A New Agreement Coefficients-Based Approach to Compare Improved FMECA Methods

Andrés A. Zúñiga, *Student Member*, *IEEE*, João F. P. Fernandes, *Member*, *IEEE* and Paulo J. Costa Branco, *Member*, *IEEE*

***Abstract*— Commonly, the efficacy of new FMECA methods is conducted through qualitative comparisons between rankings; for a small number of failure modes, this approach is suitable but can become unpractical or lead to misleading results for more extensive problems. This fact motivated us to introduce an alternative approach to compare different FMECA methods based on agreement metrics that allow the statistical comparison between rankings generated by independent raters. Despite its relevance, the application of agreement coefficients is limited in the FMECA context. The proposed approach considers the agreement assessment between different methodologies used in FMECA analysis (Risk Priority Isosurfaces RPI, VIKOR, ITWH, and Type-II Fuzzy Inference System) when applied to a study case regarding blood transfusion widely used in the literature for benchmarking and consisting of eleven failure modes. We selected the RPI methods as a reference to compare the other forenamed methods. Results show that our agreement coefficient-based comparison approach proves effective for the statistical comparison of different FMECA methods instead of the rankings qualitative comparison.**

***Index Terms*—FMECA; Risk assessment; Type-II fuzzy inference systems; Agreement Coefficient; Cohen's kappa.**

I. INTRODUCTION

FAILURE Modes Effects and Criticality Analysis (FMECA) is a qualitative risk assessment method designed to identify potential failure modes, their causes, and systems performance effects [1]. The objective of FMECA is to identify the possible ways a failure can occur, how often it occurs, how severe the failure affects the system performance, and what preventive measures should be taken to avoid the failure.

The classical FMECA analysis is based on three factors, called risk factors, to characterize each failure mode [1]: the Severity (SEV), which characterizes the effect of the failure mode qualitatively, and the Frequency of Occurrence (OCC), that characterize how likely it is it the failure mode to occur, and the Detectability (DET) that characterize how detectable is the failure mode before to occur. Each risk factor is classified into specific risk categories represented by a numerical scale, and it can be a 1 to 10 scale as used in [1] or a 1 to 5 scale as in [2].

Each failure mode is assessed through a risk priority number (RPN). The RPN results from the composition, (denoted by the symbol $\circ$), between SEV, OCC, and DET (1), being the product the generally adopted operator.

$$RPN = (SEV)\circ(OCC)\circ(DET) \tag{1}$$

Because the RPN calculation in the classical FMECA approach results from the unique product between three integers, there is no associated computational complexity. Although FMECA is a very popular qualitative method for failure analysis, computation of the RPN has some disadvantages [3,4]. The main are:

1) The RPN computation does not consider any difference degree between the three risk factors, OCC, SEV, and DET (i.e., no weight averaging each risk factor);
2) Although a higher RPN is usually associated with more critical failure modes, this is not always true [5,6];
3) The scales for the three risk factors are generally considered arbitrarily and may not accurately represent the risk characteristics in specific problems.

To deal with the classical FMECA's shortcomings, some approaches based on computational intelligence and decision-making methods have been proposed in the past years.

Bowles and Peláez [3] presented one of the first applications of fuzzy inference systems FIS to improve the FMECA analysis. Their results showed that the proposed FIS approach allowed overcoming some FMECA issues like imprecise information related to the risk factors. Recently in [5], the authors conducted a literature review about FMECA methods published between 1998 and 2018. The review shows that publications about FMECA improvements have increased in the last ten years. Methods like grey theory and fuzzy inference systems were used mostly to improve the FMECA analysis. In [6], multi-criteria decision-making (MCDM) methods and uncertainty theory are applied to model the vagueness related to FMECA processes. This book includes a broad review of academic works that apply MCDM methods to overcome FMECA issues.

In [7], the authors present a fuzzy rules base and grey relation theory to improve the FMECA analysis conducted for an ocean-going fish vessel. The proposed methodology includes linguistic terms and allows to assign weights to each risk factor. In [8], an improved prioritization method is

---

This work was supported by national funds through the Fundação para a Ciência e a Tecnologia, I.P., through IDMEC, under LAETA, Project UIDB/50022/2020, and by Secretaría Nacional de Educación Superior, Ciencia, Tecnología e Innovación (SENESCYT) of the Ecuadorian Government through fellowship CZ05-000291-2017. (Corresponding author: Andrés A. Zúñiga.).

Andrés A. Zúñiga, João F. P. Fernandes and Paulo J. Costa Branco are with the IDMEC, Instituto Superior Técnico, University of Lisbon, 1049-001 Lisbon, Portugal (e-mail: andres.zuniga@tecnico.ulisboa.pt; joao.f.p.fernandes@tecnico.ulisboa.pt; pbranco@tecnico.ulisboa.pt).

proposed based on combining the "cloud model" and a modified PROMETHEE [9] method. The authors also compared qualitatively with other approaches like IF-TOPSIS [10], Fuzzy VIKOR [11], IVIF-MULTIMOORA [12], and ITL-GRA. Their methodology was applied to prioritize the potential failure modes in the emergency department's treatment process. Results showed that the proposal could overcome the shortcomings of the traditional FMEA method, yielding more reasonable and credible risk ranking results.

In [13], the authors describe a combination of the variable precision rough set theory to represent the vagueness associated with the FMEA members, using the TODIM method to improve the failure modes' ranking. Their proposed approach was applied to a real case study of a steam valve system and compared qualitatively with the classical FMEA, the Fuzzy TODIM, the Rough TOPSIS [14], and linguistic distribution assessments using the LDA-based TODIM. The authors conclude that, for their case study, the risk priority obtained through their method is more robust than FMEA based on Rough TOPSIS and Rough Vikor.

Recently, Wang et. All in [15] proposed a hybrid FMEA framework integrating the "Gained and lost dominance score" GLSD method, Choquet integral [16], and "cloud model" for risk analysis of potential failure modes. This methodology was used to prioritize the risk of 20 failure modes in the machine tools of the manufacturing industry. When comparing the proposed approach with others like a generalized TODIM method combined with Choquet integral and a cloud model-based TOPSIS, the authors verified that their method "is more reasonable and reliable than other extended FMEA frameworks."

