of having strong interaction between the input variable that will not be shown by the average function provided by the PDP.

3 | ACCURACY, INTERPRETATION, AND VALIDATION OF THE PROPOSED FRAMEWORK

3.1 | Case-study models and earthquakes

Figure 7 shows the configurations of the case-study models. The first one is a benchmark asymmetric reinforced concrete 3D model (the details can be found in Fajfar et al., 2006), whereas the second is a five-story 2D model that is used in previous studies (e.g., Noureldin et al., 2021b). The natural periods of these models are 0.58 and 1.25 s, respectively. Three different earthquake sets corresponding to return periods of 75, 475, and 2475 years, which are mapped to immediate occupancy, life safety, and collapse prevention limit states, respectively, are used. Each level is represented by 15 earthquakes (the details of the earthquakes can be found in Noureldin et al., 2021a) as shown in Figure 8.

All analyses are carried out using an Intel Xeon(R) central processing unit (CPU) E5-2620 v3 @ 2.40 GHz 24 with Cuda-10.0 and three GEFORCE GTX TITANx 12 GB graphics processing unit (GPUs), as well as Python 3.7.7 distributed in Anaconda 4.12.0 (64-bit). A high-performance computer with double 32 GB GPU and 184 GB RAM is utilized for very intensive computation tasks.

3.2 | Important input features using ANOVA

In this section, the important input features that will be used for the framework are investigated globally. For each entry of the entire dataset, the null hypothesis (p-value) that a particular input feature has a significant effect on a particular output response is calculated. A 0.05 threshold is used such that if the null hypothesis is less than that value, the input feature has a significant effect on the response variable; otherwise, it has no significant effect. Table 2 shows the percentage of the dataset that returns p-values less than 0.05 for each input feature for the three response variables. The table shows that the natural period (T) is important for all response variables, compared to the strength ratio (SR), which showed much less effect, especially for D and V. Regarding the ground motion characteristics, the FC, PGV and PGD, and AI are considered important features and have a significant effect on the three response variables since the percentages reach almost 100%. The effect of the PGA turned out to be important for A and V; however, only 33.3% of the dataset shows that the PGA has a significant effect on D. Another interesting finding is that D is highly affected by the fault type, Vs_{30}, lowest usable frequency, and the duration. However, these input features have far less effect on A and V. Based on that, these features should be considered in training the

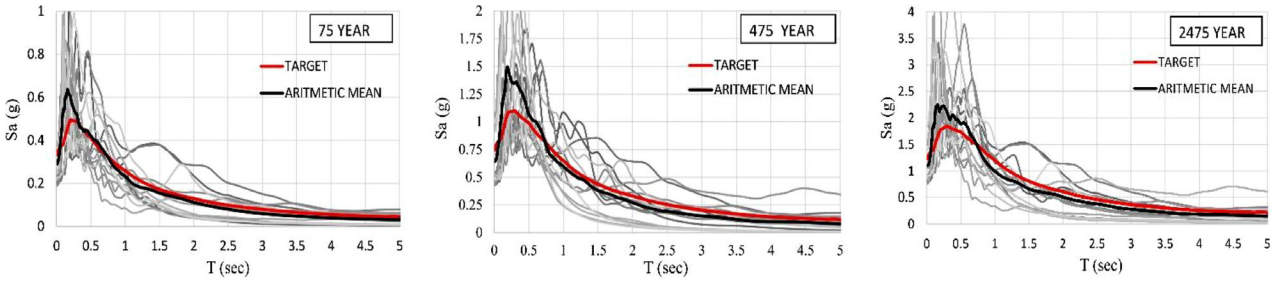


FIGURE 8 Response spectra of 75, 475, and 2475 years return period.

TABLE 2 Relative importance of input features based on analysis of variance (ANOVA) for the entire dataset (% of the dataset).

Our database	*SR*	*T*	Frequency content	PGA	PGV	PGD	Magnitude	R_{rup}	AI	Fault type	Vs_{30}	Lowest usable frequency	Duration (5%–95%)
D	30.1	97.4	100	33.3	100	100	100	82.0	100	100	100	100	100
A	75.8	98.7	100	100	100	96.4	71.3	67.9	100	30.7	24.9	53.3	40.1
V	27.5	100	100	100	100	100	75.7	62.6	100	30.9	28.1	60.2	40.7

Abbreviations: *A*, maximum acceleration; AI, Arias intensity; PGA, peak ground acceleration; PGV, peak ground velocity; PGD, peak ground displacement; R_{rup}, fault distance; *SR*, strength ratio; *V*, maximum base shear; Vs_{30}, 30-m average shear wave-velocity.

DLM in the case of *D* but may not be highly required for training the DLMs for *A* and *V*.

3.3 | Accuracy of the proposed framework

In this section, the accuracy of the proposed framework is investigated using *PICP*, *MPIW*, *CWC*, and *MAE*. Figure 9 shows the PIs provided by the DLMs (i.e., QDN1 and QDN2) for the training dataset. The green dots in the graph represent the NLTHA results, whereas the red dots represent the upper and lower bounds for each sample point.

It is found that QDN1 provides PIs with a coverage probability (*PICP*) of more than 96.8% for the investigated EDPs (i.e., *D, A*, and *V*). A similar value of the *PICP* is found for the test dataset (which is 5% of the whole dataset). This indicates that there is a slim chance (less than 3.2%) for any point to be outside of the PI provided by the DLMs. On the other hand, the mean width of the PI (*MPIW*) is found to be less than 0.45 for all the PIs. For *D, A*, and *V*, the *CWCs* turned out to be 0.729, 0.441, and 1.07, respectively. Since a less value of *CWC* or *MPIW* indicates a better PI, the PI of *A* is the best, compared to other EDPs. The "best" implies that PIs have a high coverage probability with the least width possible. All *MAE*s of the PIs for EDPs are found to be less than 0.026. For QDN2, it is found that *PICP*, *MPIW*, *CWC*, and *MAE* turned out to be 0.98, 0.088, 0.086, and 0.001, respectively, for *D*. Table 3 summarizes the values of the different accuracy metrics of the DLMs for all EDPs.

TABLE 3 Accuracy metrics of QDN1 and QDN2.

	QDN1			QDN2		
	D	***A***	***V***	***D***	***A***	***V***
PICP	0.968	0.974	0.98	0.98	0.984	0.986
MPIW	0.438	0.44	0.251	0.088	0.082	0.073
CWC	0.729	0.441	1.07	0.086	0.081	0.072
MAE	0.026	0.021	0.012	0.001	0.0001	0.0001

Abbreviations: *A*, maximum acceleration; CWC, coverage width-based criterion; MAE, mean absolute error; MPIW, mean prediction interval width; PICP, prediction interval coverage probability; QDN1, probalistic deep learning model-2; QDN2, probalistic deep learning model-2; *V*, maximum base shear.

