Assessing the Performance of a Classification-Based Vulnerability Analysis Model

Tai-Ran Wang, Vincent Mousseau, Nicola Pedroni, Enrico Zio

 DOI: 10.1111/risa.12305

Assessing the Performance of a Classification-Based Vulnerability Analysis Model

Tai-ran Wang,[1,*] Vincent Mousseau,[2] Nicola Pedroni,[1] and Enrico Zio[1,3]

In this article, a classification model based on the majority rule sorting (MR-Sort) method is employed to evaluate the vulnerability of safety-critical systems with respect to malevolent intentional acts. The model is built on the basis of a (limited-size) set of data representing (*a priori* known) vulnerability classification examples. The empirical construction of the classification model introduces a source of uncertainty into the vulnerability analysis process: a quantitative assessment of the performance of the classification model (in terms of accuracy and confidence in the assignments) is thus in order. Three different approaches are here considered to this aim: (i) a model-retrieval-based approach, (ii) the bootstrap method, and (iii) the leave-one-out cross-validation technique. The analyses are presented with reference to an exemplificative case study involving the vulnerability assessment of nuclear power plants.

KEY WORDS: Classification model; confidence estimation; MR-Sort; nuclear power plants; vulnerability analysis

1. INTRODUCTION

The vulnerability of safety-critical systems and infrastructures (e.g., nuclear power plants) is of great concern, given the multiple and diverse hazards that they are exposed to (e.g., intentional, random, natural)[(1)] and the potential large-scale consequences. This has motivated an increased attention in analyses to guide designers, managers, and stakeholders in (i) the systematic identification of the sources of vulnerability, (ii) its qualitative and quantitative assessment,[(2,3)] and (iii) the selection of proper actions to reduce it. In this article, we are concerned only with *intentional* hazards (i.e., those related to malevolent acts) and we mainly address issue (ii) mentioned above (i.e., the quantitative evaluation of vulnerability).

With respect to that, due to the specific features (low frequency but important effects) of intentional hazards (characterized by significant *uncertainties* due to behaviors of different rationality) the analysis is difficult to perform by traditional risk assessment methods.[(1,4,5)] For this reason, in this work we propose to tackle the issue of evaluating vulnerability to malevolent intentional acts by an empirical classification modeling framework. In particular, we adopt a classification model based on the majority rule sorting (MR-Sort) method[(6)] to assign an alternative of interest (i.e., a safety-critical system) to a given (vulnerability) class (or category). The MR-Sort classification model contains a group of (adjustable) parameters that have to be calibrated by means of a set of *empirical* classification examples (also called training set), that is, a set of alternatives with the corresponding preassigned vulnerability classes.

[1]Chair on Systems Science and the Energy Challenge, European Foundation for New Energy-Electricité de France, Ecole Centrale Paris and Supélec, Chatenay Malabry Cedex, France.
[2]Laboratory of Industrial Engineering, Ecole Centrale Paris, Grande Voie des Vignes, F92-295, Chatenay Malabry Cedex, France.
[3]Politecnico di Milano, Energy Department, Nuclear Section, c/o Cesnef, via Ponzio 33/A, 20133, Milan, Italy.
*Address correspondence to Tai-ran Wang, Ecole Centrale Paris and Supélec, Grande Voie des Vignes, F92-295, Chatenay Malabry, Cedex, France; tairan.wang@ecp.fr.

Due to the finite (typically small) size of the set of training classification examples usually available in the analysis of real complex safety-critical systems, the performance of the classification model is impaired. In particular, (i) the classification *accuracy* (resp., error), that is, the expected fraction of patterns correctly (resp., incorrectly) classified, is typically reduced (resp., increased); (ii) the classification process is characterized by significant uncertainty, which affects the *confidence* of the classification-based vulnerability model: in our work, we define the confidence in a classification assignment as in Ref. 10, that is, as the probability that the class assigned by the model to a given (single) pattern is the correct one. Obviously, there is the possibility that a classification model assigns correctly a very large (expected) fraction of patterns (i.e., the model is very accurate), but at the same time *each* (correct) assignment is affected by significant uncertainty (i.e., it is characterized by low confidence). It is worth mentioning that besides the scarcity of training data, there are many additional sources of uncertainty in classification problems (e.g., the accuracy of the data, the suitability of the classification technique used): however, they are not considered in this work.

The performance of the classification model (i.e., the classification accuracy—resp., error—and the confidence in the classification) needs to be quantified: this is of paramount importance for taking robust decisions in the vulnerability analyses of safety-critical systems.[7,8]

In this article, three different approaches are used to assess the performance of a classification-based MR-Sort vulnerability model in the presence of small training data sets. The first is a model-retrieval-based approach,[6] which is used to assess the expected percentage error in assigning new alternatives. The second is based on *bootstrapping* the available training set in order to build an ensemble of vulnerability models;[9] the method can be used to assess both the accuracy and the confidence of the model: in particular, the confidence in the assignment of a given alternative is given in terms of the full (probability) distribution of the possible vulnerability classes for that alternative (built on the bootstrapped ensemble of vulnerability models).[10] The third is based on the leave-one-out cross-validation (LOOCV) technique, in which one element of the available data set is (left out and) used to test the accuracy of the classification model built on the remaining data: also this approach is employed to estimate the accuracy of the classification vulnerability model as the expected percentage error, that is, the fraction of alternatives incorrectly assigned (computed as an average over the left-out data).

The contribution of this work is twofold:

- classification models have proved useful in a variety of fields including finance, marketing, environmental and energy management, human resources management, medicine, risk analysis, fault diagnosis, etc.,[11] but to the best of the authors' knowledge, this work is the first to propose a classification-based hierarchical framework for the analysis of the vulnerability to intentional hazards of safety-critical systems;
- the bootstrap method is originally applied to estimate the confidence in the assignments provided by the MR-Sort classification model, in terms of the probability that a given alternative is correctly classified.

The article is organized as follows. The next section presents the hierarchical framework for vulnerability analysis to intentional hazards. Section 3 shows the classification model applied within the proposed framework. Section 4 describes the learning process of a classification model by the disaggregation method. In Section 5 three approaches are proposed to analyze the performance of the classification model. Then, the proposed approaches are validated on the case study of a group of nuclear power plants (NPPs) in Section 6. Finally, Sections 7 and 8 present the discussion and conclusions of this research.

2. GENERAL FRAMEWORK: VULNERABILITY TO INTENTIONAL HAZARDS

Vulnerability is defined in different ways depending on the domains of application, for example, a measure of possible future harm due to exposure to a hazard,[1] the identification of weaknesses in security, focusing on defined threats that could compromise a system's ability to provide a service,[12] the set of conditions and processes resulting from physical, social, economic, and environmental factors that increase the susceptibility of a community to the impact of hazards.[13]

With the focus on the susceptibility to intentional hazards, the three-layers hierarchical model developed in Ref. 14 is considered and shown in Fig. 1. The susceptibility to intentional hazards is characterized

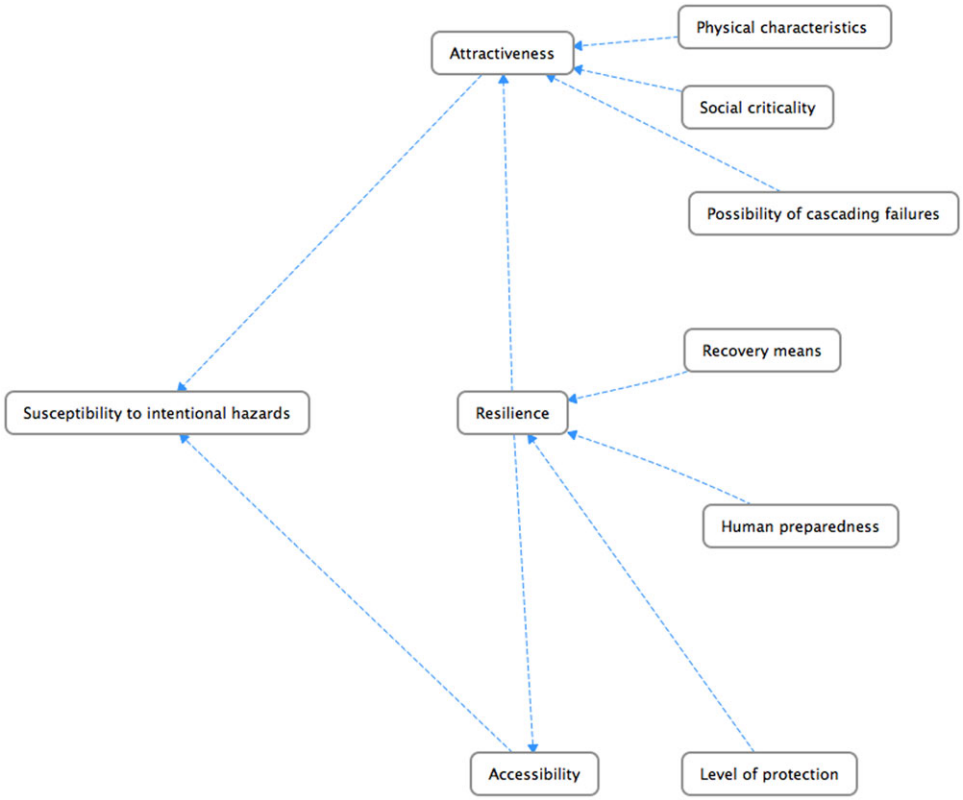

Fig. 1. Hierarchical model for susceptibility to intentional hazards.

in terms of attractiveness and accessibility. These are hierarchically broken down into factors that influence them, including resilience seen as pre-attack protection (which influences on accessibility) and post-attack recovery (which influences on attractiveness). The decomposition is made in six criteria, which are further decomposed into a layer of basic subcriteria, for which data and information can be collected. The details of the general framework of analysis are not given here for brevity; the interested reader is referred to Ref. 14 and to Appendix A.

