Martínez, M.; Ruiz, JC.; Antunes, N.; De-Andrés-Martínez, D.; Vieira, M. (2022). A Multi-Criteria Analysis of Benchmark Results With Expert Support for Security Tools. IEEE Transactions on Dependable and Secure Computing. 19(4):2151-2164. https://doi.org/10.1109/TDSC.2020.3048202

A Multi-criteria Analysis of Benchmark Results with Expert Support for Security Tools

Miquel Martínez, Juan-Carlos Ruiz, Nuno Antunes, David de Andrés and Marco Vieira

Abstract—The benchmarking of security tools is endeavored to determine which tools are more suitable to detect system vulnerabilities or intrusions. The analysis process is usually oversimplified by employing just a single metric out of the large set of those available. Accordingly, the decision may be biased by not considering relevant information provided by neglected metrics. This paper proposes a novel approach to take into account several metrics, different scenarios, and the advice of multiple experts. The proposal relies on experts quantifying the relative importance of each pair of metrics towards the requirements of a given scenario. Their judgments are aggregated using group decision making techniques, and pondered according to the familiarity of experts with the metrics and scenario, to compute a set of weights accounting for the relative importance of each metric. Then, weight-based multi-criteria-decision-making techniques can be used to rank the benchmarked tools. The usefulness of this approach is showed by analyzing two different sets of vulnerability and intrusion detection tools from the perspective of multiple/single metrics and different scenarios.

Index Terms—Benchmark Analysis, Security tools, Multiple-Criteria Decision Making, Decision Support

1 Introduction

Security tools have a growing importance nowadays to help developers in protecting their systems against security threats [1]. The usefulness of these tools is manifold, as they may be applied during development to recommend best coding practices, during verification and validation phases to disclose vulnerabilities, or after deployment to protect the system against security attacks [1]. Their lack of expertise usually leads development teams to trust the outputs of those tools, but research and practice show that their effectiveness is not always satisfactory [2], [3].

Benchmarking is the 'go to' technique when it comes to the assessment and comparison of tools according to some characteristic [4]. Although benchmarks have been traditionally used to compare the performance of systems [4], other benchmark approaches have also been proposed to evaluate different types of properties such as dependability [5]. *The key for the success of a benchmark is its adoption by the community*, and therefore, it is imperative that proposed benchmarks meet a set of properties and provide their users with useful insights. For this, *one of the most important points is the quality of the metrics used*.

In benchmarks that follow measurement-based approaches, metrics are computed from measurements obtained during the benchmark execution [4], [5]. Resulting values must be understood in relative terms, and they are mostly useful for comparison, improvement, and tuning. In addition to that, metrics should meet a set of properties to be useful to benchmark users [6], [7]. *A good metric should provide repeatable and reproducible results, be consistent (i.e. should not be open to subjectivity), be understandable by users, and be meaningful in the context where it is being applied* [7].

- *M. Martínez, D. de Andrés and J.-C. Ruiz are with the Instituto ITACA of the Universitat Politècnica de València, Spain.*
- *N. Antunes and M. Vieira are with the Centre for Informatics and Systems of the University of Coimbra, Portugal.*

In recent years, several works (e.g. [2], [3], [8], [9], [10]), and some of our previous publications [11], [12], have proposed different approaches for assessing and comparing security tools. These works characterize the effectiveness of tools in terms of true positives and false positives, from which general purpose metrics such as *Precision*, *Recall*, and *F-measure* [13] are derived. In most cases, tools are simply compared using one of these metrics. But, even when different sets of metrics were taken into account, only one of them was finally considered, while the remaining simply acted as tie-breaker or were completely disregarded [14]. This simplification *bias the conclusions by leaving out the information potentially provided by ignored metrics*.

Another transversal concern is that, although benchmarks may consider multiple metrics, they usually rely on the same set of metrics regardless the application scenario. Thus, since the same set of metrics may not be optimal for all the scenarios in a given application domain, in addition to considering multiple metrics, *benchmarks should also consider the relative importance of each metric for each analysis scenario*.

Finally, *metrics' relative importance within the context of each scenario must be weighed up by domain experts*. This applies even when no experts are explicitly involved in the analysis of benchmark results. In that case, benchmark users are implicitly adopting the role of experts when comparing and ranking the considered alternatives.

Finally, the high complexity underlying the simultaneous consideration of multiple metrics and goals makes this analysis very prone to biased conclusions. *The integration of several experts' judgements to weigh up the relative importance of metrics within the context of each scenario*, provides an interesting degree of diversity with a high potential to mitigate such bias. To combine elicited opinions, simple aggregation methods have often been used with the result that biases, inter-expert dependencies, and other factors that might affect experts' judgements are often ignored [15]. Although

different alternatives are available today, those based on the Analytic Hierarchy Process (AHP) have shown their usefulness in the engineering domain, not only for capturing experts' judgements [16], but also for measuring the level of uncertainty existing in final decision outcomes [17].

This paper addresses the aforementioned challenges by proposing **a new analysis approach suitable to weigh up the relative importance of benchmark metrics in different scenario while taking into consideration the opinion of experts**. This approach is called MABRES, which stands for *Multi-criteria Analysis of Benchmark Results using Expert Support*. Compared with existing analysis techniques, the key novelties of MABRES are three-fold. First, it enables several experts to participate in the analysis process, thus providing means to aggregate their individual judgements. Second, it allows the simultaneous consideration of multiple metrics, while enabling traceability from metrics to scores and vice-versa. And third, it considers the influence of application scenarios in the interpretation of metrics, thus resulting in a context-aware analysis process.

The outline of this paper is as follows. Section 2 details the context of this research. Section 3 presents MABRES and Section 4 exemplifies its usefulness by *benchmarking two different types of security tools*. Section 5 discusses the potential and limitations of this proposal. Finally, Section 6 presents the main conclusions of this work.

TABLE 1: Selected metrics for benchmarking security tools [14].

Recall (R)	$\frac{TP}{TP+FN}$
	Proportion of positive cases that are correctly classified with respect to existing vulnerabilities. Also called *true positive rate* or *sensitivity*.
Precision (P)	$\frac{TP}{TP+FP}$
	Proportion of positive cases that are correctly classified. Also known as *true positive accuracy*, *positive predictive value*, or *confidence*.
F-measure (F)	$2 \times \frac{precision \times recall}{precision+recall} = \frac{2 \times TP}{2 \times TP+FN+FP}$
	Harmonic mean of *precision* and *recall*.
Informedness (I)	$\frac{TP}{TP+FN} - \frac{FP}{FP+TN}$
	Quantifies how consistently the predictor predicts the outcome, i.e. how informed a predictor is for the specified condition versus chance.
Markedness (M)	$\frac{TP}{TP+FP} - \frac{FN}{FN+TN}$
	Quantifies how consistently the outcome has the predictor as a marker, i.e. how marked a condition is for the specified predictor versus chance.

Note: *TP*, *TN*, *FP*, and *FN* stands for *True Positives*, *True Negatives*, *False Positives*, and *False Negatives*, respectively.

2 Research Context

Benchmarks are standard tools that enable the evaluation and comparison of systems or components according to specific characteristics (e.g. performance, dependability, etc.) [4]. It is well known that the particular application domain of a benchmark influences the definition of its components. Although some benchmarks may include other components, the key ones usually are: **metrics**, **workload**, **procedure**, and **experimental setup**.

Above all, the usefulness of a benchmark is tied up with the metrics used to portray the characteristics of the system, and how they provide a useful insight according to the goals of benchmark users. However, research and practice show that *currently used approaches to analyze security metrics for computer benchmarks are not adequate* [12]. Most benchmarks use a **single metric**, which provides a limited view of results, or a **small set of metrics** but analyzed in a disjoint manner (e.g. TP and FP in [2], and TP and FN in [3]). This raises the need for ways of combining metrics to provide an aggregated view of the system's characteristics.

The metrics to be considered, the scenarios where they can be applied, and the need for a method to score alternative security tools attending to those metrics, are some of the relevant issues addressed in this section.

2.1 Multiple Metrics for Benchmarking Security Tools

Vulnerability and intrusion detection tools can be seen as binary classifiers, as they classify parts of the target application into two classes: vulnerable (positive) or non-vulnerable (negative). Several metrics are available to portray the effectiveness of binary classifiers, like information retrieval systems and machine learning algorithms [18]. Most of those metrics are computed from raw measures reported in a confusion matrix, which represents the possible outcomes for each classified instance [19]. Such outcomes are specified in terms of the amount of *true positive*, *true negative*, *false positive*, and *false negative* detections obtained for each evaluated tool.

There is a wide variety of metrics that can be derived from the aforementioned outcomes but, despite their distinct denominations in different areas, many of them are in practice synonyms. Attending to their precise meaning, our previous research [14] proposed a list of 5 representative metrics that characterizes security (vulnerability detection) tools for benchmarking purposes. This list can be found in Table 1. **Recall** determines the ratio of correctly reported positives with respect to existing vulnerabilities, whereas **Precision** indicates the proportion of reported positives that are correctly classified. **F-measure** is the harmonic mean between *Recall* and *Precision*. **Informedness** and **Markedness** were defined in [13] as a way to measure the accuracy of a predictor (a tool in this context) considering the chance of doing right predictions based on the number of vulnerabilities. **Informedness** combines *Recall* and *Specificity* to express how informed are the classifications of a tool in comparison to chance, whereas **Markedness** combines *Precision* and *Miss rate* to measure how marked the classifications of a recommender are in comparison to chance.

An important issue identified in this previous research is that, even though the suitability of selected metrics was verified after consulting with experts in the field, just a single metric was used for each considered scenario. Thus, although experts declared different levels of familiarity with the use of each proposed metric, and they had different preferences for using those metrics in the considered scenarios, this information was not taken into consideration to ponder

the relative importance provided to metrics. Our proposal will address this issue.