Figure 10 shows the estimated model uncertainty for the proposed framework for the three EDPs for 400 randomly sampled points of the testing dataset. Generally, the uncertainty is low, which means that the PIs bound most of the points. But it becomes high in areas that are not well represented in the training dataset (some regions in the *D* graphs). This indicates the importance of having good coverage of the input space on the overall model uncertainty performance.

3.4 | Explainability of the framework using XAI

In this section, the relationship between the input features and the predicted response is explored using XAI

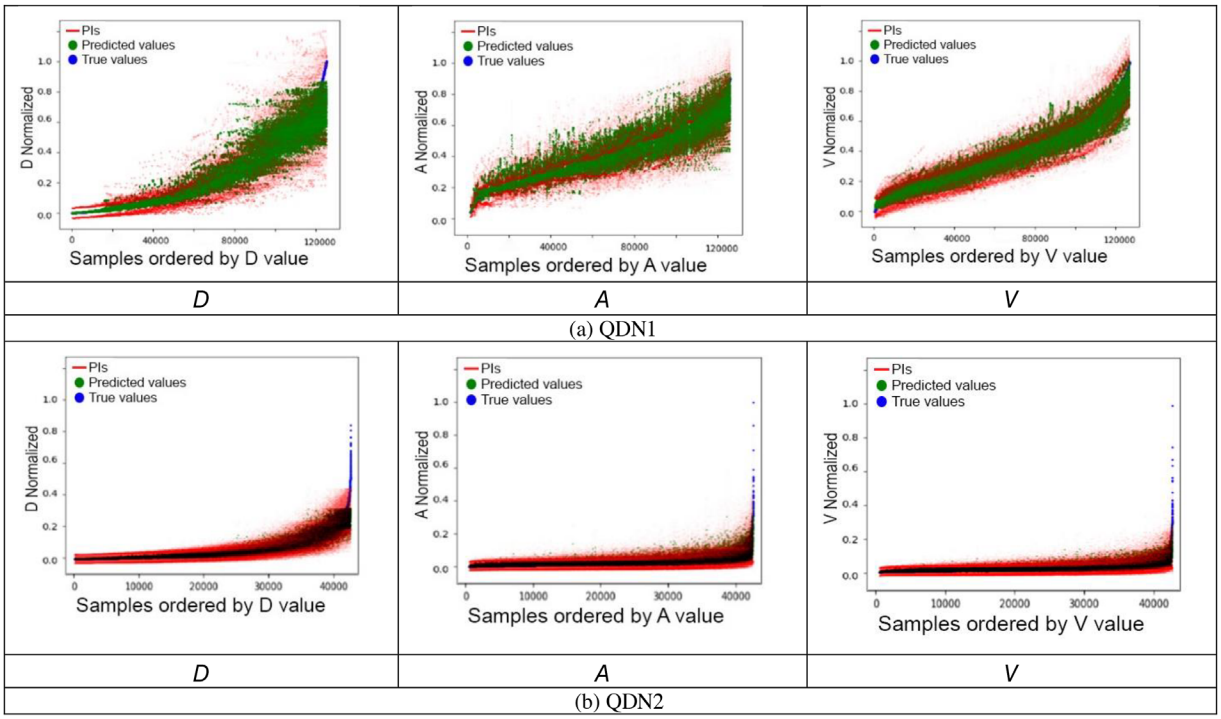


FIGURE 9 Prediction interval (PI) plots provided by the deep learning models: (a) QDN1 and (b) QDN2. PIs, prediction intervals.

for the case study models (i.e., locally). DLMs are considered, in general, as black boxes that contain complex functions that may not be understood on their own. XAI is a technique that explains the latent model input–output relationship regardless of the type of deep/machine learning model investigated (i.e., model agnostic technique). In the current study, PDPs and ICE plots are implemented using the histogram-based gradient boosting technique used for training faster decision trees. Figure 11 shows the variation in the partial dependence for all input features (PDPs) related to the predicted response. Each plot represents the relationship between the partial dependency of the average response parameter with the change of the normalized values of the input feature. The figure shows that some input features greatly impact the model's prediction response. For example, FC, PGA, and PGV highly affect the drift response. This can be distinguished by the jumps on the curves with steep slopes on the plot. Whereas, PGD, magnitude, fault distance (R_{rup}), AI, and fault type have a marginal effect on the response. This can be noticed by having an almost horizontal line on the plot, which indicates that no change occurs in the response due to the change in the input feature value. On the other hand, Vs_{30}, the lowest usable frequency, and duration, show a moderate effect on the model response, especially, for normalized input feature values less than 0.5.

Figure 12 shows ICE plots for different input features with the predicted response (for *D*). The ICE plot shows all instances related to a specific entry in the dataset for the predicted responses. The dashed line in the plot indicates the average of the responses across different values for the input feature. Most plots show a uniform scattered pattern of predicted responses. Some plots, such as FC of the 2D model and PGV of the 3D model, show a dense population of curves near the average (as if the distribution across these curves is skewed and can be fitted by lognormal distribution). This may help in providing the most probable range of the predicted response for a specific input feature range. In addition, it provides, for a particular entry of the dataset, the change of the response with a specific feature while holding other features constant.

For comparison purposes, the ANOVA method is used to highlight the relative importance of all earthquake input features for the case study models. Table 4 shows *p*-values and F_0 for the case study models for the three EDPs (i.e., *D*, *A*, and *V*). It can be observed that many *p*-values are less than 0.05, and this feature is important for a specific output response prediction. For example, PGV and AI are important input features for both case study models for all output responses. Other input features are important for a particular response but are not for others. For example, the PGD and the magnitude are not important input features for acceleration response prediction of the 3D model because their *p*-values are 0.47 and 0.68, respectively. Both *p* and F_0 values indicate the same relative importance. F_0 values may distinguish between the input features that have very small *p*-values.