In [17], the authors propose a methodology to evaluate and classify failure modes into different risk classes instead of concentrating on a global prioritization of the failure modes. Their methodology is based on a combination of Hesitant Uncertain Linguistic Z Numbers (HULZNs) [18] used to represent the FMEA experts' information and a modified Density-Based Spatial Clustering of Applications with Noise (DBSCAN) to cluster the failure modes. The authors applied the approach to a case study of twenty-five failure modes associated with a geothermal power plant and then compared it with the conventional FMEA based on RPN and the Fuzzy Vikor; they conclude that their proposed methodology is adequate to establish the criticality of the failure modes. A similar approach, also based on clustering analysis, was presented by Liu and Li in [19]. They use the K-means clustering [20] to classify the FMEA experts, a weighting scheme for the experts' risk attitudes, and a prioritization scheme based on a combination of regret theory (RT) and the PROMETHEE II method. The case study in this paper contains eleven failure modes related to the cold chain green logistics risk assessment problem. It considers the criteria of fourteen professional experts as part of the analysis. Their proposed method was then compared with the classical FMEA, TOPSIS-based FMEA, VIKOR-based FMEA, MABAC-based FMEA, COPRAS-based FMEA, MARCOS-based FMEA, RT-based FMEA, and PROMETHEE-based FMEA. The authors concluded that the proposed method is better than other FMEA methods for validity and reliability.

In [21], the authors propose the application of an FMEA based on Fuzzy Rules Base (FRB) [22] and Grey Relational Theory (GRT) [23] to improve the failure modes prioritization. Each proposed method was applied to a 500 Km pipeline system, identifying 27 failure modes. Authors conclude that their FRB-based approach is useful when only the relative ranking between failure modes is important. Their GRT-based approach can be used when the contribution of each risk factor is necessary for the optimal decision-making regarded resource allocation.

Liu et al. in [4] proposed the application of interval 2-tuple hybrid weighted distance (ITHWD) in an FMECA analysis conducted on a blood transfusion problem. In their work, eleven failure modes with RPN higher than 80 were selected to apply the ITHWD approach. The results proved to be a useful way to prioritize the failure modes in the presence of uncertainty and incomplete information. In [24], a similar approach also represented the uncertainty using interval type-2 fuzzy sets to rank failure modes, which is now in a real oil spill incident. Based on five experts, their criteria were aggregated considering a rule-based approach, with the final fuzzy set subsequently defuzzified to find the RPN value.

Reference [25] shows the application of type-2 fuzzy-based FMECA in the risk assessment of manufacturing facilities in the automotive industry. The paper includes a fuzzy extension of the ordered weighted average (OWA) to assign an importance level to each fuzzy risk factor. Although the proposed method is limited to triangular membership functions, the suggested approach offers additional flexibility to the experts in making judgments. It provides better modeling of uncertainty in terms of intra and interpersonal uncertainty.

In a more recent work [26], Qin and Pedrycz developed an approach that combines interval type-2 fuzzy sets and evidential reasoning applied to FMECA analysis of a steam valve system, considering eight failure modes. The methodology was revealed to be more precise than conventional methods like fuzzy VIKOR and fuzzy TOPSIS, reducing the probability of producing the same RPN. The weighting scheme applied to the three risk factors has made the result more comprehensive and capable of better expressing the uncertainty than type-1 fuzzy methods.

Anes et al. [27] show an FMECA approach based on two mathematical functions: the first deals with cases where the order of importance of risk variables is sufficient to prioritize failure modes. The second set of functions is an extension of the first one and considers each variable's relative weight. Here their approach was applied to the blood transfusion problem analyzed in [4] and compared with other fuzzy-based methods. The results indicate that the proposed risk isosurface method has a good potential to prioritize failure modes according to their risk.

Despite the strength of the above-described methods to improve the failure modes' prioritization in FMECA analysis, their efficacy is usually difficult to assess. Commonly, the efficacy of new FMECA methodologies is evaluated qualitatively by comparing one-to-one the rankings obtained by each method and assessing the *agreement* between methods.

When the number of failure modes is small, this qualitative approach is appropriate. However, a qualitative analysis becomes unpractical or sometimes leads to misleading results when one has to consider a great number of failure modes.

The agreement coefficients appear as a suitable statistical-based mean to determine the effectiveness of new prioritization methods in the FMECA context. The next section shows the theoretical aspects of the rank agreement problem and one of the most used agreement coefficients, the Cohen's kappa.

II. IMPORTANCE OF MEASURING THE AGREEMENT BETWEEN DIFFERENT FMECA METHODS

Agreement assessment between rankings is a well-known problem in biological and social sciences. However, the application of agreement assessment metrics in the FMECA context is limited and directed toward evaluating FMECA team members' agreement instead of different FMECA methodologies. For example, in [28] the authors applied Kendall's coefficient to determine the agreement between human experts in the medical risk analysis context. In [29], the authors show the application of FMECA's web-based three-round Delphi technique in the context of risk assessment related to the transition from paper-based records to digital-based records in the radiotherapy department; the Wilcoxon matched-pairs signed-ranks test and Kendall's coefficient of concordance were used to establishing the consensus between the FMECA's risk factor. In [30], instead of Kendall's concordance coefficient, authors have used the Cohen's kappa coefficient for the agreement assessment between raters and their risk factors evaluation. In that work, the authors developed a knowledge-based approach to improve the classical FMEA in the context of vehicle components. Two raters conducted the analysis, and the Cohen's kappa coefficient was used to evidence their level of agreement

In that context, this work is inserted when the application of agreement assessment metrics has been limited to evaluating FMECA team members instead of different FMECA methodologies. Our main goal is to introduce a methodology to conduct a statistical-based metrics comparison between different FMECA approaches as a superior alternative to the usual qualitative comparison.

The paper is organized as follows. Section III. explains the main concepts of the rank agreement problem. Section IV. introduces the use of the concordance coefficient in the FMECA context. Section V. shows the FMECA case study, Section VI. describes the implementation of fuzzy-based FMECA methods. Section VII. shows the results of the application of the proposed approach and the results of agreement between FMECA methods. Section VIII. . discusses the obtained results, and Section IX. shows the paper's conclusion and future developments.

III. REVISING THE MAIN CONCEPTS FOR RANK AGREEMENT PROBLEMS

A. The measurement of rank agreement

Consider a collection of n objects classified by a particular characteristic. Let m a finite number of judges or evaluators who made the rank of the n objects according to their appreciation of the objects' characteristics. It can be important to know the degree of agreement between the evaluators' decisions. This kind of problem has usually been known as *the problem of m ranking*, as originally stated by Kendall and Smith in [31]. They define it as: "If m persons rank n objects according to some quality of the objects, there arises the problem: does the set of m rankings of n show any evidence of a community of judgment among the m individuals?" [31]. The community of judgment is usually called an *agreement*.

The *agreement*, also known as *concordance*, *reproducibility* [32], or *interrater reliability* [33], is a concept closely related to but fundamentally different from correlation, as well asserted in [32–35]. The *agreement* can be defined as "the *degree of concordance* between two or more sets of measurements" [36], but it is common for both terms to be used as synonyms.