For the purpose of this article, only six criteria are considered: physical characteristics, social criticality, possibility of cascading failures, recovery means, human preparedness, and level of protection (Fig. 1). These six criteria are used as the basis to assess the vulnerability of a given safety-critical system of interest (e.g., an NPP). Four levels (or categories) of vulnerability are considered: satisfactory, acceptable, problematic, and serious. In this view, the issue of assessing vulnerability is here tackled within a classification framework: given the characterization of a critical system in terms of the six criteria mentioned above, a proper vulnerability category (or class) has to be selected for that system. A description of the algorithm used to this purpose is given in the following section.

It is worthy to mention that the cyber characteristics are not taken into account in this work; in future work they will be added for the criteria physical characteristics and protection.

3. CLASSIFICATION MODEL FOR VULNERABILITY ANALYSIS: THE MR-SORT METHOD

The MR-Sort method is a simplified version of ELECTRE Tri, an outranking sorting procedure in which the assignment of an alternative to a given category is determined using a complex concordance-non-discordance rule.[15,16] We assume that the alternative to be classified (in this article, a safety-critical system or infrastructure of interests, e.g., an NPP) can be described by an n-tuple of elements $x = \{x_1, x_2, ..., x_i, ..., x_n\}$, which represent the evaluation of the alternative with respect to a set of n

criteria (by way of example, in this article the criteria used to evaluate the vulnerability of a safety-critical system of interest may include its physical characteristics, social criticality, level of protection, and so on: see Section 2). We denote the set of criteria by $N = \{1, 2, ..., i, ..., n\}$ and assume that the values x_i of criterion i range in the set X_i[9] (e.g., in this article all the criteria range in [0, 1]). The MR-Sort procedure allows assigning any alternative $x = \{x_1, x_2, ..., x_i, ..., x_n\} \in X = X_1 \times X_2 \times ... \times X_i \times ... \times X_n$ to a particular predefined category (in this article, a class of vulnerability), in a given ordered set of categories, $\{A^h : h = 1, 2, ..., k\}$; as mentioned in Section 2, $k = 4$ categories are considered in this work: A^1 = satisfactory, A^2 = acceptable, A^3 = problematic, A^4 = serious.

To this aim, the model is further specialized in the following way:

- We assume that X_i is a subset of $\mathbb{R}$ for all $i \in \mathbb{N}$ and the subintervals $(X_i^1, X_i^2, ..., X_i^h, ..., X_i^k)$ of X_i are compatible with the order on the real numbers, that is, for all $x_i^1 \in X_i^1, x_i^2 \in X_i^2, ..., x_i^h \in X_i^h, ..., x_i^k \in X_i^k$, we have $x_i^1 > x_i^2 > ... > x_i^h > ... > x_i^k$. We assume furthermore that each interval $x_i^h, h = 2, 3, ..., k$ has a smallest element b_i^h, which implies that $x_i^{h-1} \geq b_i^h > x_i^h$. The vector $b^h = \{b_1^h, b_2^h, ..., b_i^h, ..., b_n^h\}$ (containing the lower bounds of the intervals X_i^h of criteria $i = 1, 2, ..., n$ in correspondence of category h) represents the lower limit profile of category A^h.
- There is a weight ω_i associated with each criterion $i = 1, 2, ..., n$, quantifying the relative importance of criterion i in the vulnerability assessment process; notice that the weights are normalized such that $\sum_{i=1}^{n} \omega_i = 1$.
  In this framework, a given alternative $x = \{x_1, x_2, ..., x_i, ..., x_n\}$ is assigned to category $A^h, h = 1, 2, ..., k$, if

$$\sum_{i \in \mathbb{N}: x_i \geq b_i^h} \omega_i \geq \lambda \text{ and } \sum_{i \in \mathbb{N}: x_i \geq b_i^{h+1}} \omega_i < \lambda, \quad (1)$$

  where λ is a threshold ($0 \leq \lambda \leq 1$) chosen by the analyst. Rule (1) is interpreted as follows. An alternative x belongs to category A^h if: (1) its evaluations in correspondence of the n criteria (i.e., the values $\{x_1, x_2, ..., x_i, ..., x_n\}$) are at least as good as b_i^h(lower limit of category A^h with respect to criterion i), $i = 1, 2, ..., n$, on a subset of criteria that has sufficient importance (in other words, on a subset of criteria that has a weight larger than or equal to the threshold λ chosen by the analyst); and at the same time (2) the weight of the subset of criteria on which the evaluations $\{x_1, x_2, ..., x_i, ..., x_n\}$ are at least as good as b_i^{h+1}(lower limit of the successive category A^{h+1} with respect to criterion i), $i = 1, 2, ..., n$, is not sufficient to justify the assignment of x to the successive category A^{h+1}. Notice that alternative x is assigned to the best category A^1 if $\sum_{i \in \mathbb{N}: x_i \geq b_i^1} \omega_i \geq \lambda$ and it is assigned to the worst category A_k if $\sum_{i \in \mathbb{N}: x_i \geq b-k-1} \omega_i < \lambda$. Finally, it is straightforward to notice that the parameters of such a model are the $k \cdot n$ lower limit profiles (n limits for each of the k categories), the n weights of the criteria $\omega_1, \omega_2, ..., \omega_i, ..., \omega_n$, and the threshold λ, for a total of $n(k+1)+1$ parameters.

4. CONSTRUCTING THE MR-SORT CLASSIFICATION MODEL

In order to construct an MR-Sort classification model, we need to determine the set of $n(k+1)+1$ parameters described in Section 2, that is, the weights $\omega = \{\omega_1, \omega_2, ..., \omega_n\}$, the lower profiles $b = \{b^1, b^2, ..., b^h, ..., b^k\}$,with $b^h = \{b_1^h, b_2^h, ..., b_i^h, ..., b_n^h\}, h = 1, 2, ..., k$, and the threshold λ; in this article, λ is considered a fixed, constant value chosen by the analyst (e.g., $\lambda = 0.9$).

To this aim, the decision maker provides a training set of classification examples $D_{TR} = \{(x_p, \Gamma_p^t), p = 1, 2, ..., N_{TR}\}$, that is, a set of N_{TR} alternatives (in this case, NPPs) $x_p = \{x_1^p, x_2^p, ..., x_i^p, ..., x_n^p\}, p = 1, 2, ..., N_{TR}$ together with the corresponding real preassigned categories (i.e., vulnerability classes) Γ_p^t (the superscript t indicates that Γ_p^t represents the true, *a priori* known vulnerability class of alternative x_p).

The calibration of the $n(k+1)$ parameters is done through the learning process detailed in Ref. 6. In extreme synthesis, the information contained in the training set D_{TR} is used to restrict the set of MR-Sort models compatible with such information, and to finally select one among them.[6] The *a priori* known assignments generate constraints on the parameters of the MR-Sort model. In Ref. 6, such constraints have a linear formulation and are integrated into a mixed integer program (MIP) that is designed to select one (optimal) set of such parameters ω^* and b^* (in other words, to select one classification model $M(\cdot|\omega^*, b^*)$) that is coherent with

the data available and maximizes a defined *objective function*. In Ref. 6, the optimal parameters ω^* and b^* are those that maximize the value of the minimal slack in the constraints generated by the given set of data D_{TR}. Once the (optimal) classification model $M(\cdot|\omega^*, b^*)$ is constructed, it can be used to assign a new alternative x(i.e., a new NPP) to one of the vulnerability classes $A^h, h = 1, 2, ..., k$: in other words, $M(x|\omega^*, b^*) = \Gamma_x^M$ where Γ_x^M is the class assigned by model $M(\cdot|\omega^*, b^*)$ to alternative x and assumes one value among $\{A^h : h = 1, 2, ..., k\}$. Further mathematical details about the training algorithm are not given here for brevity: the reader is referred to Ref. 6 and to Appendix B.