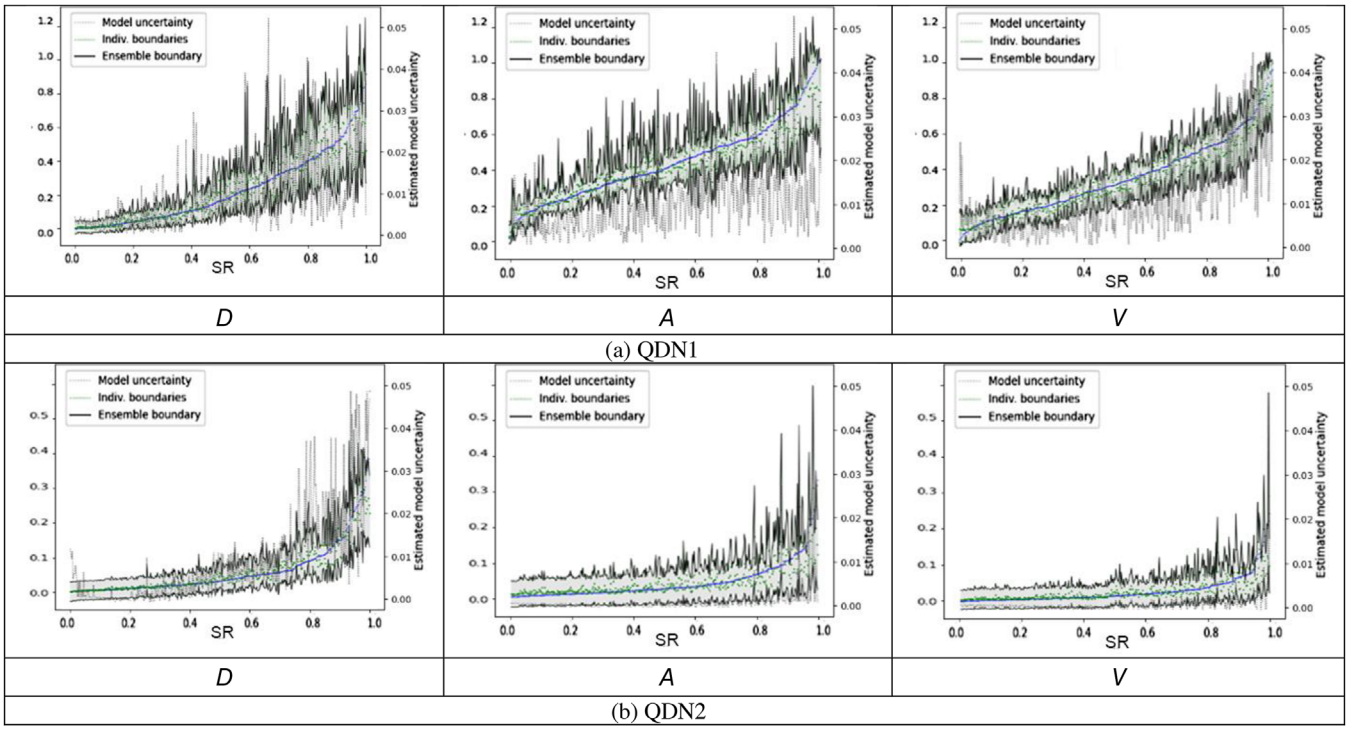


FIGURE 10 Estimated model uncertainty of the proposed framework: (a) QDN1 and (b) QDN2.

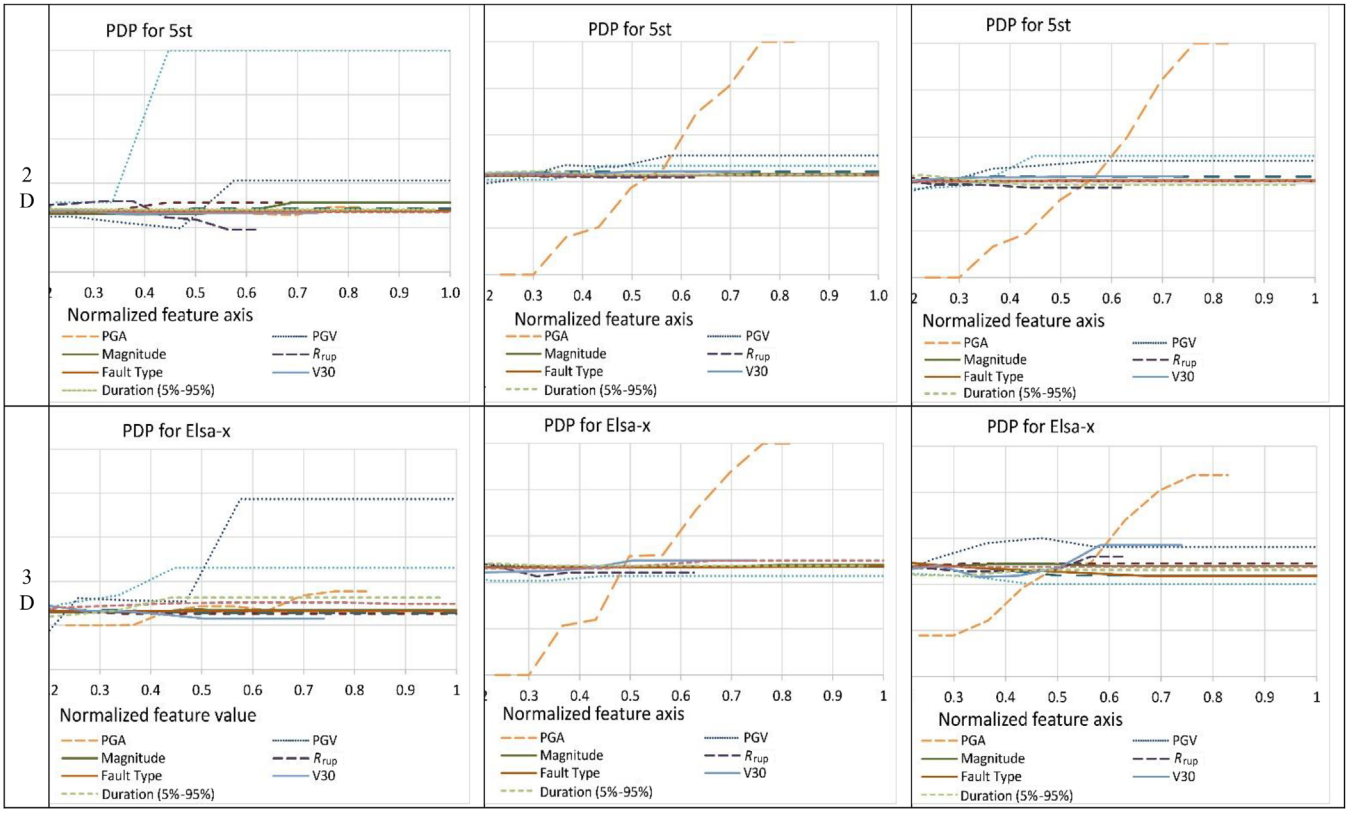


FIGURE 11 Drift partial dependence plots (PDPs) for the model structure. PGV, peak ground velocity; R_{rup}, fault distance.