The existence of agreement implies correlation, but the reciprocal may not be true, as shown in [37]. Correlation statistics are usually applied to represent the association between two or more variables that do not necessarily measure the same attribute. In contrast, agreement statistically describes the concordance measure in individuals' opinion regarding the same attribute or characteristic [32]. Therefore, the concordance or agreement can be measured between a pair or several raters. To give the reader a broader perspective concerning the different types of coefficients of agreement, reference [33] was included since it contains an exhaustive analysis of some coefficients of agreement currently used in social and biological sciences: Cohen's kappa, Scott's Pi, Krippendorf's Alpha, Gewt's AC_1, Aicking's α, Cronbach Alpha, Kendall's Tau, among others.

The Cohen's Kappa coefficient was selected to conduct the concordance analysis due to its simple formulation and it is a well proven indicator for sixty years. The next sections show details about the kappa coefficient in its unweighted and weighted version.

B. Cohen's coefficient of agreement

Cohen's coefficient, usually known as Cohen's kappa and denoted by κ, is a statistic useful for inter-rater or intra-rater reliability measures [38,39]. Cohen's coefficient compares the proportion of objects the raters agreed with and the proportion of objects for which disagreement is expected [38]. Cohen's coefficient was originally proposed to measure the agreement between two raters. However, it can be extended for more than two raters, as shown in [39]. Following, we resume how Cohen's coefficient is computed to give us a quantitative measure of concordance between a set of raters.

Let N objects $n = 1, 2, \cdots, N$, be classified independently into k categories by two separated and independent raters, observers or judges, called A and B, as shown in Table I. For example, Object 1 was rated as Category 5 by Rater A and Category 3 by Rater B. The categories can represent an intrinsic characteristic of the classified objects or a single ordinal ranking from 1 to *kk*.

Let p_{ij} be the proportion of objects that Rater A classified in the category *i*, $i = 1, 2, \cdots, k$, and Rater B classified in the category *j*, $j = 1, 2, \cdots, k$, respectively. Table II shows the proportion of classified objects between the two raters.

TABLE I
EXAMPLE OF *N* OBJECTS RANKED BY TWO RATERS

Objects	Rater A	Rater B
Object 1	Category 5	Category 3
Object 2	Category 2	Category *k*
⋮	⋮	⋮
Object n	Category *k*	Category 5
⋮	⋮	⋮
Object N	Category 1	Category 1

TABLE II
THE PROPORTION OF CLASSIFIED OBJECTS

		Rater B						
	Cat	1	2	…	*j*	…	*k*	Total
Rater A	1	p_{11}	p_{12}	…	p_{1j}	…	p_{1k}	p_{1+}
	2	p_{21}	p_{22}	…	p_{2j}	…	p_{2k}	p_{2+}
	⋮	⋮	⋮		⋮		⋮	⋮
	i	p_{i1}	p_{i2}	…	p_{ij}	…	p_{ik}	p_{i+}
	⋮	⋮	⋮		⋮		⋮	⋮
	k	p_{k1}	p_{k2}	…	p_{kj}	…	p_{kk}	p_{k+}
	Total	p_{+1}	p_{+2}	…	p_{+j}	…	p_{+k}	1

Proportions p_{i+} and p_{+j} appear in the last column and line in Table II, respectively. Here, the symbol + represents summation over the index, and p_{i+} and p_{+j} are the frequencies or marginal probabilities for an object to be assigned into category *i* for Rater A and category *j* for rater B, as shown in [33,40]. The p_{i+} and p_{+j} values can be computed by (2) and (3), repectively:

$$p_{i+} = \sum_{j=1}^{k} p_{ij}\ , \tag{2}$$

$$p_{+j} = \sum_{i=1}^{k} p_{ij}\ , \tag{3}$$

where $\sum_{i=1}^{k} p_{i+} = 1$ and $\sum_{j=1}^{k} p_{+j} = 1$.

Let p_0 be the "observed" proportion of agreement between raters [38] and expressed by (4) [33,40]:

$$p_0 = \sum_{i=1}^{k} p_{ii}\ . \tag{4}$$

One characteristic of p_0 is that it does not take into account the agreement obtained only by chance (this means not really "agreeing" at all) [41]. Therefore, to obtain the expected proportion of agreement by chance, denoted by p_e , one uses equation (5). It is based on the probability that Rater A assigns the objects in category *i* in general, and the probability that Rater B assigns the objects in the same category also in general. As a consequence, for all $i = j$, the probability p_e is computed as in (5) [33]:

$$p_e = \sum_{i=1}^{k} \left(p_{i+} \cdot p_{+i} \right) . \tag{5}$$

So, Cohen's κ coefficient can be defined as (6), as it was originally proposed in [33,38,40]:

$$\kappa = \frac{p_0 - p_e}{1 - p_e} . \tag{6}$$

The lower and upper limits for κ are -1 and 1, respectively, but usually, its value falls between 0 and 1 [41]. When the observed agreement p_0 is greater than the agreement expected by chance p_e , the coefficient κ takes positive values. When the observed agreement p_0 is less than the agreement expected by chance p_e , κ takes negative values [38].

Cohen's κ coefficient equals unity $\kappa = 1$ when (and only when) there is a *perfect agreement* between raters. For perfect agreement, there is a necessary condition where $p_{i+} = p_{+j}$ [38]. On the other side, it becomes null $\kappa = 0$ when the observed agreement is no better than that expected by chance as if the raters had *guessed* every rating [41].

Cohen's κ coefficient becomes negative $\kappa < 0$ when the *agreement "measured" is worse* than expected by chance. Because the upper limit of κ is 1, values less than 0 likely mean *poor agreement* [38].

It is important to point out that the Cohen's coefficient does not indicate whether the disagreement is due to random or systematic differences between raters [41]. Hence, as shown in Table III, the value of κ can be interpreted using labels assigned for different ranges to express the strength of agreement, as proposed in [34] .

In some circumstances, the original Cohen's κ coefficient produces unexpected results, being in these cases referred to in the literature as the *kappa paradoxes* [33]. These paradoxes have related to using marginal probabilities to quantify the expected agreement by chance p_e . As indicated in [33], two main sources of paradoxes can be pointed out:

1) When an expected observed agreement exists, an unexpected larger value appears p_e . The definition of κ in (6) shows that, for a fixed value of p_o , the smaller the value of p_e , the greater the value of κ. This occurs because p_e represents the "agreement expected by chance," a higher unexpected value p_e can mean that a

large number of objects were classified only by chance by raters.

2) The second source of paradoxes occurs if the marginal proportions p_{i+} and p_{+j} are imbalanced. In this case, this can produce higher or lower values of κ according to the symmetry of the imbalance.

