A Series of Forecasting Models for Seismic Evaluation of Dams Based on Ground Motion Meta-Features

Mohammad Amin Hariri-Ardebili[a,*], Sasan Barak[b,c]

[a] *Department of Civil, Environmental and Architectural Engineering, University of Colorado, Boulder, USA*
[b] *Department of Decision Analytics and Risk, Southampton Business School, University of Southampton, UK*
[c] *Faculty of Economics, VŠB Technical University of Ostrava, Czech Republic*

Abstract

Uncertainty quantification (UQ) due to seismic ground motions variability is an important task in risk-informed condition assessment of infrastructures. Since performing multiple dynamic analyses is computationally expensive, it is valuable to develop a series of forecasting models based on the unique ground motion characteristics.

This paper discusses the application of six different machine learning techniques on forecasting the structural behavior of gravity dams. Various time-, frequency-, and intensity-dependent characteristics are extracted from ground motion signals and used in machine learning. A large set of about 2,000 real ground motions are used, each includes about 35 meta-features. The major outcome of this study is to show the applicability of meta-modeling-based UQ in seismic safety evaluation of dams. As an intermediary result, the advantages of different machine learning algorithms, as well as meta-feature selection possibility is discussed for the current dataset. This paper proposes a feasibility study to reduce the computational costs in UQ of large-scale infra-structural systems.

Keywords: Uncertainty Quantification, Dams, Forecasting, Machine Learning, Big Data

1. Introduction

Risk-based performance assessment of structures and infrastructures has become a vital task in the last three decades. More specifically, the growing interest in performance based earthquake engineering (PBEE) [1] has made the risk-informed condition assessment a systematic method. Developing accurate material constitutive models, smart and fast solution algorithms, verification and validation, and finally quantifying the existing uncertainties are only some of the challenges that engineers and scientists face.

Many different, and sometimes diverse, techniques have been proposed for uncertainty quantification (UQ) of engineering structures. Most of them, at least for the problems with implicit limit state (LS) functions, recommend a variation of the Monte Carlo Simulation (MCS) family (i.e. Latin Hypercube Sampling (LHS), Importance Sampling, etc.). Although this technique is promising in general, a large number of required simulations (to result in a stable solution) makes its application limited to simple cases, or very important projects.

Machine learning techniques (and surrogate meta-models in general) are effective ways to reduce the computational burden, and bring UQ to daily engineering practice. Application of the surrogate meta-models in dam engineering can be summarized in two categories: 1) Structural health monitoring to compile and post-process the measured data from instrumentation and use them for future predictions [2], and 2) Machine learning-based methods to be applied on the outcome of numerical simulations to develop further surrogate models. The focus of this paper will be on the second part, and thus, the existing literature about this concept is reviewed.

*Corresponding author
Email address: mohammad.haririardebili@colorado.edu (Mohammad Amin Hariri-Ardebili)

1.1. Literature Review

Chen et al. [3] proposed an improved response surface meta-model (RSM) for linear dam-foundation systems to evaluate the probability of sliding. Concrete and rock's modulus of elasticity are assumed to be random variables (RVs). Karimi et al. [4] proposed an artificial neural network (ANN) procedure for system identification of concrete gravity dams which is coupled by a hybrid finite element (FE)-boundary element analysis. This technique is used to predict the dynamic characteristics of an empty dam. The foundation is assumed to be rigid, and the analyses are all performed in a linear elastic range. The conjugate gradient and the Levenberg-Marquardt algorithms are used to train the ANNs. Fan et al. [5] combined the RSM with a finite-step method to compute the explicit performance function of the system and reliability index. The failure path and the functionality failure mode were computed for a roller compact concrete (RCC) dam.

Gu et al. [6] used the least-squares support vector machine (LS-SVM) in back analysis of RCC dams, and determined the complex mechanical properties. First, the initial samples are uniformly designed and then, a transversely isotropic model is established to train the samples. Next, the complex nonlinear relationship between relative values of hydraulic components of dam displacements and mechanical parameters is established. Moreover, Su et al. [7] applied a similar idea on gravity dams with the extension of a criterion for optimal selection of parameters in back analysis. In this technique, the key index of optimal selection is the parameter sensitivity. The uniform design method was combined with an ANN and SVM to build the mapping relationship between multiple material parameters and dam responses at different positions.

Gaspar et al. [8] proposed a probabilistic thermal model to propagate uncertainties on some RCC's physical properties where a thermo-chemo-mechanical model was used to describe the dam behavior. A global sensitivity analysis was performed considering a bi-dimensional random field heterogeneity. Cheng et al. [9] adopted a kernel principle component analysis (KPCA) method to eliminate the effect of environmental variables and monitor the health of the dam under varying conditions. Gu et al. [10] developed a new method based on the chaos genetic optimization algorithm to inverse the actual initial zoning deformation modulus and to determine the inversion objective function using the dam displacement measured data and the FE method.

Rezaiee-Pajand and Tavakoli [11] introduced an efficient method for crack detection in concrete gravity dams using a hybrid genetic algorithm (GA) and FE methods. The GA identifies the location and magnitude of cracks in dams by minimizing the difference between the analytical responses and the measured ones. Xin and Chongshi [12] applied credibility theory into the stability failure analysis of a gravity dam. Stability was evaluated as a hybrid quantity considering both the fuzziness and randomness of the failure criterion, design parameters and measured data. Furthermore, Cao et al. [13] studied the stability of high arch dam abutments as a fuzzy random event. The instability risk ratio models were proposed based on credibility theory and were calculated using the MCS and fuzzy random post-processing.

In a series of papers, Hariri-Ardebili and Pourkamali-Anaraki [14, 15] showed the application of several machine learning techniques in multi-hazard (i.e. seismic, hydrologic, and aging) reliability analysis of gravity dams. Both simplified linear elastic and nonlinear damage-based models were used. They showed the capability of machine learning techniques in classification and regression analysis with the specific application on gravity dams. Moreover, Hariri-Ardebili [16], Hariri-Ardebili and Boodagh [17] proposed a set of design of experiment (DOE) techniques in order to develop a polynomial-based surrogate model to quantify the material uncertainty in coupled dam-reservoir-foundation systems. DOEs such as two-level and three-level factorial designs, central composite design, Teguchi design, etc. were discussed in detail and the meta-models were validated by a large MCS-based dataset.

1.2. Contributions and Organization

In this paper, the application of different machine learning techniques is discussed on forecasting the structural responses of gravity dams subjected to the impact of multiple ground motions. Various futures of ground motion signals are extracted, and used in data forecasting and prediction. To the best of the authors' knowledge, this problem has not been addressed yet in the field of structural and earthquake engineering. Various researchers' work on the concept of so-called optimal intensity measure (IM) selection for specific structural systems has been discussed [18, 19, 20, 21, 22, 23, 24, 25]; however, none of these used a very large dataset of as-recorded ground motions (e.g. ~2,000 as of this paper) to correlate the quantity of interest (QoI) and IM parameters. Therefore, the novelty of this paper can be summarized as follows:

- It is one of the few applications of machine learning techniques in concrete dam engineering. The majority of the current applications are limited to monitoring data and not FE-based data.
- It contains one of the largest datasets used in seismic analysis of an engineering structure (more specifically dams). None of the previous applications in dam engineering exceeded the use of 100-200 ground motion signals.
- It compares and contrasts up to six machine learning techniques on an identical engineering problem with aleatory uncertainty.
- It determines the efficient and optimal ground motion IM parameters using forecasting techniques. The traditional technique to identify an optimal IM uses concepts like efficiency, practicality, sufficiency, proficiency, and hazard compatibility [25].
- It contains a discussion on meta-feature selection in machine learning, and its direct application in engineering problems with aleatory uncertainty.

The rest of the paper is organized as follows: first, a list of unique ground motion signatures is provided in Sec. 2, followed by the machine learning techniques used in this paper, Sec. 3. The case study FE model is explained in Sec. 4, data preparation and specific treatments are discussed in Sec. 5, and finally the results are presented in Sec. 6. The paper concludes with the major findings and also proposes for the future works in Sec. 7.

2. Identification of the Unique Ground Motion Signatures

Biometric recognition is an acceptable tool for identification and authentication in computer science. Biometrics refers to an automatic recognition of individuals using their physiological and/or behavioral specifications. Characteristics, such as eye scan, finger print, DNA test, and voice recognition are unique for each person. Having some or all these data, different persons can be distinguished/identified.

Similarly, a recorded (i.e. real) earthquake ground motion has unique characteristics, mainly because it is generated from a specific fault rupture source at a unique location and time period, and is recorded at the specific location by a seismograph [26]. Therefore, the time- and frequency-dependent characteristics are also unique. Vector quantities, such as acceleration time history, Fourier amplitude, response spectrum, Arias intensity time history etc. are unique for a recorded ground motion. Figure 1 compares some of the human biometric characteristics and the ground motion identifiers.

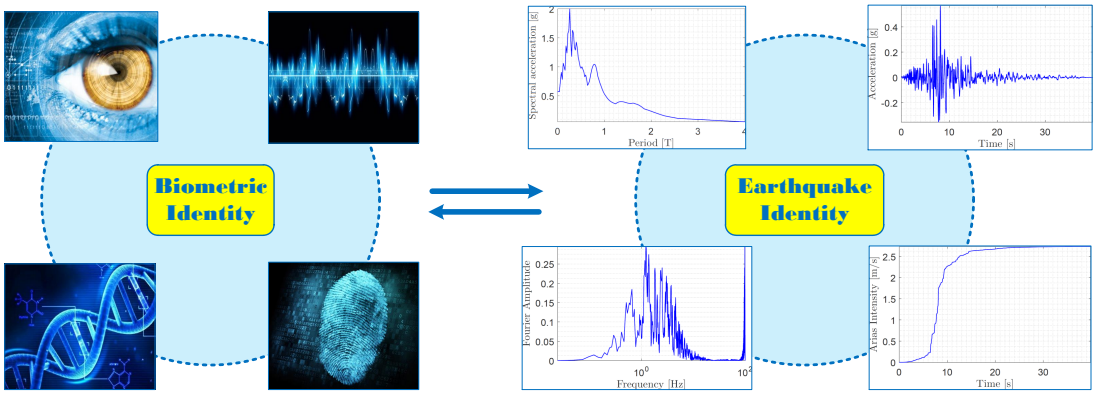


Figure 1: Comparison between the human biometric characteristics and the earthquake ground motion identities

These unique characteristics of the ground motion records make the seismic response of the structural system very complicated (and unique). Thus, the dynamic response (i.e. QoI) of an engineered structure (e.g. displacements, stresses, damage pattern) highly depends on the input ground motion signal. To establish a one-by-one relationship between the input signal and the output QoI, it is important to present the ground

motion records with one or several scalar IM parameters. A wide range of time-, frequency-, spectral- and intensity-dependent IM parameters are summarized in Table 1. This list will be used to correlate the QoIs with the recorded ground motions. For each single ground motion signal, all $N_{IM} = 35$ IM parameters can be extracted. Therefore, having N_{gm} ground motions, a matrix of $X_{N_{gm} \times N_{IM}}$ constructs the input domain.