Obviously, the number N_{TR} of available classification examples is finite and quite small in most real applications involving the vulnerability analysis of safety-critical systems. As a consequence, the model $M(\cdot|\omega^*, b^*)$ is only a partial representation of reality and its assignments are affected by uncertainty: this uncertainty, which needs to be quantified to build confidence in the decision process that follows the vulnerability assessment.

In the following section, three different methods are presented to assess the performance of the MR-Sort classification model.

5. METHODS FOR ASSESSING THE PERFORMANCE OF THE CLASSIFICATION-BASED VULNERABILITY ANALYSIS MODEL

5.1. Model-Retrieval-Based Approach

The first method is based on the model-retrieval approach proposed in Ref. 6. A fictitious set D_{TR}^{rand} of N_{TR} alternatives $\{x_p^{rand} : p = 1, 2, ..., N_{TR}\}$ is generated by random sampling within the ranges X_i of the criteria, $i = 1, 2, ..., n$. Notice that the size N_{TR} of the fictitious set D_{TR}^{rand} has to be the same as the real training set D_{TR} available, for the comparison to be fair. Also, an MR-Sort classification model $M(\cdot|\omega^{rand}, b^{rand})$ is constructed by randomly sampling possible values of the internal parameters, $\{\omega_i : i = 1, 2, ..., n\}$ and $\{b_h : h = 1, 2, ..., k-1\}$. Then, we simulate the behavior of a decision-maker (DM) by letting the (random) model $M(\cdot|\omega^{rand}, b^{rand})$ assign the (randomly generated) alternatives $\{x_p^{rand} : p = 1, 2, ..., N_{TR}\}$. In other words, we construct a learning set D_{TR}^{rand} by assigning the (randomly generated) alternatives using the (randomly generated) MR-Sort model, that is, $D_{TR}^{rand} = \{(x_p^{rand}, \Gamma_p^M) : p = 1, 2, ..., N_{TR}\}$, where Γ_p^M is the class assigned by model $M(\cdot|\omega^{rand}, b^{rand})$ to alternative x_p^{rand}, that is, $\Gamma_p^M = M(x_p^{rand}|\omega^{rand}, b^{rand})$. Subsequently, a new MR-Sort model $M'(\cdot|\omega', b')$, compatible with the training set D_{TR}^{rand}, is inferred using the MIP formulation summarized in Section 3 and in Appendix B. Although models $M(\cdot|\omega^{rand}, b^{rand})$ and $M'(\cdot|\omega', b')$ may be quite different, they coincide on the way they assign elements of D_{TR}^{rand}, by construction. In order to compare models M and M', we randomly generate a (typically large) set D_{test}^{rand} of *new* alternatives $D_{test}^{rand} = \{x_p^{test,rand} : p = 1, 2, ..., N_{Test}\}$ and we compute the percentage of assignment errors, that is, the proportion of these N_{Test} alternatives that models M and M' assign to different categories.

In order to account for the randomness in the generation of the training set D_{TR}^{rand} and of the model $M(\cdot|\omega^{rand}, b^{rand})$, and to provide robust estimates for the assignment errors ϵ, the procedure outlined above is repeated for a large number N_{sets} of random training sets $D_{TR}^{rand,j}$, $j = 1, 2, ..., N_{sets}$; in addition, for each set j the procedure is repeated for different random models $M(\cdot|\omega^{rand,l}, b^{rand,l}), l = 1, 2, ..., N_{models}$. The sequence of assignment errors thereby generated, $e_{jl}, j = 1, 2, ..., N_{sets}, l = 1, 2, ..., N_{models}$, is then averaged to obtain a robust estimate for ϵ. The procedure is sketched in Fig. 2.

Notice that this method does not make any use of the original training set D_{TR} (i.e., of the training set constituted by real-world classification examples). In this view, the model-retrieval-based approach can be interpreted as a tool to obtain an absolute evaluation of the expected error that an "average" MR-Sort classification model $M(\cdot|\omega, b)$ with k categories, n criteria, and trained by means of an "average" data set of given size N_{TR} makes in the task of classifying a new generic (unknown) alternative.

5.2. The Bootstrap Method

A way to assess *both* the accuracy (i.e., the expected fraction of alternatives correctly classified) *and* the confidence of the classification model (i.e., the probability that the category assigned to a given alternative is the correct one) is by resorting to the bootstrap method,[17] which is used to create an ensemble of classification models constructed on different data sets bootstrapped from the original one:[18] the final class assignment provided by the ensemble

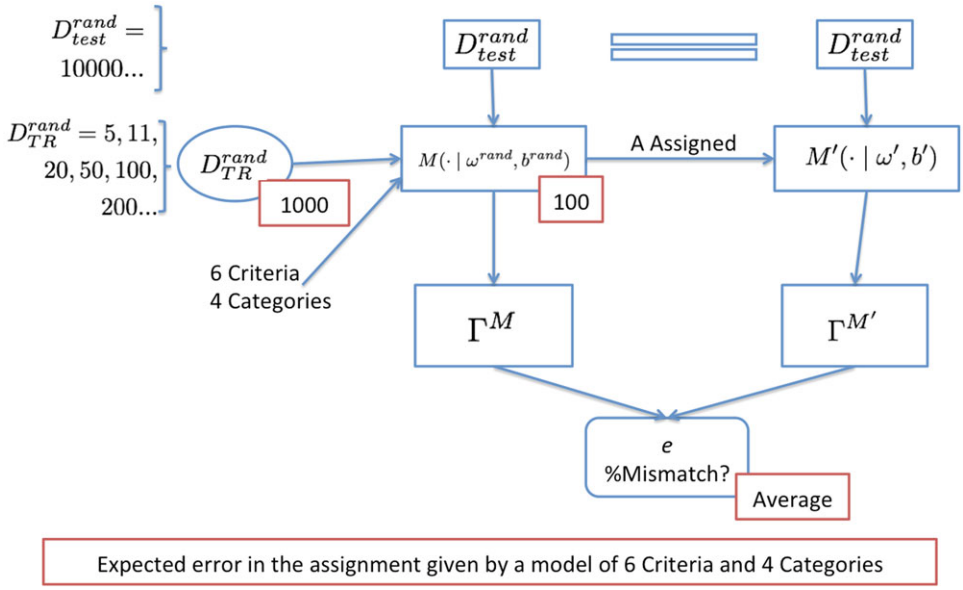

Fig. 2. The general structure of the model-retrieval approach.

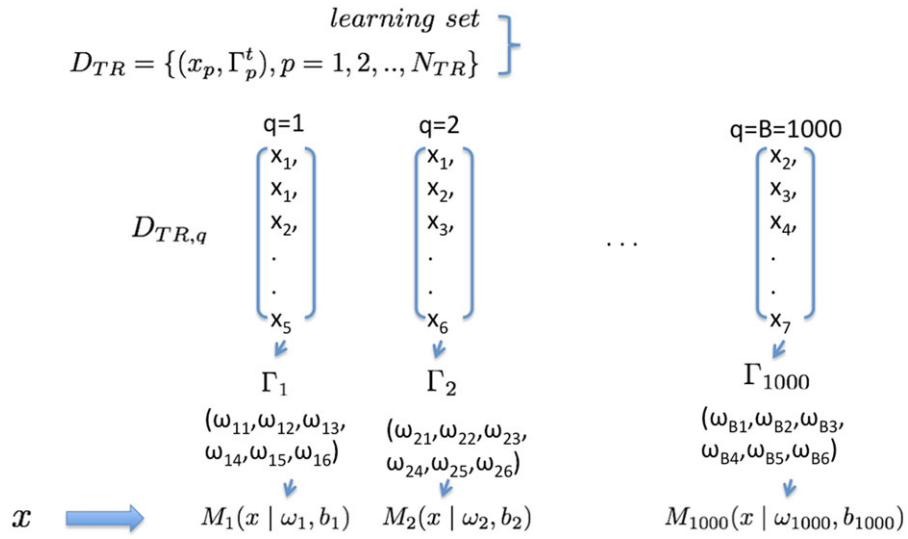

Fig. 3. The bootstrap algorithm.

is based on the combination of the individual output of classes provided by the ensemble of models.[10]