2.2 Considering Analysis Scenarios

Although the usefulness of a benchmark is strongly conditioned by the set of provided metrics, the relative importance of such metrics may vary attending to the goals of each specific benchmark scenario [13]. For instance, when selecting third party cores to be integrated into a hardware system [20], these cores are usually analyzed from the four-fold perspective of their performance, silicon area, power consumption, and robustness. Thus, a suitable core for a mobile system will offer a good balance between performance and power consumption, whereas an eligible core for an automotive system must exhibit a good level of both robustness and performance. This example shows that considered benchmark scenarios are key, not only for the selection of relevant metrics, but also for the adequate analysis of such metrics.

This contribution considers an *analysis scenario* as an scenario defining a set of requirements that conditions the interpretation of benchmark metrics. This not only refers to the set of selected metrics but also to their relative importance. Hence, benchmark rankings may vary from one analysis scenario to another even when using the same set of metrics and the judgement of the same set of experts. In contrast, a context-less analysis of benchmark results is useless in most cases, except when considering only one metric or having a clear winner for all selected metrics.

Among the few efforts deployed to face this problem in the security domain, the work presented in [14] identified four common benchmark scenarios for the use and analysis of vulnerability detection tools. These scenarios are named as i) *Business-critical* (BC), ii) *Heightened-critical* (HC), iii) *Best effort* (BE), and iv) *Minimum effort* (ME). In the *Business-critical* scenario, the best security tools are those detecting the highest number of vulnerabilities, while missing the lowest possible number. In the *Heightened-critical* scenario, detecting the highest number of vulnerabilities is also important, but it cannot be done at any cost, so it is necessary to avoid tools reporting too many false positives. Then, the *Best effort* scenario focuses on tools providing a good balance between high-level detection and false positives. Finally, in the *Minimum effort* scenario, the goal is to look for tools reporting the lowest number of false positives with a high confidence in reported vulnerabilities. Further information on these scenarios, including examples of the types of systems included in each one, is provided in Table 2.

Despite the utmost importance of considering the aforementioned scenarios, the reader must understand that their definition falls out of the scope of this publication. The contribution of this paper focuses on to what extent and how experts can take into consideration the specific requirements imposed by each scenario to establish, for each one, the relative importance of considered metrics.

Before concluding this subsection, it is worth mentioning that, even if several metrics for benchmarking vulnerability detection tools were proposed in [14], only one of them was considered for ranking purposes for each scenario, with a second metric nominated as a tie-breaker (see Table 3). The discussion here should not be directed towards the suitability and relevance of the proposed metrics for the considered scenarios but, again, towards the fact that final decisions are based on just one single metric, thus neglecting the importance of other available (and representative) metrics.

TABLE 2: Scenarios for the use of security tools [14].

Scenario	Requirements
Business-critical (BC)	The BC scenario represents systems with highly demanding security requirements, where the exploitation of a vulnerability can lead to economical or reputation losses. These are systems such as home banking, stock trading, or large-scale e-commerce. The development of this kind of systems is assumed to have enough resources to deal with all reported vulnerabilities, even if they are wrongly classified (false positives). Thus, the goal is to select a tool able to detect the highest number of vulnerabilities, leaving undetected the lowest possible number.
Heightened-critical (HC)	The HC scenario represents those systems where applications are subjected to high security requirements (but not as high as those running in BC scenarios). This could be the case of applications dealing with sensitive data, like governmental portals or large scale social networks. Here, the aim is to detect the highest number of vulnerabilities, but unlike BC, it cannot be done at any cost, so it is necessary to avoid tools reporting too many false positives.
Best effort (BE)	The BE scenario represents applications that are less exposed to attacks or are not so critical. As the resources available to fix reported vulnerabilities are limited, time or budget constraints must also be considered. Examples of these systems include big web portals where attacks represent a small direct financial loss or intranet applications that are less exposed to external attacks. Here, the goal is to look for tools able to detect a high number of vulnerabilities while reporting a low number of false positives.
Minimum effort (ME)	The ME scenario represents applications with low resources and not much criticality concerns, that might not be subjected to a lot of external attacks. Due to budget reasons, the time and money available to fix vulnerabilities are usually tight. Hence, tools reporting the lowest number of false positives with a high confidence in the reported vulnerabilities are desired for this scenario. This would be the case of content management systems for small and medium companies, and information/advertising web sites.

TABLE 3: Recommended metrics for each scenario [14].

Scenario	Main Metric	Tiebreaker
Business-critical (BC)	Recall	Precision
Heightened-critical (HC)	Informedness	Recall
Best effort (BE)	F-measure	Recall
Minimum effort (ME)	Markedness	Precision

2.3 Do We Really Need a New Analysis Approach?

The answer to the question may be *no, we don't* if we accept that i) security tools can be properly ranked and selected with just a fraction of all available metrics, and ii) the advice of experts is not relevant when analyzing such metrics.

However, accepting these assumptions attempts against the principles of accuracy and confidence that must be expected from any benchmark. As motivated so far, this is a multi-criteria decision making problem (selection of the best alternative using multiple and often conflicting criteria) that requires not only context-awareness (considering the influence of application scenarios in the analysis process), but

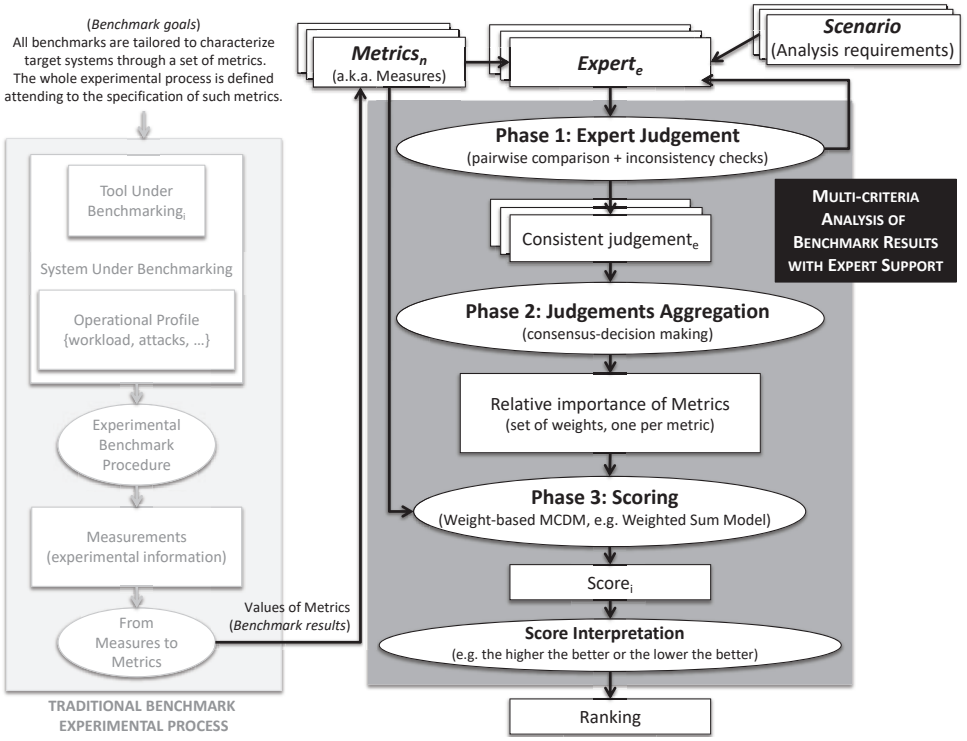


Fig. 1: The MABRES approach: a Multi-criteria Analysis of Benchmark Results with Expert Support.

also a non-negligible level of domain-expert support (due to the technical characteristics and the level of criticality of the targeted tools). To the best of our knowledge, no analysis approach has been proposed so far to address all these requirements when benchmarking security tools.

3 A Multi-criteria Analysis Approach with Expert Support

This section provides a high level overview of the proposed *Multi-criteria Analysis of Benchmark Results with Expert Support* (MABRES) methodology and its constituent phases.

3.1 Approach Overview

According to the traditional benchmarking process depicted in Figure 1, benchmark targets (security tools, in our case) are instantiated attending to an operational profile that includes, among other things, a workload and an attack/vulnerability-load. Measurements retrieved from experimentation are carefully treated to deduce the finally reported metrics. However, benchmarks seldom specify the analysis process, so it is left in the benchmark users' hands. In other words, benchmark users assume the role of experts in the analysis of benchmark metrics.

MABRES aims at complementing that benchmarking process by establishing a systematic procedure in the analysis of benchmark results to ease its use, understanding, and final explanation. At the same time, it also guarantees the traceability from metrics to scores and vice-versa during the whole analysis process.

The analysis approach supported by MABRES, detailed in Figure 1, runs just after the traditional benchmarking process. It relies on the existence of i) a set of metrics that characterizes each targeted security tool (multi-criteria component), and ii) the specification of one or several benchmarking scenarios (context-awareness component)[1]. MABRES also enables domain experts (expert support component) to participate in the analysis process. As previously pointed out, even when no expert is explicitly involved in the analysis process, it is assumed that the benchmark user implicitly becomes an expert when ranking alternatives.

The proposed methodology, supported by MABRES, works in three successive phases:

1) First, experts compare available metrics in pairs, attending to the analysis requirements imposed by considered scenarios. This is the *Expert Judgement* phase, which results in a pairwise comparison matrix per expert and scenario. Each one of these matrices is automatically processed to detect inconsistencies in comparisons. As a result, inconsistent matrices can be reviewed or discarded, whereas consistent ones are processed to obtain a set of vectors capturing experts' judgements.
2) The second phase, the so-called *Judgements Aggregation* phase, establishes a consensus among all individual judgements provided by experts. This consensus is expressed as an automatically computed vector of weights. They reflect the relative importance globally provided by experts to metrics in each scenario.
3) The third phase, named the *Scoring* phase, relies on the use of multi-criteria decision making (MCDM)

[1]The terms *benchmarking scenario*, *analysis scenario* or simply *scenario* will be indistinctly used from now on in this paper to refer to the set of requirements conditioning the analysis of benchmark metrics.

techniques to compute a final score for each benchmarked alternative. Any weight-based MCDM can be used for this purpose but, to keep things as simple as possible, our recommendation is to follow the well-known and widely used Weighted Sum Model method. With this method, each metric is treated as a different selection criterion and its influence in the final score is pondered by its attributed weight according to the consensus among experts.