Table 1: A comprehensive list of ground motion IM parameters [25]

No.	Description of IM	Symbol	Mathematical model
1	Total duration	t_{tot}	$t_{1.00I_A} - t_{0.00I_A}$
2	Significant duration	t_{sig}	$t_{0.95I_A} - t_{0.05I_A}$
3	Seismological duration	D_{5-75}	$t_{0.75I_A} - t_{0.05I_A}$
4	Peak ground acceleration	PGA	$max\left(\lvert\ddot{u}(t)\rvert\right)$
5	Peak ground velocity	PGV	$max\left(\lvert\dot{u}(t)\rvert\right)$
6	Peak ground displacement	PGD	$max\left(\lvert u(t)\rvert\right)$
7	Root-mean-square of acceleration	a_{RMS}	$\sqrt{\frac{1}{t_{tot}}\int_0^{t_{tot}}(\ddot{u}(t))^2 dt}$
8	Root-mean-square of velocity	v_{RMS}	$\sqrt{\frac{1}{t_{tot}}\int_0^{t_{tot}}(\dot{u}(t))^2 dt}$
9	Root-mean-square of displacement	u_{RMS}	$\sqrt{\frac{1}{t_{tot}}\int_0^{t_{tot}}(u(t))^2 dt}$
10	Root-square of acceleration	a_{rs}	$\sqrt{\int_0^{t_{tot}}(\ddot{u}(t))^2 dt}$
11	Root-square of velocity	v_{rs}	$\sqrt{\int_0^{t_{tot}}(\dot{u}(t))^2 dt}$
12	Root-square of displacement	u_{rs}	$\sqrt{\int_0^{t_{tot}}(u(t))^2 dt}$
13	Arias intensity	I_A	$\frac{\pi}{2g}\int_0^{t_{tot}}(\ddot{u}(t))^2 dt$
14	Specific energy density	SED	$\int_0^{t_{tot}}(\dot{u}(t))^2 dt$
15	Cumulative absolute velocity	CAV	$\int_0^{t_{tot}}\lvert\ddot{u}(t)\rvert\, dt$
16	Cumulative absolute displacement	CAD	$\int_0^{t_{tot}}\lvert\dot{u}(t)\rvert\, dt$
17	Shaking intensity rate	SIR	$(I_{A5-75})(D_{5-75})^{-1}$
18	Acceleration spectrum intensity	ASI	$\int_{0.1}^{0.5} S_a\left(T, \xi = 5\%\right) dT$
19	Velocity spectrum intensity	VSI	$\int_{0.1}^{2.5} S_v\left(T, \xi = 5\%\right) dT$
20	Displacement spectrum intensity	DSI	$\int_{2.0}^{5.0} S_d\left(T, \xi = 5\%\right) dT$
21	Effective peak acceleration	EPA	$\frac{1}{2.5\times0.4}\times\int_{0.1}^{0.5} S_a\left(T, \xi = 5\%\right) dT$
22	Effective peak velocity	EPV	$\frac{1}{2.5\times0.4}\times\int_{0.8}^{1.2} S_v\left(T, \xi = 5\%\right) dT$
23	Improved effective peak acceleration	$IEPA$	$\frac{1}{2.5\times0.4}\times\int_{T_p^a-0.2}^{T_p^a+0.2} S_a\left(T, \xi = 5\%\right) dT$
24	Improved effective peak velocity	$IEPV$	$\frac{1}{2.5\times0.4}\times\int_{T_p^v-0.2}^{T_p^v+0.2} S_v\left(T, \xi = 5\%\right) dT$
25	First-mode spectral acceleration	$S_a(T_1)$	$S_a(T_1, \xi = 5\%)$
26	First-mode spectral velocity	$S_v(T_1)$	$S_v(T_1, \xi = 5\%)$
27	First-mode spectral displacement	$S_d(T_1)$	$S_d(T_1, \xi = 5\%)$
28-31	Higher-mode spectral acceleration	$S_a(T_i)$	$S_a(T_i, \xi = 5\%),\ i = 2, ..., 5$
32	Spectral acceleration at predominant period	$S_a(T_p^{accel})$	-
33	Spectral velocity at predominant period	$S_v(T_p^{vel})$	-
34	Sustained maximum acceleration	SMA	Abs max $\ddot{u}(t)$ sustained for 3 cycles
35	Sustained maximum velocity	SMV	Abs max $\dot{u}(t)$ sustained for 3 cycles

Note: $\ddot{u}(t)$, $\dot{u}(t)$ and $u(t)$ are acceleration, velocity and displacement time histories, respectively.

So, ideally, having these meta-features a ground motion record can be isolated, and the structural responses associated with it can be identified. This is one of the objectives of this paper, and thus, a series of forecasting techniques are adopted to achieve this goal.

3. Forecasting Techniques: A Brief Overview

Forecasting techniques can be used effectively to predict the dam response, and reduce the total number of required simulations. Six forecasting techniques are used in this paper. They are briefly reviewed in this section for those engineers (and not the data scientists) who are not familiar with the fundamentals of forecasting.

3.1. Decision Tree Regression (DTR)

Decision tree regression (DTR) is a non-parametric and nonlinear machine learning technique. It takes advantage of a hierarchical structure for recursively segmenting training data, and therefore, it has great

flexibility and interpretability in data analysis. The most common strategy to induct a decision tree is greedy top-down construction which recursively partitions the data into subsets until the stopping criterion has been met. The stopping criterion is crucial so that it can prevent growing branches that does not affect the tree quality [27]. Some of the stopping rules are:

1. The number of observations in a node is less than a pre-specified threshold.
2. All observations assigned to a node belong to the same class.
3. Depth of the node is more than some pre-specified limit.
4. Nodes' purity is more than a specified threshold [28].

Decision trees are often prone to over-fitting according to high variance, and hence, methods are proposed to find the right sized tree. The most famous method is pruning trees [29], so in order to have a high quality tree, first a complete tree is built, and then inefficient sub-trees, that do not influence the cost function significantly, are removed.

The evaluation function used for splitting classification trees in the CART (classification and regression tree) method is Gini index, which describes the chance of coming up with a false node for the data if the node was chosen randomly from the nodes' distribution. The Gini index can be stated as:

$$\text{Gini}(t) = 1 - \sum \left[P\left(K|t\right)\right]^2 \tag{1}$$

where $P\left(K|t\right)$ is the proportion of finding the data class K in the node t (node purity). The objective is to minimize Gini index. From the formula it can be inferred that if the classification is done in a perfect way, the Gini index would be zero [30, 31].

3.2. Random Forest (RF)

A random forest (RF) is a kind of ensemble classifier, made by a combination of decision trees which are created by recursive partitioning. The idea behind RF is to combine the results of many different decision trees to overcome the vulnerabilities of the individual one. In order to construct a RF model, with the help of bootstrapping, new sets of training data are created, and then, RF randomly chooses the variable for each set (for better diversity in results). Next, RF starts creating decision trees for each group with respective variables. Finally, the outcome forest of trees is combined and the average of the predictions is considered as the result [32].

For testing the accuracy of a RF Out-of-Bag (OOB) data, which are the samples that were not selected in bootstrapping in the RF procedure, can be used. An error will be assigned by applying the RF model to the OOB data, and hence, the performance of the RF model can be examined. One key parameter in a RF algorithm is the importance score (IS) which is described as:

$$\text{IS} = \frac{1}{N_{tree}} \sum_{i=1}^{B} \left(Err_i - Err_i^*\right) \tag{2}$$

Where N_{tree} is the number of trees in RF model, and Err_i and Err_i^* are the errors of each tree applied on OOB and perturbed OOB data, respectively. More details can be found in Thakur and Kumar [33]. The benefit of using RF is to have a reduced variance in comparison to a single tree so that over-fitting will not happen [30].

3.3. Tree Bagging (TB)

Bagging, a short term for "Bootstrap Aggregation", is an ensemble technique that uses bootstraps to generate samples of the original data. In prediction problems, this algorithm averages the prediction over a collection of bootstrap samples and the class of a new observation is the most selected class among the number of trees constructed on bootstrap samples [34]. In this algorithm, trees are grown deep without pruning. However, by building sufficient trees, over-fitting is less probable. A pseudo-code for tree bagging is shown in algorithm 1. In this method, similar to the RF data, the OOB data can be used to justify the performance of the model [35].

Algorithm 1 Tree Bagging pseudo-code

```
for i = 1:N_tree do
    Generate bootstrapped sample of the data
    Create un-pruned decision trees on the samples
    Average on all the outcomes
end for
```

3.4. Extreme Gradient Boosting (XGBoost)

XGBoost is a statistical nonlinear machine learning algorithm used for functions such as, classification, regression and ranking [36]. Scalability and speed of XGBoost in comparison to other solutions has made it very popular among most successful algorithms used by data scientists [37]. It is an implementation of the gradient boosted trees algorithm [38]. In order to combine a set of weak learners to develop a strong learner, two ways are proposed, either first build a set of learners and then, average the result like the RF and TB methods; or sequentially add learners in order to optimize the cost function in a step-wise manner as in boosting methods [39].

In additive learning of XGBoost, the first learner is fitted on the whole data, and the next learners are fitted to the residuals of the former ones. In fact, each learner is fitted using information from previously fitted learners. The general function for the prediction at each step is presented as follows:

$$\hat{f}_j^t = \sum_{i=1}^{t} f^i(x_j) = \hat{f}_j^{t-1} + f^t(x_j) \tag{3}$$

where $f^i(x_j)$ is the learner at step i, $\hat{f}_j^t$ is prediction at step t, and x_j is the input variable.

Unlike the RF and TB, gradient boosting methods are prone to over-fitting, if the number of trees is too large. A computation procedure for preventing over-fitting can be found in Fan et al. [40]. More information about computation of XGBoost in R Statistical [41] can be found in Chen and Guestrin [37].