The basic idea is to generate different training data sets by random sampling with replacement from the original one:[17] such different training sets are used to build different individual classification models of the ensemble. In this way, the individual classifiers of the ensemble possibly perform well in different regions of the training space and thus they are expected to make errors on alternatives with different characteristics; these errors are balanced out in the combination, so that the performance of the ensemble of bootstrapped classification models is in general superior to that of the single classifiers.[18,19] This is a desirable property since it is a more realistic simulation of the real-life experiment from which our data set was obtained. In this article, the output classes of the single classifiers are combined by *majority voting*: the class chosen by most classifiers is the ensemble assignment. Finally, the

accuracy of the model is given by the fraction of the patterns correctly classified. The bootstrap-based empirical distribution of the assignments given by the different classification models of the ensemble is then used to measure the confidence in the classification of a given alternative x that represent the probability that this alternative is correctly assigned.[10,20]

In more detail, the main steps of the bootstrap algorithm are as follows (Fig. 3):

(1) Build an ensemble of B (typically of the order of 500–1,000) classification models $\{M_q(\cdot|(\omega_q, b_q) : q = 1, 2, ..., B)\}$ by random sampling with replacement from the original data set D_{TR} and use each of the bootstrapped models $M_q(\cdot|\omega_q, b_q)$ to assign a class $\Gamma_x^q, q = 1, 2, ..., B$, to a given alternative x of interest (notice that Γ_x^q takes a value in $A^h, h = 1, 2, ..., k$). By so doing, a bootstrap-based empirical probability distribution $P(A^h|x), h = 1, 2, ..., k$ for category A^h of alternative x is produced, which is the basis for assessing the confidence in the assignment of alternative x. In particular, repeat the following steps for $q = 1, 2, ..., B$:
   (i) Generate a bootstrap data set $D_{TR,q} = \{(x_p, \Gamma_p^t) : p = 1, 2, ..., N_{TR}\}$, by performing random sampling with replacement from the original data set $D_{TR} = \{(x_p, \Gamma_p^t) : p = 1, 2, ..., N_{TR}\}$ of N_{TR} input/output patterns. The data set $D_{TR,q}$ is thus constituted by the same number N_{TR} of input/output patterns drawn among those in D_{TR}, although due to the sampling with replacement some of the patterns in D_{TR} will appear more than once in $D_{TR,q}$, whereas some will not appear at all.
   (ii) Build a classification model $\{M_q(\cdot|\omega_q, b_q) : q = 1, 2, ..., B\}$, on the basis of the bootstrap data set $D_{TR,q} = \{(x_p, \Gamma_p^t) : p = 1, 2, ..., N_{TR}\}$.
   (iii) Use the classification model $M_q(\cdot|\omega_q, b_q)$ to provide a class $\Gamma_x^q, q = 1, 2, ..., B$ to a given alternative of interest, that is, $\Gamma_x^q = M_q(x|\omega_q, b_q)$.

(2) Combine the output classes $\Gamma^q, q = 1, 2, ..., B$ of the individual classifiers by majority voting: the class chosen by most classifiers is the ensemble assignment Γ_x^{ens}, i.e., $\Gamma_x^{ens} = argmax_{A^h}[card_q\{\Gamma_x^q = A^h\}]$.

(3) As an estimation of the confidence in the majority-voting assignment Γ_x^{ens}(step 2, above), we consider the bootstrap-based empirical probability distribution $P(A^h|x)$, $h = 1, 2, ..., k$, that is, the probability that category A^h is the correct category given that the (test) alternative is Ref. 6. The estimator of $P(A^h|x)$ here employed is: $P(A^h|x) = \frac{\sum_{q=1}^{B} I\{\Gamma_q = A^h\}}{B}$, where $I\{\Gamma_q = A^h\} = 1$, if $\Gamma_q = A^h$, and 0 otherwise.

(4) Finally, the error of classification is presented by the fraction of the number of the alternatives being assigned by the classification model and the total number of the alternatives. The accuracy of the classification model is defined as the complement to 1 to the error.

5.3. The LOOCV Technique

LOOCV is a particular case of the cross-validation method. In cross-validation, the original training set D_{TR} is divided into N partitions, $A_1, A_2, ..., A_N$, and the elements in each of the partitions are classified by a model trained by means of the elements in the remaining partitions (leave-p-out cross-validation).[20] The cross-validation error is, then, the average of the N individual error estimates. When N is equal to the number of elements N_{TR} in D_{TR}, the result is LOOCV, in which each instance $x_p, p = 1, 2, ..., N_{TR}$ is classified by all the instances in D_{TR} except for itself.[21] For each instance $x_p, p = 1, 2, ..., N_{TR}$ in D_{TR}, the classification accuracy is 1 if the element is classified correctly and 0 if it is not. Thus, the average LOOCV error (resp., accuracy) over all the N_{TR} instances in D_{TR} is ϵ/N_{TR} (resp., $1 - \epsilon/N_{TR}$), where ϵ(resp., $N_{TR} - \epsilon$) is the number of elements incorrectly (resp., correctly) classified. Thus, the accuracy in the assignment is estimated as $1 - \epsilon/N_{TR}$.

With respect to the leave-p-out cross-validation, the LOOCV produces a smaller bias of the true error rate estimator. However, the computational time increases significantly with the size of the data set available. This is the reason why the LOOCV is particularly useful in the case of small data sets. In addition, for *very sparse* data sets (e.g., of size lower than or equal to 10), we may be *forced* to use LOOCV in order to maximize the number of training examples employed and to generate training sets containing an amount of information that is sufficient and reasonable for building an empirical model.[22] In Fig. 4, the algorithm is sketched with reference to a training set D_{TR} containing $N_{TR} = 11$ data (like in the case study considered in the following section).

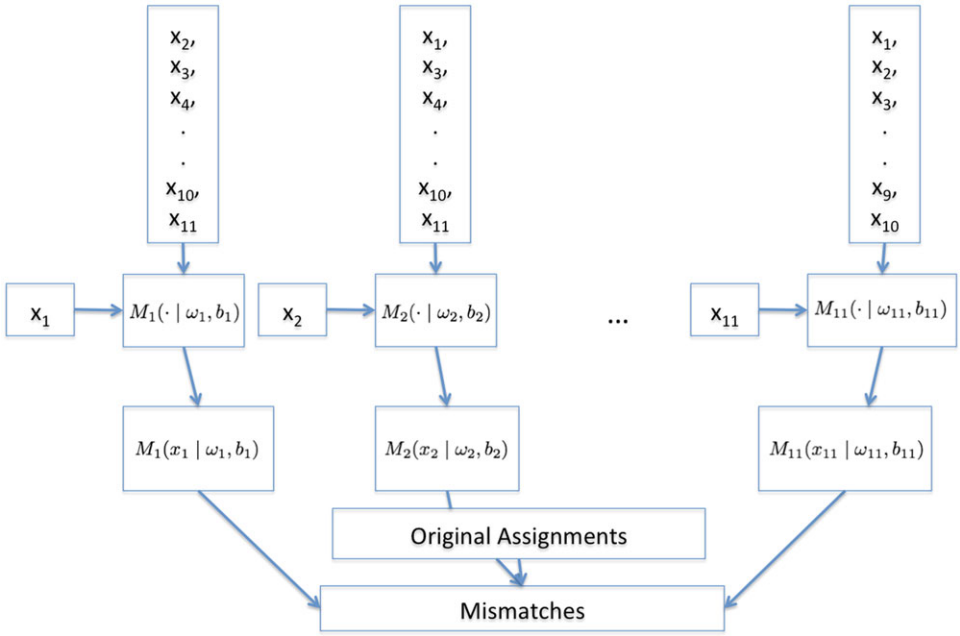

Fig. 4. Leave-one-out cross-validation study procedure.

6. APPLICATION

The methods presented in Section 5 are here applied on an exemplificative case study concerning the vulnerability analysis of NPPs.[14] We identify $n = 6$ main criteria $i = 1, 2, ..., n = 6$ by means of the hierarchical approach presented in Ref. 14 (see Section 2); x_1 = physical characteristics, x_2 = social criticality, x_3 = possibility of cascading failures, x_4 = recovery means, x_5 = human preparedness, and x_6 = level of protection. Then, $k = 4$ vulnerability categories A^h, $h = 1, 2, ..., k = 4$ are defined as: A^1 = satisfactory, A^2 = acceptable, A^3 = problematic, and A^4 = serious (Section 2). The training set D_{TR} is constituted by a group of $N_{TR} = 11$ NPPs x_p with the corresponding *a priori* known categories Γ'_p, that is, $D_{TR} = \{(x_p, \Gamma'_p) : p = 1, 2, ..., N_{TR} = 11\}$. The training set is summarized in Table I.

In what follows, the three techniques of Section 5 are applied to assess the performance of the MR-Sort classification-based vulnerability analysis model built using the training set D_{TR} of Table I.

