Machine learning modelling of structural response for different seismic signal characteristics: A parametric analysis

MACHINE LEARNING MODELLING OF STRUCTURAL RESPONSE FOR DIFFERENT SEISMIC SIGNAL CHARACTERISTICS: A PARAMETRIC ANALYSIS

M. De Iuliis[1], E. Miceli[2], P. Castaldo[3]

[1] Department of Structural, Geotechnical and Building Engineering (DISEG), Politecnico di Torino, Turin, Italy, e-mail: massimiliano.deiuliis@polito.it

[2] Department of Structural, Geotechnical and Building Engineering (DISEG), Politecnico di Torino, Turin, Italy, e-mail: elena.miceli@polito.it

[3] Department of Structural, Geotechnical and Building Engineering (DISEG), Politecnico di Torino, Turin, Italy, e-mail: paolo.castaldo@polito.it

ABSTRACT

The present study investigates the best seismic parameters for modeling the dynamic response of various non-linear structural systems by comparing different Machine Learning (ML) algorithms. A total of 400 synthetic excitations were generated and analyzed against 23 seismic parameters. These signals were used in a step-by-step numerical analysis to calculate the dynamic responses of 1000 single-degree-of-freedom (SDOF) systems with varying mechanical properties. The data obtained from these responses were processed using 20 ML algorithms, including linear regression, tree, support vector machine (SVM), boosted and bagged trees, and artificial neural network (ANN). Each ML algorithm used a single seismic parameter as input to determine the most predictive parameters for modeling structural responses, defining the high predictive seismic parameters (HPSP) set.

To validate the obtained results, the most effective model predictions have been compared with the results of the parametric step-by-step analyses performed for a new group of natural ground motions. The findings demonstrate that with a properly calibrated training phase, considering the specific site hazard and selecting seismic parameters from the HPSP set, the ML model can accurately estimate seismic responses whit a significantly reduced computational effort. This study underscores the potential of integrating ML techniques into the performance-based seismic design approach.

Keywords: Machine learning algorithm; artificial excitation; non-linear SDOF system; seismic parameter; peak relative displacement; predictive metrics.

1. INTRODUCTION

Machine Learning (ML) techniques have developed exponentially in recent decades, achieving a fundamental role in many areas of science, finance and engineering. By definition [1], ML is a set of algorithms that gives computers the ability to achieve a mathematically formalised goal without being explicitly programmed, based solely on the available data and capacity to derive useful information from the data. ML algorithms can be divided into three main categories: supervised learning, unsupervised learning and reinforcement learning. Supervised learning category uses prior information contained in a labelled dataset to determine a model that as best as possible approximates the relationship between the input and labelled output data. Differently, unsupervised learning aims to infer the natural structure present in a no labelled dataset, while reinforcement learning is based on an autonomous agent's searching for the best strategy to maximise its gain through interaction with an environmental system.

The use of these procedures in the field of civil engineering research dates to the early 2000s, with the studies by Adeli [2] having the idea of integrating soft-computing techniques such as artificial neural networks, genetic algorithms, fuzzy logic and wavelet analysis into the solution of numerical engineering problems. An extensive analysis on evolutionary computation in the context of structural design, such as topological optimal design, was conducted by Kicinger et al. [3], while Saka and Geem [4] examined various mathematical modelling schemes aimed at the optimal design of steel frames. In recent years, the application of these soft computing techniques to civil engineering has been considerably accelerated due to the availability of high-performance computing machines, at ever lower costs, capable of handling and processing large databases in an acceptable timeframe. Vadyala et al. [5] highlights the potential of physics-based ML models in solving real-world problems, while Kumar [6] discusses the diverse applications of artificial intelligence (AI) in civil engineering projects, including structural design, construction management and structural health monitoring. Ahmad et al. [7] and Beskopylny et al. [8] focus on the use of advanced ML approaches, such as bagging and artificial neural networks, in predicting the compressive strength of concrete and in various civil engineering applications. Use of AI to increase safety management of civil engineering construction is analyzed by Zhang [9], while Zhang et al [10] studied the promising perspective concerning the application of ML approach in geotechnical engineering. An interesting overview of the application of ML to building design is provided by Sun et al. [11], they show how ML model can be useful for predicting and assessing structural performance, identifying structural condition and informing preemptive and recovery decisions by extracting patterns from data collected via various sources and media. Cao et al. [12] focus on combining artificial neural network with other intelligence algorithms to obtain more accurate results in evaluating the different parameters of composite beams and shear connectors and determining the compressive strength of concrete. A review on current ML algorithms implemented in building structural health monitoring systems has been proposed by Gomez-Cabrera and Escamilla-Ambrosio [13], this approach has proven successful in predicting the level of damage in a hierarchical classification. Finally, Zhang et al. [14] discussed the main contributions related to the use of ML for building load prediction. These studies collectively underscore the growing role of ML and AI in enhancing the efficiency, accuracy and sustainability of civil engineering projects.

As for the applications of ML to earthquake engineering, several notable contributions are present in the literature. Xie et al. [15] provided an exhaustive overview of the numerous contributions, distinguishing four sectors into which the corresponding scientific literature can be divided: seismic hazard analysis, system identification and damage detection, seismic fragility assessment, structural control for earthquake mitigation.

ML techniques therefore offer a new paradigm in the modelling and prediction of the dynamic response of a structural system, no longer an evaluation based on a physical model of reality, but on the processing of databases collected from experiments, historical information or numerical simulations performed on physical models to build substitute models able to predict the structural response. Several studies confirm the applicability and effectiveness of ML methods for seismic response analysis of different structural types [16]-[19], both in the linear and non-linear field [20]-[21] as well as in the case where the interaction between soil and structure is explicitly considered [22]. Recently, Demertzis et al. [23] focused their attention on the problem of method interpretability, highlighting how a lack of expertise associated with the use of complex ML architectures can affect the performance of the intelligent model and, ultimately, reduce the algorithm's reliability and generalization which should characterize these systems. Sun [24] drew attention to problems related to the accuracy of the model and computational effort, by proposing an alternative data-driven framework to reconstruct structure responses through ML techniques from limited available sources which may potentially benefit for "real-time" application.

Regarding the use of ML techniques aimed also to effectively characterise a seismic signal, it is possible to select appropriate data from the available databases of recorded events on a global scale [25]-[26] for modelling future events from a probabilistic point of view and studying their impact on

the built environment. There are already many studies conducted in this sense dealing with neural networks [27]-[29], genetic programming techniques [30], modelling through tree structures [31] and mixed methods [32]-[33].
With reference to the seismic modelling of a structural system, the characterisation of the incoming seismic signal represents a crucial point in vulnerability assessment, particularly when the dynamic response of such systems to events with a high return period [34] is strongly non-linear. Few contributions are devoted to this specific issue in the literature [35]-[36]. The mathematical model of the structural dynamic response is generally obtained, through supervised or unsupervised ML techniques, without a preliminary study aimed at understanding which variables as best as possible characterise the properties of the seismic excitation and, therefore, allow for a more reliable seismic modelling of the structure. Similar sensitivity studies have been conducted in the past [37], however, always within a traditional numerical approach. In this context, the analysis of the information hidden in a seismic signal can be very useful to simulate the dynamic response of a structure with very important applicative advantages: it allows for the definition of more reliable structural models, it limits the amount of information to be stored and processed for a single seismic signal and, consequently, the computation time, making it possible to apply ML techniques within classical seismic design approach without high performance computer machines.
The present study describes a wide investigation to define the best seismic parameters characterising earthquake records aimed at modelling the dynamic response of several non-linear structural systems using different ML algorithms. In detail, 400 synthetic excitations have been generated and examined with respect to 23 seismic parameters. Using these signals, a parametric numerical analysis has been conducted to calculate the dynamic responses of 1000 equivalent elastic perfectly-plastic single-degree-of-freedom (SDOF) systems with different mechanical characteristics subjected to such excitations. The choice of this system is justified to balance the simplicity and interpretability of the model and its representativeness with respect to a large class of structural systems. The chosen model, in fact, is widely used in all the most advanced standards regarding, for example, non-linear static analysis methods (e.g., pushover analysis). Modelling of the dynamic response on the basis of the data obtained has followed through supervised learning techniques using 20 ML algorithms: linear regression, tree, support vector machine (SVM), boosted and bagged trees, and artificial neural network (ANN). The processing phase has been carried out by considering a single seismic parameter representative of excitations as input data for each ML algorithm to define the best seismic parameter able to characterise structural responses as well as the best type of algorithm. Finally, to validate the results obtained, the most effective model predictions have been compared with the results of the parametric step-by-step dynamic analyses developed for a new group of natural ground motions.

2. GENERATION OF ARTIFICIAL SEISMIC RECORDS

An essential need has always been to define models as realistic as possible for seismic actions to use within design procedures. This need arises from the consideration that the safety level of a structural system can be accurately defined through non-linear dynamic analyses and, consequently, the use of ML techniques within earthquake engineering strengthens this need. Indeed, any supervised learning algorithm requires a very large number of data, representative of the dynamic behaviour of the system, to effectively fine-tune the high number of parameters aiming at defining a reliable structural system model. This modelling procedure is generally denoted as "training phase" and is based on the comparison between the simulated response obtained from the numerical algorithm and actual response of the structural system to several accelerograms. To this scope, many accelerograms are required and herein generated randomly. Each artificial record can be defined as the composition of a random process with an assigned power spectral density (PSD) and of a modulation function to model the time variance of the seismic record intensity. Specifically, the model of Kanai and Tajimi [38]-[39], which is still the most widely used in this context due to its effectiveness and relative simplicity [40], suggests a PSD function coinciding with that related to the random response in terms

of acceleration of a SDOF oscillator with parameters ω_g and ξ_g, subject to a stationary process with infinite variance (i.e., white noise), of intensity S_0, as follows:

$$S_a = \frac{\omega_g^4 + 4\xi_g^2\omega_g^2\omega^2}{\left(\omega_g^2 - \omega^2\right) + 4\xi_g^2\omega_g^2\omega^2} \cdot S_0 \tag{1}$$

Depending on the choice of the parameters ω_g and ξ_g, Eq.(1) accurately represents the PSD of the random process for different soil conditions. In this model, the soil is represented by a linear filter (i.e., a SDOF oscillator) excited at the bedrock level by a process having infinite bandwidth [41]. However, this formulation has the drawback of having a soil displacement PSD which diverges in the case of frequencies approaching zero. This led Clough and Penzien [42] to add a second filter, in series with the first one, with the parameters ω_2 and ξ_2. The second filter is useful to avoid energy concentration at low frequencies and applies:

$$S_{CP} = S_a \cdot \frac{\omega^4}{(\omega_2^2 - \omega^2) + 4\xi_2^2\omega_2^2\omega^2} \tag{2}$$

By neglecting the evolution of the frequency with the time, each seismic excitation can be modelled as a Gaussian stationary process having mean value equal to zero and two-sided PSD function $S_f = S_{CP}$. Realizations of such a process can be obtained using the spectral representation method [43], which leads to the following expression:

$$X(t) = \sqrt{2} \cdot \sum_{k=1}^{N} \sqrt{2 \cdot S_f(\omega_k) \cdot \Delta\omega} \cdot \cos\left(\omega_k + \Phi_k\right) \tag{3}$$

where, being ω_u the upper limit of the frequency domain, beyond which the power density is considered to be zero, and dividing ω_u into N parts, $\Delta\omega = \frac{\omega_u}{N}$, $\omega_k = k \cdot \Delta\omega = k \cdot \frac{\omega_u}{N}$ and Φ_k represents the *k*-th realization of a uniformly distributed random variable in the interval $[0,2\pi]$.
With the aim of accounting for the non-stationarity in real seismic signals, a deterministic function modulating the stochastic process is introduced and defined in the time domain, as follows:

$$X_{NS}(t) = e(t) \cdot X(t) \tag{4}$$

where $e(t)$ is denoted as "modulation function" and should strictly depend on both time and frequency. In the following, Shinozuka and Sato proposal [44] is used and applies:

$$e(t) = \delta_1 \cdot \left[e^{-\beta_1 t} - e^{-\beta_2 t}\right] \tag{5}$$

By using the described model, 400 artificial seismic excitations with variable duration have been randomly generated. The sample size, coherent with [45], was chosen in such a way as to strike a balance between the need for a large amount of data for the model training phase and, at the same time, avoid biasing phenomena related to the presence of seismic signals with identical characteristics. The parameters ω_g and ξ_g are modeled as independent random variables, uniformly distributed in the intervals $[3\pi, 5\pi]$ (rad/s) and $[0.6,1]$, respectively. The terms ω_f and ξ_f are considered as deterministic parameters: $\omega_f = 1.6\, rad/s$ and $\xi_f = 0.6$. The other parameters are set

as follows: $\omega_u = 12\pi$ (mean value of the uniform probability density function defined between $[8\pi, 16\pi]$), $N = 250$.
The term S_0 is modeled in accordance with the formulation proposed by [46] to achieve an average peak ground acceleration (PGA) of 0.7 g over the 400 seismic records. Finally, the parameters of the modulation function are, respectively, set equal to $\delta_1 = 25.302$, $\beta_1 = 0.141$ and $\beta_2 = 0.157$. These assumptions are consistent with the hypothesis of a stiff soil site [47].