The final result of these three phases is a single score that must be subsequently interpreted following a pre-established criterion, such as the-higher-the-better or the-lower-the-better. This is how each evaluated alternative is integrated into the final ranking.

3.2 Phase 1: Individual Expert Judgement

In this initial phase, experts judge the relative importance of each pair of metrics, which should be highly conditioned by the considered scenario. For example, let us consider two metrics in the field of networking such as *throughput* (amount of information transmitted per second) and *integrity* (percentage of packets received without modifications). In a scenario where users exchange large files with public data over a network, *throughput* might be more important than *integrity*, since information may be exchanged very fast and corrupted packets could be requested again. However, if users were exchanging small files containing private data, the *integrity* of the packets would be more important. Thus, experts need to have a good insight of the *specification of the scenario* to make an informed decision on to what extent some metrics are more important than others.

Human subjectivity makes it difficult for a person to be accurate when simultaneously comparing more than 2 elements. However, humans can easily and reliably compare pairs. Indeed, it is well-known that pairwise comparison is less error-prone than considering all metrics at the same time, and it can be easily (re)checked in case of finding inconsistencies among comparisons.

The pairwise comparison method enables experts to use quantitative values to express qualitative decisions and thus weigh the relative importance of metrics. It is part of the Analytic Hierarchy Process (AHP) [21], a renown decision-making framework developed by mathematicians in the 80s and used today in many different application domains, ranging from business to engineering [22], [23], [24].

In practice, experts use a 1 to 9 scale, known as the *Fundamental Scale of Absolute Numbers* (see Table 4), to translate their qualitative decisions into quantitative values. With the assistance of this scale, experts compare metrics two-by-two and their judgements are used to fill a *pairwise comparison matrix* from which the requirements are calculated. The comparison of n metrics leads to the definition of a $n \times n$ matrix, as shown in Equation 1. Since the intensity of the importance of a metric M_i with respect to another metric M_j is represented by x_{ij}, the opposite intensity is $x_{ji} = 1/x_{ij}$, which makes the matrix reciprocal. Hence, $\forall i, j \in \{1, \ldots, n\}, x_{ij} \times x_{ji} = 1$.

TABLE 4: The fundamental scale of absolute numbers.

Definition	Description	Intensity[a]
Equal	A and B are equally important	1
Moderate	A is somewhat more important than B	3
Strong	A is much more important than B	5
Very strong	A is very much more important than B	7
Extreme	A is absolutely more important than B	9

[a] Intensities of 2, 4, 6, and 8 can be used to express intermediate values. Very close importance can be represented with 1.1–1.9.

$$\begin{array}{c} \\ M_1 \\ M_2 \\ \vdots \\ M_n \end{array} \begin{array}{c} \begin{array}{cccc} M_1 & M_2 & \cdots & M_n \end{array} \\ \begin{pmatrix} 1 & x_{12} & \cdots & x_{1n} \\ x_{21} & 1 & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & 1 \end{pmatrix} \end{array} \tag{1}$$

Figure 2 provides a concrete example of an actual pairwise comparison matrix for metrics A, B and C. The expert considers that B is moderately more important than A, B is very much more important than C, and A is much more important than C. Since the matrix is reciprocal, it can be filled with just these 3 comparisons.

Once available, this pairwise comparison matrix must be processed to determine the weights that represent the relative importance assigned by the expert to each metric. There are two main methods that can be used to compute this *priority vector*: the *eigenvalue* method [25] and the *row geometric mean* (RGM) method [26]. The work done in [27], [28] shows that the difference between their outputs is negligible. Since the RGM method requires less computational power it is the method proposed in this paper.

Equation 2 shows the procedure followed by the RGM method to compute the priority vector. It consists of three successive steps: *i)* compute the geometric mean for each row of the pairwise comparison matrix, *ii)* sum up all computed geometric means, and *iii)* divide each geometric mean by the resulting sum. The result is a priority vector $w = (w_1, \cdots, w_n)$ containing n different weights (one per metric) so that $\forall i \in \{1, \ldots, n\}, w_i \geq 0$ and $\sum_{j=1}^{n} w_j = 1$.

$$w_i = \frac{\sqrt[1/n]{\prod_{j=1}^{n} x_{ij}}}{\sum_{k=1}^{n} \left(\sqrt[1/n]{\prod_{j=1}^{n} x_{kj}} \right)} \tag{2}$$

This process is depicted in Figure 2. The resulting priority vector should be understood as the expert declaring that her relative preference for A, B, and C is 27.9%, 64.9%, and 7.2%, respectively.

The main problem with pairwise comparison matrices is that humans are involved in their definition and, consequently, they may contain inconsistencies due to (subjective) interpretation. The *consistency ratio* (CR) is a statistically reliable estimate to quantify the consistency of the resulting priority vector [29]. As Figure 2 shows, the CR is computed in three successive steps. First the Principal Eigen Vector (PEV) is calculated by multiplying the sum of the various columns of the pairwise comparison matrix ($1 \times n$ matrix) and the weights contained in the priority vector ($n \times 1$ matrix). Then, a consistency index (CI) is deduced attending to the PEV and the number of metrics under study (n). Finally, the CR can be obtained by normalizing the CI to the random

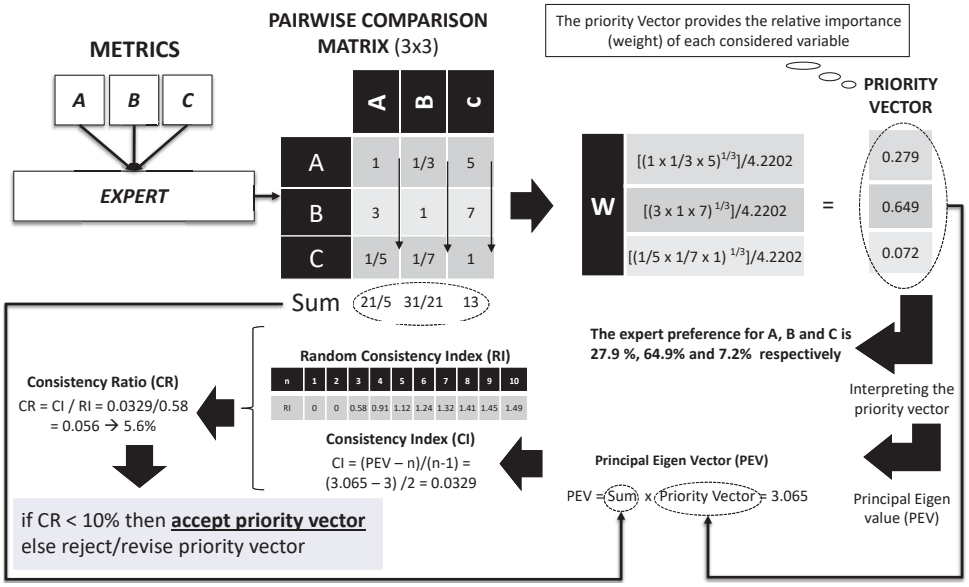


Fig. 2: Capture and automatic processing of the individual judgements of an expert.

consistency index (RI) that is directly obtained from a table defined in [30]. A $CR < 0.1$ denotes that the intensities representing the relative importance between elements of the matrix are consistent. Inconsistent matrices can be either rejected or reviewed until they become consistent. Figure 2 shows this process in action. The finally computed value for CR is 0.056 , so the priority vector is accepted.

The formal justification of this whole process falls beyond our purpose, although interested readers may refer to [31], [30] for further details. The important thing is that this process is representative for the community, since it is largely adopted by the academia and the industry [25], and it can be fully automated, which eases its use.

3.3 Phase 2: Aggregation of Individual Judgements

At this point in the methodology, e experts have determined the relative importance of n metrics considering the application scenario for the target security tools, thus providing a set of e judgements (one per expert).

The aggregation of individual judgements can be carried out in multiple ways using, for instance, consensus decision-making methods or voting theories. Nevertheless, it must be clear that we are not looking for a winner or loser expert. The goal of this proposal is to reach an agreement that accounts for the individual contributions of all experts. This does not mean that all the judgements should be treated equally. As mentioned in Section II, the relevance of each judgement should take into account the familiarity of each expert with considered metrics and/or analysis scenarios.

There are two main methods that have proven to be useful in group decision making when decisions are expressed as priority vectors: the *aggregation of individual judgements* (AIJ) and the *aggregation of individual priorities* (AIP) [32]. Despite their differences, both methods lead to the same set of group priorities, i.e. the same consensus, when the priority vector is computed using the RGM method [33], [34] (as proposed in Section 3.2).

According to this, the AIJ method seems to fit best MABRES, as it promotes a judgement aggregation approach that reuses existing pairwise comparison matrices to produce a new one called the *group comparison matrix* (GCM). Every element of a GCM is the result of computing the *weighted geometric mean* (WGM) of the elements located at the very same position in all the pairwise comparison matrices provided by experts. The required weights must be obtained attending to the familiarity of each expert with each metric and analysis scenario.

This proposal involves directly asking experts about their familiarity and then recursively aggregating the provided answers using the RGM, as shown in Figure 3. First, the familiarity reported by experts is managed separately to obtain a weight for metrics and another for the scenario. Then, these two weights are aggregated together again using the RGM. The obtained aggregated weights reflect the relative importance of experts' judgements to contribute to the GCM, and they are denoted by $\omega = \{\omega_1, \cdots, \omega_e\}$, where $\forall i \in \{1, \ldots, e\}, \omega_i \geq 0$ and $\sum_{i=1}^{e} \omega_i = 1$.