3.5. Artificial Neural Networks (ANN)

ANNs are used to map an input to a desired output like a mathematical function. They are inspired by the behavior of neurons located in the brain [42]. ANNs are non-parametric estimators that can be used for several kinds of tasks, such as forecasting, clustering, function approximation and optimization [43]. The inputs for a neural network are vectors of variables corresponding to an observation. These vectors are weighted and combined by linear filters and become the inputs of the hidden layers where the nonlinear computation is performed on the inputs. Network output will be calculated by an activation function which receives outputs of hidden layers and calculates the output of the network [44].

The mathematical explanation of a multilayer perception, which is one of the most preferred models of ANN, can be presented as [45]:

$$F_t = \beta_0 + \sum_{n=1}^{N} \beta_n W \left(\theta_{0i} + \sum_{j=1}^{m} \theta_{nj} x_j \right) \tag{4}$$

Where m is the number of input parameters (x_j), N is the number of nodes in the hidden layer, θ_{nj} is the weight of the output layer, and β_n is the weight of the hidden layer. Here the zeroth indices in the weight coefficients refer to the bias nodes in each layer.

One may use either sigmoid or hyperbolic tangent functions as a transfer function, W. The whole process of learning is achieved by adjusting the weight parameters and the network is updated each time it has been fed by a new dataset. After the parameters are updated, the desired outcome will be the classification of data. More on this can be found in Ragg et al. [46].

3.6. Support Vector Regression (SVR)

Originally, this method was proposed to handle the forecasting problems [47] in a set of data $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ where $(x_i, y_i) \in \mathbb{R}^2$ are the respective input and output. In this algorithm,

with the help of Kernel functions, the input data is mapped to a new higher dimensional space which is called the feature space [14]. In the feature space, an estimated function is considered as:

$$g(x) = \alpha x + \beta \tag{5}$$

which is, in fact, the equation for a hyper-plane.

Then, the cost function can be described as:

$$\begin{gathered} C(\lambda) = \frac{\lambda}{n} \sum_{j=1}^{n} E_r\left(f_j, y_j\right) + \frac{1}{2} \|\eta\|^2 \\ E_r\left(f_j, y_j\right) = |f - y| - r \; |f - y| \geq r \;\; \text{or} \;\; 0 \;\; \text{otherwise} \end{gathered} \tag{6}$$

where λ is the penalty parameter, $\frac{1}{2} \|\eta\|^2$ is the regularization, r is the tube size for the error function E_r, and f is the desired quantity.

Finally, with the help of Lagrange multipliers, the equation for the hyper-plane can be rewritten as:

$$g\left(x, \beta_i, \beta_i^*\right) = \sum_{i=1}^{n} \left(\beta_i - \beta_i^*\right) K\left(x, x_i\right) + c \tag{7}$$

where $K\left(x, x_i\right)$ is the Kernel function. Further details on SVR can be found in Vapnik [48].

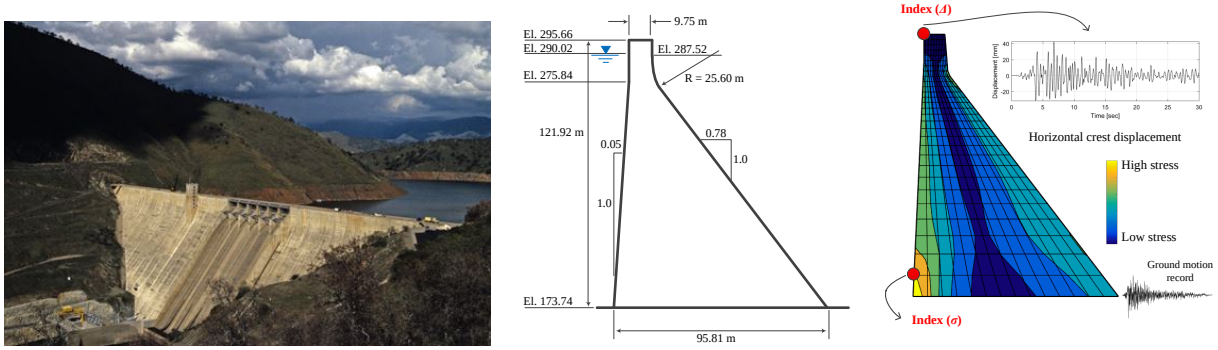
4. Case Study Description

In this paper, Pine Flat gravity dam, Figure 2(a), is selected as a case study. This dam is often used as a test-bed in the literature [49]. The dam height is 121.92 m and its length (cross-stream direction) is 560.83 m. The thickness at the base and the crest are 95.81 and 9.75 m, respectively. Figure 2(b) shows the cross-section of the tallest non-overflow monolith including the mesh in Figure 2(c). The finite element program EAGD [50] is used to analyze the dam, including the reservoir water and foundation effects. The updated version of the code includes new compliance data for the dam-foundation interaction (DFI). The foundation rock is idealized as a homogeneous, isotropic, viscoelastic half-plane. The DFI effects are included by adding the dynamic stiffness matrix for the rock region in the dam's equation of motion [51]. The reservoir water is idealized by a fluid domain of constant depth and infinite length in the upstream direction. The dissipation of hydrodynamic pressure waves by the reservoir bottom materials is accounted for by applying a boundary condition which partially absorbs the incident waves. Since the system is analyzed in a linear elastic condition, a relatively coarse mesh of 450 four-node plane strain elements are used for the dam domain. Applied loads on the system are: 1) self-weight, 2) hydrostatic pressure, and 3) seismic loads.

One may notice this is a relatively simple yet accurate enough procedure to obtain the dynamic response of a concrete dam. The modern analysis techniques adopt fine mesh for the concrete dam specially in the vicinity of the neck and heel. The Eulerian or Lagrangian fluid elements might be used for the reservoir, and a massed foundation with absorbing boundary conditions (e.g. infinite elements, perfectly matching layers) is required. The authors have already implemented all those advanced techniques in several publications [52, 25]. All of these techniques affect the dynamic response; however, their relative importance is kept nearly constant. This is the reason to adopt a simplified technique in this paper to present a framework (and not the exact results) for statistical analysis of dam response.

Standard material properties are assumed for the mass concrete and the foundation rock. Concrete properties are: modulus of elasticity, $E_c = 24$ GPa, Poisson's ratio, $\nu_c = 0.2$, mass density, $\rho_c = 2470$ kg/m^3, and constant hysteretic damping, $\eta_c = 0.06$. On the other hand, for the foundation rock the following material properties are considered: $E_f = 21.5$ GPa, $\nu_f = 0.33$, $\rho_f = 2680$ kg/m^3, and $\eta_f = 0.05$. Moreover, the wave reflection coefficient for the reservoir bottom materials, α_w, is assumed to be 0.50. Investigation of the material uncertainty is not within the scope of this paper, and studied elsewhere [16].

Since the main objective in this paper is to establish a model for the ground motion record-to-record (RTR) variability, only the horizontal component of the seismic excitations is applied at the foundation base (whereas the recorded earthquake signal is on the free-field). Therefore, a de-convolution process is required to determine the motion at the rigid base boundary.



(a) Dam and its surrendering environment (b) Geometry and dimensions (c) Mesh, index points, and QoIs

Figure 2: Pin Flat gravity dam

Displacement and stress responses are local QoIs and easy to process for the structural systems. Therefore, in this paper, crest horizontal displacement and vertical stress at the vicinity of the dam's heel are used. Based on Figure 2(c), there is a stress concentration at the dam-foundation interface which might affect the forecasting analysis. Thus, one level (or element) higher than the heel is considered as an index point for the stress response. This figure also shows the displacement time history at the crest, and the non-concurrent envelope of the vertical stresses within the dam body. Note that in the context of linear elastic analyses, crest displacement and the non-concurrent stress envelope are the most representative quantifiers for performance evaluation [53].

5. Data Preparation

5.1. Inputs and Outputs Data Mining

The applied ground motions are downloaded from PEER [54] website (only the NGA-West2). A total of 1,929 ground motion records are used. These signals are directly obtained from the first 2,000 records in the PEER database (the remaining 71 records are either unavailable or numerically did not converge for this example). Since a very large dataset of input ground motions are used for the simulations, no ground motion selection or scaling technique is required. This allows investigation of different ground motions with various time- and frequency-domain signatures. This type of wide-range analysis technique is usually referred to as "Cloud Analysis" [55]. For each input ground motion, the resulted QoI parameters are recorded in an array. It is noteworthy that one may use a smaller ground motion database which is specifically selected for the site and according to structural characteristics [56, 57]. This may reduce the variance and improve the machine learning algorithms.

Figure 3 illustrates the general trend for some of the selective IM parameters resulted from 1,929 ground motions. In most of the cases, there exists a decaying trend for the observed parameters by increasing the value of the quantity. Although it seems that a log-normal distribution [25] can be fitted to these data, it is beyond the scope of this paper.

In addition, Figure 4 illustrates the correlation among the IM parameters (Note: DSI and EPV are not shown since the precision of the initial Matlab code to extract this information was not enough).

Furthermore, Figure 5(a) compares two QoIs (i.e. displacement and stress) at the index points. There exists a linear relationship between them; however, the dispersion is still considerable. A statistical summary of the response parameters (from all data) is as follows:

- Δ^{H}_{max}: minimum = 1.08, lower adjacent = 1.08, 25th percentile = 7.14, median = 14.82, mean = 23.50, 75th percentile = 28.74, upper adjacent = 61.04, maximum = 224.74.
- σ^{yy}_{max}: minimum = -0.09, lower adjacent = -0.09, 25th percentile = 0.15, median = 0.45, mean = 0.77, 75th percentile = 0.99, upper adjacent = 2.24, maximum = 8.70.