As stated in [33], applying weights on the original κ coefficient overcomes the paradoxes. In this way, we introduce follow the weighted version of Cohen's κ coefficient.

TABLE III
LABELS FOR COHEN'S κ COEFFICIENT IN TERMS OF THE STRENGTH OF AGREEMENT

κ range	Strength of agreement
$\kappa < 0.00$	Poor agreement
$0.00 < \kappa \leq 0.20$	Slight agreement
$0.20 < \kappa \leq 0.40$	Fair agreement
$0.40 < \kappa \leq 0.60$	Moderate agreement
$0.60 < \kappa \leq 0.80$	Substantial agreement
$0.80 < \kappa \leq 1.00$	Almost perfect agreement

C. Cohen's weighted kappa

The development of Cohen's weighted kappa coefficient, denoted by κ_w, was motivated by the "appearance of some disagreements in assignments. That is, some off-diagonal cells in the $k \times k$ matrix (Table II) have greater significance than others" [42]. In resume, the weighted version of Cohen's κ coefficient allows to avoid unexpected results or the so-called *kappa paradoxes*.

Let w_{ij} be the weight for agreement assigned to the $i^{th} - j^{th}$ cell of Table II. The weighted kappa coefficient becomes defined by (7) [43]:

$$\kappa_w = \frac{p_o^w - p_e^w}{1 - p_e^w}, \tag{7}$$

where p_o^w is the weighted version of the observed proportion of agreement between raters, being defined by (8), as stated in [43].

$$p_0^w = \sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij} p_{ij}, \tag{8}$$

Term p_e^w in (7) is the weighted version of the expected proportion of agreement obtained by chance, which p_e^w is defined by (9).

$$p_e^w = \sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij} p_{i+} p_{+j}, \tag{9}$$

Considering Eq. (4), the unweighted kappa can be interpreted as a special case of weighted kappa when all disagreements are given the same weight or the value 1 [42,43].

Weights can be assigned using any judgment procedure. In many instances, they may result from a consensus of a committee of substantive experts [42]. In [33], the author proposed six weighting schemes for κ_w. Nevertheless, the linear and quadratic weighting schemes are the most applied in κ_w calculation [40–47].

The linear weighting scheme, denoted by $w^{(1)}$, considers the absolute value of the difference between categories *i* and *j*, defined by (10) [44].

$$w_{ij}^{(1)} = 1 - \frac{|i-j|}{n-1}, \tag{10}$$

The quadratic weighting scheme considers the squared difference between categories *i* and *j*, which is defined by (11) as proposed in [44]:

$$w_{ij}^{(2)} = 1 - \left(\frac{i-j}{n-1}\right)^2, \tag{11}$$

D. Cohen's weighted kappa test of significance

Let H_0 be the null hypothesis stated as *raters' agreement is no better than the agreement expected by chance*. Let H_1 be the alternative hypothesis stated as *raters' agreement is better than the agreement expected by chance*. The probability distribution of κ_w can then be approximated by a normal distribution, as stated in [44]. The estimated variance $\hat{\sigma}$, when there is no association between raters' assignments, that is, when the agreement is no better than the agreement expected by chance (H_0), can be calculated using Eq. (12) [48]:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{k}\sum_{j=1}^{k}\left(p_{i+}\, p_{+j}\left[w_{ij}\left(\overline{w}_{i+} + \overline{w}_{+j}\right)^2\right]\right) - p_e^2}{n\left(1-p_e\right)^2}, \tag{12}$$

where $\overline{w}_{i+} = \sum_{j=1}^{k} w_{ij} p_{+j}$ represents the weighted average of the weights in the i^{th} row in Table II and $\overline{w}_{+j} = \sum_{i=1}^{k} w_{ij} p_{i+}$ represents the weighted average of the weights in the j^{th} column.

Assuming that $\kappa_w / \hat{\sigma}$ follows a normal distribution, it is possible to test the hypothesis of agreement expected by chance by referencing the standard normal distribution. The test statistics is thus defined by (13):

$$z = \frac{\kappa_w}{\hat{\sigma}}. \tag{13}$$

The null hypothesis H_0 is rejected for a one-tailed test if the test statistic $|z| \geq z_\alpha$, where z_α is the critical value that leaves the alpha-level α in the upper tail of the standard normal distribution. In this work, the level of significance was selected as $\alpha = 0.05$ and the critical value $z_\alpha = 1.645$ [49]. If the test statistics results in $|z| \geq 1.645$, the null hypothesis will be rejected.

IV. THE MEASUREMENT OF RANK AGREEMENT IN THE FMECA CONTEXT

Inter-rater reliability was originally aimed to measure the agreement between human raters. However, in the last decade, inter-rater reliability methods have been applied to compare the efficiency of algorithms generally used for classification.

In [50], the authors used Cohen's κ coefficient to measure the performance between classifiers in a wrapper feature selection. Their proposed a fuzzy optimization procedure considered Cohen's κ maximization as a performance criterion, which improved the stability of the selection method by punishing any agreement obtained by chance. Similarly, the authors in [51] used machine learning-based classifiers to identify likely untreated sewage spills from wastewater treatment plants. In their problem, the agreement between different models was also assessed using Cohen's κ coefficient.

Now, in the context of Android malware detection [52], six supervised machine learning classifiers were used. Again, the criteria considered to evaluate the classifiers performance in supply correct or incorrect predictions was Cohen's κ as their performance metric. More recently, in [53], the proposal of a deep learning algorithm for classifying the severe acute respiratory syndrome coronavirus 2 (SARS CoV-2) amongst coronaviruses is presented. Cohen's κ is used among another six metrics to measure the performance between different neural network-based models.

Another recent application of Cohen's κ *coefficient* to compare algorithms is presented in [54], where the author has used machine learning-based classification models in a complex experimental physics data analysis. Cohen's κ is used as a performance metric to compare the crisp and fuzzy-based algorithms.

The papers cited in the previous paragraph have shown the potential of Cohen's κ to evaluate the agreement between machine learning-based algorithms in different applications. For this reason, we consider this coefficient as having great suitability to compare the agreement between computational algorithms used now to improve the prioritization of failure modes in the FMECA context.

Cohen's κ *coefficient* can be applied but considering the following assumptions:

1) The FMECA problem contains n failure modes to be ranked. Hence, these failure modes will represent the n classified objects in inter-rater reliability terms;
2) The FMECA methods used to improve the failure modes prioritization can be considered as the m independent raters in inter-rater reliability terms;
3) Each algorithm classifies the failure modes from the highest risk (priority 1) to the lowest risk (priority $k=n$). In this context, we can consider each priority classification as the k categories in inter-rater reliability terms.