Using these weights, the GCM can be computed by Equation 3. Here, x_{ij}^k denotes the value in the position (i, j) of the pairwise comparison matrix defined by an expert e_k. So, each value (x_{ij}^G) of the GCM is obtained after applying the WGM to the element x_{ij} of all pairwise comparison matrices defined by all experts. Figure 3 depicts a simple example, where the GCM is calculated from the pairwise comparison matrices of two experts and their relative familiarity to the metrics and the considered scenario.

$$x_{ij}^G = \prod_{k=1}^{e} \left(x_{ij}^k\right)^{\omega^k} \tag{3}$$

The group comparison matrix is, in essence, a pairwise comparison matrix. So, the RGM can be applied to deduce the associated priority vector (see Figure 3). If the GCM is consistent, then the resulting priority vector, let us call it the *consensus priority vector*, can be accepted.

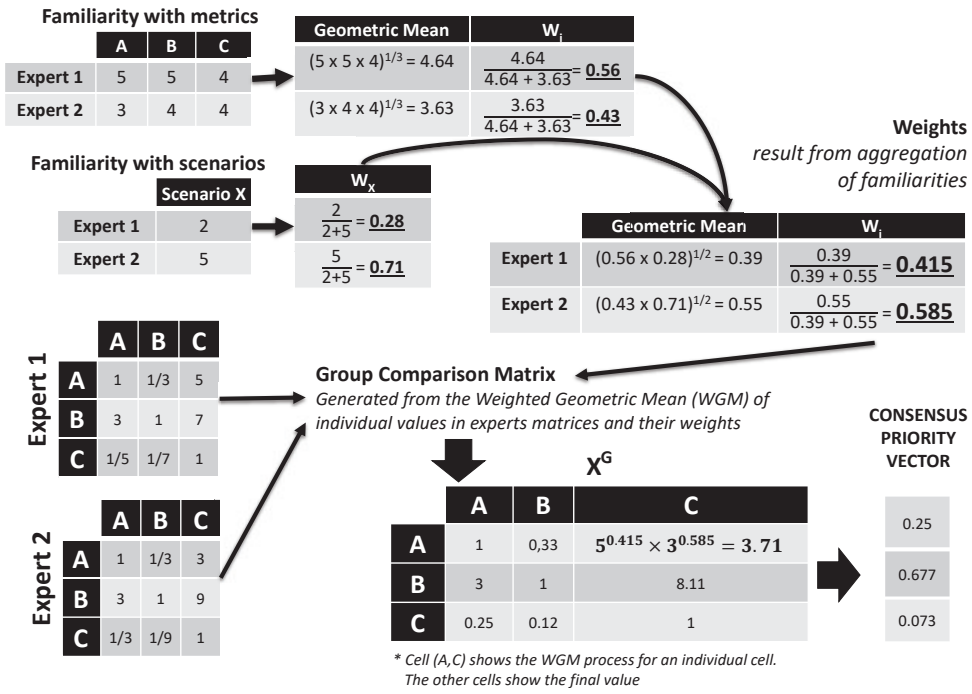


Fig. 3: Weighting the individual judgments of experts.

The advantage of this proposal is that it scales up with the number of considered metrics, scenarios, and experts, and at the same time it let us adjust experts' judgements according to their level of expertise.

3.4 Phase 3: Scoring

The inputs to this phase are the computed *consensus priority vector* and the set of values obtained by each alternative under evaluation for the n considered metrics. The result is the final score that will be later interpreted to rank all benchmark targets.

This phase may be the most controversial in the whole approach, since one of the most challenging issues when aggregating metrics is to properly capture in a single score information of the benchmark target [35]. The goal is not only to compute a single score, but rather to use the most simple, easy to use, and understandable method. At the same time, this process must be traceable to clearly identify how metrics are transformed into scores and what is the precise contribution of each metric. These are mandatory requirements to keep the analysis sound and representative to the community of potential benchmark users.

Although there exist several methodologies to cope with these requirements, they all present some drawbacks. For instance, the mathematical addition cannot be directly applied to all metrics, central tendency methods often hide underlying distributions, wealth inequality and distribution fitting techniques are hard to interpret and their results are usually difficult to trace back to the original metrics, and custom formulae are hard to validate [36]. We propose to score benchmarked alternatives using the Weighted Sum Model attending to the following features:

- First, it is the best known and simplest multi-criteria decision making (MCDM) method for comparing and ranking a number of alternatives in terms of a complex set of criteria (metrics and their relative importance).
- Second, it adapts well to the considered metrics (see Table 1) since i) they are expressed in the same units, and ii) they are interpreted following the same benefit criteria, i.e. the-higher-the-better. It is worth mentioning that the values for all metrics range between 0 and 1, except for *Informedness* and *Markedness* that range between -1 and 1. This problem can be easily solved by rescaling these two metrics between 0 and 1 by adding 1 and dividing by 2.
- Third, scores can be directly carried out using the set of inputs already available at this phase.
- Fourth, these scores are very easy to interpret and also follow a benefit criteria.
- And last, but not least, benchmark users have been using it for years, so it is meaningful for everyone in the benchmarking domain.

The mathematical notation of the scoring process for an alternative A_i in an scenario S_x is shown in Equation 4. Here, wc_k denotes the priority calculated in the consensus priority vector for the k^{th} metric, and m_k refers to the value that alternative A_i has obtained in the k^{th} metric.

$$Score(A_i, S_x) = \sum_{k=1}^{n} (m_k \times wc_k) \tag{4}$$

By following this process, one score is finally attributed to each alternative in each scenario. Accordingly, they can be used to rank the evaluated alternatives for each scenario to support the decision making process.

TABLE 5: Benchmarked vulnerability detection tools.

Label	Tool	Technique	Provider
VDT01	RAD-WS$_l$[a]	Anomaly detection + penetration testing	Univ. Coimbra
VDT02	FindBugs	Static code analysis	Univ. Maryland
VDT03	Yasca	Static code analysis	SourceForge
VDT04	JetBrains	Static code analysis	IntelliJ IDEA
VDT05	RAD-WS$_t$[b]	Anomaly detection + penetration testing	Univ. Coimbra
VDT06	WebInspect	Penetration testing	HP
VDT07	Rational AppScan	Penetration testing	IBM
VDT08	Web Vuln. Scanner	Penetration testing	Acunetix
VDT09	IPT-WS	Penetration testing	Univ. Coimbra
VDT10	Sign-WS	Penetration testing	Univ. Coimbra

[a] Results are reported in terms of lines.
[b] Results are reported in terms of tested inputs.

TABLE 6: Benchmarked IDS tools.

Label	Tool	Configuration	Monitoring level	Detection technique
IDT1	ACD[a] [37]	ACD1	Application	Anomaly-based
IDT2		ACD3		
IDT3		ACD10		
IDT4		ACD30		
IDT5		ACD100		
IDT6	Green SQL [38]	v1.2.2	Database	Signature-based
IDT7	Apache Scalp [39]	Scalp sqlia configured for SQL injection	Application	Signature-based
IDT8		Scalp xss configured for Cross Site Scripting		
IDT9	Snort [40]	v2.8 using a set of custom rules	Network	Anomaly-based
IDT10	DB IDS [41]	DB IDS	Database	Anomaly-based
IDT11	ModSecurity [42]	ModSec	Application	Signature-based

[a] ACD stands for Anomalous Character Distribution, a tool with a configurable threshold defining the minimum deviation from the legitimate profile distribution to be considered malicious. ACD30, for instance, indicates an instantiation of the tool with a threshold of 30.

4 Case Studies

This section illustrates how the proposed analysis approach can be applied to two different sets of security tools. The first set consists of 10 vulnerability detection tools, labelled from *VDT01* to *VDT10* in Table 5. Results obtained after benchmarking these tools are fairly homogeneous, as the same 5 tools consistently get the best scores for each considered metric with a large difference with respect to their competitors. The second set consists of 11 intrusion detection system (IDS) tools for SQL injection attacks in web applications, labelled from *IDT01* to *IDT11* in Table 6. Results from benchmarking IDS tools present a higher variability of best and worst option depending on the considered metric. Interested readers can find further details about the considered vulnerability detection tools in [12] and about the IDS tools in the references included in Table 6.

As already mentioned in Section 2, MABRES can consider any metrics and scenarios, but their definition falls out of the scope of this contribution. MABRES focuses on reaching a consensus among experts on how to aggregate considered metrics to obtain a score that quantifies the goodness of security tools for a given scenario. This is why MABRES will analyze both case studies through a set of *state-of-the-art* metrics (see Table 1) and analysis scenarios (see Table 2) that have been already discussed by the community and published in related conferences and journals.

Finally, remember that *Informedness* and *Markedness* are rescaled to make them compatible with the rest of metrics during the scoring phase, as explained in Section 3.4.

4.1 Capturing Experts' Priorities

A total of 34 researchers with previous work and publications related to the use of metrics derived from binary classifiers (like the one from the University of British Columbia), security (like the one from the University of Maryland), and/or benchmarking (like those from the University of Florence), were asked to participate in this case study as experts in the domain. This approach was inspired by [43], where experts were selected among professionals with expertise in one, but not necessarily in all, the dimensions of the problem under study to reduce the intrinsic uncertainty existing in multi-criteria decision making problems.

The breakdown of experts into their country and affiliation is listed in Table 7. As they were selected among our main contacts in the domain, they are primarily located in Europe (79%), North America (12%) and Latin America (9%). Although the selection of experts may seem to be biased, this is not a great concern for this case study. Our main goal is to show the usefulness and feasibility of the proposed approach to obtain a score that quantifies the goodness of security tools for a given scenario by integrating the expertise of a set of experts and the information provided by a whole set of metrics, rather than providing a conclusive formula for computing that score. This is not a great concern in our case, as the judgement of new experts can be seamlessly integrated into the resulting consensus priority vector as soon as they are available.