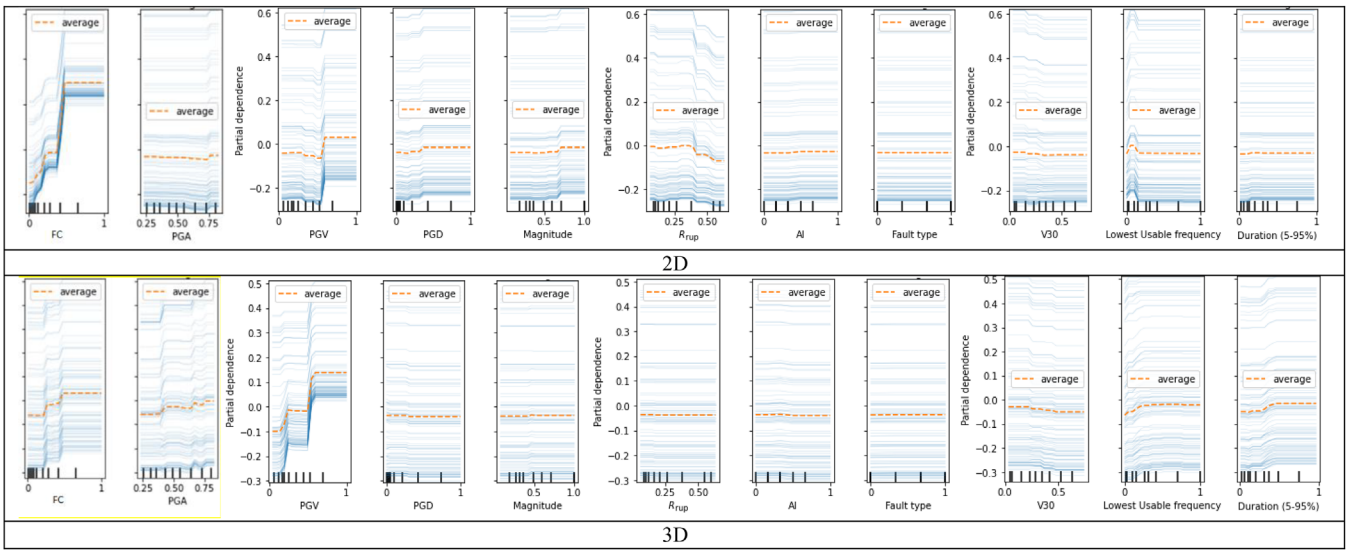


FIGURE 12 Individual conditional expectation plots for different input features with the predicted response for *D*. PGV, peak ground velocity; R_{rup}, fault distance.

TABLE 4 ANOVA results for the case study models.

			PGA	PGV	PGD	Magnitude	R_{rup}	AI	Fault type	FC	Lowest usable frequency	Duration (5%–95%)
	***p*-value**											
2D	*D*	0	0.12	0	0	0	0.01	0	0	0	0	0
	A	0	0	0	0	0	0.02	0	0.18	0.09	0.03	0.11
	V	0	0	0	0	0	0.02	0	0.11	0.10	0.02	0.07
3D	*D*	0	0	0	0	0	0.02	0	0	0	0	0
	A	0.07	0	0	0.47	0.68	0	0	0.14	0.02	0.28	0.14
	V	0	0	0	0	0	0.48	0	0.21	0	0.01	0.02
	F_0											
2D	*D*	4827.79	2.45	1036.74	531.39	151.16	6.65	369.19	39.75	13.28	99.61	150.09
	A	61.76	351.40	101.10	29.38	8.82	5.90	91.14	1.82	2.98	4.62	2.62
	V	60.68	304.83	104.59	31.11	10.37	5.16	95.97	2.63	2.72	5.72	3.28
3D	*D*	353.87	12.74	597.78	158.59	108.68	5.19	334.68	33.15	10.69	73.57	84.76
	A	3.39	501.93	14.84	0.53	0.17	11.80	19.12	2.23	5.28	1.19	2.20
	V	19.73	108.38	58.83	13.87	12.20	0.51	55.07	1.60	10.54	7.86	5.26

Abbreviations: *A*, maximum acceleration; AI, Arias intensity; FC, frequency content; PGA, peak ground acceleration; PGV, peak ground velocity; PGD, peak ground displacement; R_{rup}, fault distance; *V*, maximum base shear; Vs_{30}, 30-m average shear wave-velocity.

4 | VALIDATION AND COMPARISON WITH OTHER METHODS

In this section, the superiority of the proposed framework is validated by two different criteria. The first criterion is related to the accuracy of uncertainty quantification; whereas the second is related to the reliability of the framework in enveloping system responses under real earthquake scenarios at different intensity levels. First, the proposed framework's uncertainty quantification accuracy metrics and PIs are compared with those obtained from one of the most commonly used methods in practice for uncertainty quantification, which is the Bayesian method. After that, the responses obtained from NLTHAs are compared with the upper and lower bounds (envelopes) provided by the proposed framework

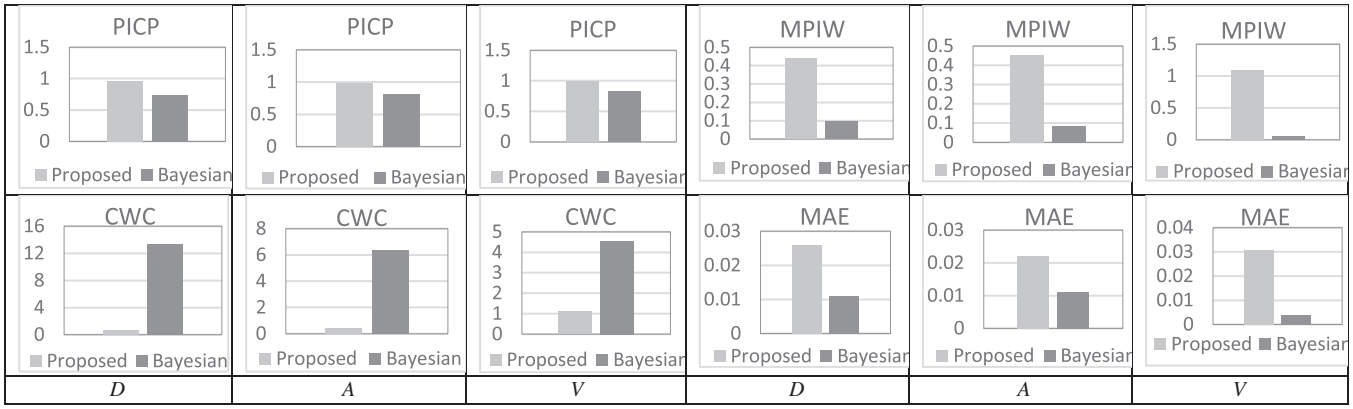


FIGURE 13 Comparison of the proposed and Bayesian methods. CWC, coverage width-based criterion; MAE, mean absolute error; MPIW, mean prediction interval width; PICP, prediction interval coverage probability.

for different damage state levels and different earthquake severity.

4.1 | Uncertainty quantification comparison

Generally, the distribution of weights is used for network training in the Bayesian method. In Bayes' theorem, a prior distribution is assumed to optimize weights of a posterior distribution to obtain the predictive distribution such that the total variance of the NN training is as in Equation (18).

$$\sigma_i^2 = \frac{1}{\beta} + \nabla_{w^{MP}}^T \hat{y}_i (H^{MP})^{-1} \nabla_{w^{MP}} \hat{y}_i \tag{18}$$

where β is a hyperparameter of the cost function w^{MP}, the most probable value of the NN parameters; $\hat{y}_i$ is the ith future sample; $(\nabla_{w^{MP}}^T \hat{y}_i)$ is the NN output gradient related to its parameters w^{MP}; and H^{MP} is the Hessian matrix of the sum of squares of the network weights. Knowing the ith future sample total distribution, the $(1-a)$% PI can be estimated as

$$(1-\alpha)\%PI = \hat{y}_i \pm z^{1-\frac{\alpha}{2}} \left(\frac{1}{\beta} + \nabla_{w^{MP}}^T \hat{y}_i \left(H^{MP}\right)^{-1} \nabla_{w^{MP}} \hat{y}_i \right)^{\frac{1}{2}} \tag{19}$$

where $(z^{1-\frac{\alpha}{2}})$ is the percentage quantile $(1-\frac{\alpha}{2})$ of an assumed distribution.