We identified two possible approaches to apply Cohen's kappa to assess the agreement between different FMECA methods:

1) Assess the agreement between all the m FMECA methods when applied to the same problem and without considering a reference one, and;
2) Assess the one-to-one agreement between the m FMECA methods when applied to the same problem and an FMECA method selected as the reference.

The second approach allows evaluating the effectiveness of the FMECA methods when a reference is available. We consider it very helpful to assess the efficacy of new FMECA methods. However, the main concern with this approach is the selection of an FMECA methodology as the reference.

To illustrate the application of agreement coefficients in the FMECA context, we show the simple computation of Cohen's kappa for two FMECA rankings reported in [27], RPI(SC_4) and RPI(SC_5), which rankings are listed in Table IV, including the classical RPN. In terms of interrater reliability, $N=11$ objects (failure modes), $k = 11$ categories (rankings) and $m = 2$ raters (FMECA methods, RPI(SC_4) and RPI(SC_5).

TABLE IV
RANKINGS FOR FMECA METHODS RPI(SC_4) AND RPI(SC_5)

Failure mode	RPN Rank	RPI(SC_4)	RPI(SC_5)
FM1	5	4	5
FM2	4	5	7
FM3	1	2	4
FM4	8	7	9
FM5	9	11	11
FM6	10	6	3
FM7	10	9	6
FM8	2	1	1
FM9	6	8	8
FM10	3	3	2
FM11	6	10	10

Table V shows the proportion of failure modes rated in each ranking by methods RPI(SC_4) and RPI(SC_5), and marginal proportions p_{i+} computed using (2) and p_{+j} computed as in (3). We have chosen the quadratic weighting scheme for this example, with the weights between rankings computed as shown in (8). Table VI shows the quadratic weights computed. For example, the quadratic weight $w_{35}^{(2)}$ between ranking $i = 3$ and ranking $j = 5$ was obtained by applying (8) as:

$$w_{35}^{(2)} = 1-\left(\frac{3-5}{11-1}\right)^2 = 1-\left(\frac{-2}{10}\right)^2 = 1-\left(-0.2\right)^2 = 0.96$$

That is, the weight for category 3 and category 5 is 0.96, as shown in Table VI (row 3 and column 5).

The observed proportion of agreement between raters p_o^w was computed as in (5), corresponding to the sum of the respective multiplication between elements of Table V and Table VI as follows:

$$\begin{aligned} p_0^w &= 0.0909\times 1+0.0909\times 0.96+0.0909\times 0.99+0.0909\times 0.99 \\ &+0.0909\times 0.96+0.0909\times 0.91+0.0909\times 0.96+0.0909\times 1 \\ &+0.0909\times 0.91+0.0909\times 1+0.0909\times 1 = 0.9709 \end{aligned}$$

TABLE V
THE PROPORTION OF FAILURE MODES RANKED BY RPI(SC$_4$) AND RPI(SC$_5$).

		RPI(SC$_5$)											
	Ranking	1	2	3	4	5	6	7	8	9	10	11	p_{i+}
RPI(SC$_4$)	1	0.0909	0	0	0	0	0	0	0	0	0	0	0.0909
	2	0	0	0	0.0909	0	0	0	0	0	0	0	0.0909
	3	0	0.0909	0	0	0	0	0	0	0	0	0	0.0909
	4	0	0	0	0	0.0909	0	0	0	0	0	0	0.0909
	5	0	0	0	0	0	0	0.0909	0	0	0	0	0.0909
	6	0	0	0.0909	0	0	0	0	0	0	0	0	0.0909
	7	0	0	0	0	0	0	0	0	0.0909	0	0	0.0909
	8	0	0	0	0	0	0	0	0.0909	0	0	0	0.0909
	9	0	0	0	0	0	0.0909	0	0	0	0	0	0.0909
	10	0	0	0	0	0	0	0	0	0	0.0909	0	0.0909
	11	0	0	0	0	0	0	0	0	0	0	0.0909	0.0909
	p_{+j}	0.0909	0.0909	0.0909	0.0909	0.0909	0.0909	0.0909	0.0909	0.0909	0.0909	0.0909	1

The observed proportion of agreement between raters p_o^w was computed as in (5), corresponding to the sum of the respective multiplication between elements of Table V and Table VI as follows:

$$p_0^w = 0.0909\times1+0.0909\times0.96+0.0909\times0.99+0.0909\times0.99$$
$$+0.0909\times0.96+0.0909\times0.91+0.0909\times0.96+0.0909\times1$$
$$+0.0909\times0.91+0.0909\times1+0.0909\times1 = 0.9709$$

The expected proportion of agreement obtained by chance p_e^w was computed by (6) considering the values and weights from Table VI as follows:

$$p_e^w = w_{11}p_{1+}\ p_{+1} + w_{12}p_{1+}\ p_{+2} + w_{13}p_{1+}\ p_{+3} + w_{14}p_{1+}\ p_{+4}$$
$$+ w_{15}p_{1+}\ p_{+5} + \cdots + w_{11\,10}p_{11+}\ p_{+10} + w_{11\,11}p_{11+}\ p_{+11}$$

The value for weighted kappa is computed as in (4):

$$\kappa_w = \frac{0.9709-0.800}{1-0.800} = 0.8545$$

The agreement equals to 0.8545 between FMECA methods RPI(SC$_4$) and RPI(SC$_5$) can be interpreted qualitatively as an agreement "almost perfect," as suggested in Table III.

TABLE VI
QUADRATIC WEIGHTS FOR THE EXAMPLE.

Ranking	1	2	3	4	5	6	7	8	9	10	11
1	1	0.99	0.96	0.91	0.84	0.75	0.64	0.51	0.36	0.19	0
2	0.99	1	0.99	0.96	0.91	0.84	0.75	0.64	0.51	0.36	0.19
3	0.96	0.99	1	0.99	0.96	0.91	0.84	0.75	0.64	0.51	0.36
4	0.91	0.96	0.99	1	0.99	0.96	0.91	0.84	0.75	0.64	0.51
5	0.84	0.91	0.96	0.99	1	0.99	0.96	0.91	0.84	0.75	0.64
6	0.75	0.84	0.91	0.96	0.99	1	0.99	0.96	0.91	0.84	0.75
7	0.64	0.75	0.84	0.91	0.96	0.99	1	0.99	0.96	0.91	0.84
8	0.51	0.64	0.75	0.84	0.91	0.96	0.99	1	0.99	0.96	0.91
9	0.36	0.51	0.64	0.75	0.84	0.91	0.96	0.99	1	0.99	0.96
10	0.19	0.36	0.51	0.64	0.75	0.84	0.91	0.96	0.99	1	0.99
11	0	0.19	0.36	0.51	0.64	0.75	0.84	0.91	0.96	0.99	1

V. FMECA CASE STUDY

The selected FMECA case study corresponds to a risk assessment in the blood transfusion process analyzed using the classical FMECA in [55] and subsequently analyzed using fuzzy-based and MCDM-based FMECA approaches in [4,6,27,56]. According to [55], 19 failure modes were originally identified, and the 11 failure modes with RPN higher than 80 were selected for further analysis. We considered two main reasons to choose this FMECA case study to apply the proposed agreement assessment approach:

1) This study case was already used for benchmarking in some studies, and;
2) Because the study case has only 11 failure modes, comparing the different methods used to improve the FMECA prioritization in terms of the influence of the risk factors becomes more intuitive.