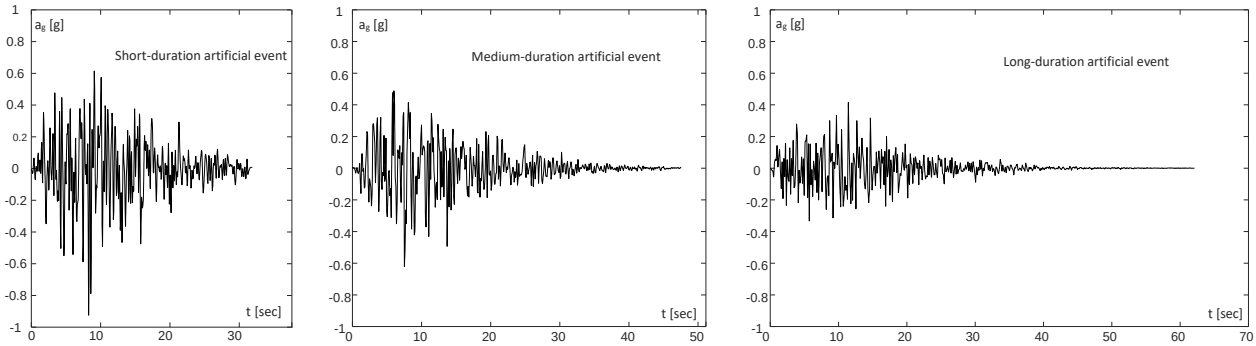

Figure 1: Examples of three artificial seismic records with different durations.

Figure 1 shows the time-histories of three artificial seismic signals having short, medium and long duration, respectively. Meanwhile, Figure 2 represents the elastic pseudo-acceleration spectra S_a[g] for all the artificial seismic records including also the mean spectrum and ASCE design spectrum [48] corresponding to class A site (i.e., stiff soil - $S_{ds} = 1.7g$, $S_{d1} = 0.8g$, $T_L = 1.5\,sec$), for a damping factor of 5%.

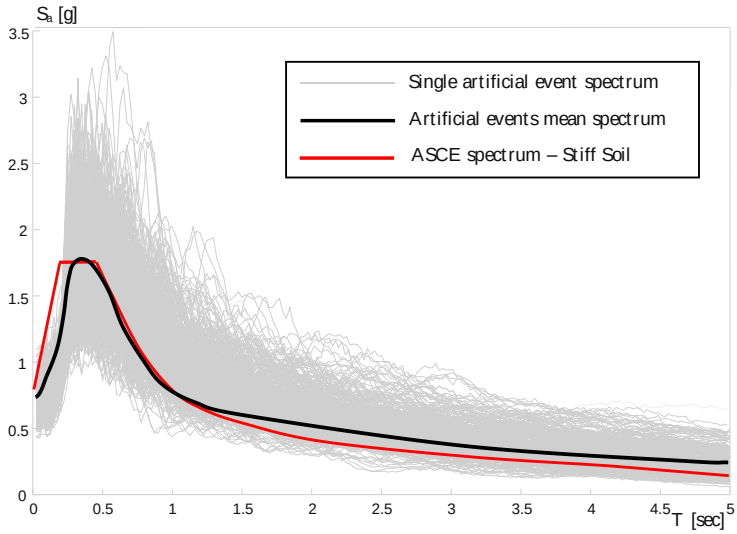

Figure 2: Elastic pseudo-acceleration spectra of the 400 artificial seismic records.

3. DYNAMIC RESPONSE OF SEVERAL NON-LINEAR SDOF SYSTEMS

The assessment of seismic performance of non-linear structural systems with respect to the characteristics of seismic inputs represents one of the primary objectives within earthquake engineering. This issue requires an accurate modeling of the non-linear behavior of the system, a suitable representation of expected collapse modes, and an effective probabilistic modeling of both material properties and input seismic signal features. In this context, the adoption of ML models can represent a turning point in the numerical analysis of the non-linear response of a system. However, the use of such algorithms should never neglect the physical understanding of the obtained results. This suggests the need of classifying the numerous seismic parameters, regarding the input motions,

involved in numerical analysis, to identify the most significant ones with respect to the seismic response.
As for the structural system, the elastic perfectly-plastic SDOF model has been adopted for evaluating the seismic response in terms of displacement. This model, introduced by Shibata and Sozen [49], is widely used in both scientific literature and design provisions [50]-[52] and is characterized by three parameters: the elastic stiffness (k), the yield strength (f_y) and the peak relative displacement capacity (x_m) (Figure 3a). The dynamic equation governing the seismic response of a non-linear SDOF system with mass m, inherent viscous coefficient c and resisting force $f_S(x)$ can be written, as follows:

$$\ddot{x} + 2\xi\omega_n\dot{x} + \omega_n^2 x_y \overline{f}_s(x) = -\ddot{u}_g \tag{6}$$

where $\omega_n = \sqrt{\frac{k}{m}}$, $\xi = \frac{c}{2m\omega_n}$, $\overline{f}_s(x) = \frac{f_S(x)}{f_y}$. In the following, the value of the yield strength has been defined in each case as the product between the peak value of the earthquake-induced resisting force in the corresponding linear system f_0 and yield strength factor R_y, as follows: $f_y = R_y \cdot f_0$. During the elastic response $\overline{f}_s(x)$ represents the restoring force and depends on the stiffness k. The corresponding elastic vibration period is denoted as T_0.

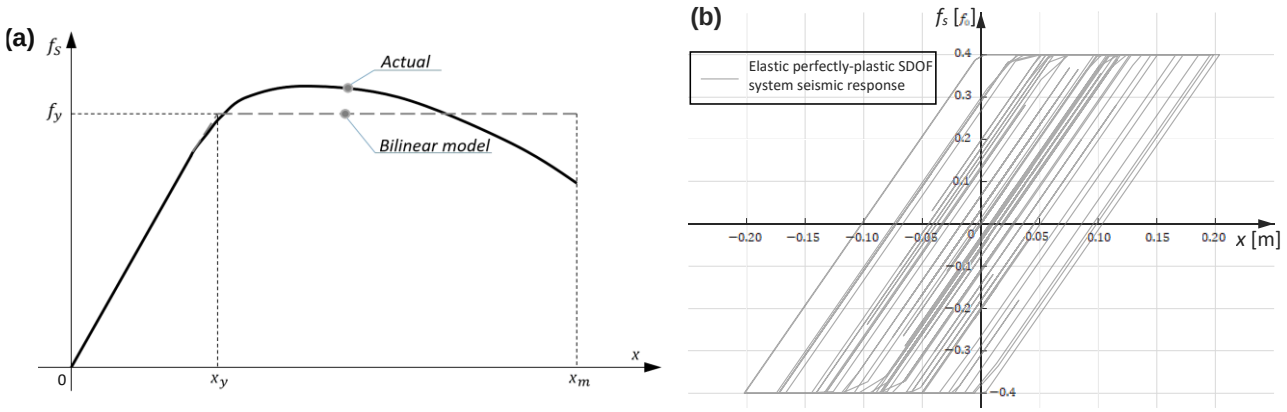

Figure 3: Force-relative displacement curve (a), force-relative displacement time-history (b).

In Figure 3b, a typical dynamic response in terms of force-relative displacement (i.e., $f_s{-}x$) for a non-linear SDOF system considered within the parametric analysis is showed. The yield strength is the same in both directions of the response, and the unloading phase from a point of maximum deformation occurs along a path parallel to the elastic branch. The deformation cycle, besides exhibiting a non-linear trend, is clearly dependent on the previous loading history, as well as the effect of seismic action at a specific instant is a function of the condition of the oscillator.
A comprehensive parametric analysis, aimed at determining the relative displacement demand and, thus, the maximum ductility required for the 400 generated seismic excitations, has been carried out. Specifically, for each seismic excitation, 1000 different elastic perfectly-plastic SDOF systems have been considered. The initial stiffness k has been varied such that the elastic vibration period of the system varies in the range $[0.05\ sec, 3.00\ sec]$. Similarly, the yield strength has been varied so that the parameter R_y ranges from a minimum value of 0.05 to a maximum value of 1, corresponding to a linear dynamic response of the system.
The results of the parametric analysis for the considered set of artificial seismic excitations are represented in Figures 4-6. The figures show, respectively, the peak relative displacement, x_{peak}, peak relative velocity, $\dot{x}_{peak}$ and peak absolute acceleration, $\ddot{u}_{peak}$, to a single record and their mean values to all the 400 records on varying elastic and inelastic mechanical parameters of the SDOF systems.
In the following, the peak relative displacement is the response parameter of the non-linear SDOF systems to predict through ML models.

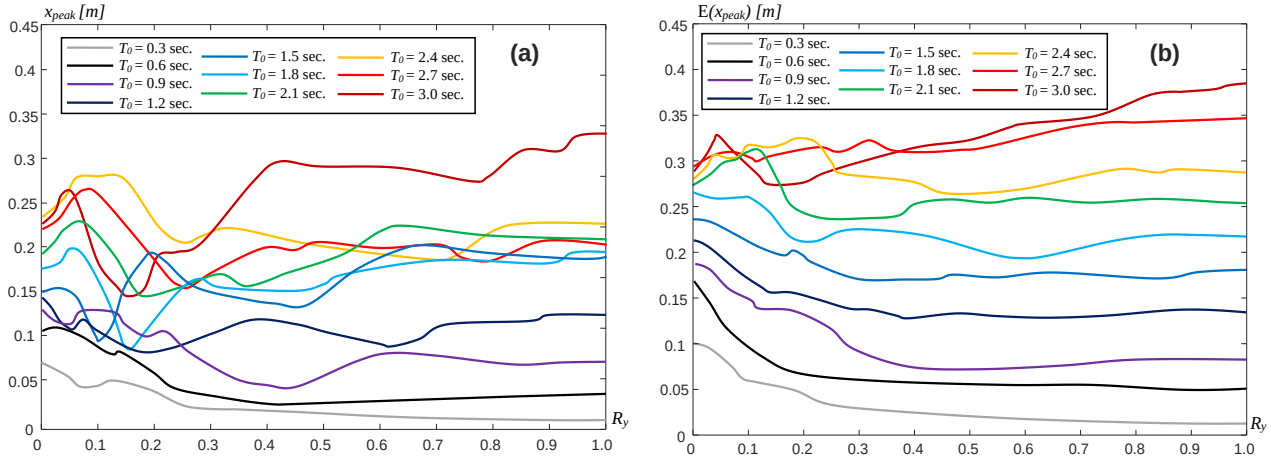

Figure 4: Peak relative displacements of the SDOF systems: (a) single artificial record, (b) mean over the 400 records.

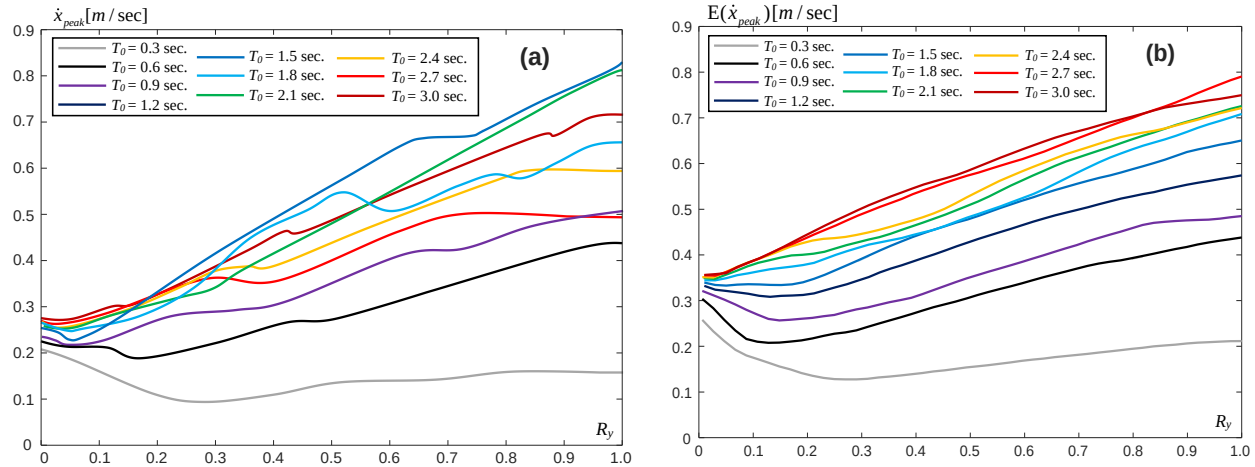

Figure 5: Peak relative velocity of the SDOF systems: (a) single artificial record, (b) mean over the 400 records.