Figure 13 compares the proposed and the Bayesian methods in terms of PICP, MPIW, and CWC for all EDPs. The vertical axis represents the value of each accuracy metric indicated on each plot. The figure shows that the proposed framework provides a higher coverage probability (PICP more than 97%), compared to the Bayesian method (74%~83%), for all EDPs. Although the Bayesian method provides less MPIW, compared to the proposed method, its PI has poor coverage. This means that it does not include all predicted points within the generated PI. The overall PI quality can be judged well by investigating the CWC metric, which combines both the PICP and the MPIW. The figure shows that CWC values of the PI generated by the proposed method (0.73, 0.44, and 1.1, respectively, for *D*, *A*, and *V*) are considerably lower than that of the Bayesian method (13.41, 6.4, and 4.5, respectively, for the same EDPs), which shows that the proposed method works better than the Bayesian method in terms of this uncertainty quantification metric (i.e., CWC). In addition, unlike the Bayesian method, the proposed method can be easily implemented since it does not require the formation of the Hessian matrix of the sum of squares of the network weights, which is highly computationally demanding, especially for large datasets (Kabir et al., 2018). *MAE* in this context is not providing much information about the bounds of the PI like *CWC*. Also, the *MAE* does not provide information about the coverage probability; rather, it provides information about how close the predicted value is from the target value, which is not very important in evaluating PIs. The Bayesian method showed low *MAE* with high *CWC*, which indicates less accuracy of the total PI. On the other hand, the proposed framework showed the opposite, which indicates high accuracy of the total PI. Figure 14 compares the PI boundaries of the test dataset between the proposed and the Bayesian methods. In the figure, the green dots represent the predicted output, whereas the PI boundaries are represented by light red lines. The figure shows that PIs provided by the proposed method get wider with the increased value of the EDP so that they encompass more data. On the other hand, Bayesian PIs are much narrower, compared to the

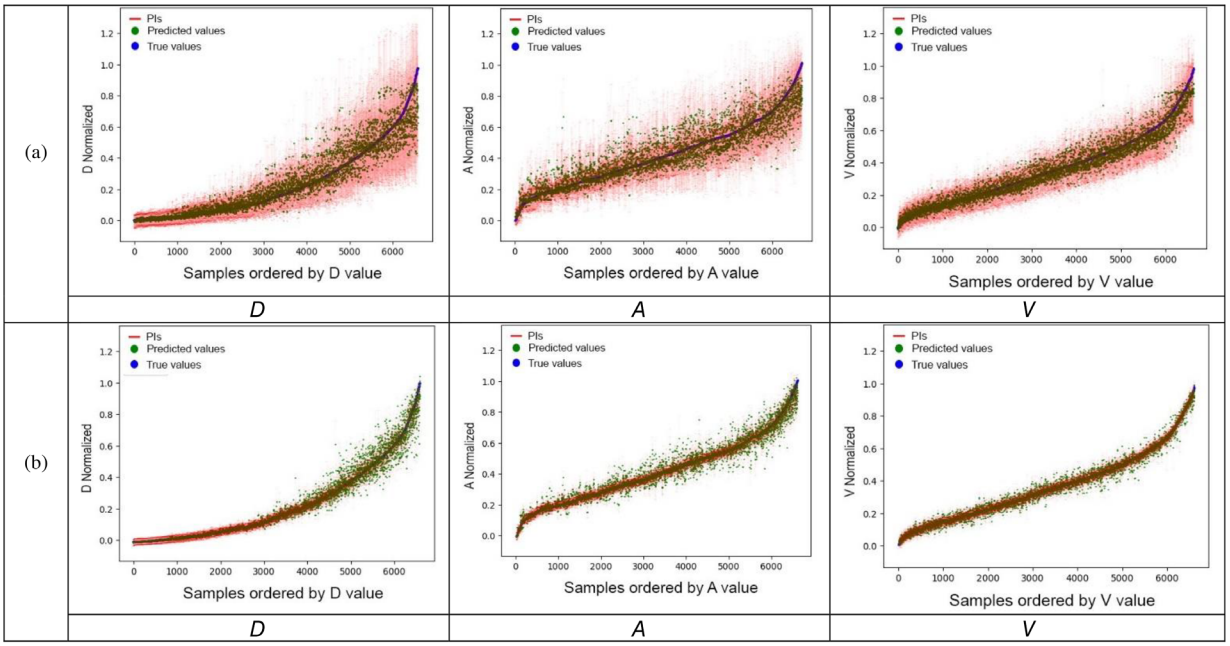


FIGURE 14 Comparison of prediction interval (PI): (a) proposed framework and (b) Bayesian method.

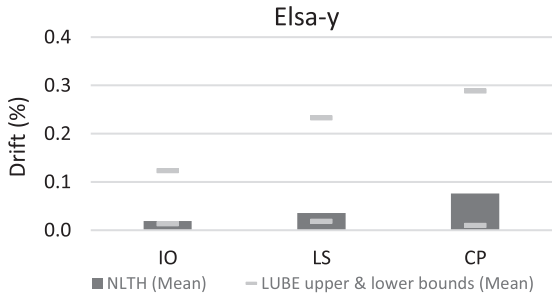


FIGURE 15 Mean lower and upper bounds compared with mean responses at different damage limit states. NLTH, nonlinear time history analysis.

PIs obtained from the proposed method, especially at the large values (i.e., inelastic range) of the EDP, leaving many points not included inside the PI boundaries.

4.2 | Comparison with nonlinear dynamic methods

Figure 15 shows the mean lower and upper bounds, compared with the mean drift values obtained using NLTHAs at different limit states. It can be observed that for different limit states, NLTHA results are within the proposed PI bounds. The same is shown in Figure 16 for all 45 earthquakes, where the proposed PIs serve as an envelope for all responses obtained from earthquakes with different intensity levels.

4.3 | Seismic response assessment of structures

Some important applications in earthquake engineering are highly related to uncertainty, such as seismic fragility, LCC, and the degree of resiliency of the structure, which is measured by the RI. These applications are estimated using the NLTHA of the 3D model and are compared with the upper and lower bounds provided by the proposed framework.

The seismic fragility of a structure can be estimated as (Eldin et al., 2020)

$$P = 1 - \Phi\left[\frac{\ln(\hat{C}/\hat{D})}{\beta_{TOT}}\right] \tag{20}$$

where P is the probability of reaching or exceeding a specific limit state, $\hat{C}$ and $\hat{D}$ are the median capacity of the structure and the median seismic demand, respectively. β_{TOT} is the structural damage/limit state uncertainty, and $\Phi[.]$ is the normal distribution function.