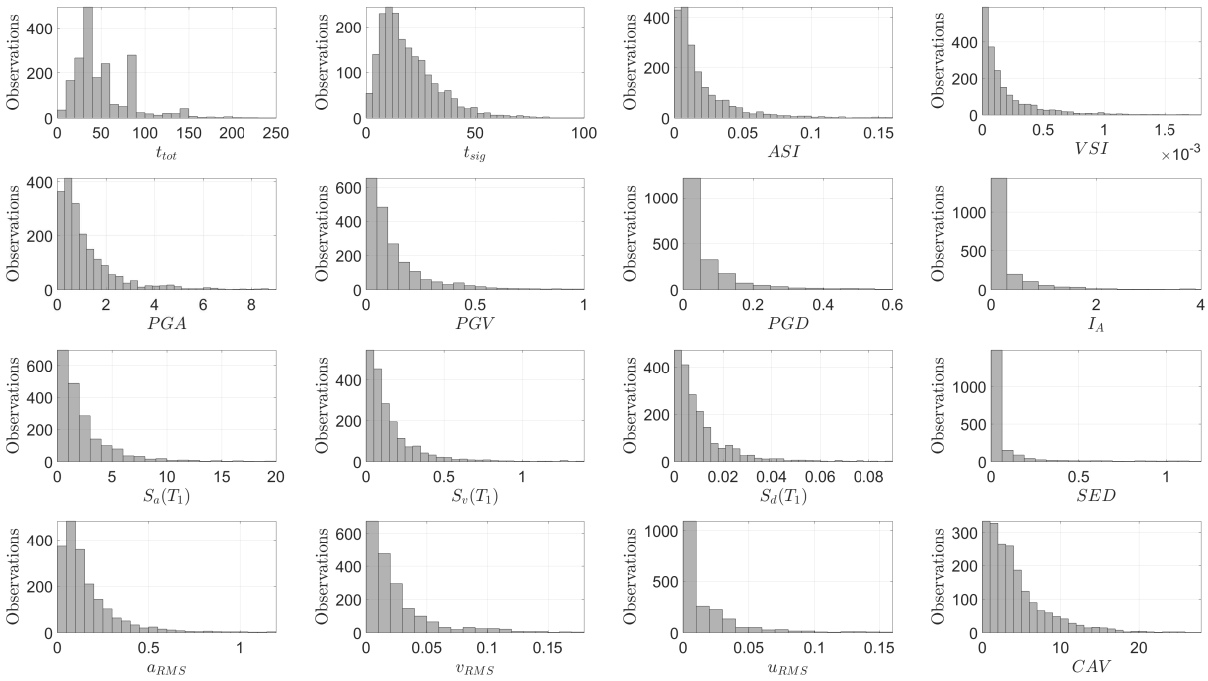


Figure 3: Distribution of the selective input parameters for 1929 ground motion signals

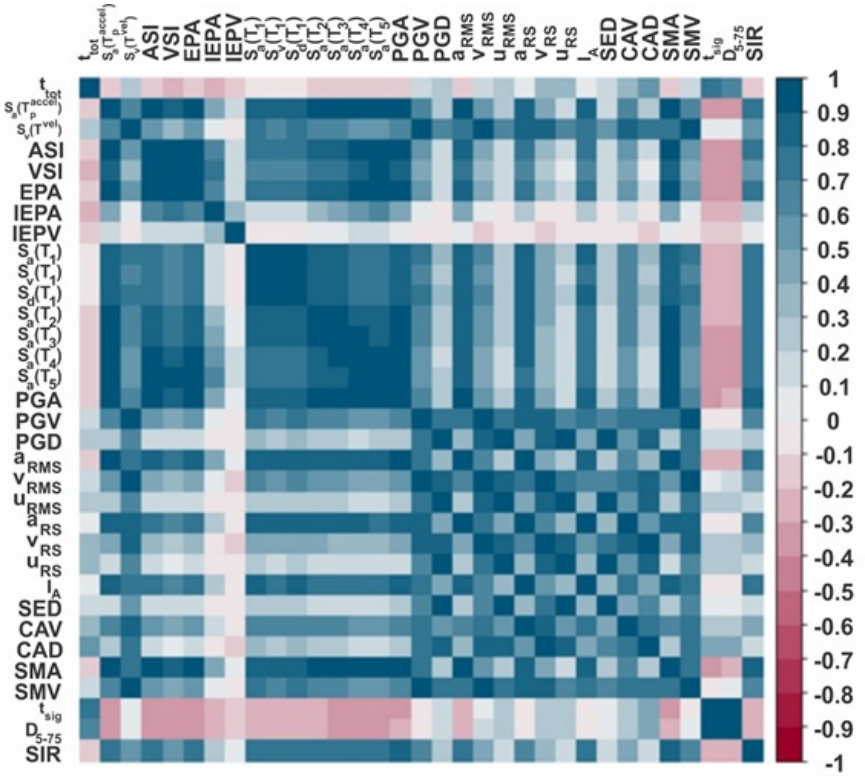
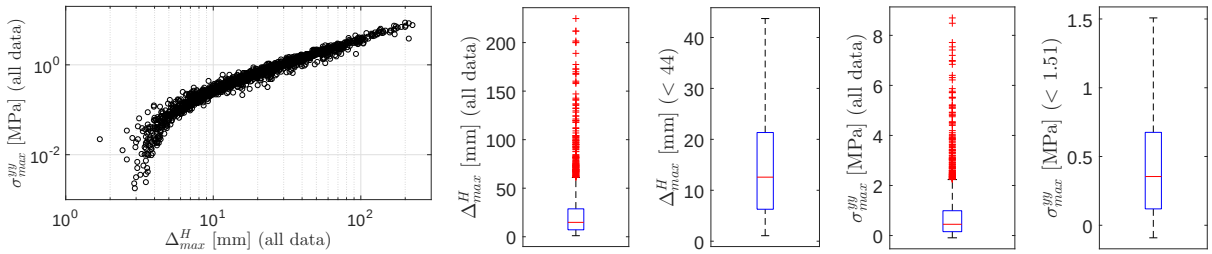


Figure 4: Correlation map between input IM parameters

For each response quantity, two sets of boxplots are shown in Figures 5(b) and 5(c). The left one (labeled as "all data") is based on all the observations. For a more accurate prediction, some of the large data are

eliminated and thus, updated boxplots (the right plots specified with a limit) are also prepared in each case. The truncation values are shown along the vertical axis.



(a) Displacement vs. stress (b) Displacements boxplots (c) Stresses boxplots

Figure 5: Relationship between two response quantities resulted from 1929 ground motion signals

5.2. Error Functions

Regression error functions are essential to calculate the prediction error of the meta-methods. Although there are many metrics, the root-mean-square-error (RMSE) and the symmetric mean absolute percent error (SMAPE) functions are adopted in this paper:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N} (\hat{y}_i - y_i)^2}{N}} \tag{8}$$

$$\text{SMAPE} = \frac{1}{N} \sum_{i=1}^{N} \frac{|\hat{y}_i - y_i|}{\frac{|\hat{y}_i| + |y_i|}{2}} = \frac{2}{N} \sum_{i=1}^{N} \frac{|\hat{y}_i - y_i|}{|\hat{y}_i| + |y_i|} \tag{9}$$

where the y_i and $\hat{y}_i$ are real and predicted response parameters, and N is the number of observations.

5.3. Cross Validation

To improve the training models, a particular sample of train-dataset is reserved with the aim of validating the trained model before finalizing it. By cross validation which validates the model on different subsets of data, its effectiveness is improved to predict the target value for test-dataset. In order to add the cross validation, the "traincontrol" function [58] is used with method of "cv" for 10 different divisions of the train-dataset (this method is called "10-fold" cross validation).

In the k-fold cross validation, the train data set is partitioned randomly into k equal size subsets. For tuning model parameters $k-1$ subsamples are used as training data, and the remaining one is adopted for validation. This process is continued until each of the k parts are used exactly once as a validation set. The final estimation could be an average of k produced estimation. Figure 6 illustrates a scheme of a 10-fold cross validation process.

5.4. Applied Forecasting Models

To predict the QoIs, the machine learning techniques in Sec. 3 are applied on the dataset. Following is a short description on the applied methods for those want to reproduce the results. Note that the raw data can be provided by the first author upon request.

DTR: In this method, a simple DTR is used to divide the data into subgroups based on different exogenous variables. As a regression model, predicting value will be the average target value of each subgroup. For implementing this method, an "rpart" function [59] with 10-fold cross-validation, and a maximum depth of 10 for the tree is used.

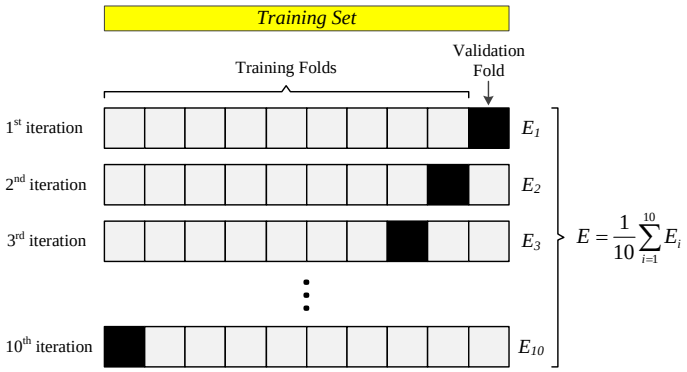


Figure 6: Cross validation with 10 folds

RF: In the RF method, the prediction is based on the bagging simple decision trees with the difference that in each splitting section just a subset of predicting variables are used. Randomly choosing a subset of predicting variables helps the model to have less computation and to avoid the over-fitting problem. For implementing this method, the "randomForest" function [60] with "ntree = 500" (number of trees) is used (which indicates the number of trees to be grown).

TB: In the tree bagging, some of the weak CART models are aggregated to create a more accurate predictor. Besides, this method helps to reduce variance and avoid over-fitting. The bagging function from "ipred" library [61], with "nbag = 25" (number of bag) tuning parameters are used (which indicates the number of bootstrap replications). Also, the 10-fold cross validation is used as a training control.

XGBoost: The extreme gradient boosting model is an optimized implementation of a boosting method which has a good performance with fast calculation. This model is implemented by the "xgboost" function [36] with "nround = 100" that indicates the maximum number of iterations and for this number the learning rate is set to 0.2.

ANN: In this part, a common neural network is used to predict the target value. The predicting variables' values feed into the neural network, then by passing through two hidden layers the predicted values are estimated. For this method, an "H2O" deep learning function [62] with a rectifier activation function is used. For the number of hidden layer nodes, a convenient rule is used which is based on the mean value of input and output variables. Therefore, two hidden layers are used in which the first one has 15 nodes and the second one includes 7 nodes. Besides, the number of epochs (iterations) is set to 200.

SVR: In support vector regression model, each sample is mapped into a higher dimensional space with predicting features, and then a hyperplane is found which divides the samples into two subsets. For this method, "SVM" function from "e1071" library [63] is used. In this function, the "type = eps-regression" parameter is adopted to indicate that the problem is regression with a default radial kernel function. With the aim of better differentiation, the "epsilon" tuning parameter is set to 0.01.