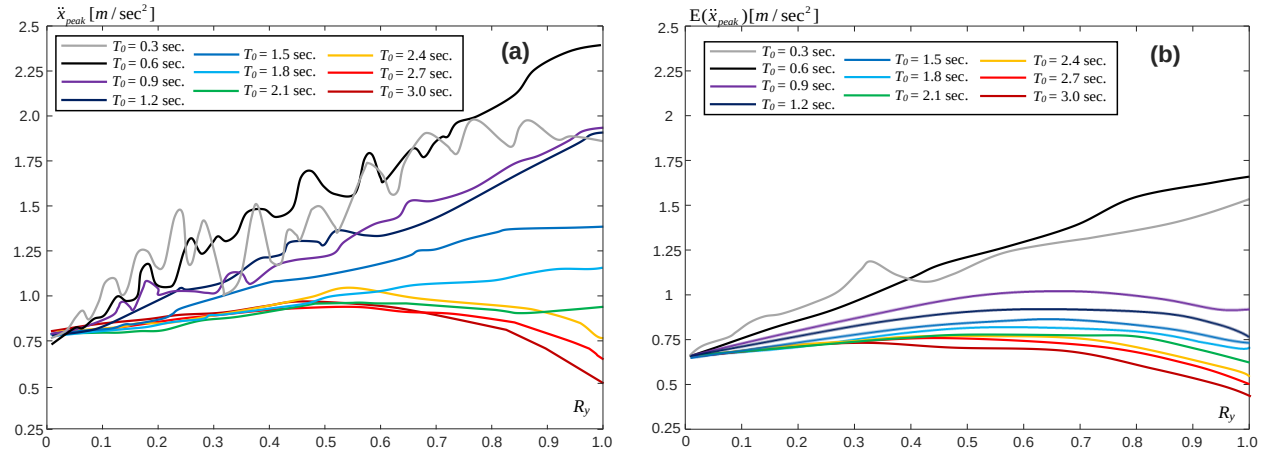

Figure 6: Peak absolute acceleration of the SDOF systems: (a) single artificial record, (b) mean over the 400 records.

4. SUPERVISED MACHINE LEARNING MODELS

The outcomes obtained from the parametric analysis, discussed in the previous section, have been utilized as input and output data in an extensive numerical analysis, by considering 20 ML algorithms with the aim of studying the relationship between the mechanical parameters characterizing the SDOF system, characteristics of the seismic excitations and dynamic responses of the systems in terms of the peak relative displacements. Specifically, the two parameters characterizing the system, namely k and R_y, together with a single specific seismic parameter, representing the seismic signal, have been provided as input data for the specific ML algorithm defining the corresponding ML model. The

aim of the numerical analysis is to identify the seismic parameters suitable for modeling non-linear dynamic response of a wide class of structural systems using supervised ML techniques. Therefore, 23 seismic parameters have been considered [53], which encompass all relevant physical aspects of a seismic excitation, including those related to kinematic quantities such as acceleration, velocity and displacement, as well as the energy carried by the signal, quantities pertaining to the entire frequency spectrum, and time duration terms. These parameters are listed and described in Table 1. The supervised ML algorithms adopted in the numerical analysis belong to five macro-categories:

- Linear Regression: this category represents the classical approach aiming to model the relationship between input and output variables through linear functions [60]. Four different versions of algorithms have been considered:
    - Linear: the algorithm minimizes the sum of squared differences between predicted and observed values;
    - Interaction linear: the algorithm introduces interactive terms capturing non-linear relationships or conditional dependencies between independent variables;
    - Robust: the algorithm introduces a loss function to mitigate the impact of outliers during training, ensuring reduced sensitivities to errors from anomalous values;
    - Stepwise: the algorithm sequentially adds or removes variables basing on statistical criteria, this simplifies the model by considering only the most significant variables.
- Tree: this category is based on decision trees, which iteratively split data into homogeneous subsets based on their most significant features. Three versions of algorithms have been utilized: fine, medium and coarse, based on the number of terminal nodes expected for the tree structure. Finer structures capture more details from input data but are more susceptible to overfitting, creating a model specific to the training set rather than generalizing to the entire dataset [61].
- Support Vector Machine (SVM): the goal of this category is to find a hyperplane in an N-dimensional space that optimally classifies input data. Six versions of algorithms have been tested: three versions named as linear, quadratic and cubic SVM, utilizing polynomial kernel functions of the first, second and third degree, respectively. These kernels transform data from the original space to a higher-dimensional space to define more complex decision boundaries. The other three versions, named as SVM with fine or medium or coarse Gaussian kernel, present a kernel function between two vectors defined by a Gaussian function, with its width depending on parameter γ. This parameter varies in the three cases, influencing the algorithm's ability to adapt to complex patterns [62].
- Ensemble: this category combines different approaches to enhance overall performance. Two distinct versions of algorithms are considered:
    - Boosted Trees: this algorithm combines several simpler approaches, often shallow decision trees or weak learners, to create a more robust predictive ML model. The iterative training process focuses on correcting errors made by preceding ML models.
    - Bagged Trees: this algorithm combines multiple instances of decision trees using bootstrap sampling with replacement. Each tree is trained on a different subset of the data, introducing diversity to the ensemble and improving overall robustness [63].
- Neural Networks: this category is inspired by the functioning of the human brain, consisting of layers of interconnected artificial neurons (i.e., nodes). Each node receives input, undergoes a weighted transformation and produces an output. Training involves optimizing connection weights, enabling the network to learn complex patterns in the data. The various algorithms derive from the number of nodes considered and number of hidden layers. Specifically, five versions of algorithms are considered: three with a single hidden layer of different sizes: narrow (i.e., 10 nodes), medium (i.e., 25 nodes) and wide (i.e., 100 nodes); two with two and three hidden layers, respectively [64].

Table 1: Seismic parameters in numerical analysis.

S_p	Seismic parameter	Description
1	Peak Acceleration	Peak acceleration of the seismic input time-history: $\boldsymbol{a_{g,peak}}$
2	Peak Velocity	Peak velocity of the seismic input time-history: $\boldsymbol{v_{g,peak}}$
3	Peak Displacement	Peak displacement of the seismic input time-history: $\boldsymbol{x_{g,peak}}$
4	AV ratio	Peak ground acceleration-to-velocity ratio: $\boldsymbol{AV} = a_{g,peak}/v_{g,peak}$
5	Acceleration RMS	Square root of the mean of acceleration squares: $\boldsymbol{a_{RMS}} = \sqrt{\frac{\int_0^{T_t} a^2 dt}{T_t}}$, where T_t is the time duration of the signal
6	Velocity RMS	Square root of the mean of velocity squares: $\boldsymbol{v_{RMS}} = \sqrt{\frac{\int_0^{T_t} v^2 dt}{T_t}}$, where T_t is the time duration of the signal
7	Displacement RMS	Square root of the mean of displacement squares: $\boldsymbol{x_{RMS}} = \sqrt{\frac{\int_0^{T_t} x^2 dt}{T_t}}$, where T_t is the time duration of the signal
8	Arias Intensity	Time-integral of the square of the acceleration: $\boldsymbol{I_A} = \frac{\pi}{2g}\int_0^{T_d} a(t)^2 dt$, where T_d is the time duration of the signal above a threshold equal to $0.05g$ [54]
9	Characteristic Intensity	$\boldsymbol{I_C} = {a_{RMS}}^{1.5} \cdot {t_s}^{0.5}$, where t_S is the significant duration [55]
10	Specific Energy Density	Energy of the velocity-time history: $\boldsymbol{SED} = \int_0^{T_d} v(t)^2 dt$, where T_d is the time duration of the signal above a threshold equal to $0.05g$
11	Cumulative Absolute Velocity	Area under the accelerogram over a significant duration: $\boldsymbol{CAV} = \int_0^{T_t} \lvert a(t) \rvert dt$, where T_t is the time duration of the signal
12	Acceleration Spectrum Intensity	Area under the acceleration response spectrum between periods of 0.1 sec and 0.5 sec: $\boldsymbol{ASI} = \int_{0.1}^{0.5} PSA(T,\xi) dT$, where PSA is the pseudo spectral acceleration; T is the natural period and ξ=0.05 is the damping coefficient
13	Velocity Spectrum Intensity	Area under the velocity response spectrum between periods of 0.1 sec and 0.5 sec: $\boldsymbol{VSI} = \int_{0.1}^{0.5} PSV(T,\xi) dT$, where PSV is the pseudo spectral velocity; T is the natural period and ξ=0.05 is the damping coefficient [56]
14	Housner Intensity	$\boldsymbol{SI_H} = \int_{0.1}^{2.5} S_v(T,\xi) dT$, where S_v is the pseudo-velocity response spectrum; T is the natural period and ξ=0.05 is the damping coefficient [57]
15	Sustained Maximum Acceleration	Third-highest acceleration absolute value largest peak in the time history record: $\boldsymbol{SMA}$
16	Sustained Maximum Velocity	Third largest velocity peak in the time history record: $\boldsymbol{SMV}$
17	Effective Design Acceleration	Peak acceleration that remains after filtering out seismic signals above 9 Hz: $\boldsymbol{EDA}$
18	A95 parameter	Peak value of acceleration corresponding to 95% of Arias Intensity value: $\boldsymbol{A95}$ [57]
19	Predominant Period	Period corresponding to the maximum spectral acceleration occurs in a 5% damped acceleration response spectrum: $\boldsymbol{T_p}$
20	Mean Period	Average frequency content: $\boldsymbol{T_m} = \sum_i \left(\frac{C_i^2}{f_i}\right) / \sum_i C_i^2$, where C_i are the Fourier amplitudes, and f_i represent the discrete Fourier transform frequencies between 0.25 and 20 Hz [58]
21	Uniform Duration	Total time of the earthquake motion in which the amplitudes exceed a specified threshold (0.025g): $\boldsymbol{D_u}$ [59]
22	Bracketed Duration	Elapsed time between the first and last acceleration excursions greater than a specified threshold (0.025g): $\boldsymbol{D_b}$ [59]
23	Significant Duration	Time required for the release of accumulated strain energy by rupture along the fault. It is herein defined as the duration between 5% and 95% of the Arias intensity: $\boldsymbol{D_S}$ [59]

Therefore, 20 different algorithms have been analyzed with respect to the 23 seismic parameters to model the responses of the 1000 non-linear SDOF systems to the 400 artificial records. In this way, 460 ML models have been implemented for the 400 signals to reproduce the response of the 1000 non-linear SDOF systems. No preliminary analyses based on feature selection techniques were carried out, as these may discard parameters whose predictive role may be significant. This can happen due to limitations of the model or possible interactions between the parameters present in the model [65]. For this reason, a comprehensive analysis involving all seismic parameters available in the literature was preferred, rather than reducing the number of parameters through techniques that could limit the generality of the results.

As for ML model validation, the method called "*n*-fold cross-validation" has been employed. This implies to divide the original dataset, composed of the 400000 cases, into *n* equal parts, denoted as "validation folds". The training and validation process is repeated n times, each time using the *j*-th fold (with *j*=1… *n*) as validation set and the remaining $n-1$ folds as the training set. This way, each observation in the dataset has the opportunity to be in both the training and validation sets. During each iteration, the model is trained on the $n-1$ training folds and evaluated on the basis of the single validation fold. In this study, considering the overall number of cases, *n* was set to 10, dividing the dataset into 10 subsets for each analyzed case (i.e., each algorithm with each seismic parameter). Two evaluation metrics are then computed: specifically, the root mean squared error (RMSE) and the Mean Absolute Error (MAE) have been considered. Furthermore, for each algorithm and each considered model, the variance of performance across the 10 different folds used in the cross-validation process has been evaluated.

The RMSE provides a measure of the dispersion between the model-predicted values and actual values in the validation dataset, as expressed by Eq.(7):

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(x_i - x_{i,\,pred}\right)^2} \tag{7}$$

in which N represents the total number of observations in the validation dataset (i.e., N=400000), x_i is the target parameter value for the model, represented by the peak relative displacement of the non-linear SDOF system obtained from the parametric analysis of Section 3, and $x_{i,pred}$ is the predicted value by using the considered ML model.

The MAE is a measure of the accuracy of a prediction model, it is the average of the absolute errors, which is the absolute value of the difference between the model's predictions and the observed actual values, Eq.(8):

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|x_i - x_{i,\,pred}\right| \tag{8}$$

in which N represents the total number of observations in the validation dataset (i.e., N=400000).

The combined use of both metrics allows for a more comprehensive evaluation of the model's performance. In fact, while the RMSE reveals the presence of outliers and the level of dispersion in the predictions, the MAE provides a simple and straightforward measure of the average error, being less sensitive to high-intensity prediction errors. This approach is particularly useful when dealing with limited datasets, aiming to obtain reliable performance estimates. It also helps to identify the variability in model performance across different partitions of the dataset, providing a more robust assessment of its ability to capture the relationship between input and output data.

To show some results obtained from the validation process, Fig. 7 reports the RMSE mean and standard deviation values among the 10 considered folds for each one of the 20 ML algorithms, assuming the peak acceleration (Table 1) as the seismic parameter for the ML models for all 400

signals and 1000 non-linear SDOF systems. Figure 8 shows the same values for the MAE predictive metric.