LCC of structures can be calculated as follows (Gencturk, 2013; Noureldin & Kim, 2021, 2023):

$$LCC = C_o + \int_0^L E\left[C_{SD}\right]\left(\frac{1}{1+\lambda}\right)^t dt \tag{21}$$

where C_o is the initial construction cost, L is the service life of the structure, λ is the annual discount rate, and $E[C_{SD}]$ is the annual expected seismic damage cost.

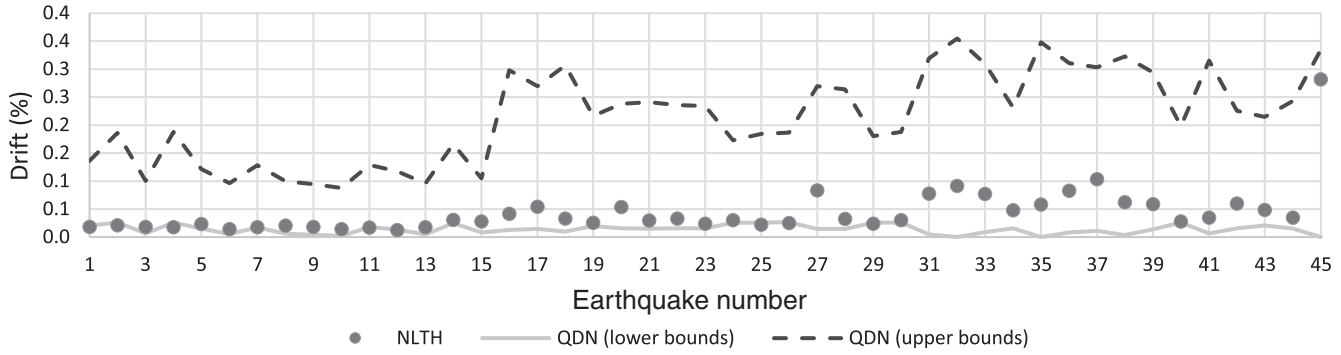


FIGURE 16 PI bounds for all response prediction values for different earthquake severity levels. NLTH, nonlinear time history analysis.

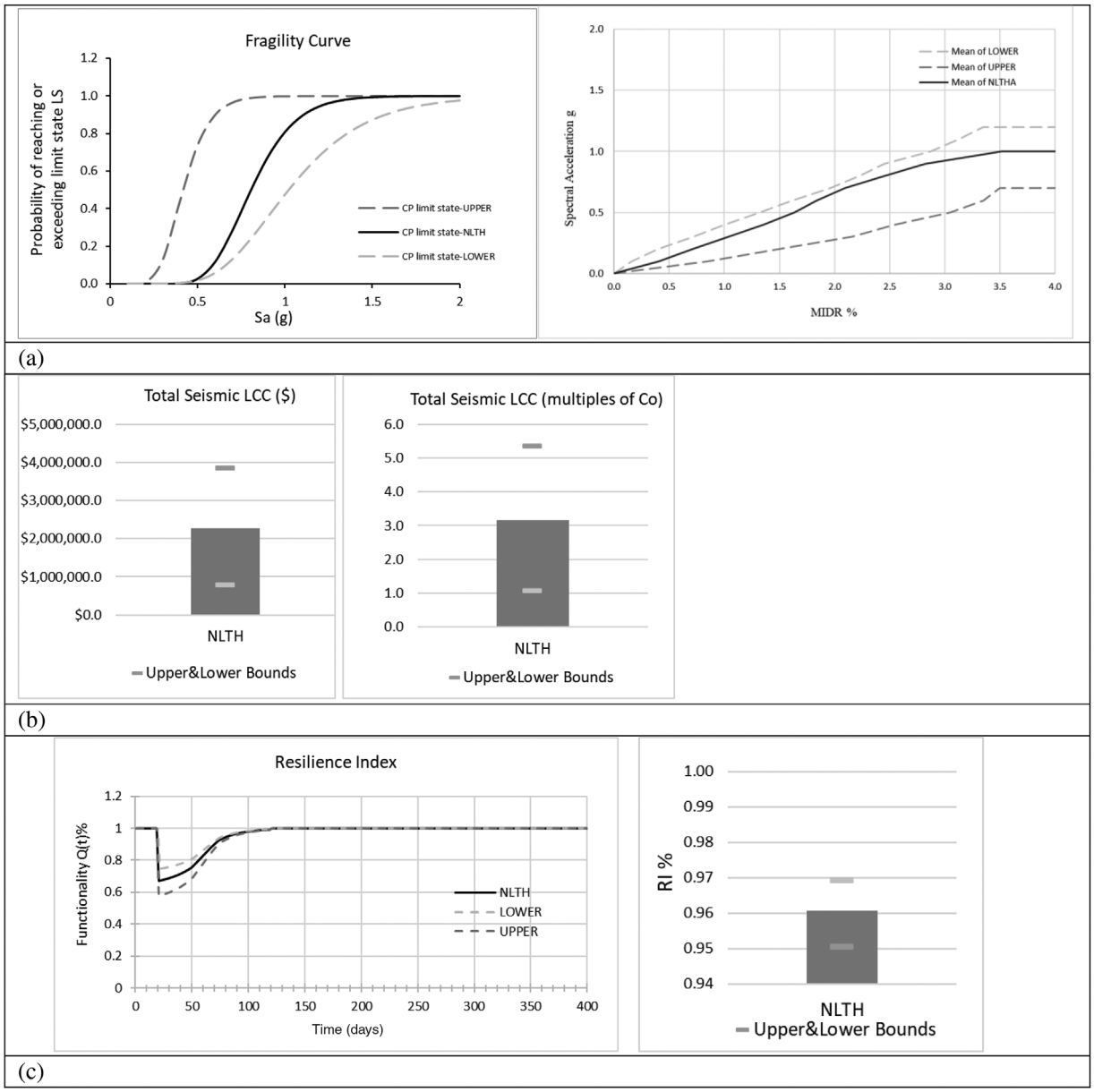


FIGURE 17 PI bounds for (a) Fragility and median incremental dynamic analysis curve, (b) life cycle cost (LCC), and (c) resilience index (RI). LS, life safety; NLTH, nonlinear time history analysis.

RI of a structure can be estimated by Equation (22):

$$R = \int_{T_{OE}}^{T_{OE}+T_{LC}} Q(t)/T_{LC}\,dt \tag{22}$$

$$Q(t) = [1 - L(I, T_{RE})]\,[H(t - T_{OE}) - H(t - (T_{OE} + T_{RE}))] \times f_{REC}(t, T_{OE}, T_{RE}) \tag{23}$$

where T_{OE} is the time when an event E occurs; T_{RE} is the time it takes to recover from event E (recovery time); T_{LC} is the control time; $Q(t)$ (Equation 23) is the functionality function; $L(I,\ T_{RE})$ is the loss function; I is the earthquake intensity; $H(.)$ is the Heaviside step function; and $f_{REC}(t, T_{OE}, T_{RE})$ is the recovery function. For the case study model, T_{RE} and T_{LC} are assumed 120 and 400 days, respectively.