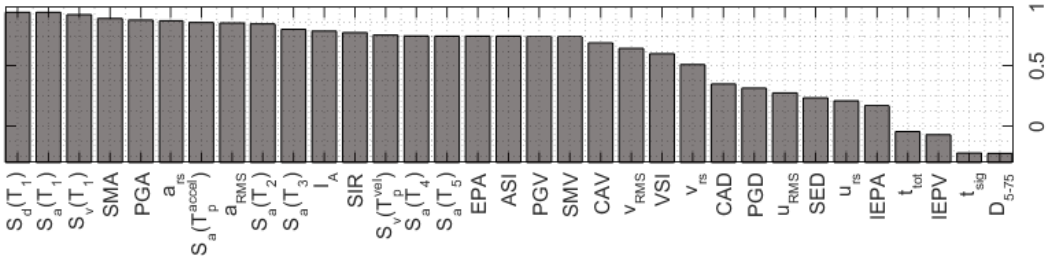
6. Results and Discussion

So far, all the fundamental background information as well as the data preparation and post-processing techniques are explained. In this section, the results are presented in two groups: 1) using all the ground motion meta-features in Table 1 (See Sec. 6.1), and 2) using only a subset of selective meta-features (See Sec. 6.2). Detailed information about meta-feature selection is presented in Appendix A.

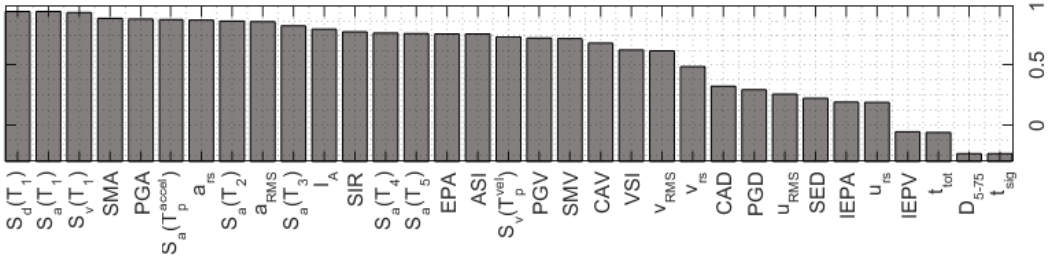
6.1. Employing All the Meta-Features

This section investigates the quality of prediction based on different techniques. All the ground motion-dependent meta-features are employed. First, the raw correlation among the input IMs and two QoIs is shown in Figure 7. This is a simple direct correlation among each of 35 IM parameters and one of the system outputs (i.e. displacement or stress) which varies $[-1, +1]$. The major observations are:

- There is a high similarity between the displacement- and stress-based correlations.
- First-mode spectral ordinates have the highest correlation with response quantities. This is also confirmed with traditional structural dynamics [64].
- $S_d(T_1)$ and $S_a(T_1)$ have the highest positive correlation (> 0.95).
- Total and effective durations have the lowest (and negative) correlation with QoIs. This is consistent with the physics of the linear dynamic analysis, where the ground motion duration does not directly play an important role. Its effect is dominant in nonlinear analysis of brittle materials.



(a) Displacements boxplots



(b) Stresses boxplots

Figure 7: Features correlation with two response quantities

Table 2: GOF for the displacement with all the meta-features

	Train		Test	
	RMSE	SMAPE	RMSE	SMAPE
Decision Tree Regression	34.45	0.198	36.72	0.209
Tree Bagging	30.25	0.188	32.48	0.194
Random Forest	11.22	0.052	26.51	0.121
XG-Boost	1.72	0.014	27.88	0.126
Neural Network	23.35	0.143	25.53	0.139
SVR	21.58	0.103	29.02	0.128

Table 3: GOF for the stress with all the meta-features

	Train		Test	
	RMSE	SMAPE	RMSE	SMAPE
Decision Tree Regression	1.306	0.614	1.513	0.540
Tree Bagging	1.119	0.497	1.229	0.430
Random Forest	0.434	0.125	1.115	0.237
XG-Boost	0.068	0.062	1.199	0.249
Neural Network	0.933	0.278	1.202	0.278
SVR	0.893	0.243	1.241	0.274

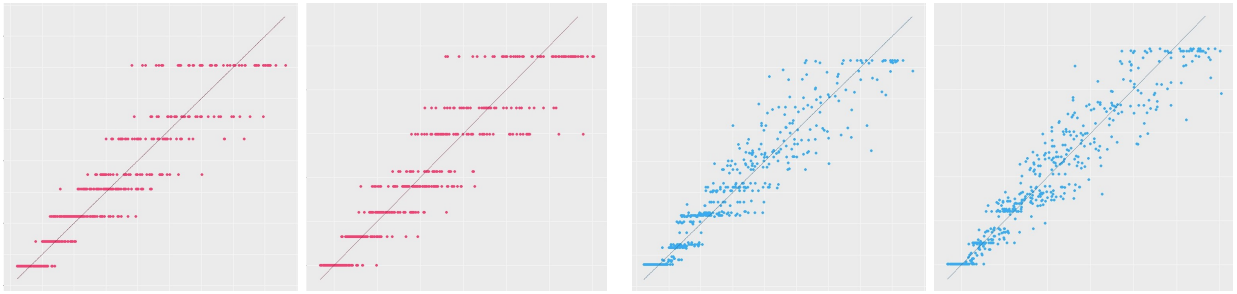
Tables 2 and 3 provide a general overview on the Goodness-of-fit (GOF) by comparing the error functions resulted from six techniques used in this paper. For each response, the GOF is provided for the train and test data separately. The major observations are:

- For displacement response:
  - For training data set the order of models from lowest to highest error functions is:
    - Based on RMSE (or SMAPE): XGB, RF, SVR, ANN, TB, DTR.
  - For test data set the order of models from lowest to highest error functions is:
    - Based on RMSE: ANN, RF, XGB, SVR, TB, DTR.
    - Based on SMAPE: RF, XGB, SVR, ANN, TB, DTR.
- For stress response:
  - For training data set the order of models from lowest to highest error functions is:
    - Based on RMSE (or SMAPE): identical to the displacement.
  - For test data set the order of models from lowest to highest error functions is:
    - Based on RMSE: RF, XGB, ANN, TB, SVR, DTR.
    - Based on SMAPE: RF, XGB, SVR, ANN, TB, DTR.
- The order of models for the training dataset is identical for displacement and stress (either RMSE or SMAPE).
- For all the data sets and error functions, the DTR provides the worst prediction.
- For the training data set and any error function, XGBoost is the best model.
- For the test date set, the RF is the best model. Again, the second rank belongs to XGBoost.
- Figure 8 illustrates the quality of the prediction for all six models and two QoIs. Deficiency of DTR and to some extent Tree Bagging is obvious. On the other hand, RF has a uniform trend along the equality axis.

In addition to the above mentioned general discussion, the following model-based intermediary findings are important from the engineering and scientific points of view:

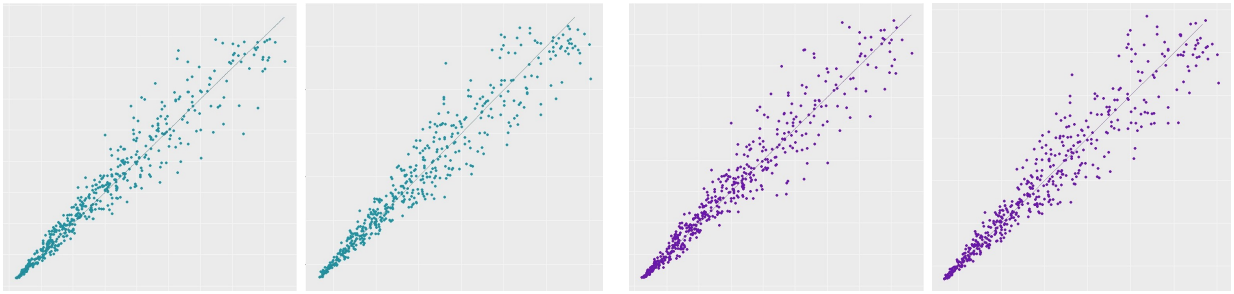
- DTR charts for the displacement and stress responses are shown in Figure 9. In a decision tree, several nodes are connected until a result is reached. Each leaf node is presented as an if/then rule. Cases that satisfy the if/then statement are placed in the node. In a DTR, the output of each node is binary (i.e. yes/no) with a specific probability. The probabilities are narrowed down until the lowest level. Summation of all information gain at each level should be equal to 100%.

  Displacement and stress-based trees, in this example, have four levels with 8 final leafs. First of all, Figure 9 shows that among many meta-features $S_d(T_1)$, $S_v(T_1)$, $S_a(T_1)$ and SMA contribute to form a model with a major probability of occurrence. Second, the lighter color of a box/leaf shows the importance of that rule. Finally, these 8 leafs correspond to 8 layers of data available in Figure 8(a). Clearly, DTR provides a discrete prediction of the results.
- Decision tree-based importance factors are obtained as:
  - For displacement: $S_d(T_1) = 3.015$, $S_a(T_1) = 2.843$, SMA $= 1.872$, $S_v(T_1) = 1.864$, ...
  - For stress: $S_v(T_1) = 2.657$, $S_d(T_1) = 2.648$, $S_a(T_1) = 2.636$, SMA $= 2.418$, PGA $= 1.252$, ...
- Figures 10(a) and 10(b) present the evolution of the error function (RMSE) with respect to number of trees incorporated in the RF method. As seen, for any practical purposes, 100 trees for displacement response and 200 for the stress might be enough. The displacement response is converged earlier than the stress one.



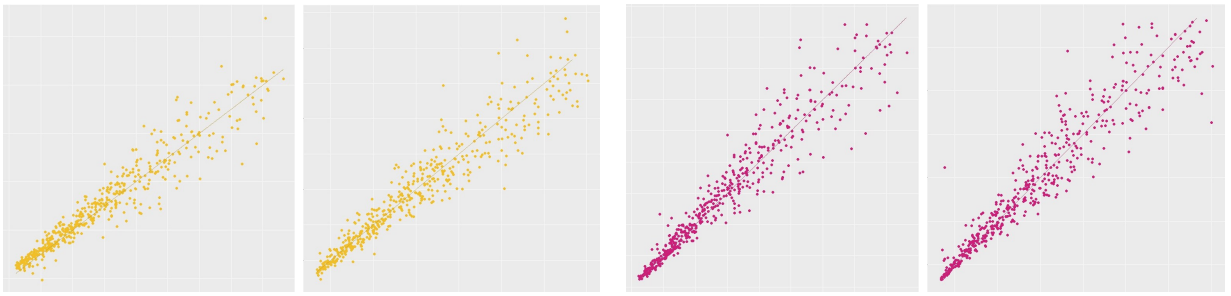
(a) Decision tree regression

(b) Tree bagging method



(c) Random forest method

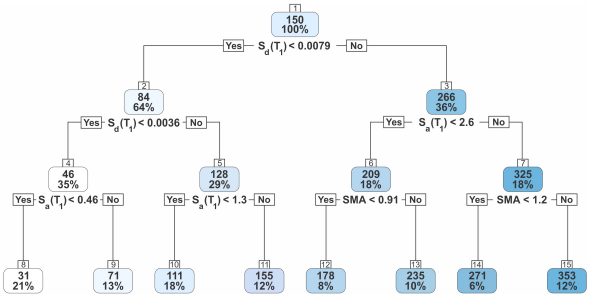
(d) XG-Boost model



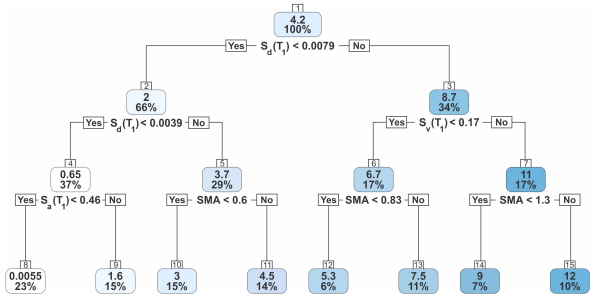
(e) Neural Network model

(f) Support vector regression

Figure 8: Quality of prediction based on different forecasting techniques and QoIs. In each sub-figure the left plot belongs to the displacement and the right one shows the stress variation. In each plot, the horizontal axis is the computed finite element model (i.e. true value) and the vertical axis is the predicted one based on machine learning. The solid black line presents the equality line.