Table I. Training Set with $N_{TR} = 11$ Assigned Alternatives

Alternatives, x_p	Vulnerability Class Γ'_p
$x_1 = \{0.61, 0.6, 0.75, 0.86, 1, 0.94\}$	A^1
$x_2 = \{0.33, 0.27, 0, 0.575, 0.4, 0.72\}$	A^3
$x_3 = \{0.55, 0.33, 0.5, 0.725, 0.7, 0.71\}$	A^2
$x_4 = \{0.55, 0.33, 0.75, 0.8, 0.7, 0.49\}$	A^3
$x_5 = \{0.39, 0.23, 0.5, 0.6, 0.6, 0.62\}$	A^3
$x_6 = \{0.39, 0.27, 0.75, 0.725, 0.7, 0.68\}$	A^2
$x_7 = \{0.61, 0.7, 0.5, 0.725, 0.9, 0.94\}$	A^2
$x_8 = \{0.16, 0.1, 0.5, 0.475, 0.3, 0.59\}$	A^4
$x_9 = \{0.1, 0, 0.25, 0.5, 0.6, 0.61\}$	A^4
$x_{10} = \{0.1, 0, 0, 0.3, 0.3, 0.43\}$	A^4
$x_{11} = \{0.61, 0.7, 0.75, 1, 1, 0.94\}$	A^1

6.1. Application of the Model-Retrieval-Based Approach

We generate $N_{sets} = 1,000$ different training sets $D_{TR}^{rand,j}$, $j = 1, 2, ..., N_{sets}$, and for each set j, we randomly generate $N_{models} = 100$ models $M(\cdot|\omega^{rand,l}, b^{rand,l})$, $l = 1, 2, ..., N_{models} = 100$. By so doing, the expected accuracy (1-ϵ) of the corresponding MR-Sort model is obtained as the average of $N_{sets} \cdot N_{models} = 1,000 \cdot 100 = 100,000$ values $(1 - \epsilon_{jl})$, $j = 1, 2, ..., N_{sets}$, $l = 1, 2, ..., N_{models}$ (see Section 5.1). The size N_{test} of the random test set D_{TR}^{rand} is $N_{test} = 10,000$. Finally, we perform the procedure of Section 5.1 for different sizes N_{TR} of the random training set D_{TR}^{rand} (even if the size of the real training set available is $N_{TR} = 11$; see Table I): in particular, we choose $N_{TR} = 5, 11, 20, 50, 100,$ and 200. This analysis serves the purpose of outlining the behavior

of the accuracy $(1-\epsilon)$ as a function of the amount of classification examples available.

The results are summarized in Fig. 5 where the average percentage assignment error ϵ is shown as a function of the size N_{TR} of the learning set (from 5 to 200). As expected, the assignment error ϵ tends to decrease when the size of the learning set N_{TR} increases: the higher the cardinality of the learning set, the higher (resp., lower) the accuracy (resp., the expected error) in the corresponding assignments. Comparing these results with those obtained by Leroy *et al.* [6] using MR-Sort models with $k=2$ and 3 categories and $n=3$–5 criteria, it can be seen that for a given size of the learning set, the error rate (resp., the accuracy) grows (resp., decreases) with the number of model parameters to be determined by the training algorithm $= n(k+1)+1$. It can be seen that for our model with $n=6$ criteria and $k=4$ categories, in order to guarantee an error rate inferior to 10% we would need training sets consisting of more than $N_{TR}=100$ alternatives. Typically, for a learning set of $N_{TR}=11$ alternatives (like that available in the present case study), the average assignment error ϵ is around 30%; correspondingly, the accuracy of the MR-Sort classification model trained with the data set D_{TR} of size $N_{TR}=11$ available in the present case is around $(1-\epsilon)=70\%$: in other words, there is a probability of 70% that a new alternative (i.e., a new NPP) is assigned to the correct category of vulnerability.

In order to assess the randomness intrinsic in the procedure used to obtain the accuracy estimate mentioned above, we have also calculated the 95% confidence intervals for the average assignment error ϵ of the models trained with $N_{TR}=11, 20,$ and 100 alternatives in the training set. The 95% confidence interval for the error associated to the models trained with 11, 20, and 100 alternatives as learning set are [25.4%, 33%], [22.2%, 29.3%], and [10%, 15.5%], respectively. For illustration purposes, Fig. 6 shows the distribution of the assignment mismatch built using the $N_{sets}\cdot N_{models}=100,000$ values ϵ_{jl}, $j=1,2,...,N_{sets}=1,000, l=1,2,...,N_{models}=100$, generated as described in Section 5.1 for the example of 11 alternatives.

6.2. Application of the Bootstrap Method

A number B (= 1,000) of bootstrapped training sets $D_{TR,q}, q=1,2,...,1,000$ of size $N_{TR}=11$ is built by random sampling with replacement from D_{TR}. The sets $D_{TR,q}$ are then used to train $B=1,000$ different classification models $\{M_1, M_2, ..., M_{1000}\}$.

Table II. Number of Patterns Classified with Confidence Value

Confidence range	(0.4, 0.5]	(0.5, 0.6]	(0.6, 0.7]
Number of patterns	1	2	0
Confidence range	(0.7, 0.8]	(0.8, 0.9]	(0.9, 1]
Number of patterns	1	2	5

This ensemble of models can be used to classify new alternatives. Fig. 7 shows the probability distributions $P(A_h|x_p), h=1,2,...,k=4, p=1,2,...,N_{TR}=11$, empirically generated by the ensemble of $B=1,000$ bootstrapped MR-Sort classification models in the task of classifying the $N_{TR}=11$ alternatives of the training set $D_{TR}=\{x_1, x_2, ..., x_{N_{TR}}\}$. The categories highlighted by the rectangles are those selected by the majority of the classifiers of the ensemble: it can be seen that the assigned classes coincide with the original categories of the alternatives of the training set (Table I), that is, the accuracy of the inferred classification model based on the given training set (with 11 assigned alternatives) is 1.

In order to investigate the confidence of the algorithm in the classification of the test patterns, the results achieved testing one specific pattern taken in turn from the training set are analyzed. For each test of a specific pattern x_i, the distribution of the assignments by the $B=1,000$ classifiers shows the confidence of the assignment of the classification model on this specific pattern. By way of example, it can be seen that alternative x_3 is assigned to Class A^2 (the correct one) with a confidence of $P(A^2|x_3)=0.81$, whereas alternative x_6 is assigned to the same class A^2, but with a confidence of only $P(A^2|x_6)=0.56$.

Notice that the most interesting information regards the confidence in the assignment of the test pattern to the class with the highest number of votes, that is, the class actually assigned by the ensemble system according to the majority voting rule adopted.[10] In this respect, Table II reports the distribution of the confidence values associated to the class to which each of the 11 alternatives has been assigned.

Thus, a $10/11 \approx 91\%$ of all class assignments with confidence bigger than 0.5 are correct.

6.3. Application of the LOOCV Method

Based on the original training set D_{TR} of size $N_{TR}=11$, we generate 11 "new" training sets $D_{TR,i}, i=1,2,...,11$ (each containing $N_{TR}-1=1-$ assigned alternatives) by taking out each time one of the alternatives from D_{TR}. These 11 training

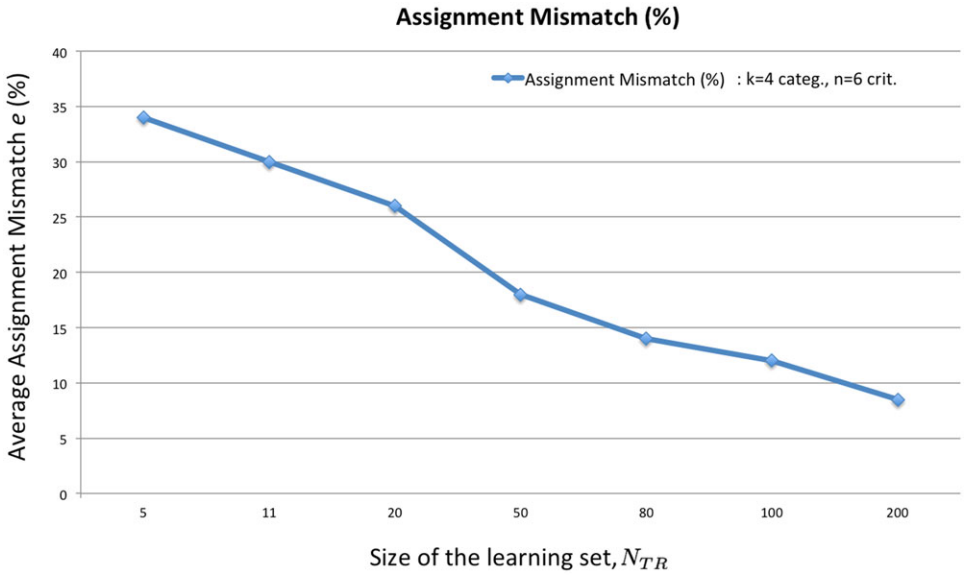

Fig. 5. Average assignment error ϵ (%) as a function of the size N_{TR} of the learning set according to the model-retrieval-based approach of Section 5.1.