Figure 17 shows the bounds of the PIs predicted by the developed machine learning framework and the results obtained by NLTHA, which indicates that the results obtained from the proposed framework could provide reliable lower and upper bounds for the seismic fragility curve, LCC, and RI of the 3D model, compared to those obtained by NLTH.

5 | CONCLUSION

The aim of the proposed framework is to provide upper and lower bounds for seismic responses of low- to mid-rise buildings under earthquake excitations using distribution-free PIs. These PIs can be used for various seismic assessment applications such as seismic fragility, LCC, resilience assessment, and so forth. The framework consists of two probabilistic deep ensemble learning models combined with XAI techniques. To achieve the highest accuracy of the framework, ensemble deep NNs were used where the optimum hyperparameters were obtained using a Bayesian optimizer. The predictors (input parameters) associated with output responses were investigated using XAI techniques such as PDPs, ICE, and ANOVA. NLTHA and Bayesian methods were used to validate the framework. The followings are the main findings of the current study:

1. The proposed architecture is simple and easy to implement and provides reliable PIs with high accuracy without pre-assumed data distribution.
2. XAI results showed that the most influential input features on output responses (i.e., EDPs) are the fundamental period, FC, PGA, PGV, and PGD. Sometimes, other input features may have a relatively high association with a specific EDP but not for all EDPs.
3. The proposed framework has much less *CWC* (approximately 1.1 or less), compared to the Bayesian method (ranging from 4.5 to 13.41), for all EDPs. This highlights the advantage of the framework in terms of both the coverage probability and the mean width of the PIs.
4. The obtained PIs provide reliable bounds for the seismic fragility curve, LCC, and RI for the benchmark model subjected to different earthquake severity levels.
5. The proposed framework is most suitable for low- to mid-rise regular buildings.

The main advantages of the proposed framework can be summarized in several points. First, the framework requires significantly fewer computational resources than conducting NLTHAs, making it more efficient and practical to use for performance assessments of large-scale buildings. Second, it provides reliable bounds for important seismic assessment metrics, making it a desirable and reliable tool for decision-makers. Third, it allows for the identification and interpretation of the most important factors influencing building response, which can be useful for understanding the behavior of the structure and support informed design decisions. On the other hand, the main limitation of the proposed framework is that it is developed for low- to mid-rise regular buildings and may not be as effective when applied to irregular or tall buildings.

ACKNOWLEDGMENTS

This study was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean Government (MSIT; No. 2021R1A2C1011198) and (No. 2021R1A2C2006631).

REFERENCES

Alam, K. M. R., Siddique, N., & Adeli, H. (2020). A dynamic ensemble learning algorithm for neural networks. *Neural Computing and Applications*, *32*, 8675–8690.

Ali, T., Eldin, M. N., & Haider, W. (2023). The effect of soil-structure interaction on the seismic response of structures using machine learning, finite element modeling and ASCE 7–16 Methods. *Sensors*, *23*(4), 2047.

ASCE/SEI 41. (2017). ASCE standard, seismic evaluation and retrofit of existing buildings. American Society of Civil Engineers.

Celarec, D., & Dolšek, M. (2013). The impact of modelling uncertainties on the seismic performance assessment of reinforced concrete frame buildings. *Engineering Structures*, *52*, 340–354.

D'Angela, D., Magliulo, G., Celano, F., & Cosenza, E. (2021). Characterization of local and global capacity criteria for collapse assessment of code-conforming RC buildings. *Bulletin of Earthquake Engineering*, *19*(9), 3701–3743.

Eldin, M. N., Dereje, A. J., & Kim, J. (2020). Seismic retrofit of framed buildings using self-centering PC frames. *Journal of Structural Engineering*, *146*(10), 04020208.

Eltouny, K. A., & Liang, X. (2021). Bayesian-optimized unsupervised learning approach for structural damage detection. *Computer-Aided Civil and Infrastructure Engineering*, *36*(10), 1249–1269.

Fajfar, P., Dolšek, M., Marušić, D., & Stratan, A. (2006). Pre-and post-test mathematical modelling of a plan-asymmetric reinforced concrete frame building. *Earthquake Engineering & Structural Dynamics*, *35*(11), 1359–1379.

FEMA 440. (2005). *Improvement of nonlinear static seismic analysis procedures*. Federal Emergency Management Agency.

FEMA P-58. (2018). Seismic performance assessment of buildings. *Federal Emergency Management Agency*.

Feng, N., Zhang, G., & Khandelwal, K. (2021). *On the application of data-driven deep neural networks in linear and nonlinear structural dynamics*. https://arxiv.org/abs/2111.02784

Franchin, P., & Cavalieri, F. (2015). Probabilistic assessment of civil infrastructure resilience to earthquakes. *Computer-Aided Civil and Infrastructure Engineering*, *30*(7), 583–600.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*(5), 1189–1232.

Gencturk, B. (2013). Life-cycle cost assessment of RC and ECC frames using structural optimization. *Earthquake Engineering & Structural Dynamics*, *42*(1), 61–79.

Gharehbaghi, S., Yazdani, H., & Khatibinia, M. (2020). Estimating inelastic seismic response of reinforced concrete frame structures using a wavelet support vector machine and an artificial neural network. *Neural Computing and Applications*, *32*(8), 2975–2988. https://doi.org/10.1007/s00521-019-04075-2

Ghiasi, R., Noori, M., Altabey, W. A., Silik, A., Wang, T., & Wu, Z. (2021). uncertainty handling in structural damage detection via non-probabilistic meta-models and interval mathematics, a data-analytics approach. *Applied Sciences*, *11*(2), 770. https://doi.org/10.3390/app11020770

Goldstein, A., Kapelner, A., Bleich, J., & Pitkin, E. (2015). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *journal of Computational and Graphical Statistics*, *24*(1), 44–65.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems*, *27*, Montreal, Quebec, Canada.

Kabir, H. D., Khosravi, A., Hosen, M. A., & Nahavandi, S. (2018). Neural network-based uncertainty quantification: A survey of methodologies and applications. *IEEE Access*, *6*, 36218–36234.

Kendall, A., & Gal, Y. (2017). What uncertainties do we need in Bayesian deep learning for computer vision? *Conference on Neural Information Processing Systems (NIPS)*, Long Beach, CA.

Kim, T., Kwon, O. S., & Song, J. (2019). Response prediction of nonlinear hysteretic systems by deep neural networks. *Neural Networks*, *111*, 1–10.

Kim, T., Song, J., & Kwon, O. S. (2020). Probabilistic evaluation of seismic responses using deep learning method. *Structural Safety*, *84*, 101913.