Table VII shows the FMECA analysis for the case study, including the ranking obtained using the classical RPN [55]. The classical FMECA shortcoming related to failure modes with identical risk factors and RPN is evident between failure modes FM9 and FM11, which have the same RPN equal to 112 and both ranked as priority 6; the shortcoming related to failure modes with different risk factors but identical RPN is evident between and failure modes FM6 and FM7, having RPN 80 and both ranked as priority 10.

TABLE VII
CLASSICAL FMECA TABLE FOR THE CASE STUDY ABOUT RISK ASSESSMENT IN THE BLOOD TRANSFUSION PROCESS [46].

Failure mode	Failure mode	S	O	D	RPN	RANK
FM1	Insufficient and/or incorrect clinical information on request form	7	6	3	126	5
FM2	Blood plasma abuse	6	6	5	180	4
FM3	Insufficient preoperative assessment of the blood product requirement	7	5	7	245	1
FM4	Blood group verification incomplete	7	5	3	105	8
FM5	Delivery of blood sample and/or request form delayed	5	3	6	90	9
FM6	Incorrect blood components issued	10	1	8	80	10
FM7	Quality checks not performed on blood products	8	2	5	80	10
FM8	Preparation time before infusion >30 min	8	6	5	240	2
FM9	Transfusion cannot be completed within the appropriate time	7	4	4	112	6
FM10	Blood transfusion reaction occurs during the transfusion process	8	4	7	224	3
FM11	Bags of blood products are improperly disposed of bags	7	4	4	112	6

This case study was analyzed in the literature using different methods to improve the prioritization of failure modes. Table VIII shows the ranking using the classical FMECA analysis [55], the FMECA ranking using Fuzzy VIKOR [57], the ranking using interval 2-tuple hybrid weighted distance measure [4], and the FMECA RPI(SC_4) and RPI(SC_5) [27]. Since our proposed approach requires an FMECA ranking tobe considered as the reference one, we selected the FMECA method denoted as RPI(SC_4) as the reference. The reasons that lead us to select this method as the reference one are based on the authors' claims described in [27]:

1) The selected FMECA, RPI(SC_4), does not require additional previous knowledge about the problem;
2) The method considers a weighting importance related to the risk factors, and;
3) The failure modes prioritization agrees with the expectation made for the risk scenario.

VI. TYPE-II FUZZY-BASED FMECA METHODS

In addition to the FMECA methods listed in Table VIII, this paper proposes the application of a Mamdani Type-II Fuzzy Inference System to improve the classical FMECA, and a comparison with the reference FMECA method RPI(SC_4) by using the Cohen's kappa.

Five categories were initially attributed for each risk factor, as shown in Table IX, each represented by a Type-II membership function and a Fuzzy Inference System FIS, described in the following subsection.

TABLE VIII
RANKINGS FOR DIFFERENT FMECA IMPROVEMENT METHODS APPLIED TO THE CASE STUDY IN [46].

Failure mode	RPN Rank	Fuzzy VIKOR	ITHWD	RPI(SC_4)	RPI(SC_5)
FM1	5	4	4	4	5
FM2	4	7	6	5	7
FM3	1	2	1	2	4
FM4	8	8	10	7	9
FM5	9	11	11	11	11
FM6	10	1	3	6	3
FM7	10	6	9	9	6
FM8	2	5	5	1	1
FM9	6	10	7	8	8
FM10	3	3	2	3	2
FM11	6	9	8	10	10

TABLE IX
CATEGORIES FOR THE RISK FACTORS IN THE FUZZY FMECA

Severity category	Occurrence Category	Detection Category	RPN Category	Rating
Hazardous – SHA	Frequent – OF	Absolutely impossible – DAI	Extreme – RE	9,10
Very High – SVH	Probable – OP	Low – DL	High – RH	7, 8
Moderate – SM	Occasional – OO	Moderate – DM	Moderate – RM	4, 5, 6
Low – SL	Very unlikely – OVU	High – DH	Low – RL	2, 3
Minor – SMI	Remote – OR	Almost certain – DAC	Minor – RMI	1

A. Membership functions of Type-II Fuzzy Inference System

We considered four types of membership functions: triangular, trapezoidal, g-bell, and Gaussian membership functions.

The Type-II triangular membership function is represented by its upper limit, *triU*, and its lower limit, *triL*. The upper triangular membership function is defined in terms of parameters a^+, $b^{+,}$ and c^+ as shown in (9) [58]. The lower triangular membership function is defined in terms of parameters a^-, b^-, c^- and an additional term called *scl* that represents the maximum membership value, as suggested in [58] and shown in (10),

$$triU\left(x;a^+,b^+,c^+\right)=\begin{cases}0, & x<a^+\\ \left(x-a^+\right)/\left(b^+-a^+\right), & a^+\le x\le b^+\\ \left(c^+-x\right)/\left(c^+-b^+\right), & b^+\le x\le c^+\\ 0, & x>c^+\end{cases}, \tag{9}$$

$$triL\left(x;scl,a^{-},b^{-},c^{-}\right)=\begin{cases}0, & x<a^{-}\\ scl\cdot\left(x-a^{-}\right)/\left(b^{-}-a^{-}\right), & a^{-}\leq x\leq b^{-}\\ scl\cdot\left(c^{-}-x\right)/\left(c^{-}-b^{-}\right), & b^{-}\leq x\leq c^{-}\\ 0, & x>c^{-}\end{cases}, \quad (10)$$

Table X lists the parameters for each triangular membership function representing the fuzzy categories of the three risk factors and those associated with the Fuzzy RPN. Fig. 1 shows the Type-II triangular membership functions defined in Table X.