As defined in Section 3.2, experts must compare all metrics in pairs to determine their relative importance for each given scenario. As 5 different metrics are considered in this case study (*Precision*, *Recall*, *F-measure*, *Informedness*, *Markedness*), a total of 10 comparisons must be carried out for each of the 4 considered scenarios (*Business-critical*, *Heightened-critical*, *Best effort*, *Minimum effort*), for a total of 40 comparisons.

TABLE 7: Invited experts by country and affiliation.

Country	Affiliation	Invited	Accepted
Portugal	University of Coimbra	6	4
	University of Lisbon	3	2
Italy	Federico II University of Naples	3	2
	University of Florence	2	2
	University of Padua	1	1
UK	City University of London	2	1
	Newcastle University	1	0
	University of Kent	1	0
Brazil	University of Campinas	2	2
	Universidade Federal de Alagoas	1	1
Spain	Universitat Politècnica de València	3	3
USA	University of Maryland	1	1
	Netflix	1	0
	IBM Research	1	0
Germany	ERNW GmbH	1	1
	University of Würzburg	1	0
Canada	University of British Columbia	1	1
Hungary	Budapest University of Technology and Economics	1	0
France	Laboratory for Analysis and Architecture of Systems	1	0
Belgium	KU LEUVEN	1	0
	Total	34	21

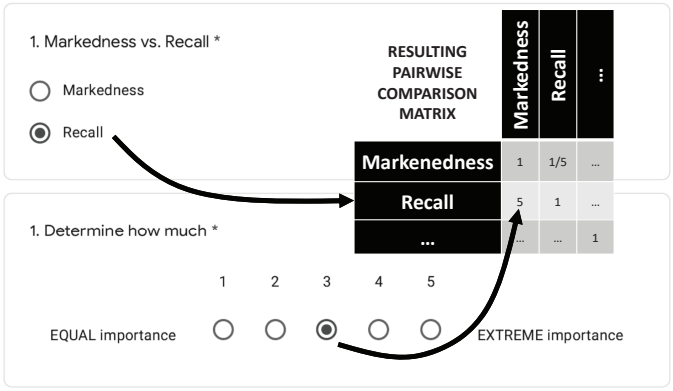


Fig. 4: Example of how answers provided by experts directly map to pairwise comparison matrices.

TABLE 8: Consistency Ratio (CR) of pairwise comparison matrices for each considered scenario and expert.

	Consistency Ratio (CR)			
Experts	*BC*	*HC*	*BE*	*ME*
E01	0.29	0.17	0.11	0.09
E02	0.25	0.29	0.24	1.26
E03	0.14	0.15	0.10	0.10
E04	0.21	0.37	0.68	0.29
E05	0.13	0.10	0.03	0.14
E06	0.06	0.07	0.01	0.07
E07	0.05	0.04	0.02	0.03
E08	0.14	0.20	0.14	0.13
E09	0.06	0.03	0.00	0.02
E10	0.03	0.51	0.18	0.04
E11	0.12	0.31	0.25	0.04
E12	0.30	0.07	0.02	0.20
E13	0.24	0.22	0.14	0.72
E14	0.11	0.18	0.08	0.35
E15	0.19	0.00	0.12	0.29
E16	0.80	0.28	0.15	0.09
E17	0.06	0.06	0.06	0.06
E18	0.23	0.09	0.16	0.12
E19	0.23	0.27	0.00	0.36
E20	0.10	0.06	0.05	0.09
E21	0.42	0.08	0.25	0.16
High consistency	5	9	9	9
Tolerable consistency	7	4	8	5

To reduce the burden of fulfilling this task, an online questionnaire[2] was prepared to support experts in performing their work. It must be noted that this is not an ad-hoc questionnaire that should be statistically processed to extract information from experts' answers, but a tool that guides and simplifies the work of completing the required pairwise comparisons. Two questions, as depicted in Figure 4, were defined for each pairwise comparison: i) which is the preferred metric between the two presented, and ii) which is the intensity of this preference (following the scale presented in Table 4). The example displayed in Figure 4 shows that the expert rates *Recall* as *much more important* than *Markedness* for a given scenario. In such a way, all the required information can be easily collected and processed.

As Table 7 shows, 21 out of 34 (62%) researchers accepted our invitation to collaborate in this work and completed the online questionnaire. The Consistency Ratio (CR) of resulting comparisons matrices was computed to determine whether provided answers were consistent, and could be used in this case study, or they should be rejected or rebuilt. Table 8 lists the computed CR for each expert and scenario. A $CR < 0.10$ denotes a highly consistent matrix but, following the experience reported in [44], matrices with a $CR < 0.20$ present a tolerable consistency and can also be accepted. Despite the simplicity of the approach and the guidelines included in the questionnaire, 7 experts provided up to 3 or 4 inconsistent matrices. This reveals one of the interests of the proposal, since it shows that inconsistencies in the subjective judgement of experts can be detected early in the analysis process, which provides the opportunity to fix or discard them and avoid incorrect interpretations later. Due to experts' time constraints, and not to burden them with excessive workload, it was decided that inconsistent matrices (33%) would be discarded and those with a tolerable consistency would be accepted.

[2]Available at https://goo.gl/forms/EEmkUmLIj20nMJS33

TABLE 9: Experts' declared familiarity with metrics and resulting weight for the consensus priority vector. Empty cells denote inconsistent (rejected) matrices.

	Declared familiarity[a]					Weight			
Experts	*R*	*P*	*F*	*I*	*M*	*BC*	*HC*	*BE*	*ME*
E01	4	4	3	2	2		0.09	0.07	0.08
E02	3	3	3	3	3				
E03	1	2	2	1	1	0.05	0.04	0.03	0.04
E04	5	5	4	3	3				
E05	3	3	3	3	3	0.11	0.09	0.08	0.09
E06	5	5	5	5	5	0.18	0.14	0.12	0.14
E07	2	2	2	2	2	0.07	0.06	0.06	0.06
E08	5	5	3	1	1	0.09		0.06	0.07
E09	3	3	3	2	2	0.10	0.08	0.06	0.08
E10	3	3	2	1	1	0.07		0.04	0.05
E11	1	1	1	1	1	0.04			0.03
E12	5	5	5	1	1		0.08	0.06	
E13	5	5	5	1	1			0.06	
E14	3	3	3	1	1	0.07	0.06	0.05	
E15	5	5	5	1	1	0.10	0.08	0.06	
E16	5	5	5	1	1		0.08	0.06	0.08
E17	2	3	2	1	1	0.06	0.05	0.04	0.05
E18	5	5	5	1	1			0.06	0.08
E19	4	4	3	1	1			0.05	
E20	3	3	1	1	1	0.06	0.05	0.04	0.05
E21	4	4	4	3	3		0.10		0.10

[a] Scale 1-5: 1 - Low familiarity, 5 - High familiarity

TABLE 10: Consensus priority vector for each scenario.

		Weight				
Scenario	**#Experts**	*R*	*P*	*F*	*I*	*M*
BC	12	0.58	0.08	0.12	0.14	0.08
HC	13	0.30	0.14	0.22	0.23	0.11
BE	17	0.11	0.15	0.31	0.22	0.21
ME	14	0.07	0.31	0.14	0.21	0.27

Another important influential aspect is the experts' familiarity with considered metrics, which is listed in Table 9. *Recall*, *Precision*, and *F-measure*, with an average familiarity of 3.6 ± 1.4, 3.7 ± 1.2, and 3.2 ± 1.3, respectively, are easier to understand and use than *Informedness* and *Markedness*, with an average familiarity of 1.7 ± 1.1 and 1.7 ± 1.1, respectively. The declared familiarity with metrics for experts that provided a consistent pairwise comparison matrix is used to compute that expert's contribution to the consensus priority vector for each scenario. Resulting weights are listed in Table 9. The judgement of experts $E06$, $E21$, and $E05$, which declared the highest familiarity with metrics,

TABLE 11: Benchmark results from case study 1 [14].

Tool	*R*	*P*	*F*	*I*	*M*
VDT01	0.793	1.000	0.885	0.793	0.953
VDT02	0.552	0.923	0.691	0.541	0.825
VDT03	1.000	0.640	0.780	0.864	0.640
VDT04	0.149	0.325	0.205	0.075	0.144
VDT05	0.753	1.000	0.859	0.753	0.903
VDT06	0.323	0.455	0.378	0.156	0.195
VDT07	0.241	0.388	0.297	0.076	0.105
VDT08	0.019	1.000	0.037	0.019	0.702
VDT09	0.241	0.567	0.338	0.161	0.304
VDT10	0.741	1.000	0.851	0.741	0.899

TABLE 12: Case study 1: Single Metric (SM) vs. MABRES.

Business-Critical		
SM	MABRES	Score
VDT03	VDT03	0.897
VDT01	VDT01	0.833
VDT05	VDT05	0.797
VDT10	VDT10	0.787
VDT02	VDT02	0.618
VDT06	VDT06	0.307
VDT07	VDT09 (↑ 1)	0.273
VDT09	VDT07 (↓ 1)	0.226
VDT04	VDT04	0.159
VDT08	VDT08	0.154

Heightened-Critical		
SM	MABRES	Score
VDT03	VDT03	0.749
VDT01	VDT01	0.734
VDT05	VDT05	0.701
VDT10	VDT10	0.693
VDT02	VDT02	0.546
VDT09	VDT06 (↑ 1)	0.244
VDT06	VDT09 (↓ 1)	0.225
VDT07	VDT07	0.172
VDT04	VDT04	0.127
VDT08	VDT08	0.109

Best Effort		
SM	MABRES	Score
VDT01	VDT01	0.886
VDT05	VDT05	0.854
VDT10	VDT10	0.847
VDT03	VDT03	0.772
VDT02	VDT02	0.706
VDT06	VDT09 (↑ 1)	0.316
VDT09	VDT08 (↑ 3)	0.315
VDT07	VDT06 (↓ 2)	0.296
VDT04	VDT07 (↓ 1)	0.216
VDT08	VDT04 (↓ 1)	0.175

Minimum Effort		
SM	MABRES	Score
VDT01	VDT01	0.913
VDT05	VDT05	0.885
VDT10	VDT10	0.879
VDT02	VDT02	0.758
VDT08	VDT03 (↑ 1)	0.732
VDT03	VDT08 (↓ 1)	0.510
VDT09	VDT09	0.356
VDT06	VDT06	0.302
VDT04	VDT07 (↑ 1)	0.223
VDT07	VDT04 (↓ 1)	0.195

will contribute the most to the consensus priority vector. However, the judgment of experts $E04$ and $E02$, which also declared a high familiarity, will not be taken into account as all their matrices were inconsistent. The consensus priority vector for each scenario is listed in Table 10.