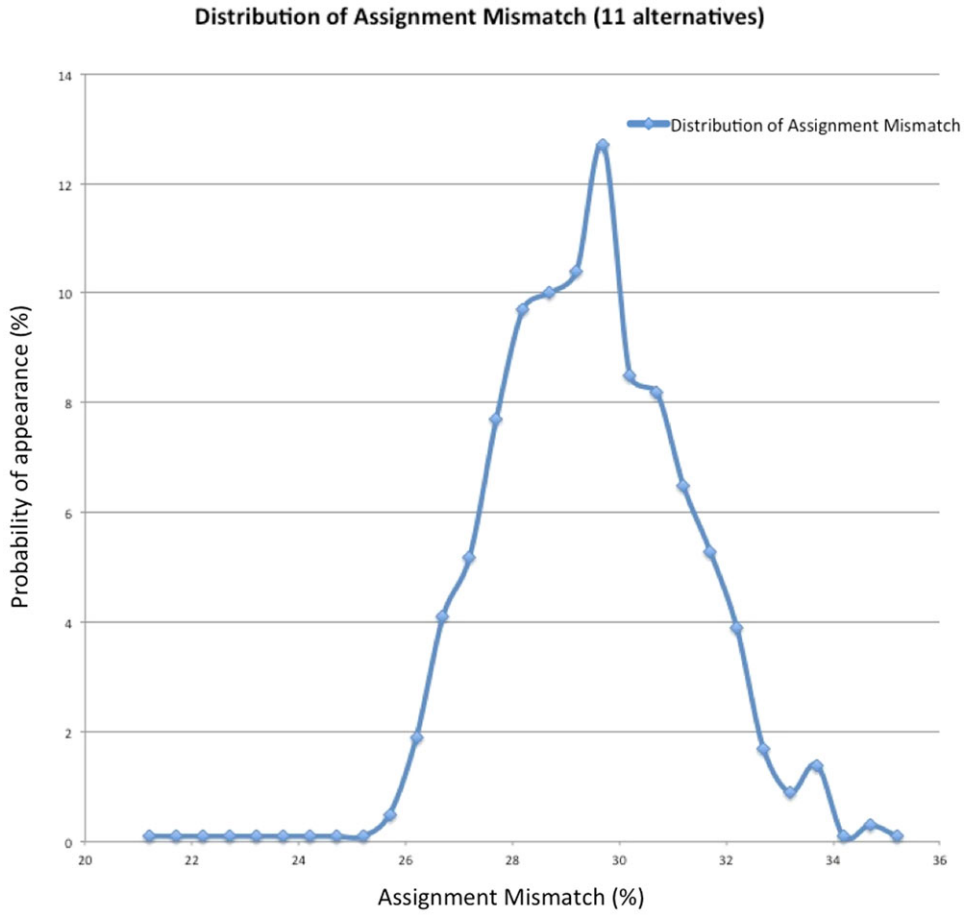

Fig. 6. Distribution of the assignment mismatch for an MR-Sort model trained with $N_{TR} = 11$ alternatives (%).

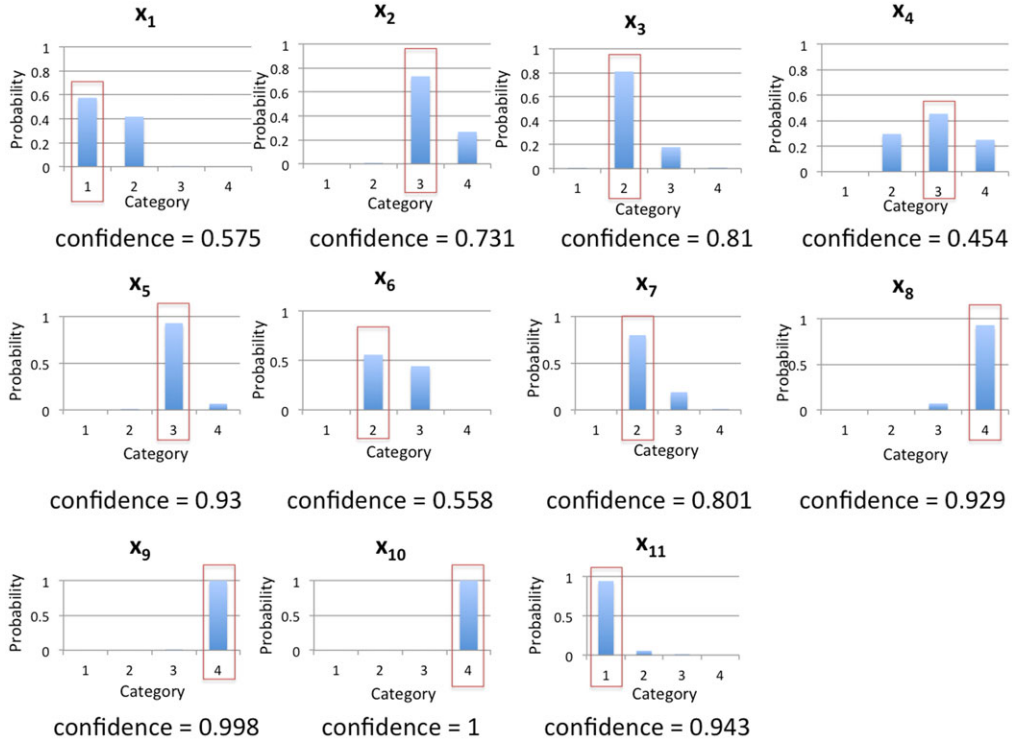

Fig. 7. Probability distributions $P(A_h|x_p)$, $h = 1, 2, ..., k = 4$, $p = 1, 2, ..., N_{TR} = 11$ obtained by the ensemble of $B = 1{,}000$ bootstrapped MR-Sort models in the classification of the alternatives x_p contained in the training set D_{TR}.

Table III. Comparison Between the Real Categories and the Assignments Provided by the LOOCV Models

Alternative	Real Categories, Γ_p^t	Assignments by LOOCV Method
x_1	1	1
x_2	3	3
x_3	2	2
x_4	3	2
x_5	3	3
x_6	2	3
x_7	2	2
x_8	4	4
x_9	4	4
x_{10}	4	4
x_{11}	1	1

sets are then used to train 11 different classification models $M_1, M_2, ..., M_{11}$. Each of these 11 models is used to classify the alternative correspondingly taken out. Table III shows the comparison between the real classes Γ_p^t of the alternatives of the training set and the categories assigned by the trained models.

It can be seen that $\epsilon = 2$ out of the $N_{TR} = 11$ alternatives are assigned incorrectly (alternatives x_4 and x_6). Thus, the accuracy in the classification is given by the complement to 1 of the average error rate, that is, $1 - \epsilon/N_{TR} = 1 - 2/11 = 1 - 0.182 = 0.818$. Notice that the 95% confidence interval for this recognition rate is [0.5901, 1].

7. DISCUSSION OF THE RESULTS

The three proposed methods provide conceptually and practically different estimates of the performance of the MR-Sort classification model.

The model-retrieval-based approach provides a quite general indication of the classification capability of a vulnerability model with given characteristics. Actually, in this approach the only constant, fixed parameters are the size N_{TR} of the training set (given by the number of real-world classification examples available), the number of criteria n, and the number of categories k (given by the analysts according to the characteristics of the systems at hand). On this basis, the space of all possible training sets of size N_{TR} and the space of all possible models with the

above-mentioned structure (n criteria and k categories) are randomly explored (again, notice that no use is made of the original real training set): the classification performance is obtained as an average over the possible random training sets (of fixed size) and random models (of fixed structure). Thus, the resulting accuracy estimate is a realistic indicator of the expected classification performance of an "average" model (of given structure) trained with an "average" training set (of given size). In the case study considered, the average assignment error (resp., accuracy) is around 30% (resp., 70%).