(a) Displacements

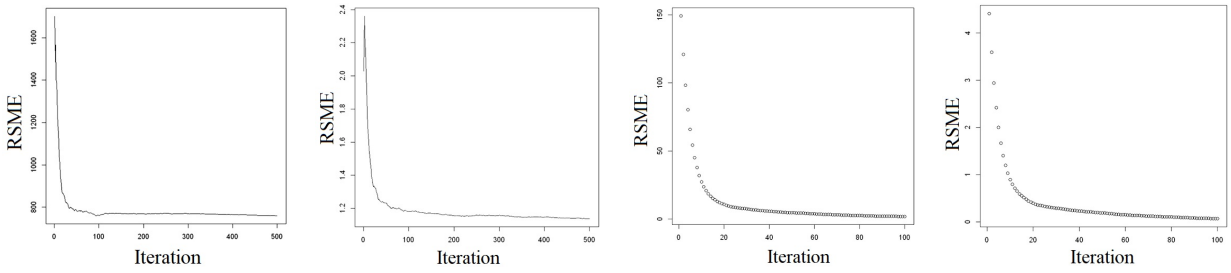


(b) Stresses

Figure 9: Decision Tree regression chart with two response quantities

- Figures 10(c) and 10(d) illustrate the RMSE error function with respect to a number of iterations in

the XGBoost model. Again, it seems that for any practical purposes, using 100 iterations is enough to get a stable result. The optimal number of iterations is important to minimize the computational cost in a forecasting problem.



(a) RF; Displacements (b) RF; Stresses (c) XGBoost; Displacements (d) XGBoost; Stresses

Figure 10: RMSE in train section for Random Forest and XGBoost models with two response quantities

- Finally, according to the ANN model, the most important relative meta-features can be summarized as:
  - For displacement: SMA $>$ $S_v(T_p^{vel})$ $>$ PGV $>$ PGA $>$ $S_d(T_1)$.
  - For stress: PGV $>$ $S_a(T_1)$ $>$ v_{rs} $>$ CAV $>$ SED.

  As seen, nearly all the meta-features in the stress response, and the top ones in the displacement response are velocity-dependent IMs.

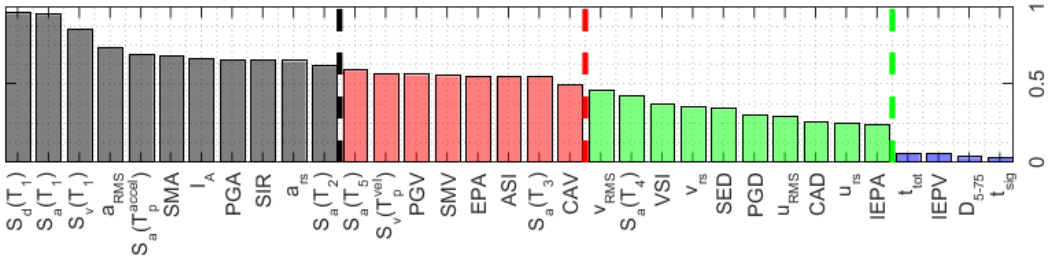
6.2. Employing the Selective Meta-Features

So far, all the existing meta-features (from Table 1) were used as input parameters of the meta-models to predict the target values. In this section, by implementing feature weighting methods, a subset of meta-features are selected which are labeled to be more significant than others. Six filtering methods have been applied from "FSelector" library [65] including: 1) Information Gain, 2) Information Gain Ratio, 3) Chi-square, 4) OneR, 5) Relief, and 6) Symmetrical Uncertainty. A brief summary of these techniques is presented in Appendix A for those readers interested in mathematical theories behind those names [66].

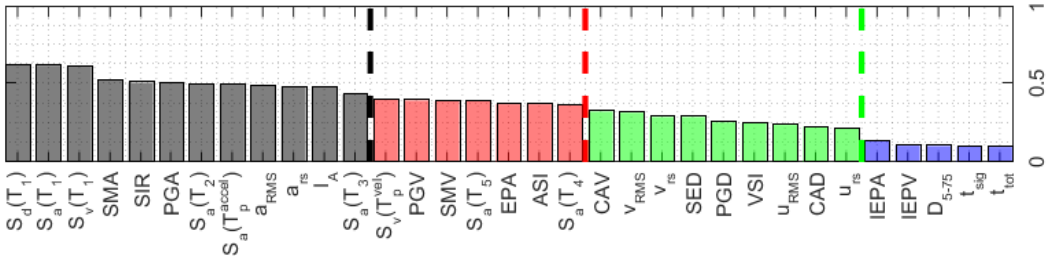
The outcome of the meta-feature selection is shown in Figure 11 in which three subsets with 28, 19 and 12 meta-features are chosen for displacement, and three subsets of 29, 19 and 11 meta-features are defined for stress response. The authors simply selected N top meta-features where there was a sudden slope change in the weight factors between the meta-feature N and $N+1$. This can be easily visualized from Figure 11 in which the selected meta-features have a darker color.

Tables 4 and 5 compare the error function for different numbers of effective meta-features. Only the test dataset is used, and the discussion on the train dataset is ignored. Major observations are:

- Comparing three subsets among each other, it reveals that subset 2 is slightly better compared to subset 1, while subset 3 has the highest error terms (in an average sense). Thus, one may conclude that reducing the initial meta-features to 2/3 might be an effective way to reduce computational efforts.
- DTR is not sensitive at all to the subset selection.
- Tree Bagging also shows a negligible reaction to the subset selection.
- The general trend based on RMSE and SMAPE is not fully consistent.
- Variation of the displacement-based error function is higher than the stress-based one for different subsets.



(a) Displacements



(b) Stress

Figure 11: Feature selection using different subsets for two response quantities; black bars (all IMs left to the black line) = subset 3; black + red bars (all the IMs left to the red line) = subset 2; green + red + black bars (all IMs left to the green line) = subset 1

Table 4: Error function for the displacement QoI with selective meta-features; only test dataset

	RMSE			SMAPE		
	Subset 1: 28 meta-features	Subset 2: 19 meta-features	Subset 3: 12 meta-features	Subset 1: 29 meta-features	Subset 2: 19 meta-features	Subset 3: 11 meta-features
Decision Tree Regression	36.72	36.72	36.72	0.2095	0.2095	0.2095
Tree Bagging	32.55	32.55	32.74	0.1939	0.1939	0.1940
Random Forest	26.55	26.33	29.18	0.1207	0.1198	0.1314
XG-Boost	27.65	27.48	30.49	0.1262	0.1270	0.1366
Neural Network	28.96	29.36	28.90	0.1447	0.1603	0.1639
SVR	28.81	27.99	30.03	0.1264	0.1282	0.1308

Table 5: Error function for the stress QoI with selective meta-features; only test dataset

	RMSE			SMAPE		
	Subset 1: 28 meta-features	Subset 2: 19 meta-features	Subset 3: 12 meta-features	Subset 1: 28 meta-features	Subset 2: 19 meta-features	Subset 3: 12 meta-features
Decision Tree Regression	1.51	1.51	1.51	0.5402	0.5402	0.5402
Tree Bagging	1.23	1.23	1.24	0.4303	0.4303	0.4310
Random Forest	1.11	1.12	1.16	0.2403	0.2378	0.2426
XG-Boost	1.19	1.17	1.24	0.2538	0.2531	0.2575
Neural Network	1.36	1.33	1.20	0.3071	0.3239	0.2558
SVR	1.16	1.17	1.17	0.2549	0.2547	0.2488

However, the most interesting findings might be related to the comparison of feature selection subsets with the references one (i.e. full model without feature selection). To achieve this goal, the error parameters in Tables 4 and 5 are normalized with the corresponding values in Tables 2 and 3, respectively. Again, the results are only presented for the test dataset using three subset selection groups. Findings are summarized in Figure 12 with the following major observations:

- The closer the normalized error values to the unit, means that the feature selection does not have any

impact on the accuracy of forecasting.

- Those normalized errors with values less than unit (shown with blue dashed line) represent the cases in which the feature selection is succeeded in overall error reduction.
- Among six forecasting techniques, four of them (i.e. DTR, TB, RF and XGB) are tree-based techniques and have a sort of inherent feature selection capability. In these techniques, the meta-features are automatically pruning the results to make them condense. As seen, those forecasting techniques are somehow neutral to feature selection. On the other hand, ANN and SVR do not have such an inherent feature selection capability and their results are highly affected.
- Among the tree-based methods:
  - DTR and tree bagging are completely neutral to feature selection using any of three subsets.
  - RF and XGBoost are neutral to feature selection in subsets 1 and 2 (i.e. large and medium size features), while in subset 3 (i.e. small size subset) they always reduce the accuracy.
- Among the none tree-based methods:
  - A neural network reduces the accuracy.
  - In general, support vector regression improves the accuracy.
- In the case of SVR (which is so far the most successful model):
  - The stress-based prediction is better than the displacement-based one.
  - There is no meaningful differences between the RMSE- and SMAPE-based evaluations.
  - The displacement-based evaluation with the smallest subset (i.e. subset 3) increases the normalized error by about 4%, all the other subsets, and also stress-based values show a positive reaction to feature selection.
- Last but not least, even for the cases in which feature selection does not improve the accuracy, this technique is promising because it shows that the same accuracy can be achieved with a smaller set of meta-features. This is important especially for the practitioners who want to run simpler models with less computational time to achieve more or less good accuracy.