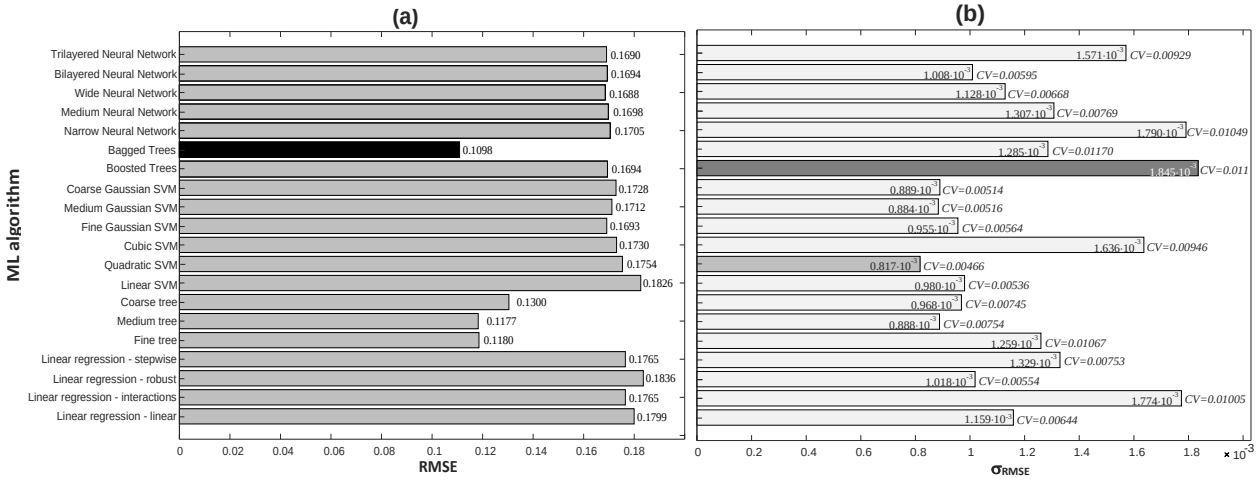

Figure 7: RMSE mean (a) and standard deviation (b) values for each considered algorithm assuming the peak acceleration as seismic parameter.

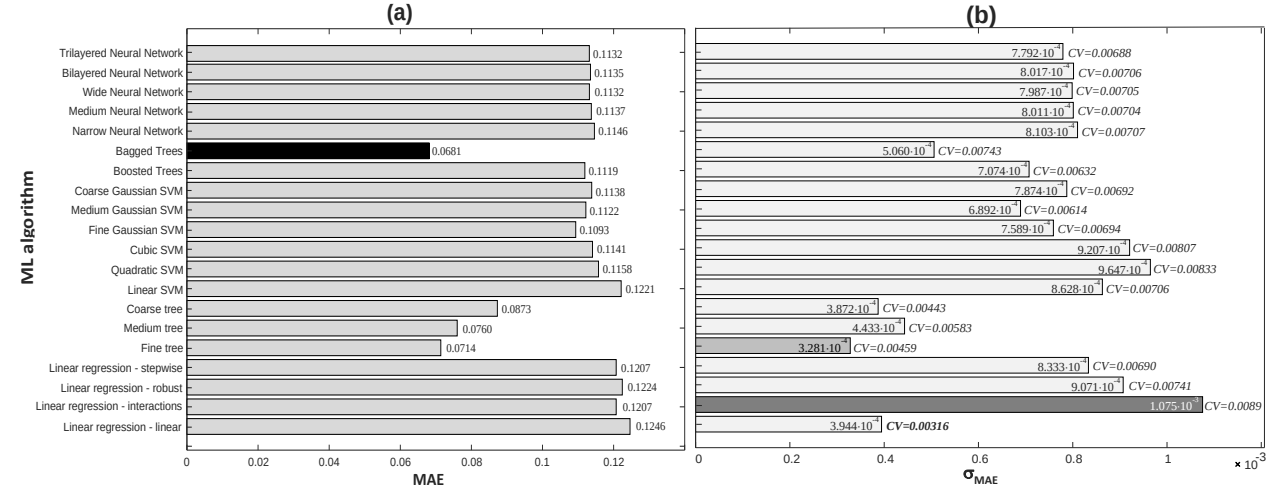

Figure 8: MAE mean (a) and standard deviation (b) values for each considered algorithm assuming the peak acceleration as seismic parameter.

Figures 7 and 8 highlight that, considering the peak acceleration as single seismic parameter, the Bagged Trees algorithm leads to the best predictive model in terms of both the considered performance metrics. Moreover, the numerical analyses aimed at evaluating the level of dispersion within the cross-validation procedure lead to very limited standard deviations for the different algorithms considered (coefficient of variation between 0.003 and 0.01), proving significant stability in predictive performance and indicating the significance of the values obtained through the considered performance metrics. In addition, for such optimal algorithm (i.e., Bagged Trees), Figure 9 shows the partial dependent plot where the structural response in terms of peak relative displacement averaged over the 1000 SDOF systems predicted by the ML model is illustrated with respect to the peak acceleration of the 400 signals. This representation allows understanding the relationship between an input variable and the predicted output of a ML model, taking into account the combined effect of other mechanical characteristics. Specifically, it shows how the prediction changes as one input variable varies while keeping the others constant to their mean value.

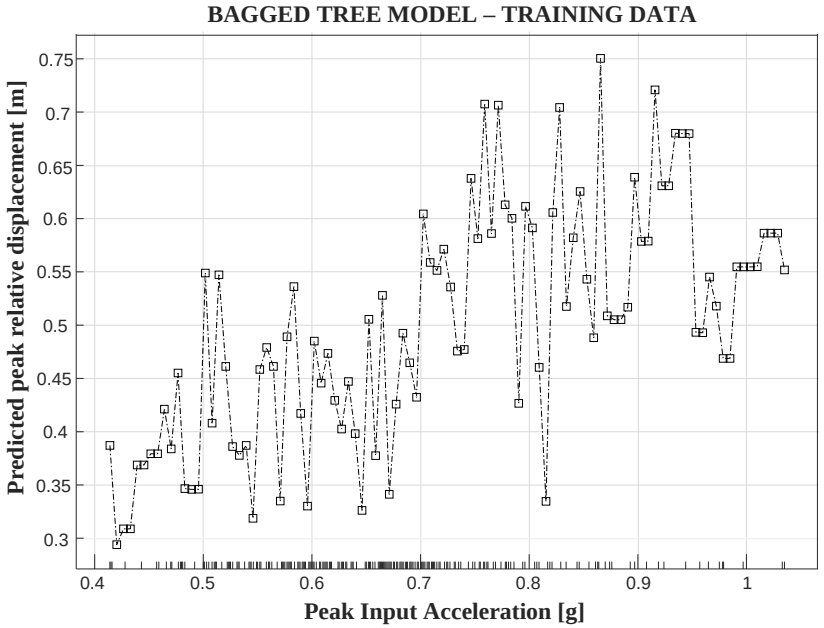

Figure 9: Partial dependent plot for bagged trees model in terms of peak relative displacement.

5. COMPARISON BETWEEN THE ML MODELS

The performance of all the 20 considered ML algorithms has been assessed with respect to the 23 seismic parameters through numerical analyses, where a single seismic parameter, denoted as S_P (Table 1), is provided as input data together with 400 signals and dynamic characteristics of the 1000 elastic perfectly-plastic SDOF systems. The main idea is to numerically test and compare the significance of the seismic parameters within a modeling approach based on ML techniques, aiming to understand which ones are more effective or contain more useful information for estimating the non-linear dynamic response of the structural systems.

To achieve this goal, a total of 460 ML models has been implemented (i.e., the 23 seismic parameters S_P and 20 different ML algorithms). For each of these models, the effectiveness in terms of model prediction has been analyzed by using both the RMSE and MAE indexes. The results obtained are summarized in Figures 10-17.

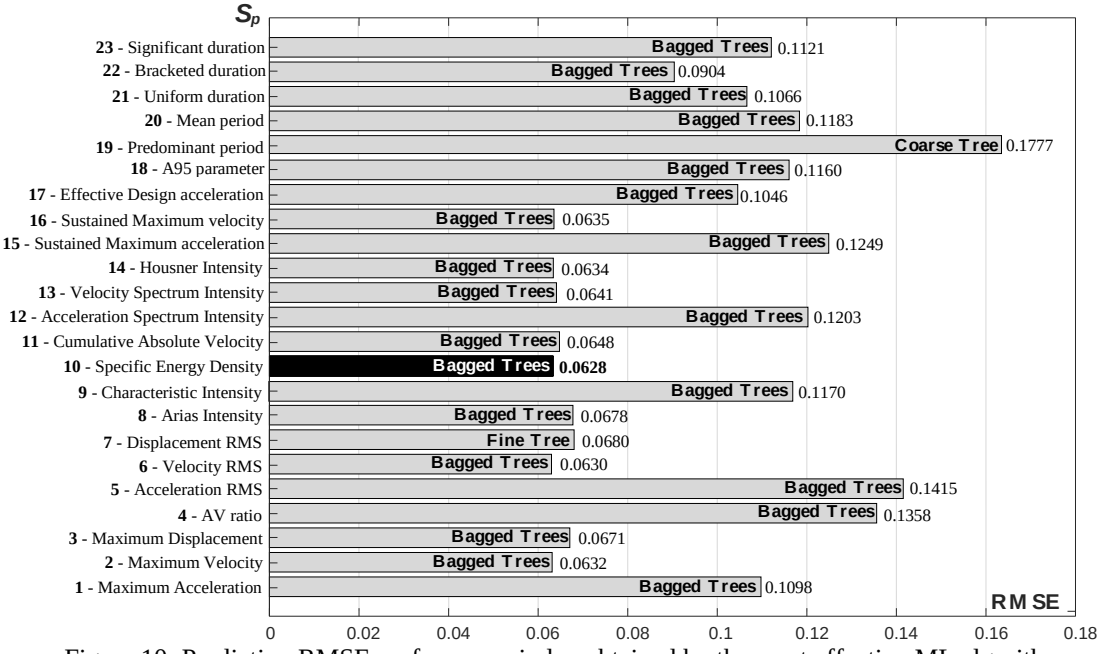

Figure 10: Prediction RMSE performance index obtained by the most effective ML algorithm.

In Figures 10-11, for each seismic parameter S_P, the ML algorithm that proved to have the best performance by considering, respectively, the RMSE and MAE performance index is pointed out, along with the corresponding values, for all the signals and non-linear SDOF systems. Similarly to the previous section, the numerical analyses also show significant stability and statistical robustness of the predictive performance obtained within the cross-validation process leading to state how the most effective algorithm in achieving the prediction objective is quite always Bagged Trees.

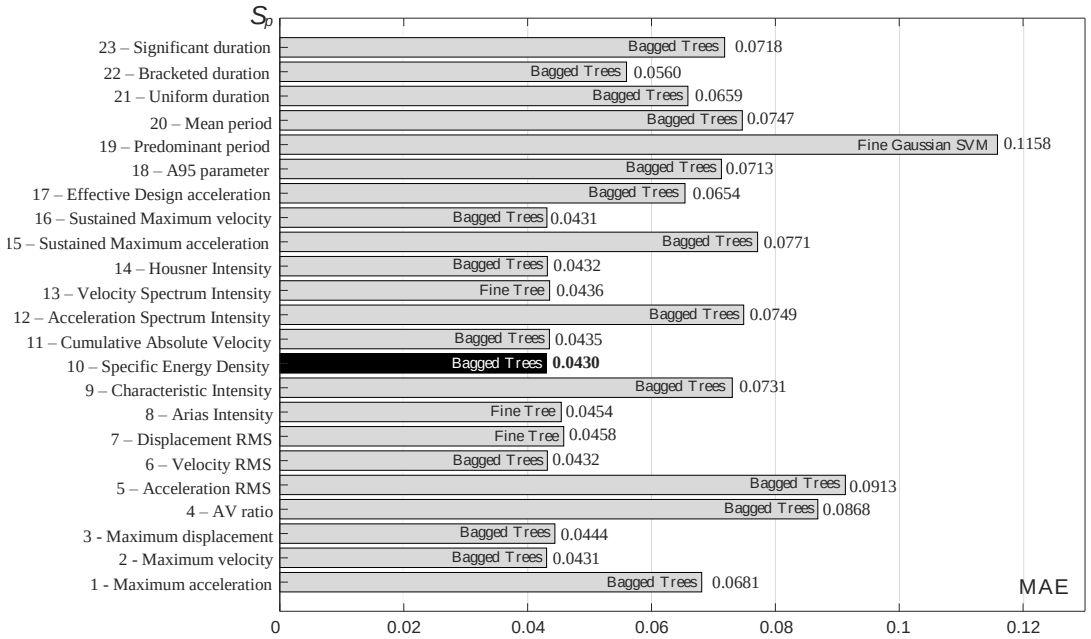

Figure 11: Prediction MAE performance index obtained by the most effective ML algorithm.