On the contrary, the bootstrap method uses the real training set available to build an ensemble of models compatible with the data set itself. In this case, we do not explore the space of all possible training sets as in the model-retrieval-based approach, but rather the space of all the classification models compatible with that particular training set constituted by real-world examples. In this view, the bootstrap approach serves the purpose of quantifying the uncertainty intrinsic in the particular (training) data set available when used to build a classification model of given structure (i.e., with given numbers n and k of criteria and categories, respectively). In this case study, the accuracy evaluated by the bootstrap method is much higher (equals to one) than that estimated by the model-retrieval-based approach: this is reasonable because the latter evaluates the accuracy on a wider (i.e., in a broad sense, more uncertain) space of possible models and training sets; on the other hand, in the former method the training set adopted is given and it represents possibly only one of those randomly generated within the model-retrieval-based approach. In addition, notice that differently from the model-retrieval-based approach, the bootstrap method does not provide only the global classification performance of the vulnerability model, but also the confidence that for each test pattern a class assigned by the model is the correct one: this is given in terms of the full probability distribution of the vulnerability classes for each alternative to be classified.

Finally, also the LOOCV method has been used to quantify the expected classification performance of the model trained with the particular training data set available. In order to maximally exploit the information contained in the training set D_{TR}, $N_{TR} = 1$ "reduced" (training) sets are built, each containing $N_{TR} - 1 = 10$ assigned alternatives: each "reduced" set is used to build a model whose classification performance is evaluated on the element correspondingly left out. The average error rate (resp., accuracy) turns out to be 18.2% (resp., 72.8%). The 95% confidence interval for the error rate (resp., accuracy) is approximately [0, 0.4099] (resp., [0.5901, 1]).

8. CONCLUSIONS

In this article, the issue of quantifying the vulnerability of safety-critical systems (in the example, NPPs) with respect to intentional hazards has been tackled within an empirical classification framework. To this aim an MR-Sort model has been trained by means of a small-sized set of data representing *a priori* known classification examples. The performance of the MR-Sort model has been evaluated with respect to: (i) its classification *accuracy* (resp., error), that is, the expected fraction of patterns correctly (resp., incorrectly) classified; (ii) the *confidence* associated to the classification assignments (defined as the probability that the class assigned by the model to a given [single] pattern is the correct one). The performance of the empirically constructed classification model has been assessed by resorting to three approaches: a model-retrieval-based approach, the bootstrap method, and the LOOCV technique. To the best of the authors' knowledge, it is the first time that:

- A classification-based hierarchical framework is applied for the analysis of the vulnerability of safety-critical systems to intentional hazards;
- The confidence in the assignments provided by an MR-Sort classification model is quantitatively assessed by the bootstrap method in terms of the probability that a given alternative is correctly classified.

From the results obtained it can be concluded that although the model-retrieval-based approach may be useful for providing an upper bound on the error rate of the classification model (obtained by exploring the space of all possible random models and training sets), the bootstrap method seems to be advisable for the following reasons: (i) it makes use of the training data set available from the particular case study at hand, thus characterizing the uncertainty intrinsic in it; (ii) for each alternative (i.e., safety-critical system) to be classified, it is able to assess the confidence in the classification by providing the probability that the selected vulnerability class is

the correct one. This is of paramount importance in the decision-making processes involving the vulnerability assessment of safety-critical systems, since it provides a metric for quantifying the "robustness" of a given decision.

APPENDIX A:

As described in Section 2, the hierarchical model developed in Ref. 14 is considered to analyze the vulnerability of NPPs to intentional hazards. The susceptibility to intentional hazards (first layer) is characterized in terms of attractiveness and accessibility (second layer). These are hierarchically broken down into factors that influence them, including resilience seen as preattack protection (which influences on accessibility) and postattack recovery (which influences on attractiveness); this decomposition is made in six criteria: physical characteristics, social criticality, possibility of cascading failures, recovery means, human preparedness, and level of protection (third layer). These six third-layer criteria are further decomposed into a layer of basic subcriteria, for which data and information can be collected (fourth layer) (see Table A1). The criteria of the layers are assigned preference directions for treatment in the decision-making process. The preference direction of a criterion indicates toward which state it is desirable to lead it to reduce susceptibility, that is, it is assigned from the point of view of the defender of an attack who is concerned with protecting the system. Although only the six criteria of the third level of the hierarchy are considered in the NPPs vulnerability analysis considered in this article, examples of evaluation of the basic subcriteria of the fourth layer are proposed in what follows for exemplification purposes: in particular, we describe an example of the procedure employed to calculate the numerical values of the third-layer criteria on the basis of the characteristics of the fourth-layer subcriteria.

In extreme synthesis, the subcriteria of the fourth layer can be characterized by crisp numbers or linguistic terms, depending on the nature of the subcriterion. These descriptive terms and/or values of the fourth-layer subcriteria are then scaled into numerical categories. The influence to the corresponding third-layer criterion of each of the subcriteria is analyzed.

To get the values of the six main third-layer criteria, (i) we assign arbitrary weights to each subcriterion and (ii) we apply a simple weighted sum to the categorical values of the constituent subcriteria.

A.1 Illustrative Example: Evaluation of the Criterion Physical Characteristics

The criterion "physical characteristics" is taken as an illustrative example. It is constituted by the subcriteria "number of workers," "nominal power production," and "number of production" or "service units." The description and category scales are presented as follows.

Number of Workers

This criterion can be seen to contribute to the attractiveness for an attack from various points of view, for example: (1) the more workers, the more work injuries and deaths from an attack; (2) the more workers, the easier for the attackers to sneak into the system; (3) the more workers, the higher the possibility that one of them can be turned into an attacker. Limiting the number of workers can, then, contribute to the security of the plant and, thus, reduce its attractiveness for an attack. Table A2 reports some reference values typical of NPPs.

Nominal Capacity

The higher the production capacity, the larger the potential consequences of lost production or security in case of an attack. Then, it is preferable to have a site with low capacity. Of course, for a fixed amount of total capacity needed, this would lead to its distribution on multiple sites, with an increase in the number of multiple targets, though each of them would lead to milder consequences if attacked. Table A3 shows some reference values of power generation capacity at NPP sites.

Number of Production or Service Units

Locally, within a single site, this criterion represents the number of potential attack points. Preference would go toward having a small number of targets on a site. Table A4 gives some reference values for NPPs.

We choose NPP x_1 as an example to show the calculation of the numerical value associated to the main criterion "physical characteristics" starting from the data relative to the three corresponding subcriteria (i.e., number of workers, nominal power production, and number of production or service units). The original data of the three subcriteria of x_1 are listed in Table A5.

Table A1. Criteria, Subcriteria, and Preference Directions

Criterion	Physical Characteristics	Social Criticality	Possibility of Cascading Failures
Subcriteria	Number of workers	Percentage of contribution to the welfare	Connection distance
	Nominal power production	Size of served cities	
	Number of production units		
Preference direction	Min	Min	Min
Criterion	Recovery Means	Human Preparedness	Level of Protection
Subcriteria	Number of installed backup components	Training	Physical size of the system
	Duration of backup components	Safety management	Number of accesses
	Duration of repair and recovery actions		Entrance control
	External emergency measures		Surveillance
Preference direction	Max	Max	Max

Table A2. Number of Workers

Level	Number of Workers
1	500
2	1,000
3	1,500
4	2,000
5	2,500

Table A3. Nominal Power Production

Level Level	Nominal Power Production
1	1,000 MWe
2	3,000 MWe
3	5,000 MWe
4	7,000 MWe
5	10,000 MWe

Table A4. Number of Production or Service Units

Level	Number of Production or Service Units
1	2
2	4
3	6

In scaling them onto corresponding category, we obtain the categorical value of alternative x_1 (Table A6).

Then, the numerical values of Table A6 are normalized (i.e., rescaled Between 0 and 1 based on the predefined scales) as shown in Table A7.

Table A5. Corresponding Subcriteria Original Data of Main Criterion Physical Characteristics of x_1

Alternative	Number of Workers	Nominal Power Production (MWe)	Number of Production or Service Units
x_1	600	1,000	2

Table A6. Categorical Value for the Subcriteria Corresponding to the Main Criterion "Physical Characteristics" of Nuclear Power Plant x_1

Alternative	Number of Workers	Nominal Power Production	Number of Production or Service Units
x_1	2	2	1

Table A7. Normalized Categorical Value for Corresponding Subcriteria of Main Criterion Physical Characteristics of x_1

Alternative	Number of Workers	Nominal Power Production	Number of Production or Service Units
x_1	0.4	0.4	0.33

Using the weights of these three subcriteria (arbitrarily assigned by the authors) in Table A8, we can apply a simple weighted sum to calculate the cumulative value for main criterion "physical characteristics": $0.4 \times 0.3 + 0.4 \times 0.5 + 0.33 \times 0.2 = 0.386$.