4.2 Ranking Vulnerability Detection Tools

The data set for this case study is extracted from [14], where 10 different vulnerability detection tools (see Table 5) were analyzed by using a single metric and a tie-breaker per scenario (see Table 3). These same data, listed in Table 11, have been processed by MABRES, using the computed consensus priority vector for each scenario (see Table 10), to obtain another ranking that takes into account the judgements of collaborating experts. The rankings obtained by using the single metric approach and MABRES are listed in Table 12.

It is important to clarify that our goal is not to corroborate or redo that previous work, but to use a data set with highly homogenous values for all metrics (tools *VDT01*, *VDT02*, *VDT03*, *VDT05*, and *VDT10* consistently rank as the top 5 tools in all scenarios) to show that including experts' judgements in the analysis does not contradict but enriches what previous research dictates.

Recall is the primary metric and *Precision* the tie-breaker in the *Business-Critical* scenario. Experts also agree on the importance of *Recall* (weight of 0.58 in the consensus priority vector), but *F-measure* and *Markedness* are given a greater weight than *Precision* (it does not act as tie-breaker). However, benchmarked values are so homogenous among tools and the weight of *Recall* is so great, that only *VDT07* and *VDT09* exchange their positions (7^{th} and 8^{th}).

In the *Heightened-Critical* scenario, *Informedness* is the primary metric and *Recall* acts as tie-breaker. Experts agree that these are the most important metrics but not in this order. *Recall* and *Informedness* are assigned a weight of 0.30 and 0.23 respectively, with *F-measure* close behind with a weight of 0.22. Even though following different criteria, only *VDT06* and *VDT09* exchange their positions (6^{th} and 7^{th}). This is also a result of having highly homogeneous values among tools.

F-measure is the primary metric and *Recall* the tie-breaker for the *Best Effort* scenario. Experts agree on providing *F-measure* the greatest weight (0.31), but *Informedness* and *Marknedness* also get a great influence on the final score (0.22 and 0.21, respectively). *Recall*, however, presents the smallest weight and cannot be considered as tie-breaker. Although these considerations do not affect the ranking of the top 5 tools, the rest of tools exchange their positions, with *VDT08* scaling from the 10^{th} to the 7^{th} position and *VDT06* sinking from the 6^{th} to the 8^{th} position.

Finally, the *Minimum Effort* scenario considers *Markedness* as primary metric and *Precision* as tie-breaker. Experts agree on providing the greatest weights to these metrics, but in reversed order (0.31 for *Precision* and 0.27 for *Markedness*). As previously commented, this barely affects the final rankings, and only *VDT03* and *VDT08*, and *VDT04* and *VDT07*, exchange their respective positions (5^{th} and 6^{th}, 9^{th} and 10^{th}).

Although this does not exhaustively validate the proposal, it is a first best effort towards showing its usefulness. Section 4.3 will study a data set with greater variability in values for all metrics to show that by taking into account the information provided by all these metrics, not just one of them, may help in taking more informed decisions.

TABLE 13: Benchmark results from case study 2.

Tool	*R*	*P*	*F*	*I*	*M*
IDT01	0.790	0.876	0.831	0.275	0.210
IDT02	0.350	0.862	0.498	0.091	0.059
IDT03	0.275	0.993	0.431	0.266	0.221
IDT04	0.193	1.000	0.324	0.193	0.211
IDT05	0.089	1.000	0.163	0.089	0.192
IDT06	0.538	1.000	0.700	0.538	0.795
IDT07	0.162	0.966	0.277	0.135	0.167
IDT08	0.180	0.940	0.302	0.127	0.140
IDT09	0.794	0.615	0.693	0.517	0.478
IDT10	0.882	0.493	0.633	0.377	0.376
IDT11	0.293	1.000	0.453	0.293	0.327

4.3 Ranking Intrusion Detection Systems (IDS) Tools

IDS tools for SQL injection attacks in web applications aim at detecting attackers trying to read, alter, or destroy the content of databases. Since detection rates are also described in terms of *TP*, *TN*, *FP*, and *FN*, then the same set of metrics and scenarios can be used to compare and rank these tools. Accordingly, the consensus priority vector for each scenario is listed in Table 10 and the discussion from Section 4.2, regarding the differences between the single metric approach and MABRES, still holds.

TABLE 14: Case study 2: Single Metric (SM) vs. MABRES.

Business-Critical		
SM	MABRES	Score
IDT10	IDT10	0.710
IDT09	IDT09	0.704
IDT01	IDT01	0.683
IDT06	IDT06	0.615
IDT02	IDT11 (↑ 1)	0.371
IDT11	IDT02 (↓ 1)	0.349
IDT03	IDT03	0.346
IDT04	IDT04	0.275
IDT08	IDT08	0.245
IDT07	IDT07	0.237
IDT05	IDT05	0.179

Heightened-Critical		
SM	MABRES	Score
IDT06	IDT09 (↑ 1)	0.571
IDT09	IDT06 (↓ 1)	0.541
IDT10	IDT10	0.539
IDT11	IDT01 (↑ 1)	0.518
IDT01	IDT11 (↓ 1)	0.305
IDT03	IDT03	0.277
IDT04	IDT02 (↑ 3)	0.254
IDT07	IDT04 (↓ 1)	0.210
IDT08	IDT08	0.178
IDT02	IDT07 (↓ 2)	0.172
IDT05	IDT05	0.118

Best Effort		
SM	MABRES	Score
IDT01	IDT06 (↑ 1)	0.711
IDT06	IDT09 (↑ 1)	0.601
IDT09	IDT01 (↓ 2)	0.581
IDT10	IDT10	0.529
IDT02	IDT11 (↑ 1)	0.456
IDT11	IDT03 (↑ 1)	0.418
IDT03	IDT04 (↑ 1)	0.358
IDT04	IDT02 (↓ 3)	0.355
IDT08	IDT07 (↑ 1)	0.313
IDT07	IDT08 (↓ 1)	0.311
IDT05	IDT05	0.270

Minimum Effort		
SM	MABRES	Score
IDT06	IDT06	0.773
IDT09	IDT09	0.581
IDT10	IDT01 (↑ 4)	0.558
IDT11	IDT11	0.544
IDT03	IDT03	0.502
IDT04	IDT10 (↓ 3)	0.484
IDT01	IDT04 (↓ 1)	0.466
IDT05	IDT07 (↑ 1)	0.423
IDT07	IDT08 (↑ 1)	0.411
IDT08	IDT05 (↓ 2)	0.410
IDT02	IDT02	0.396

Table 13 shows the results obtained from benchmarking the 11 IDS tools listed in Table 6. This set of tools presents a greater variability in the values obtained for each metric and tool than the set considered in the first case study. *IDT10* outperforms the rest of tools when considering *Recall*, and *IDT01* and *IDT09* appear as second options. *IDT04*, *IDT05*, *IDT06*, and *IDT11* get a perfect *Precision*, with *IDT03*, *IDT07*, and *IDT08* close behind. *IDT01* is the best alternative for *F-measure*, with *IDT06* and *IDT09* as second options. These tools (*IDT06* and *IDT09*) are those obtaining the best *Informedness* values (despite being quite low). Finally, *IDT06* is the best choice from the perspective of *Markedness*, while the rest of tools get a very low value. Both the single metric (SM) approach and MABRES have been applied to study the effect of neglecting information from unconsidered metrics, and how rankings may vary (see Table 14) when this information and experts' judgements are taken into account.

In the *Business-Critical* scenario both rankings are nearly identical (only *IDT02* and *IDT11* exchange the 5th and 6th position), due to the great weight assigned to *Recall*.

The first important difference can be observed in the *Heightened-Critical* scenario, in which the 1st and 2nd ranking tools (*IDT06* and *IDT09*) exchange their positions. This is due to the effect of experts providing a greater weight to *Recall* than to *Informedness* and the greater variability in the results (the tools with greater *Informedness* are not those also obtaining a greater *Recall*). Likewise, *IDT11* and *IDT01* also exchange the 4th and 5th positions, and *IDT04* falls from the 6th to the 7th position. It is to note that *IDT02* climbs up from the 9th to the 6th position because, although it presents the second worst *Informedness*, the rest of metrics contribute to improve its score. The contrary effect can be observed for *IDT07*, which falls from the 8th to the 10th position because it presents poor results for both *Recall* and *Informedness*.

The effect of the joint contribution of all measures is more pronounced in the *Best Effort* scenario. Even though experts agreed on *F-measure* being the most important metric, the combined weight of *Informedness* and *Markedness* exceeds that of *F-measure*. This causes that *IDT01*, the top tool for the SM approach, falls downto to the 3rd position. Likewise, *IDT02* falls from the 5th to the 8th position. Finally, *IDT08* and *IDT07* exchange the 9th and 10th position.

This variability is also pronounced in the *Minimum Effort* scenario where, although experts assigned a greater weight to *Precision* than to *Marknedness*, the first two tools keep their positions. However, *IDT10* falls down from the 3rd to the 6th position (it presents the 3rd best *Informedness* but with a very poor value, which is not compensated by the rest of metrics), and is replaced by *IDT01* (its very poor *Informedness* is overcome by the rest of combined metrics), which climbs up from the 8th position. Likewise, *IDT07* drops from the 8th to the 10th position (very low *Recall*), and *IDT02* climbs up from the 10th to the 7th position (its very poor *informedness* is compensated by the rest of metrics, but not much as it also presents the worst *Markedness*).