In order to support this statement, a Welch's t-test was conducted to compare the RMSE index values obtained during the 10-fold cross-validation procedure for Bagged Tree and its best alternative among the other algorithms, for each of the models obtained by varying the input seismic parameter excluding cases SP=7 and SP=19, for which the Bagged Trees algorithm does not represent the best choice in terms of predictive performance. This test aims to compare the means of two data groups and represents a more flexible version of the standard t-test, used when the variances of the groups are different. Eq.s(9)-(10), respectively, provide the t value used for the test and the number of degrees of freedom, ν, characterizing the distribution.

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{9}$$

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}} \tag{10}$$

where $\bar{X}_1$ and $\bar{X}_2$ are the RMSE performance index values for the considered model, $n_1 = n_2 = 10$ are the sample sizes, s_1^2 and s_2^2 are the sample unbiased variances. In the Welch's t-test, the number of degrees of freedom calculated using Eq.(10) is generally a non-integer value; it will subsequently

be approximated to the nearest lower integer, ν'. This approach is conservative and ensures that the test is more stringent, reducing the risk of Type I error (false positive).

Table 2: Welch's t-test p –value for the top predictive algorithm and its best alternative.

	RMSE PREDICTIVE PERFORMANCE INDEX									
	SP = 1		SP = 2		SP = 3		SP = 4		SP = 5	
	Bagged trees	Fine tree	Bagged trees	Fine tree	Bagged trees	Fine tree	Bagged trees	Medium tree	Bagged trees	Medium tree
$\overline{X}$	0.1098	0.1180	0.0632	0.0644	0.0671	0.0673	0.1358	0.1372	0.1415	0.1423
s^2	$1.65\cdot10^{-6}$	$1.59\cdot10^{-6}$	$8.36\cdot10^{-7}$	$7.41\cdot10^{-7}$	$9.54\cdot10^{-7}$	$3.98\cdot10^{-7}$	$1.06\cdot10^{-6}$	$7.55\cdot10^{-7}$	$1.07\cdot10^{-6}$	$1.07\cdot10^{-6}$
ν'	17		17		15		17		17	
t	14.391		3.072		0.299		3.167		1.77	
$p-value$	$2.986\cdot10^{-11}$		$3.468\cdot10^{-3}$		0.384		$2.817\cdot10^{-2}$		$4.733\cdot10^{-2}$	
	SP = 6		SP = 8		SP = 9		SP = 10		SP = 11	
	Bagged trees	Fine tree	Bagged trees	Fine tree	Bagged trees	Medium tree	Bagged trees	Fine tree	Bagged trees	Fine tree
$\overline{X}$	0.0630	0.0647	0.0679	0.0681	0.1170	0.1214	0.0628	0.0642	0.0648	0.0654
s^2	$9.82\cdot10^{-7}$	$8.41\cdot10^{-7}$	$4.36\cdot10^{-7}$	$7.60\cdot10^{-7}$	$2.77\cdot10^{-6}$	$1.62\cdot10^{-6}$	$3.48\cdot10^{-7}$	$3.43\cdot10^{-7}$	$7.58\cdot10^{-7}$	$1.29\cdot10^{-6}$
ν'	17		16		16		17		16	
t	7.621		0.781		6.685		5.250		1.127	
$p-value$	$3.511\cdot10^{-7}$		0.223		$2.620\cdot10^{-6}$		$3.258\cdot10^{-5}$		0.138	
	SP = 12		SP = 13		SP = 14		SP = 15		SP = 16	
	Bagged trees	Medium tree	Bagged trees	Fine tree	Bagged trees	Fine tree	Bagged trees	Medium tree	Bagged trees	Fine tree
$\overline{X}$	0.1203	0.1262	0.0641	0.0646	0.0634	0.0640	0.1249	0.1284	0.0635	0.0643
s^2	$8.72\cdot10^{-7}$	$4.52\cdot10^{-7}$	$2.60\cdot10^{-7}$	$2.44\cdot10^{-7}$	$4.29\cdot10^{-7}$	$4.28\cdot10^{-7}$	$2.07\cdot10^{-6}$	$1.84\cdot10^{-6}$	$4.46\cdot10^{-7}$	$1.25\cdot10^{-6}$
ν'	16		17		17		17		14	
t	16.268		2.137		1.980		5.515		2.012	
$p-value$	$1.125\cdot10^{-11}$		$2.237\cdot10^{-2}$		$3.206\cdot10^{-2}$		$1.893\cdot10^{-5}$		$3.193\cdot10^{-2}$	
	SP = 17		SP = 18		SP =20		SP = 21		SP = 22	
	Bagged trees	Fine tree	Bagged trees	Medium tree	Bagged trees	Medium tree	Bagged trees	Fine tree	Bagged trees	Fine tree
$\overline{X}$	0.1046	0.1120	0.1160	0.1232	0.1183	0.1230	0.1066	0.1133	0.0904	0.0948
s^2	$1.73\cdot10^{-6}$	$1.92\cdot10^{-6}$	$9.31\cdot10^{-7}$	$1.88\cdot10^{-6}$	$9.83\cdot10^{-7}$	$7.05\cdot10^{-7}$	$1.11\cdot10^{-6}$	$1.64\cdot10^{-6}$	$1.42\cdot10^{-6}$	$1.48\cdot10^{-6}$
ν'	17		16		17		17		17	
t	12.128		13.657		11.441		12.735		8.231	
$p-value$	$4.277\cdot10^{-10}$		$1.546\cdot10^{-10}$		$1.040\cdot10^{-9}$		$2.016\cdot10^{-10}$		$1.236\cdot10^{-7}$	
	SP = 23									
	Bagged trees	Fine tree								
$\overline{X}$	0.1121	0.1191								
s^2	$2.37\cdot10^{-7}$	$1.44\cdot10^{-6}$								
ν'	11									
t	17.271									
$p-value$	$1.281\cdot10^{-9}$									

Given μ_1 and μ_2 as the mean values of the two performance metrics to be tested, the hypotheses of the test are stated in Eq.(11), and the test results, consistent with a significance level of 0.05, are summarized in Table 2.

$$\begin{cases} H_0 : \mu_1 - \mu_2 \leq 0 \\ H_1 : \mu_1 - \mu_2 > 0 \end{cases} \tag{11}$$

The Bagged Trees algorithm appears to be statistically better than the other considered algorithms (rejecting the null hypothesis H_0) in 18 out of 23 cases. In two of the five cases, it does not represent the best predictive algorithm based on the chosen metrics (i.e., SP=7, SP=19). In the remaining three cases, although the algorithm exhibits a better average performance, the sample statistics do not allow for a statistically significant validation of its superiority over at least one of the remaining algorithms (i.e., SP=3, SP=8, SP=11). It is also worth highlighting that the algorithms competitive with Bagged Trees are all decision tree-based algorithms, which appear to exhibit superior predictive performance compared to other types. Similar results are obtained if the MAE instead of the RMSE index is considered.

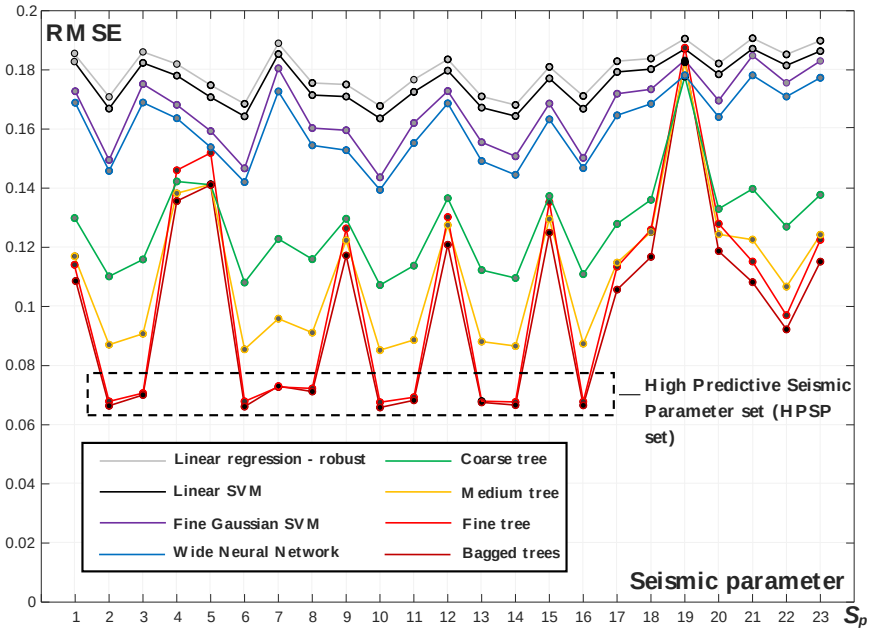

Figure 12: RMSE comparison between the different seismic parameters for some regression algorithms.

Results shown in Figures 10-11 lead to the conclusion that the parameter S_p plays a significant role in prediction accuracy. With the aim of studying this relationship, Figures 12 and 13 display the values of the RMSE and MAE performance indices for different algorithms for the various seismic parameters used as input data. Specifically, it is possible to split the set of the 23 seismic parameters into two groups: a group of ten elements where the RMSE index value ranges between 6.5% and 7% and MAE index between 0.04 and 0.05 (i.e., Highly Predictive Seismic Parameter, HPSP set), and a second group where this index is significantly greater. The HPSP set includes the following seismic parameters: Peak velocity, Peak displacement, Velocity RMS, Displacement RMS, Arias Intensity, Specific Energy Density, Cumulative Absolute Velocity, Velocity Spectrum Intensity, Housner Intensity and Sustained Maximum Velocity. It appears to be relevant to underline how in the HPSP set, the seismic parameters involving velocity and displacement, and the parameters having energetic nature obtained through integration of the seismic signal, are present. The use of parameters directly related to acceleration or time appears to be less effective in this context.

In Figure 14, the actual response in terms of peak relative displacement and response obtained using the Bagged Trees model (i.e., x_p), for all the signals and non-linear SDOF systems, are compared in two cases: one where the seismic parameter is the peak acceleration and another one where the seismic parameter is represented by the peak velocity (i.e., a HPSP). The difference between the predictions obtained with the two ML models is evident, highlighting how the use of peak velocity values allows

for a more efficient modeling. A similar comparison has been conducted, presenting identical results, for any pair of the seismic parameters S_p belonging to the two distinct groups.

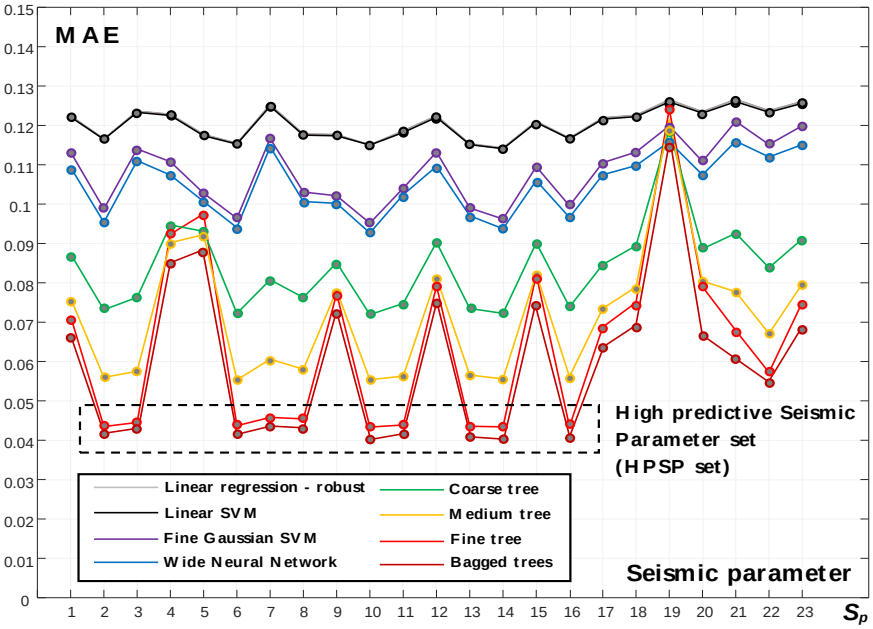

Figure 13: MAE comparison between the different seismic parameters for some regression algorithms.