Finally, considering the preference directions of Table A1 (i.e., minimization for criterion "physical characteristics") and setting for each main criteria the value "0" as the worst case and "1" as the best

Table A8. Weights of Subcriteria for Physical Characteristics

Main Criterion: Physical Characteristics	Number of Workers	Nominal Power Production	Number of Production or Service Units
Weights	0.3	0.5	0.2

one, we convert the cumulative weighed value obtained earlier to its complement to "1," that is, $1 - 0.386 = 0.614$.

For the other five main third-layer criteria, the process of calculation is the same as for criterion "physical characteristics."

APPENDIX B: MATHEMATICAL DETAILS ABOUT THE ALGORITHM OF DISAGGREGATION OF AN MR-SORT CLASSIFICATION MODEL

We consider the case involving k categories that are, thus, separated by $(k-1)$ frontier denoted $b = \{b^1, b^2, ..., b^h, ..., b^{k-1}\}$, where $b^h = \{b_1^h, b_2^h, ..., b_i^h, ..., b_n^h, h = 1, 2, ..., k\}$, n is the number of criteria that are taken into account. Let $D_{TR} = \{(x_p, \Gamma_p^t), p = 1, 2, ..., N_{TR}\}$ be the training set, where N_{TR} is the number of alternatives, and $(A^1, A^2, ..., A^k)$ be the partition of the training set, ordered from the best to worst alternatives.

For each alternative $x_p \in D_{TR}$, in category A^h of the learning set D_{TR} (for $h = 2, 3, ..., k-1$), let us define $2n$ binary variables δ_{ip}^h and δ_{ip}^{h-1}, for $p = 1, 2, ..., N_{TR}$, such that δ_{ip}^l equals to 1 iff $g_i(x_p) \geq b_i^l$ for $l = h-1, h$ and $\delta_{ip}^h = 0 \Leftrightarrow g_i(x_p) < b_i^h$. We introduce $2n$ continuous variables $c_{ip}^l (l = h-1, h)$ constrained to be equal to ω_i if $\delta_{ip}^l = 1$ and to *0* otherwise.

We consider an objective function that describes the robustness of the assignment. We introduce two more continuous variables, y_p and z_p, for each $x_p \in D_{TR}$ and α. In maximizing α, we maximize the value of the minimal slack in the constraints.

We resume all the constraints in the following mathematical program:

$$max\ \alpha, \tag{A1}$$

$$\alpha \leq y_p, \alpha \leq z_p, \forall x_p \in D_{TR}, \tag{A2}$$

$$\sum_{i,p \in \mathbb{N}} c_{ip}^l + y_p + \epsilon = \lambda, \forall x_p \in A^{l-1}, \tag{A3}$$

$$\sum_{i,p \in \mathbb{N}} c_{ip}^l = \lambda + z_p, \forall x_p \in A^l, \tag{A4}$$

$$c_{ip}^l \leq \omega_i, \forall x_p \in D_{TR}, \forall i \in \mathbb{N}, \tag{A5}$$

$$c_{ip}^l \leq \delta_{ip}^l, \forall x_p \in D_{TR}, \forall i \in \mathbb{N}, \tag{A6}$$

$$c_{ip}^l \geq \delta_{ip}^l - 1 + \omega_i, \forall x_p \in D_{TR}, \forall i \in \mathbb{N}, \tag{A7}$$

$$M\delta_{ip}^l + \epsilon \geq g_i(x_p) - b_i^l, \forall x_p \in D_{TR}, \forall i \in \mathbb{N}, \tag{A8}$$

$$M(\delta_{ip}^l - 1) \leq g_i(x_p) - b_i^l, \forall x_p \in D_{TR}, \forall i \in \mathbb{N}, \tag{A9}$$

$$\sum_{i,p \in \mathbb{N}} \omega_i = 1, \lambda \in [0.5, 1], \tag{A10}$$

$$\omega_i \in [0, 1], \forall i \in \mathbb{N}, \tag{A11}$$

$$c_{ip}^l \in [0, 1], \delta_{ip}^l \in \{0, 1\}, \forall x_p \in D_{TR}, \forall i \in \mathbb{N}, \tag{A12}$$

$$y_p, z_p \in \mathbb{R}, \forall x_p \in D_{TR}, \tag{A13}$$

$$\alpha \in \mathbb{R}, \tag{A14}$$

M is an arbitrary large positive value, and ϵ an arbitrary small positive quantity.

The case in which x_p belongs to one of the extreme categories (A^1 and A^k) is simple. It requires the introduction of only n binary variables and n continuous variables. In fact, if x_p belongs to A^1 we just have to express that the subset of criteria on which x_p is at least as good as b_1 has sufficient weight. In a dual way, when x_p lies in A^k, the worst category, we have to express that it is at least as good as b_k on a subset of criteria that has not sufficient weight.

REFERENCES

1. Kröger W, Zio E. Vulnerable Systems. London: Springer, 2001.
2. Aven T. Foundations of Risk Analysis. NJ: Wiley, 2003.
3. Aven T. Some reflections on uncertainty analysis and management. Reliability Engineering and System Safety, 2010; 95: 195–201.
4. Aven, T. Misconceptions of Risk. Chichester, UK: Wiley, 2010.
5. Aven T, Heide B. Reliability and validity of risk analysis. Reliability Engineering and System Safety, 2009; 94:1862–1868.
6. Leroy A, Mousseau V, Pirlot M. Learning the parameters of a multiple criteria sorting method. Pp. 219–233 in Brafman RI, Roberts F, Tsoukias A (eds). The Second International Conference on Algorithmic Decision Theory, Algorithmic Decision Theory. ADT 2011, LNAI 6992. Berlin: Springer,2011.
7. Aven T, Flage R. Use of decision criteria based on expected values to support decision-making in a production assurance and safety setting. Reliability Engineering and System Safety, 2009; 94:1491–1498.

8. Milazzo MF, Aven T. An extended risk assessment approach for chemical plants applied to a study related to pipe ruptures. Reliability Engineering and System Safety, 2012; 99:183–192.
9. Rocco C, Zio E. Bootstrap-based techniques for computing confidence intervals in Monte Carlo system reliability evaluation. Pp. 303–307 in Proceedings of the Annual Reliability and Maintainability Symposium. IEEE, 2005.
10. Baraldi P, Razavi-Far R, Zio E. A Method for Estimating the Confidence in the Identification of Nuclear Transients by a Bagged Ensemble of FCM Classifiers. Las Vegas, NV:NPIC&HMIT, 2010.
11. Doumpos M, Zopounidis C, Multricriteria Decision Aid Classification Methods. Netherlands: Kluwer Academic Publishers, 2002.
12. NWRA. N. W. R. A. Risk Assessment Methods for Water Infrastructure Systems. Kingston, RI: Rhode Island Water Resources Center, University of Rhode Island, 2012.
13. Hofmann M, Kjølle G, Gjerde O. Development of indicators to monitor vulnerabilities in power systems. Presented at the 2012 International Conference on Probabilistic Safety Assessment and Management (PSAM 11) & European Safety and RELiability Conference (ESREL 2012), Helsinki, Finland, 2012.
14. Wang T-R, Mousseau V, Zio E. A hierarchical decision making framework for vulnerability analysis. Pp. 1–8 in ESREL2013, Amsterdam, The Netherlands, 2013.
15. Roy B. The outranking approach and the foundations of ELECTRE methods. Theory and Decision, 1991; 31:49–73.
16. Mousseau V, Slowinski R. Inferring an ELECTRE TRI model from assignment examples. Journal of Global Optimization, 1998; 12:157–174.
17. Efron B, Thibshirani RJ. An Introduction to the Bootstrap. Monographs on Statistics and Applied Probability, Vol. 57. New York: Chapman and Hall, 1993.
18. Zio E. A study of the bootstrap method for estimating the accuracy of artificial neural networks in predicting nuclear transient processes. IEEE Transactions on Nuclear Science, 2006; 53(3):1460–1470.
19. Cadini F, Zio E, Kopustinskas V, Urbonas R. An empirical model based bootstrapped neural networks for computing the maximum fuel cladding temperature in a RBMK-1500 nuclear reactor accident. Nuclear Engineering and Design, 2008; 238: 2165–2172.
20. Baraldi P, Razavi-FraR, Zio E. Bagged ensemble of fuzzy C means classifiers for nuclear transient identification. Annals of Nuclear Energy, Elsevier Masson, 2011; 38(5):1161–1171.
21. Wilson R, Martinez TR. Combining cross-validation and confidence to measure fitness. Pp. 1409–1416 in Proceedings of the International Joint Conference on Neural Networks (IJCNN'99). Washington, DC: IEEE.
22. Gutierrez-Osuna R. Pattern analysis for machine olfaction: A review. IEEE Sensors Journal, 2002; 2(3).