The Type-II trapezoidal membership function is represented by its upper limit, *trapU*, and its lower limit denoted by *trapL*. The upper trapezoidal membership function is defined in terms of parameters a^{+}, b^{+}, c^{+}, and d^{+}, as shown in (11) [58]. The lower trapezoidal membership function is defined in terms of parameters a^{-}, b^{-}, c^{-}, d^{-}, and an additional term called *scl* that represents the maximum membership value, as shown in (12) and suggested in [58].

$$trapU\left(x;a^{+},b^{+},c^{+},d^{+}\right)=\begin{cases}\left(x-a^{+}\right)/\left(b^{+}-a^{+}\right), & a^{+}\leq x\leq b^{+}\\ 1 & b^{+}\leq x\leq c^{+}\\ \left(d^{+}-x\right)/\left(d^{+}-c^{+}\right), & c^{+}\leq x\leq d^{+}\\ 0, & otherwise\end{cases}, \quad (11)$$

$$trapL\left(x;scl,a^{-},b^{-},c^{-},d^{-}\right)=\begin{cases}scl\cdot\left(x-a^{-}\right)/\left(b^{-}-a^{-}\right), & a^{-}\leq x\leq b^{-}\\ scl & b^{-}\leq x\leq c^{-}\\ scl\cdot\left(d^{-}-x\right)/\left(d^{-}-c^{-}\right), & c^{-}\leq x\leq d^{-}\\ 0, & otherwise\end{cases}, \quad (12)$$

Table XI lists the parameters for the trapezoidal membership functions. These represent the categories of each risk factor and the Fuzzy RPN categories considered in this work. Fig. 2 shows the Type-II trapezoidal functions defined in Table XI.

The Type-II generalized bell membership function (g-bell) can be represented by its upper limit denoted by *gbellU* and its lower limit denoted by *gbellL*. The upper g-bell membership function can be defined in terms of the parameters a^{+}, b^{+} and c^{+} as shown in (13) [58], and the lower g-bell membership function can be defined in terms of the parameters a^{-}, b^{-}, c^{-} and an additional term called *scl* that represents the maximum membership value, as shown in (14) and suggested in [58].

$$gbellU\left(x;a^{+},b^{+},c^{+}\right)=\frac{1}{1+\left|\dfrac{x-c^{+}}{a^{+}}\right|^{2b^{+}}}, \quad (13)$$

$$gbellL\left(x;a^{-},b^{-},c^{-}\right)=\frac{1}{1+\left|\dfrac{x-c^{-}}{a^{-}}\right|^{2b^{-}}}, \quad (14)$$

Table XII shows the parameters for the g-bell membership functions for the three risk factors and the Fuzzy RPN considered in this work. Fig. 3 shows the Type-II g-bell membership functions defined in Table XII.

The Type-II gaussian membership function can be represented by its upper limit denoted by *gaussU* and its lower limit denoted by *gaussL*. The upper gaussian membership function can be defined using the parameters c^{+} and σ^{+} as shown in (15) [58]. The lower gaussian membership function can be defined in terms of the parameters c^{-}, σ^{-} and an additional term called *scl* representing the maximum membership value, as shown in (16) and suggested in [58].

$$gaussU\left(x;c^{+},\sigma^{+}\right)=e^{-\frac{1}{2}\left(\frac{x-c^{+}}{\sigma^{+}}\right)^{2}}, \quad (15)$$

$$gaussL\left(x;scl,c^{-},\sigma^{-}\right)=scl\cdot e^{-\frac{1}{2}\left(\frac{x-c^{-}}{\sigma^{-}}\right)^{2}}, \quad (16)$$

Table XIII shows the parameters for the gaussian membership functions that represent the risk categories of the three risk factors and the Fuzzy RPN considered in this work.

Fig. 4 shows the Type-II gaussian membership functions defined in Table XIII.

TABLE X

TRIANGULAR FUZZY MEMBERSHIP FUNCTIONS FOR THE TYPE-II FUZZY INFERENCE SYSTEM

Category	Severity	Occurrence	Detection	FuzzyRPN
9,10	*triU* (x;0,1.5,2.5)	*triU* (x;0,1.5,2.5)	*triU* (x;0,1.5,2.3)	*triU* (x;0.4,2.5,3.2)
9,10	*triL* (x;0.9,0.6,1.5,2.1)	*triL* (x;0.9,0.6,1.5,2.1)	*triL* (x;0.9,0.6,1.5,1.98)	*triL* (x;0.9,1.12,2.2,2.8)
7,8	*triU* (x;0.6,2.9,3.5)	*triU* (x;0.8,2.8,4.2)	*triU* (x;1.1,2.97,4.3)	*triU* (x;1.2,3.5,4.9)
7,8	*triL* (x;0.9,1.52,2.90,3.26)	*triL* (x;0.9,1.6,2.8,3.64)	*triL* (x;0.9,1.85,2.97,3.77)	*triL* (x;0.9,2.12,3.5,4.34)
4,5,6	*triU* (x;2.5,4.2,8.3)	*triU* (x;3.2,5.4,7.4)	*triU* (x;2.5,5,7.5)	*triU* (x;3.1,5.5,8.1)
4,5,6	*triL* (x;0.9,3.18,4.20,6.66)	*triL* (x;0.9,4.08,5.4,6.6)	*triL* (x;0.9,3.5,5.0,6.5)	*triL* (x;0.9,4.06,5.5,7.06)
2,3	*triU* (x;4.8,7.5,10.4)	*triU* (x;6.36,7.5,9.6)	*triU* (x;4.8,7.52,10.4)	*triU* (x;5.5,8,10.4)
2,3	*triL* (x;0.9,5.88,7.50,9.24)	*triL* (x;0.9,6.82,7.5,8.76)	*triL* (x;0.9,5.91,7.52,9.25)	*triL* (x;0.9,6.5,8.0,9.44)
1	*triU* (x;7.6,9.3,12.4)	*triU* (x;8.7,9.3,11.4)	*triU* (x;7.64,9.32,12.4)	*triU* (x;7.1,9.1,13.2)
1	*triL* (x;0.9,8.82,9.30,11.16)	*triL* (x;0.9,8.94,9.3,10.56)	*triL* (x;0.9,8.31,9.32,11.17)	*triL* (x;0.9,7.9,9.1,11.56)