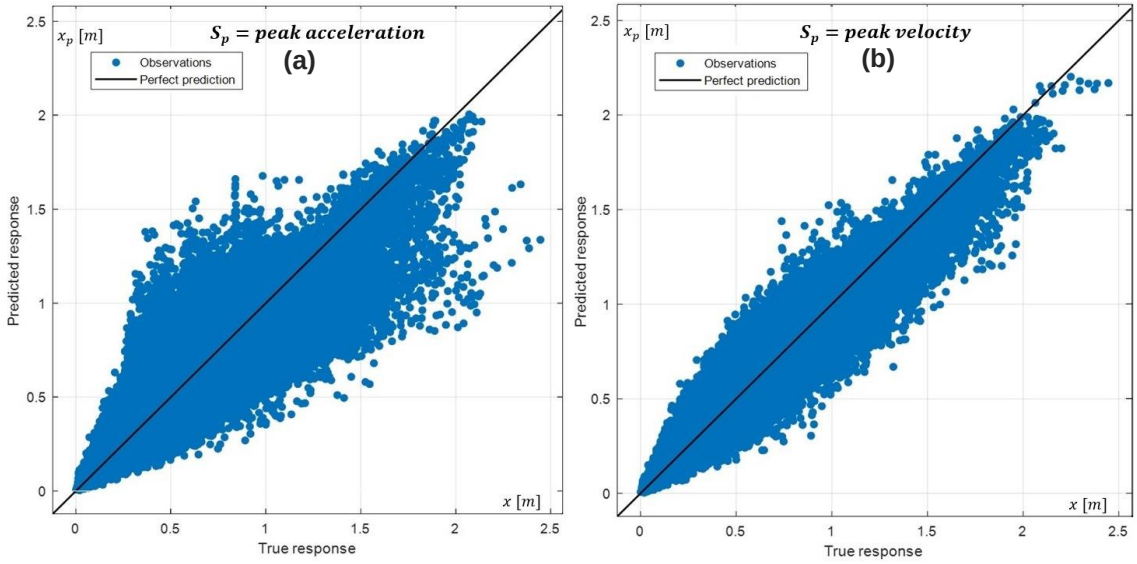

Figure 14: Comparison between the predicted response x_p and true response in Bagged Trees models, considering SP=1 (a) and SP=2 (b).

By considering the two seismic parameters for which the best prediction results have been obtained, namely the Specific Energy density and Velocity RMS, partial dependence plots are represented in Figures 15-17. Specifically, Figures 15a shows how the elastic period, T_0, of the structural system influences the peak relative displacement response predicted by the ML model averaging over the 400 signals and other structural characteristics in terms of R_y. Figure 15b shows how the yield strength factor, R_y, influences the peak relative displacement response predicted by the ML model averaging over the 400 signals and other properties in terms of T_0. Instead, Figure 16 shows how the two considered seismic parameters influence, respectively, the structural response in terms of peak relative displacement predicted by the ML model averaging over the 400 signals and 1000 SDOF systems.

Figure 15 is related to both the Specific Energy density and Velocity RMS as seismic parameters, being the two corresponding figures identical. The mentioned figures highlight how the Bagged Trees

model, without any knowledge of the time-history of the input signal but simply using a single seismic parameter characterizing such a signal, effectively manages to capture, for all parameters belonging to the HPSP set, the dependence of the non-linear response on the mechanical characteristics of the SDOF system. In particular, the relative displacement demand is greater in the case of lower stiffness levels (Fig. 15a), while the trend of the peak relative displacement as a function of the elastic resistance of the equivalent SDOF system is consistent with the various existing formulations for defining yield strength factors within the displacement-based design [66] (Fig. 16).

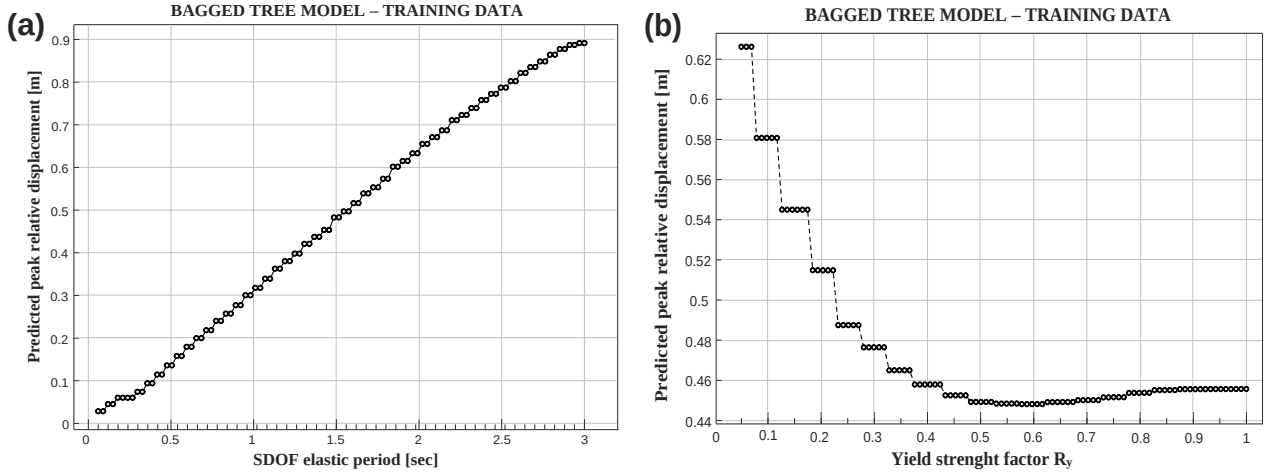

Figure 15: Partial dependence plot: SDOF system elastic period vs model prediction (a), SDOF system yield strength factor vs model prediction (b).

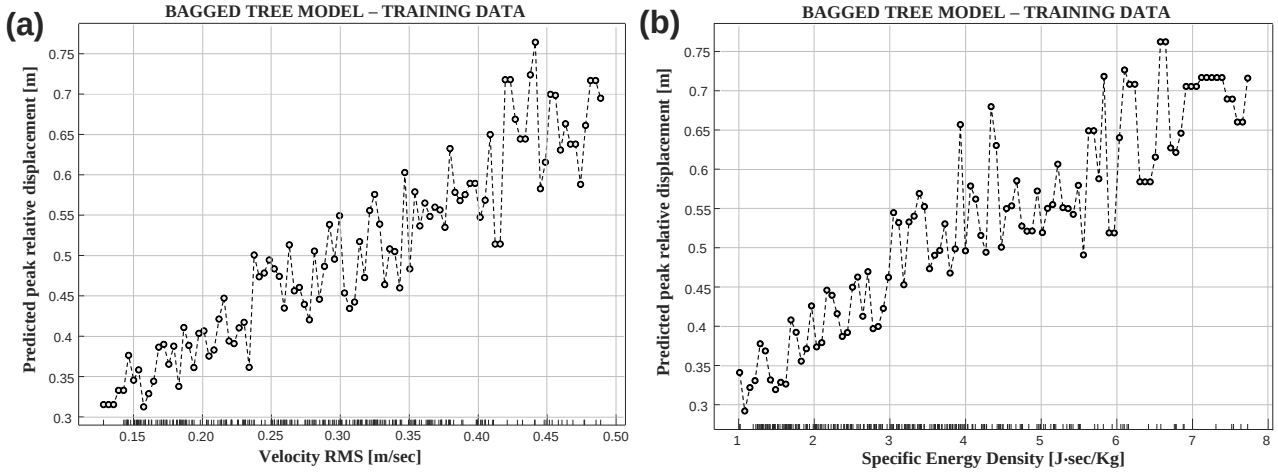

Figure 16: Partial dependence plot: seismic parameter ($\boldsymbol{v_{RMS}}$) vs model prediction (a), seismic parameter ($\boldsymbol{SED}$) vs model prediction (b).

Figure 16 relates the seismic parameter to the output represented by the peak relative displacement of the structural system. The analysis of this relationship is particularly useful for understanding the physical role that each seismic parameter plays in the non-linear response of the system. In detail, the Bagged Trees model predicts a rather regular increasing trend, highlighting a significant linear correlation between both input data and response. This regularity of correlation, despite the substantial equivalence in predictive performance of the ML model, is observed in quite all the ML models related to the HPSP set, as can be seen in Figure 17, where the same graph is reported for the other seismic parameters belonging to the HPSP set.

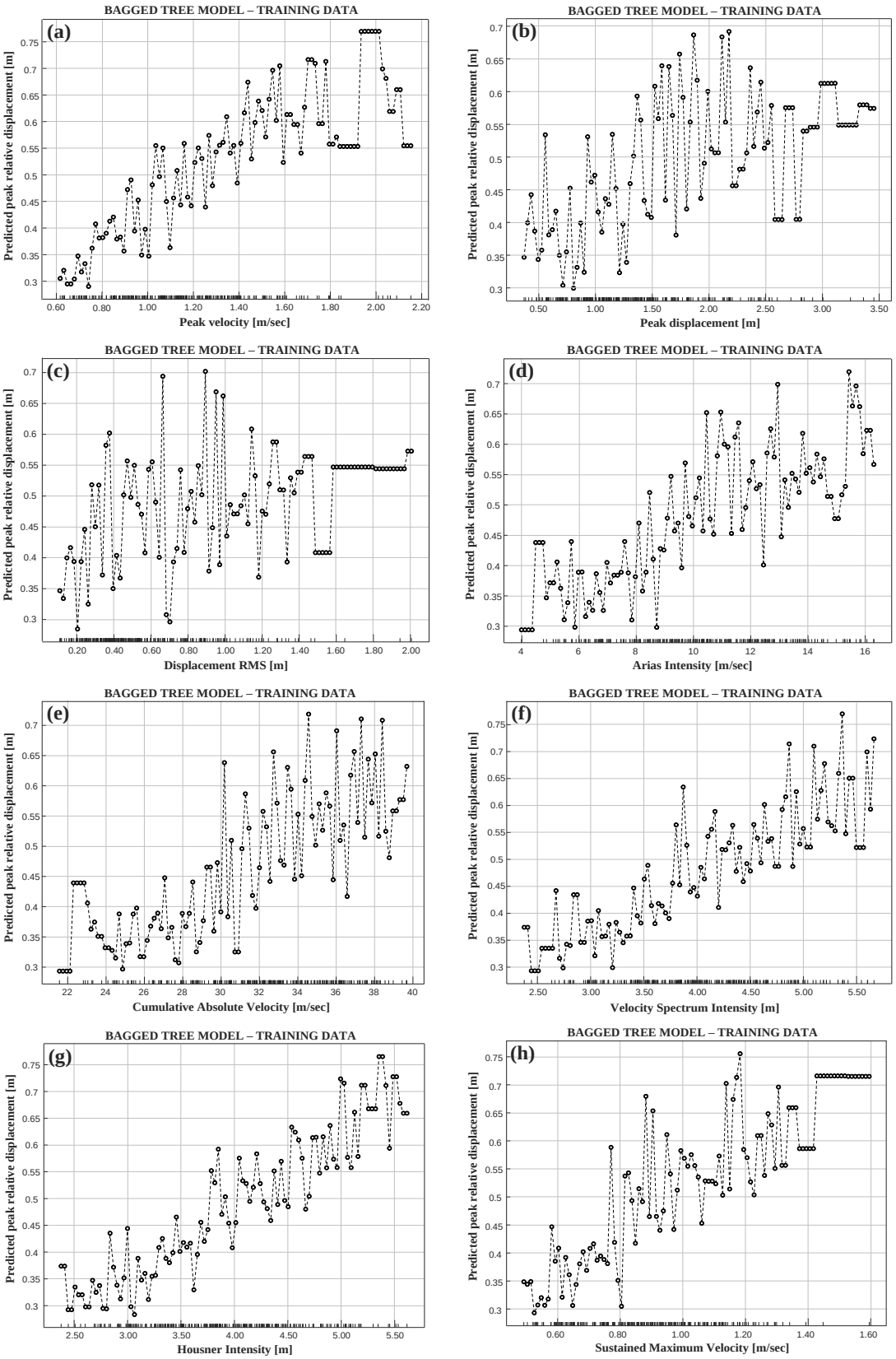

Figure 17: Partial dependence plots: seismic parameter in the HPSP set vs model prediction.

6. MACHINE LEARNING MODELS WITHIN PERFORMANCE-BASED SEISMIC DESIGN

The most advanced seismic regulations foresee the possibility, in specific cases (e.g., the design of safety-critical structures, highly irregular buildings, base-isolated structures and structures designed

for a high level of ductility), to perform non-linear dynamic analyses considering suites of accelerograms. Although this approach is theoretically the most general, it is characterized by numerical difficulties and long processing times. After identifying a seismic scenario in terms of magnitude, source-to-site distance and site geology, a perspective evolution can be the definition of a spectral shape on the basis of specific site hazard studies [67] and a generation of several seismic excitations compatible with this spectrum. These seismic signals can be used for the training phase of a ML algorithm capable of predicting the seismic performance of a structure on varying one or more seismic parameters and the main mechanical characteristics of the structural system. The advantage of such an approach is relevant: the estimation of the system performance would be carried out through numerically stable functions with dramatically reduced computational times, allowing for a preliminary real-time assessment of the safety level of the structure.
To this aim, the proposals of this study are useful to understand which seismic parameters are the most suitable to capture the information necessary for estimating the relative displacement demand of an event. Secondly, it is important to investigate the level of reliability that the ML model presents in terms of its predictive objective. Although this second issue is not the subject of the present work, it is useful to compare the predicted response through the obtained Bagged trees model with the response that the structural system exhibits in the case of natural ground motions having similar dynamic characteristics with respect to the set used for the ML algorithm training phase.

Table 3: Selected ground motion events.