These results show how final rankings, and thus the decision on which are the best tools for a given scenario, may change according to considered metrics and their relative weight. Observed differences arise from the judgement of the set of collaborating experts that, in some cases, does not match the priorities defined by the single metric approach. Section 5 will discuss about this proposal and its limitations.

5 Discussion

The analysis of the results drawn from the considered case study have shown how the proposed methodology can be useful to take into account the information provided by all considered metrics when benchmarking security tools. The proposal relies on the expertise of a set of multidisciplinary experts, which is a must nowadays when considering the integration of more complex and heterogenous systems, like Internet of Things and cyberphysical systems. This section discusses several questions and limitations that can come to the mind of users about the applicability of this approach.

5.1 Why should I use this approach?

Current benchmarking procedures for security tools based on binary classifiers offer a simple way to compare alternative solutions for different scenarios. Just one metric, from all those that can be derived from a confusion matrix (at least 14 were identified in [14]), is considered sufficient to rank target tools. This approach is widely accepted, as it makes the decision process very straightforward and easy to understand. Nevertheless, this oversimplification may counterbalance its benefits. For example, Table 11 shows that all vulnerabilities detected by *VDT08* are correctly classified, so it has a perfect *Precision* (1.00). However, this same tool presents a very low *Recall* (0.019), so it only detects a 2% of existing vulnerabilities. Accordingly, even though *VDT08* is useless, it would be the selected alternative if *Precision* was considered as the single metric to rank the tools.

This is why, in the dependability/security benchmarking domain, researchers usually require several different metrics to obtain a holistic view of the system's capabilities and draw better informed conclusions. Obviously, this process is more complex and cumbersome, in direct conflict with industry needs in terms of simplicity.

Our proposal offers an alternative to deal with this dichotomy between the points of view of academy and industry. On the one hand, the decision making process can be enriched by considering that any number of metrics can contribute to the final score of each target. Final weights are determined from the aggregation of each expert's judgement, so scores can be analysed and tracked back to the values of each metric, thus leading to better informed decisions. On the other hand, targets are ranked according to a single score, thus keeping the decision making process simple. In fact, the underlying complexity of this approach is hidden from benchmark users, as it resides in how this score is computed and not in how it is interpreted.

Even though this proposal has the potential to satisfy both academy and industry requirements, its internal complexity may prevent its adoption by the community, because relying on a complex non-familiar procedure instead of following the common and well-understood (although imperfect) path is not easy. However, the adoption of this type of approaches seems unavoidable in the near future, as systems will only be useful if secure and the evaluation of such security will require the intervention of multiple experts (due to the intrinsic complexity of solutions) and the consideration of the economical, functional and non-functional requirements imposed by each context of use.

5.2 Can you prove that this approach is better?

Any existing methodology to rank security tools taking into account several metrics, including the Single Metric (plus tie-breaker) approach and the proposed one (MABRES), can be considered as a multiple-criteria decision making (MCDM) approach. Hence, according to [45], "*it is impossible to determine precisely the best decision-making method, for to do so one needs to use the best decision-making method!*"

Although no actual proof can be provided to demonstrate that one approach is better than the other, we would like to highlight that MABRES is backed by investigation in the operational research domain, like the Analytic Hierarchy Process (AHP) [21], the aggregation of individual judgement [32], and WSM [45]. The proposed methodology scales to support any number of metrics, any number of experts, and any MCDM.

Furthermore, the judgement of more experts can be integrated into the procedure as soon as they are available. The resulting consensus priority vector can be automatically updated and rankings recomputed (a simple script can take charge of the whole process). This enables the expansion of the experts' knowledge database when more collaborators are willing to contribute.

We would like to remark that the Single Metric approach [14] was defined by their authors just solely based on its own expertise and it was validated by just 6 experts, whereas 21 experts participated in this work. Obviously, this does not invalidate the Single Metric approach which, as it has been previously discussed, constitutes a really simple and well understood approach, but the proposal presented in this work has the potential to overcome that simplicity in exchange for greater flexibility and expressiveness.

5.3 Are results trustworthy?

Obtained rankings are a direct result of experts' pairwise comparison matrices and their declared familiarity with considered metrics. This means that the subjective judgement of experts will obviously affect the final ranking. However, there are different elements that can be used to limit that effect.

The *Consistency Ratio* (CR) of pairwise comparison matrices have been checked to prevent inconsistent matrices from being used to compute the consensus priority vector. Although we have opted to include all those matrices with a tolerable consistency ($CR < 0.20$), a stricter threshold can be considered ($CR < 0.10$) to accept only highly consistent matrices. Experts may be asked to rebuild inconsistent matrices. It must be noted that pairwise comparisons can be highly consistent but completely wrong (an expert may consistently declare that *Precision* is extremely more important than all the other metrics for a given scenario, even though it is obviously not true). Although this issue cannot be prevented, it should rarely occur as experts must be selected due to their expertise in the considered domain.

The familiarity of experts with metrics is taken into account to compute each expert's contribution to the consensus priority vector. Other factors can also be integrated into this approach to limit the influence of *false experts*. For instance, experts in metrology, benchmarking, and security participated in this work, but not all were experts in all three domains. Only their familiarity with the metrics was considered, as it was our primary concern, but also their familiarity with the four considered scenarios or their domain of expertise could have been included in the questionnaire.

5.4 You have your experts and I have mine

As previously discussed, resulting rankings completely depend on the subjective judgement of experts. Thus, different sets of experts (with different domains of expertise or currents of thought) may lead to different consensus priority vectors and then different (and maybe conflicting) rankings.

This does not mean that results are not reproducible, as rankings are computed following a precise mathematical process which draws the same conclusions from the same input data. However, changing the input data (which includes pairwise comparison matrices) may lead to different conclusions. Accordingly, in order for researchers to be able to share and compare results, it would be advisable to also share all the information provide by experts, including pairwise comparison matrices and considered factors.

It must be noted that this work focuses on defining an approach to consider a set of metrics to benchmark security tools, rather than determining the exact weight that denote the contribution of each metric towards the score for each tool. The considered case study obtained a consensus priority vector for each scenario thanks to the collaboration of 21 researchers that acted as experts, but a different set of experts may agree on different weights. This does not invalidate the proposal, as it is finally a matter of acceptance by the community who should adopt similar approaches and define a set of experts, that could be periodically increased, to obtain weights for different kind of tools and scenarios.

6 Conclusions

Benchmarking security tools is a process of prime importance not only to determine the most suitable tool for a given application domain or context of use, but also to assess the effectiveness of any improvement deployed on existing tools. A large set of different metrics have stemmed from reported results in terms of true/false positives/negatives. These metrics are supposed to help benchmark users in better understanding the particular capabilities of each tool and, thus, reach a better decision. The truth is that, in practice, having to ponder and balance so many metrics, usually with conflicting goals, leads to multiple-objective optimization problems that are difficult to solve without any explicit guidelines. Accordingly, existing approaches usually opt to oversimplify the problem by considering just a single metric for each application scenario. Although this is an accepted practice by the industry, it is also understood that the final decision may be biased by not considering all the subtleties accounted by the rest of neglected metrics.

This paper has proposed a novel and fully automated approach, backed by investigation in the operational research domain, to alleviate the problem of simultaneously considering the contribution of all existing metrics towards the selection of the best security tool for a given application scenario. This approach relies on experts to judge the relative importance of each pair of metrics for each scenario. This process results in the quantification of the contribution of each metric to the particular requirements of the selected scenario. Expert's individual judgements can also be weighed according to their familiarity with considered metrics and application scenarios, thus enabling the collaboration of multidisciplibary experts to deal with the requirements of modern complex systems. The result of this whole process is a single score for each target tool that can be used for ranking and selection purposes.

Accordingly, this approach not only simplifies the decision process for benchmark users by considering a single score, but also allows for a better informed decision as the contribution of all considered metrics towards the scenario requirements is taken into account. Any bias that could be introduced by subjective judgements is minimized by considering the expertise of participants before reaching an agreement, and errors due to human intervention are also minimized by rejecting inconsistent comparisons.

Future work relates to the use of expert systems and machine learning algorithms to complement in the medium term, and replace in the long term, the work currently assumed by human experts.

Acknowledgments

This work has been partially supported by the project **EUBra-BIGSEA** (www.eubra-bigsea.eu), funded by the European Commission under the Cooperation Programme, Horizon 2020 grant agreement no `690116`, the *"Programa de Ayudas de Investigación y Desarrollo" (PAID) de la Universitat Politècnica de València* and the project **DINAMOS** (dinamos.webs.upv.es), funded by the Ministerio de Economía, Industria y Competitividad de España, grant agreement no `TIN2016-81075-R`.

References

[1] D. Stuttard and M. Pinto, *The web application hacker's handbook: discovering and exploiting security flaws*. Wiley Publishing, Inc., 2007.

[2] S. Wagner, J. Jürjens, C. Koller, and P. Trischberger, "Comparing bug finding tools with reviews and tests," *Testing of Communicating Systems*, pp. 40–55, 2005.

[3] A. Doupé, M. Cova, and G. Vigna, "Why johnny can't pentest: An analysis of black-box web vulnerability scanners," in *Detection of Intrusions and Malware, and Vulnerability Assessment*, 2010, no. 6201, pp. 111–131.

[4] J. Gray, *Benchmark Handbook: For Database and Transaction Processing Systems*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1992.

[5] K. Kanoun and L. Spainhower, *Dependability Benchmarking for Computer Systems*. John Wiley & Sons, 2008.

[6] G. Schröder, M. Thiele, and W. Lehner, "Setting goals and choosing metrics for recommender system evaluations," in *UCERSTI2 Workshop at the 5th ACM Conference on Recommender Systems, Chicago, USA*, vol. 23, 2011, p. 53.