7. Summary

The current research is only targeted at the ground motion record-to-record variability. Since a comprehensive seismic uncertainty assessment would require many dynamic analyses, one of the objectives in this paper is to reduce the number of simulations using forecasting tools. These machine learning techniques provide an interesting environment to train a meta-model and use it to predict the other potential cases.

To achieve this goal, it is necessary to present the stochastic nature of the ground motion signals in the form of several scalar quantities. Thus, in this paper, the concept of "unique ground motion signatures" is introduced, which acts similarly to the biometric recognition in human beings. Each ground motion can then be distinguished by these unique identifiers. In total, 35 characteristics (intensity measures in earthquake engineering or meta-features in computer science) are extracted and applied in forecasting models.

About 2,000 real ground motion records are used to evaluate the applicability of the proposed idea. A tall gravity dam is used as a vehicle for numerical simulations. Two widely used responses of the dam under seismic motions are extracted as output parameters (i.e. displacement and stress).

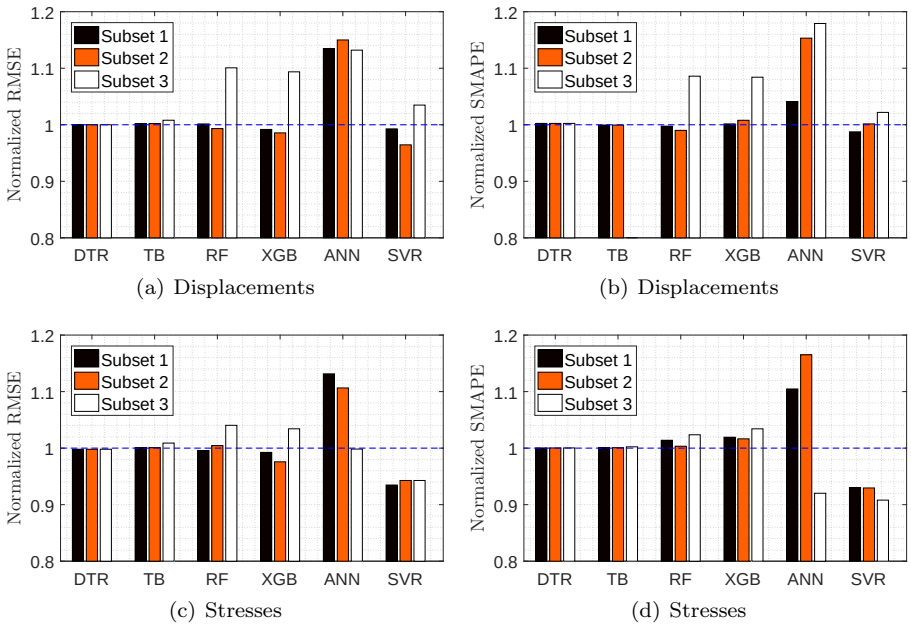


Figure 12: Normalized GOF for different subsets of feature selection compared to all features assumption

7.1. Concluding Remarks

Six forecasting models, i.e. DTR, tree bagging, RF, XGBoost, artificial neural network, and support vector regression are applied to the dataset (which is divided into train and test groups). Detailed results can be found in the paper under different sections and sub-sections; however, in general, the authors can confirm that these forecasting models are successfully implemented in the dam engineering problem. However, not all those six techniques have provided a similar accuracy. DTR was the worst model, while XGBoost was the most promising one. XGboost is an ensemble method, which builds on a big number of weak classifiers (i.e. high bias, low variance models). Decision trees are individual weak learners and sometimes even as small as trees with two leaves. The idea of boosting is to add a classifier (e.g. a decision tree) at a time, so that the next classifier is trained to improve the already trained ensemble. This idea reduces error, mainly by reducing bias (and also to some extent variance) by means of aggregating the output from many models. Therefore, it provides better results than a weak learner like a decision tree.

The authors also examined the concept of feature selection to forecast the dam responses. The idea is to use an initial subset of 35 ground motion meta-features (and not all of them) for regression analysis. Two major findings are: 1) apart from the ANN and SVR, other tree-based models are practically neutral to feature selection; and 2) the SVR provides very good performance with a small set of features. Even for those meta-models which do not affect the accuracy, employing a smaller set of input meta-features is beneficial from a computational point of view.

7.2. Directions to Future Work

Last but not least, this paper proposed a general idea of employing ground motion unique signatures in meta-modeling and advanced forecasting of the dam responses from finite element simulation. The extension of this work can be directed in the following major scopes:

- Employing the proposed idea for 3D dynamic analysis of dams including different “dam classes”.
- Extending this work for embankment/rockfill dams.
- Forecasting the nonlinear response of the dams.
- Applying the multi-component ground motion records, and combining the signal meta-features in three-dimensions.

- Introducing the material/modeling uncertainty and performing a hybrid uncertainty quantification problem.
- Combining this idea with ground motion selection techniques to be used only for a particular seismic hazard level.

8. Acknowledgment

The first author would like to express his sincere appreciation to his former advisor (and the current mentor), Professor Victor E. Saouma at the University of Colorado Boulder for his enthusiastic guidance and advice throughout this research. The second author would like to knowledge the support for this research by the Czech Science Foundation (GACR Project GA 17-22662S) and Operational Program Education for Competitiveness (Project No. CZ.1.07/2.3.00/20.0296).

References

[1] K. Porter, An overview of PEER's performance-based earthquake engineering methodology, in: Proceedings of the 9th International Conference on Applications of Statistics and Probability in Civil Engineering (ICASP9), San Francisco, CA, 2003.

[2] F. Salazar, R. Morán, M. Á. Toledo, E. Oñate, Data-based models for the prediction of dam behaviour: A review and some methodological considerations, Archives of Computational Methods in Engineering (2015) 1–21.

[3] J.-y. Chen, Q. Xu, J. Li, S.-l. Fan, Improved response surface method for anti-slide reliability analysis of gravity dam based on weighted regression, Journal of Zhejiang University-SCIENCE A 11 (2010) 432–439.

[4] I. Karimi, N. Khaji, M. Ahmadi, M. Mirzayee, System identification of concrete gravity dams using artificial neural networks based on a hybrid finite element–boundary element approach, Engineering structures 32 (2010) 3583–3591.

[5] S.-l. Fan, J.-y. Chen, J. Li, F. Wu-qiang, Roller compacted concrete gravity dams reliability analysis based on response surface approach, in: Earth and Space 2010: Engineering, Science, Construction, and Operations in Challenging Environments, 2010, pp. 3355–3367.

[6] C. Gu, B. Li, G. Xu, H. Yu, Back analysis of mechanical parameters of roller compacted concrete dam, Science China Technological Sciences 53 (2010) 848–853.

[7] H. Su, Z. Wen, S. Zhang, S. Tian, Method for choosing the optimal resource in back-analysis for multiple material parameters of a dam and its foundation, Journal of Computing in Civil Engineering 30 (2016).

[8] A. Gaspar, F. Lopez-Caballero, A. Modaressi-Farahmand-Razavi, A. Gomes-Correia, Methodology for a probabilistic analysis of an rcc gravity dam construction. modelling of temperature, hydration degree and ageing degree fields, Engineering Structures 65 (2014) 99–110.

[9] L. Cheng, J. Yang, D. Zheng, B. Li, J. Ren, The health monitoring method of concrete dams based on ambient vibration testing and kernel principle analysis, Shock and Vibration 2015 (2015).

[10] H. Gu, Z. Wu, X. Huang, J. Song, Zoning modulus inversion method for concrete dams based on chaos genetic optimization algorithm, Mathematical Problems in Engineering 2015 (2015).

[11] M. Rezaiee-Pajand, F. H. Tavakoli, Crack detection in concrete gravity dams using a genetic algorithm, Proceedings of the Institution of Civil Engineers-Structures and Buildings 168 (2015) 192–209.

[12] C. Xin, G. Chongshi, Risk analysis of gravity dam instability using credibility theory monte carlo simulation model, SpringerPlus 5 (2016) 778.

[13] X. Cao, C. Gu, E. Zhao, Uncertainty instability risk analysis of high concrete arch dam abutments, Mathematical Problems in Engineering 2017 (2017).

[14] M. A. Hariri-Ardebili, F. Pourkamali-Anaraki, Support vector machine based reliability analysis of concrete dams, Soil Dynamics and Earthquake Engineering 104 (2018) 276–295.

[15] M. A. Hariri-Ardebili, F. Pourkamali-Anaraki, Simplified reliability analysis of multi hazard risk in gravity dams via machine learning techniques, Archives of Civil and Mechanical Engineering 18 (2018) 592–610.

[16] M. A. Hariri-Ardebili, Mcs-based response surface metamodels and optimal design of experiments for gravity dams, Structure and Infrastructure Engineering 14 (2018) 1641–1663.

[17] M. Hariri-Ardebili, P. Boodagh, Taguchi design-based seismic reliability analysis of geostructures, Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards (2018) 1–19.

[18] L. Barroso, S. Winterstein, Probabilistic seismic demand analysis of controlled steel moment-resisting frame structures, Earthquake Engineering and Structural Dynamics 31 (2002) 2049–2066.

[19] S. Jankovic, B. Stojadinovic, Probabilistic performance-based seismic demand model for {R/C} frame buildings, in: Proceeding of the 13th World Conferance on Earthquake Engineering, Vancouver, B.C., Canada, 2004.

[20] S. Ramamoorthy, P. Gardoni, J. Bracci, Probabilistic demand models and fragility curves for reinforced concrete frames, Journal of Structural Engineering 132 (2006) 1563–1572.

[21] Y. Tang, J. Zhang, Probabilistic seismic demand analysis of a slender {RC} shear wall considering soil–structure interaction effects, Engineering Structures 33 (2011) 218–229.

[22] V. Bisadi, P. Gardoni, M. Head, Probabilistic demand models and fragility estimates for bridges elevated with steel pedestals, Journal of Structural Engineering 139 (2012) 1515–1528.

[23] N. Tondini, B. Stojadinovic, Probabilistic seismic demand model for curved reinforced concrete bridges, Bulletin of Earthquake Engineering 10 (2012) 1455–1479.

[24] F. Berahman, F. Behnamfar, Probabilistic seismic demand model and fragility estimates for critical failure modes of un-anchored steel storage tanks in petroleum complexes, Probabilistic Engineering Mechanics 24 (2009) 527–536.

[25] M. A. Hariri-Ardebili, V. Saouma, Probabilistic seismic demand model and optimal intensity measure for concrete dams, Structural Safety 59 (2016) 67–85.