#	Earthquake Name	Year	Station Name	Magnitude	Mechanism
1	Northwest Calif-02	1941	Ferndale City Hall	6.6	strike slip
2	Imperial Valley-03	1951	El Centro Array #9	5.6	strike slip
3	Southern California	1952	San Luis Obispo	6.0	strike slip
4	Imperial Valley-05	1955	El Centro Array #9	5.4	strike slip
5	San Fernando	1971	Cedar Springs_ Allen Ranch	6.61	reverse
6	San Fernando	1971	Pacoima Dam (upper left abut)	6.61	reverse
7	San Fernando	1971	Pasadena - CIT Athenaeum	6.61	reverse
8	Managua_ Nicaragua-01	1972	Managua_ ESSO	6.24	strike slip
9	Managua_ Nicaragua-02	1972	Managua_ ESSO	5.2	strike slip
10	Point Mugu	1973	Port Hueneme	5.65	reverse

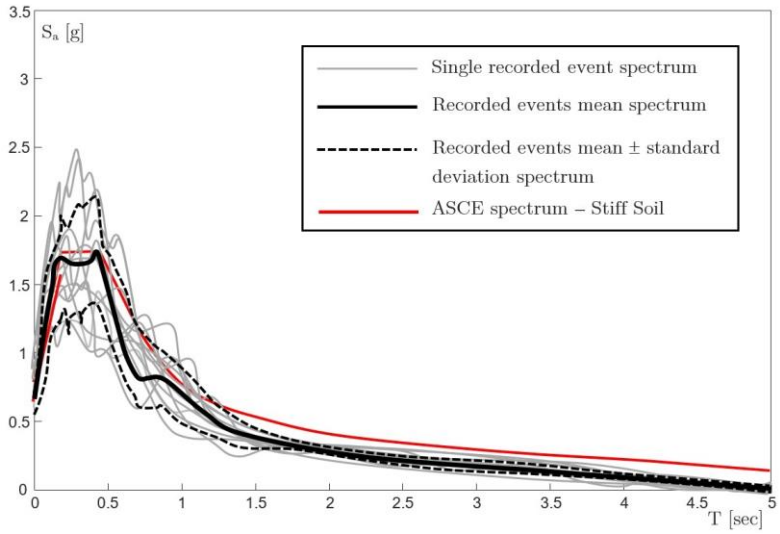

Figure 18: Elastic response spectra of the ground motions compared with ASCE spectrum for stiff soil.

From the PEER database [68], ten scaled seismic events (no-frequent records [69]) have been identified, whose spectra are compatible with the ASCE [48] spectrum. This latter one has been used as a reference for generating the artificial excitations adopted in the training phase of the ML algorithms to predict the structural system response. Table 3 reports the selected earthquake events, while in Figure 18, the ASCE reference spectrum is compared with the elastic response spectra of these natural ground motions, indicating also the average spectrum and spectra characterized by a unitary standard deviation.

In this context, it is important to emphasize that the goal of ML modeling is to calibrate the algorithm based on the specific hazard characteristics of the reference site where the structure will be built. In this perspective, the present work has highlighted that ML model becomes more efficient when specific characteristics, represented by the seismic parameters within the HPSP set, are taken into consideration. This means that a meaningful verification should focus on natural excitations whose seismic parameters, belonging to the HPSP set, have values close to those of the artificial records used for the training. In this regard, for sake of simplicity, the results of two events, namely earthquakes #6 and #10, are shown by varying the characteristics of the structural system. Such analyses involved the calculation of non-linear dynamic responses with two different approaches: i) step-by-step solution of the non-linear motion equations and ii) seismic peak response prediction using the ML techniques.

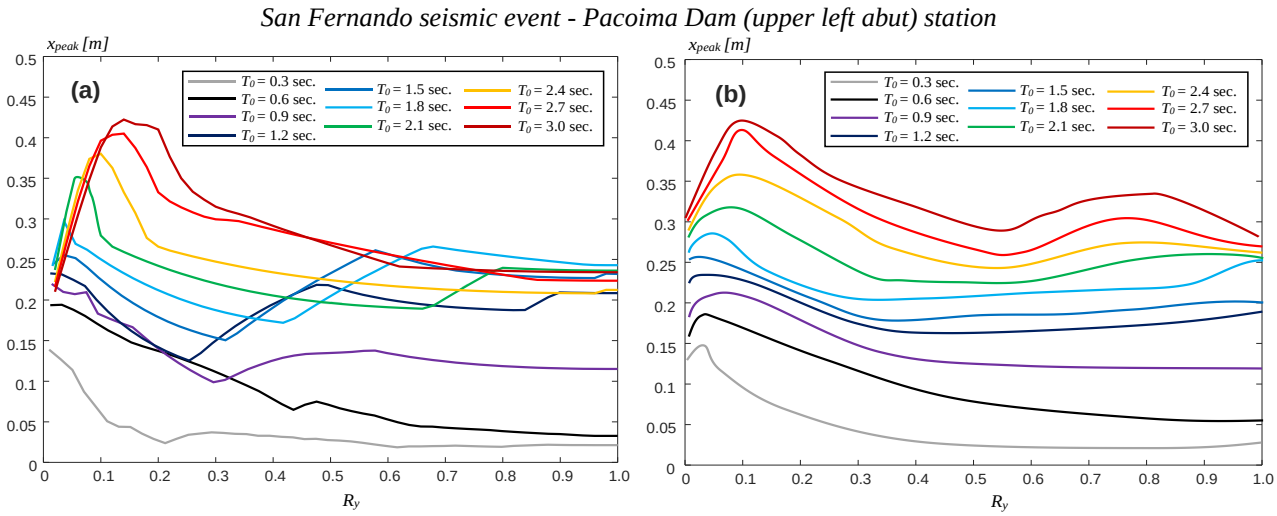

Figure 19: Comparison between seismic non-linear response obtained by the (a) dynamic analysis and (b) ML model.

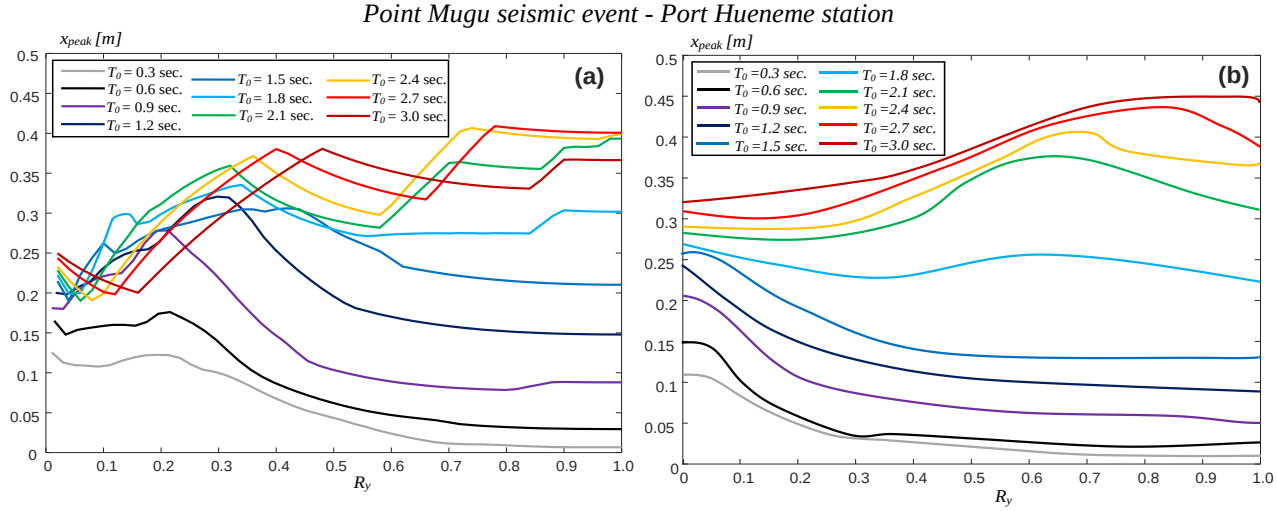

Figure 20: Comparison between seismic non-linear response obtained by the (a) dynamic analysis and (b) ML model.

The results of these analyses are summarized in Figures 19 and 20, respectively, for the two ground motions. The left graphs represent the peak relative displacement obtained from the non-linear

dynamic analyses as a function of the elastic period, T_0, and yield strength factor of the structural system, R_y. The right graphs relate to results predicted using the Bagged Trees model, considering the Cumulative Absolute Velocity as seismic parameter.
The comparison between the figures shows how, without having any information about the non-linear dynamic model, the ML model is capable of capturing the relationship between input and output data. It leads to predictions with maximum errors within 15-20% with respect to the traditionally calculated seismic response. These errors are greater for lower values of both peak relative displacement and yield strength, while they decrease when more relevant design cases are taken into account (i.e., structural systems having lower values of both stiffness and resistance).
Similar results are obtained when considering the Bagged Trees model trained with other seismic parameters from the HPSP set. Conversely, less promising results are observed when using other algorithms or when the training is conducted with the suboptimal seismic parameters, confirming the relevance of the present proposals.
The results of the described numerical analyses represent a promising initial step in integrating ML techniques into the context of the performance-based seismic design. If the training phase is appropriately calibrated with respect to the site hazard and seismic parameters selected within the HPSP set, the model is capable of estimating the performance level achieved by an equivalent structural system, reducing significantly the computational effort. This highlights the potential for enhancing the efficiency of the performance-based seismic design through the judicious incorporation of the ML methodologies.

7. CONCLUSIONS

In the field of structural engineering, the use of supervised learning algorithms, and in general, the potential offered by ML, is a topic that has recently attracted the interest of the scientific community. While these techniques appear to be promising in terms of numerical results, they pose challenges due to being black-box models, compromising the physical understanding of the problem and the control of information necessary for a correct design approach.
This contribution presents the results of a comprehensive investigation with a dual objective: firstly, to study the ML algorithms that can be effectively used in the context of a performance-based seismic design procedure, and secondly, to identify the seismic parameter, among those defined in the literature as representative variables of the seismic excitations, leads to the definition of the ML model with the best capability to predict the non-linear response of the structural system. This seismic parameter can be interpreted as the physical quantity with the most complete information to enable the model for a reliable prediction of the system response.
The results show that the family of algorithms called "Tree", specifically the "Bagged Tree", leads to models capable of making more reliable predictions of the non-linear dynamic response of a wide class of equivalent SDOF systems. The results also highlight that ten out of the twenty-three considered seismic parameters, provided as input data for the ML model, can be classified as highly predictive seismic parameters (i.e., HPSP) for the superior quality of the performance predictions. It is also noteworthy that the HPSP set is composed of the seismic parameters involving velocity, displacement and parameters with an energetic nature. Instead, the use of seismic parameters directly related to acceleration or time appears to be less effective, suggesting the need to reconsider the seismic hazard definition if ML models will be considered as an option in the context of performance-based design.
Finally, the seismic response in terms of peak relative displacement, obtained through the step-by-step resolution of the non-linear dynamic equations, is compared to the behavior predicted through the best ML model, selecting a seismic parameter from the HPSP set. This comparison was conducted considering natural ground motions properly scaled to meet the seismic parameter values similar to those of the artificial records used for the ML algorithm training phase. The results obtained show that the ML model is capable of estimating the performance level achieved by an equivalent SDOF

system, significantly reducing the computational effort, and can represent a promising approach to integrate ML models into performance-based design procedures.
The application of soft computing procedures within the field of earthquake engineering is a very broad and underexplored research area. Focusing on the proposed decision problem (i.e., selection of the most predictive seismic parameter), a future research direction should consider other advanced optimization algorithms (e.g., hybrid heuristics and metaheuristics, adaptive algorithms, self-adaptive algorithms, etc.). These algorithms should be tested, discussed and compared with the proposed approach to identify the most suitable solution for providing practical support in defining site-specific seismic hazard and a more reliable probabilistic description of the response of structural systems.

ACKNOWLEDMENTS

This study was carried out within the «Data fusion based digital twins for structural safety assessment» project – funded by European Union – Next Generation EU within the PRIN 2022 program (D.D. 104 - 02/02/2022 Ministero dell'Università e della Ricerca). This manuscript reflects only the authors' views and opinions and the Ministry cannot be considered responsible for them.

This study was also carried out within the RETURN Extended Partnership and received funding from the European Union Next-GenerationEU (National Recovery and Resilience Plan—NRRP, Mission 4, Component 2, Investment 1.3—D.D. 1243 2/8/2022, PE0000005).