[7] A. Jaquith, *Security metrics: replacing fear, uncertainty, and doubt*. Upper Saddle River, NJ: Addison-Wesley, 2007.

[8] I. Pashchenko, S. Dashevskyi, and F. Massacci, "Delta-bench: Differential benchmark for static analysis security testing tools," in *ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*, 2017, pp. 163–168.

[9] G. Hao, F. Li, W. Huo, Q. Sun, W. Wang, X. Li, and W. Zou, "Constructing benchmarks for supporting explainable evaluations of static application security testing tools," in *International Symposium on Theoretical Aspects of Software Engineering*, 2019, pp. 65–72.

[10] A. Arusoaie, S. Ciobâca, V. Craciun, D. Gavrilut, and D. Lucanu, "A comparison of open-source static analysis tools for vulnerability detection in c/c++ code," in *19th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)*, 2017, pp. 161–168.

[11] I. A. Elia, J. Fonseca, and M. Vieira, "Comparing sql injection detection tools using attack injection: An experimental study," in *21st IEEE International Symposium on Software Reliability Engineering (ISSRE 2010)*. IEEE Computer Society, 2010, pp. 289–298.

[12] N. Antunes and M. Vieira, "Assessing and Comparing Vulnerability Detection Tools for Web Services: Benchmarking Approach and Examples," *IEEE Transactions on Services Computing*, vol. 8, no. 2, pp. 269–283, 2015.

[13] D. M. Powers, "Evaluation: from Precision, Recall and F-measure to ROC, Informedness, Markedness and Correlation," *Journal of Machine Learning Technologies*, Dec. 2011.

[14] N. Antunes and M. Vieira, "On the Metrics for Benchmarking Vulnerability Detection Tools," in *The 45th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN 2015)*. Rio de Janeiro, Brazil: IEEE, 2015.

[15] E. Zio, "On the use of the analytic hierarchy process in the aggregation of expert judgments," *Reliability engineering & systems safety*, vol. 53, no. 2, pp. 127–138, 1996.

[16] H. Li, M. Lu, and Q. Li, "Software reliability metrics selecting method based on analytic hierarchy process," in *2006 Sixth International Conference on Quality Software*, Oct 2006, pp. 337–346.

[17] J. Krejčí, D. Petri, and M. Fedrizzi, "From measurement to decision with the analytic hierarchy process: Propagation of uncertainty to decision outcome," *IEEE Transactions on Instrumentation and Measurement*, vol. 66, no. 12, pp. 3228–3236, Dec 2017.

[18] T. Fawcett, "An introduction to roc analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, Jun. 2006.

[19] S. V. Stehman, "Selecting and interpreting measures of thematic classification accuracy," *Remote Sensing of Environment*, vol. 62, no. 1, pp. 77–89, 1997.

[20] I. Tuzov, D. de Andrés, and J. Ruiz, "Dependability-aware design space exploration for optimal synthesis parameters tuning," in *2017 47th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*, vol. 00, June 2017, pp. 121–132.

[21] T. Saaty, "What is the analytic hierarchy process?" in *Mathematical Models for Decision Support*, ser. NATO ASI Series, G. Mitra, H. Greenberg, F. Lootsma, M. Rijkaert, and H. Zimmermann, Eds. Springer Berlin Heidelberg, 1988, vol. 48, pp. 109–121.

[22] M.-K. Chen and S.-C. Wang, "The critical factors of success for information service industry in developing international market: Using analytic hierarchy process (ahp) approach," *Expert Systems with Applications*, vol. 37, no. 1, pp. 694–704, 2010.

[23] X. Sun and X. Fang, "Construction and operation of analytic hierarchy process about moral education evaluation in colleges and universities." *Advances in Information Sciences & Service Sciences*, vol. 3, no. 11, 2011.

[24] İ. Ertuğrul and N. Karakaşoğlu, "Performance evaluation of turkish cement firms with fuzzy analytic hierarchy process and topsis methods," *Expert Systems with Applications*, vol. 36, no. 1, pp. 702–715, 2009.

[25] T. Saaty and L. Vargas, "The seven pillars of the analytic hierarchy process," in *Models, Methods, Concepts & Applications of the Analytic Hierarchy Process*, ser. International Series in Operations Research & Management Science. Springer US, 2001, vol. 34, pp. 27–46.

[26] G. Crawford and C. Williams, "A note on the analysis of subjective judgment matrices," *Journal of Mathematical Psychology*, vol. 29, no. 4, pp. 387 – 405, 1985.

[27] Y. Dong, Y. Xu, H. Li, and M. Dai, "A comparative study of the numerical scales and the prioritization methods in ahp," *European Journal of Operational Research*, vol. 186, no. 1, pp. 229 – 242, 2008.

[28] M. W. Herman and W. W. Koczkodaj, "A monte carlo study of pairwise comparison," *Information Processing Letters*, vol. 57, no. 1, pp. 25 – 29, 1996.

[29] J. Malczewski, *GIS and Multicriteria Decision Analysis*. Wiley, 1999.

[30] T. L. Saaty, "Decision-making with the ahp: Why is the principal eigenvector necessary," *European Journal of Operational Research*, vol. 145, no. 1, pp. 85 – 91, 2003.

[31] J. A. Alonso and M. T. Lamata, "Consistency in the analytic hierarchy process: A new approach," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 14, no. 04, pp. 445–459, 2006.

[32] E. Forman and K. Peniwati, "Aggregating individual judgments and priorities with the analytic hierarchy process," *European Journal of Operational Research*, vol. 108, no. 1, pp. 165 – 169, 1998.

[33] Y. Dong, G. Zhang, W.-C. Hong, and Y. Xu, "Consensus models for ahp group decision making under row geometric mean prioritization method," *Decision Support Systems*, vol. 49, no. 3, pp. 281 – 289, 2010.

[34] J. Barzilai and B. Golany, "Ahp rank reversal, normalization and aggregation rules," *INFOR: Information Systems and Operational Research*, vol. 32, no. 2, pp. 57–64, 1994.

[35] Michele, S. Marchesi, N. Pinna, and G. C. Serra, "Power-laws in a large object-oriented software system," *IEEE Transactions on Software Engineering*, vol. 33, pp. 687–708, 2007.

[36] T. L. Alves, J. P. Correia, and J. Visser, "Benchmark-based aggregation of metrics to ratings." in *IWSM/Mensura*, K. Matsuda, K. ichi Matsumoto, and A. Monden, Eds. IEEE Computer Society, 2011, pp. 20–29.

[37] C. Kruegel and G. Vigna, "Anomaly detection of web-based attacks," in *Proceedings of the 10th ACM Conference on Computer and Communications Security*, 2003, pp. 251–261.

[38] GreenSQL, "An open source database firewall used to protect databases from sql injection attacks," https://sourceforge.net/projects/greensql (last check October 2018), 2018.

[39] Apache-Scalp, "Apache log analyzer for security," http://code.google.com/p/apache-scalp (last check October 2018), 2018.

[40] M. Roesch, "Snort - lightweight intrusion detection for networks," in *Proceedings of the 13th USENIX Conference on System Administration*, 1999. Tool available at: https://www.snort.org (last check October 2018), pp. 229–238.

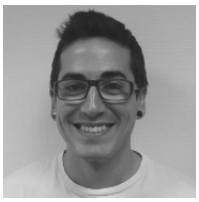
[41] J. Fonseca, M. Vieira, and H. Madeira, "Detecting malicious sql," in *Proceedings of the 4th International Conference on Trust, Privacy and Security in Digital Business*, ser. TrustBus'07. Berlin, Heidelberg: Springer-Verlag, 2007, pp. 259–268.

[42] Trustware, "Modsecurity: Open source web application firewall." https://www.modsecurity.org (software downloaded on March 2016), 2016.

[43] F. Zhu, P.-a. Zhong, and Y. Sun, "Multi-criteria group decision making under uncertainty," *Environ. Model. Softw.*, vol. 100, no. C, pp. 236–251, Feb. 2018.

[44] W. C. Wedley, "Consistency prediction for incomplete ahp matrices," *Mathematical and Computer Modelling*, vol. 17, no. 4/5, pp. 151–161, 1993.

[45] E. Triantaphyllou and H. Stuart, "An examination of the effectiveness of multi-dimensional decision-making methods: A decision-making paradox," *Decision Support Systems*, vol. 5, pp. 303–312, 1989.



Miquel Martínez received his Ph.D. degree in computer science at the Universitat Politècnica de València, Spain, in 2018. His research interests include ad hoc networks, fault and attack injection for robustness assessment and dependability/security evaluation, dependability benchmarking, and the analysis of results for multi-criteria decision problems.

Juan Carlos Ruiz received his Ph.D. degree from the Institut National Polytechnique of Toulouse in 2002. He joined the Fault-Tolerant Systems Research Group with the Institute ITACA, Universitat Politècnica de València (UPV), in 2003. He is an Associate Professor with the Computer Engineering Department at UPV. His research interests include the specification of dependability benchmarks, definition of fault-injection techniques, and design of fault-tolerance mechanisms for embedded systems.

Nuno Antunes is an Assistant Professor at the University of Coimbra, where he received his PhD in Information Science and Technology in 2014. He has been with the Centre for Informatics and Systems of the University of Coimbra (CISUC) since 2008, working in Security and Dependability topics. His expertise includes testing techniques, fault injection, vulnerability injection, experimental dependability and security evaluation, and security benchmarking. He is a member of the IEEE Computer Society.

David de Andrés received his Ph.D. degree from the Universitat Politècnica de València (UPV)in 2007. He is currently an Associate Professor with the Department of Computer Engineering at UPV. He is also with the Fault-Tolerant Systems Research Group from the Institute ITACA. His current research interests include run-time reconfigurable hardware systems, fault and attack injection for robustness assessment, and dependability benchmarking.

Marco Vieira is a Full Professor at the University of Coimbra, Portugal. His interests include dependability and security assessment and benchmarking, fault injection, software processes, and software quality assurance. Marco received the PhD degree from the University of Coimbra.