Kingma, D. P., & Welling, M. (2014). Auto- encoding variational Bayes. *Proceedings of the 2nd International Conference on Learning Representations (ICLR)*, Banff, Alberta, Canada.

Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems*, *30*, Long Beach, CA.

Li, C., Li, H., & Chen, X. (2021). A framework for fast estimation of structural seismic responses using ensemble machine learning model. *Smart Structures and Systems*, *28*(3), 425–441. https://doi.org/10.12989/sss.2021.28.3.425

Mangalathu, S., & Burton, H. (2019). Deep learning-based classification of earthquake-impacted buildings using textual damage descriptions. *International Journal of Disaster Risk Reduction*, *36*, 101111.

Mangalathu, S., Karthikeyan, K., Feng, D. C., & Jeon, J. S. (2022). Machine-learning interpretability techniques for seismic performance assessment of infrastructure systems. *Engineering Structures*, *250*, 112883.

MATLAB version 9.9.0.1718557 (R2020b) Update 6, (2020).

McKenna, F., Fenves, G., & Scott, M. (2011). Open system for earthquake engineering simulation. [Software]. OpenSees. University of California, Berkeley, CA.

Molnar, C. (2018). A guide for making black box models explainable [Software and training videos]. https://christophm.github.io/interpretable-ml-book

Montgomery, D. C. (2013). *Design and analysis of experiments* (9th ed.). Wiley.

Mu, H. Q., & Yuen, K. V. (2016). Ground motion prediction equation development by heterogeneous Bayesian learning. *Computer-Aided Civil and Infrastructure Engineering*, *31*(10), 761–776.

Nabian, M. A., & Meidani, H. (2018). Deep learning for accelerated seismic reliability analysis of transportation networks. *Computer-Aided Civil and Infrastructure Engineering*, *33*(6), 443–458.

Noureldin, M., Ali, A., Nasab, M. S. E., & Kim, J. (2021). Optimum distribution of seismic energy dissipation devices using neural network and fuzzy inference system. *Computer-Aided Civil and Infrastructure Engineering*, *36*(10), 1306–1321. https://doi.org/10.1111/mice.12673

Noureldin, M., Memon, S. A., Gharagoz, M., & Kim, J. (2021b). Performance-based seismic retrofit of RC structures using concentric braced frames equipped with friction dampers and disc springs. *Engineering Structures*, *243*, 112555.

Noureldin, M., Ali, A., Sim, S., & Kim, J. (2022). A machine learning procedure for seismic qualitative assessment and design of structures considering safety and serviceability. *Journal of Building Engineering*, *50*, 104190.

Noureldin, M., Ali, T., & Kim, J. (2023). Machine learning-based seismic assessment of framed structures with soil-structure interaction. *Frontiers of Structural and Civil Engineering*, *17*, 205–223. https://doi.org/10.1007/s11709-022-0909-y

Noureldin, M., & Kim, J. (2021). Parameterized seismic life-cycle cost evaluation method for building structures. *Structure and Infrastructure Engineering*, *17*(3), 425–439.

Noureldin, M., & Kim, J. (2023). Simplified life cycle cost estimation of low-rise steel buildings using fundamental period. *Sustainability*, *15*(3), 2706.

Pacific Earthquake Engineering Research (PEER) Center. (2021). *NGA Database*. https://peer.berkeley.edu

Payán-Serrano, O., Bojórquez, E., Bojórquez, J., Chávez, R., Reyes-Salazar, A., Barraza, M., López-Barraza, A., Rodríguez-Lozoya, H., & Corona, E. (2017). Prediction of maximum story drift of MDOF structures under simulated wind loads using artificial neural networks. *Applied Sciences*, *7*(6), 563.

Pearce, T., Brintrup, A., Zaki, M., & Neely, A. (2018). High-quality prediction intervals for deep learning: A distribution-free, ensembled approach. *Proceedings of the 35th International Conference on Machine Learning*, Stockholm, Sweden.

Rafiei, M. H., & Adeli, H. (2017a). A new neural dynamic classification algorithm. *IEEE Transactions on Neural Networks and Learning Systems*, *28*(12), 3074–3083.

Rafiei, M. H., & Adeli, H. (2017b). NEEWS: A novel earthquake early warning model using neural dynamic classification and neural dynamic optimization. *Soil Dynamics Earthquake Engineering*, *100*, 417–427.

Rafiei, M. H., Khushefati, W. H., Demirboga, R., & Adeli, H. (2017). Supervised deep restricted Boltzmann machine for estimation of concrete. *ACI Materials Journal*, *114*(2), 237–244.

Stoffel, M., Gulakala, R., Bamer, F., & Markert, B. (2020). Artificial neural networks in structural dynamics: A new modular radial basis function approach vs. convolutional and feedforward topologies. *Computer Methods in Applied Mechanics and Engineering*, *364*, 112989.

Sun, H., Burton, H. V., & Huang, H. (2021). Machine learning applications for building structural design and performance assessment: State-of-the-art review. *Journal of Building Engineering*, *33*, 101816.

Wu, R. T., & Jahanshahi, M. R. (2019). Deep convolutional neural network for structural dynamic response estimation and system identification. *Journal of Engineering Mechanics*, *145*, 04018125.

Xi, E., Bing, S., & Jin, Y. (2017). Capsule network performance on complex data. arXiv preprint arXiv:1712.03480.

Xie, Y., Sichani, M. E., Padgett, J. E., & DesRoches, R. (2020). The promise of implementing machine learning in earthquake engineering: A state-of-the-art review. *Earthq. Spectra*, *36*(4), 1769–1801. https://doi.org/10.1177/8755293020919419

Xu, Y., Lu, X., Cetiner, B., & Taciroglu, E. (2021). Real-time regional seismic damage assessment framework based on long short-term memory neural network. *Computer-Aided Civil and Infrastructure Engineering*, *36*(4), 504–521.

Yang, T. Y., Moehle, J., Stojadinovic, B., & Der Kiureghian, A. (2009). Seismic performance evaluation of facilities: Methodology and implementation. *Journal of Structural Engineering*, *135*(10), 1146–1154.

Zhang, R., Chen, Z., Chen, S., Zheng, J., Büyüköztürk, O., & Sun, H. (2019). Deep long short-term memory networks for nonlinear structural seismic response prediction. *Computers & Structures*, *220*, 55–68.

Zou, D., Zhang, M., Bai, Z., Liu, T., Zhou, A., Wang, X., Cui, W., & Zhang, S. (2022). Multicategory damage detection and safety assessment of post-earthquake reinforced concrete structures using deep learning. *Computer-Aided Civil and Infrastructure Engineering*, *37*(9), 1188–1204.

How to cite this article: Noureldin, M., Abuhmed, T., Saygi, M., & Kim, J. (2023). Explainable probabilistic deep learning framework for seismic assessment of structures using distribution-free prediction intervals. *Computer-Aided Civil and Infrastructure Engineering*, *38,* 1677–1698. https://doi.org/10.1111/mice.13015