REFERENCES

[1] A.L. Samuel, Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development* 1959; **3**(3): 535‑554.
[2] H. Adeli H, Neural networks in civil engineering: 1989-2000. *Computer-Aided Civil and Infrastructure Engineering* 2001; **16**(2): 126‑142.
[3] R. Kicinger, T. Arciszewski and K. De Jong, Evolutionary computation and structural design: A survey of the state-of-the-art. *Computers and Structures* 2005; **83**(23‑24): 1943‑1978.
[4] M.P. Saka and Z.W. Geem, Mathematical and metaheuristic applications in design optimization of steel frame structures: An extensive review. *Mathematical Problems in Engineering*, 2013; Article ID: 271031.
[5] S.R. Vadyala, S.N. Betgeri, J.C. Matthews, E. Matthews, A review of physics-based machine learning in civil engineering. *Results in Engineering*, 2022; **13**, 100316.
[6] M. Kumar, Application of Artificial Intelligence in Civil Engineering Projects. *Mathematical Statistician and Engineering Applications* 2022, **70**(1), 660-667.
[7] W. Ahmad, A. Ahmad, K.A. Ostrowski, F. Aslam, P. Joyklad and P. Zajdel, Application of advanced machine learning approaches to predict the compressive strength of concrete containing supplementary cementitious materials. *Materials*, 2021; **14** (19), 5762.
[8] A. N. Beskopylny, S.A. Stel'makh, E.M. Shcherban', L.R. Mailyan, B. Meskhi, I. Razveeva, A. Chernil'nik, N. Beskopylny, Concrete strength prediction using machine learning methods CatBoost, k-nearest neighbors, support vector regression. *Applied Sciences*, 2022: **12** (21), 10864.
[9] Y. Zhang, Safety management of civil engineering construction based on artificial intelligence and machine vision technology. *Advances in civil engineering*, 2021(1), 3769634.
[10] W. Zhang, H. Li, Y. Li, H. Liu, Y. Chen and X. Ding, Application of deep learning algorithms in geotechnical engineering: a short critical review. *Artificial Intelligence Review* 2021. 1-41.
[11] H. Sun, H.V. Burton and H. Huang, Machine learning applications for building structural design and performance assessment: State-of-the-art review. *Journal of Building Engineering* 2021, 33, 101816.
[12] Y. Cao, Y. Zandi, A.S. Agdas, Q. Wang, X. Qian, L. Fu, K. Wakil, A. Selmi, A. Issakhov, A.

Roco-Videla, A review study of application of artificial intelligence in construction management and composite beams. *Steel and Composite Structures, An International Journal*, 2021; **39**(6), 685-700.
[13] A. Gomez-Cabrera and P.J. Escamilla-Ambrosio, Review of machine-learning techniques applied to structural health monitoring systems for building and bridge structures. *Applied Sciences*, 2022, **12**(21), 10754.
[14] L. Zhang, J. Wen, Y. Li, J. Chen, Y. Ye, Y. Fu, and W. Livingood, A review of machine learning in building load prediction. *Applied Energy* 2021, 285, 116452.
[15] Y. Xie, M. Ebad Sichani, J.E. Padgett and R. DesRoches, R., The promise of implementing machine learning in earthquake engineering: A state-of-the-art review. *Earthquake Spectra*, 2020; **36** (4), 1769-1801.
[16] O.R. De Lautour and P. Omenzetter, Prediction of seismic-induced structural damage using artificial neural networks. *Engineering Structures* 2009; **31**(2), 600-606.
[17] B.K. Oh, Y. Park and H.S. Park, Seismic response prediction method for building structures using convolutional neural network. *Structural Control and Health Monitoring* 2020; **27**(5), e2519.
[18] S. Gharehbaghi, H. Yazdani and M. Khatibinia, Estimating inelastic seismic response of reinforced concrete frame structures using a wavelet support vector machine and an artificial neural network. *Neural Computing and Applications* 2020; **32**, 2975-2988.
[19] P. Hait, A. Sil and S. Choudhury, Seismic damage assessment and prediction using artificial neural network of RC building considering irregularities. *Journal of Structural Integrity and Maintenance* 2020: **5**(1), 51-69.
[20] T. Kim, O.S. Kwon and J. Song, Deep learning based seismic response prediction of hysteretic systems having degradation and pinching. *Earthquake Engineering & Structural Dynamics* 2023; **52**(8), 2384-2406.
[21] Y. Liao, H. Tang, R. Li, L. Ran, and L. Xie, Response Prediction for Linear and Nonlinear Structures Based on Data-Driven Deep Learning. *Applied Sciences* 2023; **13**(10), 5918.
[22] J. Won, J and J. Shin, Machine learning-based approach for seismic damage prediction method of building structures considering soil-structure interaction. *Sustainability* 2021; **13**(8), 4334.
[23] K. Demertzis, K. Kostinakis, K. Morfidis and L. Iliadis, An interpretable machine learning method for the prediction of R/C buildings' seismic response. *Journal of Building Engineering* 2023; **63**, 105493.
[24] H. Sun, A data-driven building seismic response prediction framework: From simulation and recordings to statistical learning. University of California, Los Angeles, 2019.
[25] L. Luzi, R. Puglia, E. Russo & ORFEUS WG5, Engineering Strong Motion Database, version 1.0. *Istituto Nazionale di Geofisica e Vulcanologia, Observatories & Research Facilities for European Seismology* 2016
[26] S. M. Mousavi, Y. Sheng, Y., W. Zhu, G.C. Beroza G.C., STanford EArthquake Dataset (STEAD): A Global Data Set of Seismic Signals for AI, *IEEE Access* 2019
[27] H. Bakhshi, A. Bagheri, G. Ghodrati Amiri and M.A. Barkhordari, Estimation of spectral acceleration based on neural networks. *Proceedings of the Institution of Civil Engineers - Structures and Buildings* 2014; **167**(8): 457–468.
[28] F. Khosravikia, P. Clayton and Z. Nagy, Artificial neural network-based framework for developing ground-motion models for natural and induced earthquakes in Oklahoma, Kansas, and Texas. *Seismological Research Letters* 2019; **90**(2A): 604‑613.
[29] P. Zhang, Y. Li, Y. Lin, H. Jiang, Time-Frequency Feature-Based Seismic Response Prediction Neural Network Model for Building Structures. *Applied Science* 2023; **13**(5), 2956
[30] A.F. Cabalar AF and A. Cevik, Genetic programming-based attenuation relationship: An application of recent earthquakes in turkey. *Computers and Geosciences* 2009; **35**(9): 1884‑1896.
[31] S.M. Hamze-Ziabari and T. Bakhshpoori T, Improving the prediction of ground motion

parameters based on an efficient bagging ensemble model of M5# and CART algorithms. *Applied Soft Computing Journal* 2018; **68**: 147‑161.
[32] A.K. Mohammadnejad, S.M. Mousavi, M. Torabi, M. Mousavi and A.H. Alavi, Robust attenuation relations for peak time-domain parameters of strong ground motions. *Environmental Earth Sciences* 2012; **67**(1): 53–70.
[33] M. Akhani, A.R. Kashani, M. Mousavi and A.H. Gandomi, A hybrid computational intelligence approach to predict spectral acceleration. *Measurement* 2019; **138**: 578–589.
[34] O.S. Kwon and A. Elnashai, The effect of material and ground motion uncertainty on the seismic vulnerability curves of RC structure. *Engineering structures* 2006; **28**(2), 289-303.
[35] F. Kazemi, N. Asgarkhani, R. Jankowski, Machine learning-based seismic fragility and seismic vulnerability assessment of reinforced concrete structures. Soil Dynamics and Earthquake Engineering,Volume 166, 2023, 107761
[36] Akbarnezhad M., Salehi M., DesRoches R., Application of machine learning in seismic fragility assessment of bridges with SMA-restrained rocking columns. Structures, Volume 50, 2023, 1320-1337.
[37] A. Massumi and F. Gholami, The influence of seismic intensity parameters on structural damage of RC buildings using principal components analysis. *Applied Mathematical Modelling* 2016; **40**(3), 2161-2176.
[38] K. Kanai, Semiempirical formula for the seismic characteristics of the ground. *Bulletin of Earthquake Research Institute* 1957; **35**, 309–325.
[39] H. Tajimi, A statistical method of determining the maximum response of a building structure during an earthquake. Proc., 2nd World Conference on Earthquake Engineering 1960; Vol. II, 781-798.
[40] H. Wang, Y. Cheng, Z. Lu, R. Huan, Q. Lü and Z. Liu, Z., Stochastic Response of Composite Post Insulators under Seismic Excitation. *Buildings* 2024, 14(6), 1539.
[41] P. Castaldo, G. Amendola, B. Palazzo, Seismic fragility and reliability of structures isolated by friction pendulum devices: seismic reliability-based design (SRBD). *Earthquake Engineering and Structural Dynamics* 2017; **46**, 425–446.
[42] R.W. Clough, J. Penzien, Dynamics of Structures (2nd edn). McGraw-Hill: New York, 1993.
[43] M. Shinozuka, G. Deodatis, Simulation of stochastic processes by spectral representation. *American Society of Mechanical Engineers*, 1991: 191-203.
[44] M. Shinozuka, Y. Sato, Simulation of nonstationary random process. Journal of Engineering Mechanical Division 1967; **93**(1): 11–40.
[45] Kundu, A., & Chakraborty, S., Deep learning-based metamodeling technique for nonlinear seismic response quantification. *IOP Conference Series: Materials Science and Engineering* 2020, Vol. 936, No. 1, pp. 12-42. IOP Publishing.
[46] C. Nuti, I. Vanzi, Influence of earthquake spatial variability on the differential displacements of soil and single degree of freedom structures, Rapporto Tecnico n.1, Dipartimento di Strutture, DIS, Università degli Studi di Roma Tre, Luglio 2004.
[47] X.L. Jin, Z. L. Huang and A.Y.T. Leung, Nonstationary seismic responses of structure with nonlinear stiffness subject to modulated Kanai–Tajimi excitation. *Earthquake engineering & structural dynamics* 2012, 41(2), 197-210.
[48] ASCE/SEI, Minimum Design Loads for Buildings and Other Structures, American Society of Civil Engineers, 2013.
[49] A. Shibata, M. Sozen, Substitute structure method for seismic design in reinforced concrete, *Journal of Structural Division, ASCE*, 1976, **102**(1): 1-18
[50] Applied Technology Council (ATC), Seismic evaluation and retrofit of concrete buildings, *Report ATC-40, Applied Technology Council*, 1996 Redwood City, CA, USA.
[51] A.K. Chopra, R.K. Goel, A modal pushover analysis procedure for estimating seismic demands for buildings, *Earthquake Engineering & Structural Dynamics* 2002, **31**(3), 561–582.
[52] G. Proietti, L. Pedone, S. D'Amore, S. Pampanin, Inelastic response spectra for an integrated

displacement and energy-based seismic design (DEBD) of structures. *Frontiers in Built Environment* 2023, **9**:1264033.
[53] Seismosoft, "SeismoSignal - A computer program for signal processing of time- histories", 2016.
[54] A. Arias, A measure of earthquake intensity. Seismic Design for Nuclear Power Plants, MIT Press, Cambridge, MA 1970, 438-469.
[55] Park Y.J., Ang A.H.-S., Wen Y.K., Damage-limiting seismic design of buildings. Earthquake Spectra, 1987, **3**(1), 1-26
[56] G.W.Housner, Spectrum intensities of strong-motion earthquakes. Proceedings of the Symposium on Earthquake and Blast Effects on Structures. EERI, Berkeley, CA, 1952
[57] J.R. Benjamin, A criterion for determining exceedance of the operating basis earthquake. EPRI Report NP-5930, Electric Power Research Institute, Palo Alto, California, 1988.
[58] R. Garg, J.P. Vemuri, K.V.L. Subramaniam, Correlating peak ground A/V Ratio with ground motion frequency content. Recent Advances in Structural Engineering, 2019, Volume 2 - Edition: 1 Publisher: Springer, Singapore
[59] J.J Bommer, A, Martínez-Pereira, The effective duration of earthquake strong motion. Journal of Earthquake Engineering, 1999, **3**(2), 127–172.
[60] G. James, D. Witten, T. Hastie, R. Tibshirani, R., An introduction to statistical learning. New York: Springer, 2013.
[61] T. Hastie, R. Tibshirani, J.H. Friedman, The elements of statistical learning: data mining, inference, and prediction. New York: Springer, 2009.
[62] L. Hamel, Knowledge discovery with support vector machines. Wiley Series on Methods and Applications in Data Mining, 2009, Daniel T. Larose, Series Editor
[63] Z.H. Zhou, Ensemble methods: foundations and algorithms. CRC press, 2012.
[64] Y. Bengio, A. Courville, I.J. Goodfellow, Deep learning: adaptive computation and machine learning. Series: Adaptive Computation and Machine Learning series, 2016, MIT Press
[65] I. Guyon, A. Elisseeff, An introduction to variable and feature selection. *Journal of machine learning research* 2003, 3(Mar), 1157-1182
[66] M.J.N. Priestley, G.M. Calvi, Concepts and procedures for direct displacement-based design and assessment. Seismic Design Methodologies for the next generation of codes, 1997, Taylor & Francis Group
[67] J.W. Baker, Conditional mean spectrum: Tool for ground motion selection. Journal of structural Engineering – ASCE, 2011, **137**:322-331
[68] Pacific Earthquake Engineering Research Center (2005) PEER Strong Motion Database on Line. Berkley, https://peer.berkeley.edu/peer-strong-ground-motion-databases.
[69] P. Castaldo, E. Miceli, Optimal single concave sliding device properties for isolated multi-span continuous deck bridges depending on the ground motion characteristics, *Soil Dynamics and Earthquake Engineering* 2023, **173**: 108128.