[26] M. Mello, H. Bhat, A. Rosakis, H. Kanamori, Identifying the unique ground motion signatures of supershear earthquakes: Theory and experiments, Tectonophysics 493 (2010) 297–326.

[27] S. K. Murthy, Automatic construction of decision trees from data: A multi-disciplinary survey, Data mining and knowledge discovery 2 (1998) 345–389.

[28] F. Esposito, D. Malerba, G. Semeraro, J. Kay, A comparative analysis of methods for pruning decision trees, IEEE transactions on pattern analysis and machine intelligence 19 (1997) 476–491.

[29] G. Rizzo, C. d'Amato, N. Fanizzi, F. Esposito, Tree-based models for inductive classification on the web of data, Web Semantics: Science, Services and Agents on the World Wide Web 45 (2017) 1–22.

[30] J. Friedman, T. Hastie, R. Tibshirani, The elements of statistical learning, volume 1, Springer series in statistics New York, 2001.

[31] R. Kumar, Decision tree for the weather forecasting, International Journal of Computer Applications 76 (2013) 0975–8887.

[32] Z. Wang, Y. Wang, R. Zeng, R. S. Srinivasan, S. Ahrentzen, Random forest based hourly building energy prediction, Energy and Buildings 171 (2018) 11–25.

[33] M. Thakur, D. Kumar, A hybrid financial trading support system using multi-category classifiers and random forest, Applied Soft Computing 67 (2018) 337–349.

[34] L. Breiman, Bagging predictors, Machine learning 24 (1996) 123–140.

[35] M. Kuhn, K. Johnson, Applied predictive modeling, volume 26, Springer, 2013.

[36] T. Chen, T. He, M. Benesty, et al., Xgboost: extreme gradient boosting, R package version 0.4-2 (2015) 1–4.

[37] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, ACM, 2016, pp. 785–794.

[38] L. Torlay, M. Perrone-Bertolotti, E. Thomas, M. Baciu, Machine learning–xgboost analysis of language networks to classify patients with epilepsy, Brain informatics 4 (2017) 159.

[39] A. Natekin, A. Knoll, Gradient boosting machines, a tutorial, Frontiers in neurorobotics 7 (2013) 21.

[40] J. Fan, X. Wang, L. Wu, H. Zhou, F. Zhang, X. Yu, X. Lu, Y. Xiang, Comparison of support vector machine and extreme gradient boosting for predicting daily global solar radiation using temperature and precipitation in humid subtropical climates: A case study in china, Energy Conversion and Management 164 (2018) 102–111.

[41] R Statistical, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2015. URL: https://www.R-project.org.

[42] K. L. Priddy, P. E. Keller, Artificial neural networks: an introduction, volume 68, SPIE press, 2005.

[43] K. Mehrotra, C. K. Mohan, S. Ranka, Elements of artificial neural networks, MIT press, 1997.

[44] N. A. Gershenfeld, N. Gershenfeld, The nature of mathematical modeling, Cambridge university press, 1999.

[45] G. Zhang, B. E. Patuwo, M. Y. Hu, Forecasting with artificial neural networks:: The state of the art, International journal of forecasting 14 (1998) 35–62.

[46] T. Ragg, H. Braun, H. Landsberg, A comparative study of neural network optimization techniques, in: Artificial Neural Nets and Genetic Algorithms, Springer, 1998, pp. 341–345.

[47] V. Vapnik, The support vector method of function estimation, in: Nonlinear Modeling, Springer, 1998, pp. 55–85.

[48] V. Vapnik, The nature of statistical learning theory, Springer science & business media, 2013.

[49] A. Løkke, A. K. Chopra, Response spectrum analysis of concrete gravity dams including dam-water-foundation interaction, Journal of Structural Engineering 141 (2014) 04014202.

[50] G. Fenves, A. Chopra, EAGD-84: A computer program for earthquake analysis of concrete gravity dams, University of California, Earthquake Engineering Research Center, 1984.

[51] G. Fenves, A. Chopra, Earthquake analysis of concrete gravity dams including reservoir bottom absorption and dam-water-foundation rock interaction, Earthquake Engineering and Structural Dynamics 12 (1984) 663–680.

[52] M. A. Hariri-Ardebili, M. R. Kianoush, Integrative seismic safety evaluation of a high concrete arch dam, Soil Dynamics and Earthquake Engineering 67 (2014) 85–101.

[53] Y. Ghanaat, Failure modes approach to safety evaluation of dams, in: Proceedings of the 13th World Conference on Earthquake Engineering, Vancouver, BC, Canada, 2004.

[54] PEER, Ground motion database, http://ngawest2.berkeley.edu/, 2017. Last viewed June 2017.

[55] F. Jalayer, Direct probabilistic seismic analysis: implementing non-linear dynamic assessments, Ph.D. thesis, Stanford University, Stanford, Palo-Alto, CA, 2003.

[56] A. Morales-Esteban, J. L. de Justo, F. Martínez-Álvarez, J. Azañón, Probabilistic method to select calculation accelerograms based on uniform seismic hazard acceleration response spectra, Soil Dynamics and Earthquake Engineering 43 (2012) 174–185.

[57] N. Jayaram, T. Lin, J. Baker, A computationally efficient ground-motion selection algorithm for matching a target response spectrum mean and variance, Earthquake Spectra 27 (2011) 797–815.

[58] M. Kuhn, et al., Caret package, Journal of statistical software 28 (2008) 1–26.

[59] T. Therneau, B. Atkinson, B. Ripley, M. B. Ripley, Package 'rpart', Available online: cran. ma. ic. ac. uk/web/packages/rpart/rpart. pdf (2018). Last viewed October 2018.

[60] L. Breiman, randomforest: Breiman and cutler's random forests for classification and regression. r package version 4.6-12, 2015.

[61] A. Peters, T. Hothorn, M. T. Hothorn, Package 'ipred', 0.8-7. The R Foundation for Statistical Computing (2009).

[62] A. Arora, A. Candel, J. Lanford, E. LeDell, V. Parmar, Deep learning with h2o, H2O. ai, Mountain View (2015).

[63] E. Dimitriadou, K. Hornik, F. Leisch, D. Meyer, A. Weingessel, Misc functions of the department of statistics (e1071), tu wien, R package 1 (2008) 5–24.

[64] A. Chopra, Dynamics of Structures: Theory and Applications to Earthquake Engineering, Prentice-Hall, Englewood Cliffs, NJ, 1995.

[65] P. Romanski, L. Kotthoff, Package 'FSelector', 2018. Last viewed October 2018.

[66] S. Barak, M. Modarres, Developing an approach to evaluate stocks by forecasting effective features with data mining methods, Expert Systems with Applications 42 (2015) 1325–1339.

[67] R. C. Holte, Very simple classification rules perform well on most commonly used datasets, Machine learning 11 (1993) 63–90.

[68] K. Kira, L. A. Rendell, The feature selection problem: Traditional methods and a new algorithm, in: Aaai, volume 2, 1992, pp. 129–134.

Appendix A. Feature Selection Techniques

Appendix A.1. Information Gain

One popular feature selection technique is to calculate the information gain (IG) of features. The IG of a feature indicates the amount of information that can be obtained with regards to the target value of classification of the feature. Entropy as a measure of uncertainty of a RV (which is the basis of IG) can be written as:

$$H(X) = -\sum_i P(x_i).\log P(x_i) \tag{A.1}$$

Where X is RV and $P(x)$ is probability mass function.

Similarly, the conditional entropy of X after observation of Y is defined as:

$$H(X|Y) = -\sum_j P(y_j).\sum_i P(x_i|y_j).\log P(x_i|y_j) \tag{A.2}$$

Information gain is defined as a decrease in uncertainty of X after observing Y:

$$IG(X;Y) = H(X) - H(X|Y) \tag{A.3}$$

The attributes which contribute more information are then selected, and the rest, with lower IG scores, are removed.

Appendix A.2. Information Gain Ratio

Information gain has a bias on attributes that have a large number of values. In a subset, as the number of the attributes increased, the entropy may be boosted. To reduce this bias, an information gain ratio (IGR) is introduced as a normalized IG. The IGR is formulated as:

$$IGR(X;Y) = \frac{H(X) - H(X|Y)}{H(X)} \tag{A.4}$$

Appendix A.3. Symmetrical Uncertainty

Similar to the IGR, and based on the concept of uncertainty, symmetrical uncertainty (SU) is a symmetrical normalized version of IG which is formulated as:

$$SU(X|Y) = \frac{2 \times IG(X;Y)}{H(X) + H(Y)} \tag{A.5}$$

The features with a larger SU value, get a higher weight. Those with a lower weight can be dropped from the feature's list, and marked as unnecessary attributes.

Appendix A.4. Chi-Square

The chi-square, χ^2, is a statistical measure to identify the dependency of two variables. This dependency of feature x and target value y could be calculated from a two-way contingency table of them. The χ^2 could be written as:

$$\chi^2(x;y) = \frac{N \times (AD - CB)^2}{(A+C)(B+D) \times (A+B)(C+D)} \tag{A.6}$$

where A the number of co-occurrence of x and y; B the number of x occurrence without y; C the number of y occurrence without x; D the number of times neither x nor y occurs.

The average χ^2 score of each feature x among target values y_i can be obtained with the following formula:

$$\chi^2_{avg}(x) = \sum_j P(y_j)\chi^2(x;y_j) \tag{A.7}$$

The higher value of χ^2 implies that the independence hypothesis will be rejected more significantly, and also shows a stronger relationship between the feature and target values.

Appendix A.5. OneR

OneR determines the weights of the features based on a simple rule in the process that just one feature is considered in the conditional situation. A simple pseudo-code of OneR is shown in algorithm 2. More details about OneR can be found in Holte [67].

Algorithm 2 OneR pseudo-code

```
for each feature x do
    for each value v of x do
        Compute class distribution based on feature value
        Set C = the most frequent class
        Create a rule regards to: if x = v then class = C
    end for
    Calculate the error rate of the rule based on whole data set
end for
Select the rule with highest accuracy
```

Appendix A.6. Relief

Kira and Rendell [68] introduced the relief filtering method to estimate the relevance weight for features. The weight estimation is based on the ability to distinguish differences between instances which belong to separate classes. Features' initial weights are set to zero, and then are updated iteratively. A pseudo-code for the relief is shown algorithm in 3.

Algorithm 3 Relief pseudo-code

```
Set weights for all features to zero
for all instances do
    Find k nearest hits (closest neighbors in the same class)
    Find k nearest miss (nearest neighbors on the dissimilar class)
    for For all features do
        Update weight according to the distance of the instances to its Hit and Nearest Miss
    end for
end for
```