TABLE XI

TRAPEZOIDAL FUZZY MEMBERSHIP FUNCTIONS FOR THE TYPE-II FUZZY INFERENCE SYSTEM

Category	Severity	Occurrence	Detection	FuzzyRPN
9,10	*trapU* (x;0.1,0.6,1.5,2.4)	*trapU* (x;0.1,0.5,1.1,2.7)	*trapU* (x;0,2,1.0,1.64,2.1)	*trapU* (x;1.0,1.0,1.64,2.5)
9,10	*trapL* (x;0.9,0.3,0.75,1.23,2.04)	*trapL* (x;0.9,0.26,0.62,0.62,2.06)	*trapL* (x;0.9,0.52,1.24,1.5,1.92)	*trapL* (x;0.9,1.0,1.0,1.38,2.16)
7,8	*trapU* (x;0.9,2.1,2.8,3.5)	*trapU* (x;1.2,1.9,3.1,4.7)	*trapU* (x;1.1,2.0,3.0,3.8)	*trapU* (x;0.84,2.41,3.2,4.1)
7,8	*trapL* (x;0.9,1.38,2.46,2.59,3.22)	*trapL* (x;0.9,1.48,2.11,2.62,4.06)	*trapL* (x;0.9,1.46,2.27,2.8,3.48)	*trapL* (x;0.9,1.47,2.88,2.9,3.74)
4,5,6	*trapU* (x;2.7,4.3,5.7,7.8)	*trapU* (x;3.4,3.9,6.1,7.2)	*trapU* (x;2.67,4.0,6.0,7.83)	*trapU* (x;2.9,4.21,5.5,7.6)
4,5,6	*trapL* (x;0.9,3.34,4.78,5.07,6.96)	*trapL* (x;0.9,3.6,4.05,5.77,6.76)	*trapL* (x;0.9,3.20,4.39,5.45,7.1)	*trapL* (x;0.9,3.42,4.6,4.87,6.76)
2,3	*trapU* (x;5.1,6.9,8.1,9.5)	*trapU* (x;6.3,6.9,8.3,9.3)	*trapU* (x;5.7,7.0,8.0,9.13)	*trapU* (x;5.5,7.0,8.0,9.5)
2,3	*trapL* (x;0.9,5.82,7.44,7.68,8.94)	*trapL* (x;0.9,6.54,7.08,8.0,8.9)	*trapL* (x;0.9,6.2,7.39,7.66,8.68)	*trapL* (x;0.9,6.1,7.45,7.55,8.90)
1	*trapU* (x;7.6,8.7,9.8,12.2)	*trapU* (x;8.1,8.9,9.9,11.2)	*trapU* (x;7.67,9.06,10.0,10.0)	*trapU* (x;7.67,9.06,10,10)
1	*trapL* (x;0.9,8.04,9.03,9.08,11.5)	*trapL* (x;0.9,8.42,9.14,9.51,10.7)	*trapL* (x;0.9,8.23,9.5,10.0,10.0)	*trapL* (x;0.9,8.23,9.5,10.0,10.0)

TABLE XII

G-BELL FUZZY MEMBERSHIP FUNCTIONS FOR THE TYPE-II FUZZY INFERENCE SYSTEM

Category	Severity	Occurrence	Detection	FuzzyRPN
9,10	*gbellU* (x;0.86,2.61,1.42)	*gbellU* (x;0.81,1.68,1.44)	*gbellU* (x;0.73,3.26,1.44)	*gbellU* (x;0.73,3.26,1.44)
9,10	*gbellL* (x;0.9,0.34,2.35,4.42)	*gbellL* (x;0.9,0.32,1.51,1.44)	*gbellL* (x;0.9,0.29,2.93,1.44)	*gbellL* (x;0.9,0.29,2.93,1.44)
7,8	*gbellU* (x;0.96,2.41,2.84)	*gbellU* (x;0.87,2.46,2.63)	*gbellU* (x;0.99,2.70,2.82)	*gbellU* (x;0.99,2.70,2.82)
7,8	*gbellL* (x;0.9,0.38,2.17,2.84)	*gbellL* (x;0.9,0.35,2.21,2.63)	*gbellL* (x;0.9,0.39,2.43,2.82)	*gbellL* (x;0.9,0.39,2.43,2.82)
4,5,6	*gbellU* (x;0.86,2.61,4.42)	*gbellU* (x;1.15,2.18,5.30)	*gbellU* (x;1.06,3.91,5.01)	*gbellU* (x;1.06,3.91,4.41)
4,5,6	*gbellL* (x;0.9,0.34,2.35,4.42)	*gbellL* (x;0.9,0.46,1.96,5.30)	*gbellL* (x;0.9,0.42,3.52,5.01)	*gbellL* (x;0.9,0.42,3.52,4.41)
2,3	*gbellU* (x;1.17,2.05,7.36)	*gbellU* (x;0.64,5.39,7.64)	*gbellU* (x;1.36,3.17,7.54)	*gbellU* (x;1.36,3.17,7.54)
2,3	*gbellL* (x;0.9,0.47,1.85,7.36)	*gbellL* (x;0.9,0.26,4.85,7.64)	*gbellL* (x;0.9,0.54,2.85,7.54)	*gbellL* (x;0.9,0.54,2.85,7.54)
1	*gbellU* (x;1.06,2.67,9.36)	*gbellU* (x;0.62,2.19,9.39)	*gbellU* (x;1.04,3.62,9.37)	*gbellU* (x;1.04,3.62,9.37)
1	*gbellL* (x;0.9,0.42,2.40,9.36)	*gbellL* (x;0.9,0.25,1.97,9.39)	*gbellL* (x;0.9,0.42,3.26,9.37)	*gbellL* (x;0.9,0.42,3.26,9.37)

TABLE XIII

GAUSSIAN FUZZY MEMBERSHIP FUNCTIONS FOR THE TYPE-II FUZZY INFERENCE SYSTEM

Category	Severity	Occurrence	Detection	FuzzyRPN
9,10	*gaussU* (x;1.44,0.62)	*gaussU* (x;1.44,0.43)	*gaussU* (x;1.5,0.51)	*gaussU* (x;1.30,0.62)
9,10	*gaussL* (x;0.9,1.44,0.28)	*gaussL* (x;0.9,1.44,0.19)	*gaussL* (x;0.9,1.50,0.23)	*gaussL* (x;0.9,1.30,0.28)
7,8	*gaussU* (x;2.84,0.82)	*gaussU* (x;2.63,0.74)	*gaussU* (x;3.5,0.84)	*gaussU* (x;2.80,0.84)
7,8	*gaussL* (x;0.9,2.84,0.37)	*gaussL* (x;0.9,2.63,0.33)	*gaussL* (x;0.9,3.5,0.38)	*gaussL* (x;0.9,2.80,0.38)
4,5,6	*gaussU* (x;4.42,0.86)	*gaussU* (x;5.3,0.98)	*gaussU* (x;5.10,0.90)	*gaussU* (x;5.10,0.91)
4,5,6	*gaussL* (x;0.9,4.42,0.39)	*gaussL* (x;0.9,5.3,0.44)	*gaussL* (x;0.9,5.10,0.41)	*gaussL* (x;0.9,5.10,0.41)
2,3	*gaussU* (x;7.54,0.85)	*gaussU* (x;7.64,0.54)	*gaussU* (x;7.54,0.85)	*gaussU* (x;8.11,0.85)
2,3	*gaussL* (x;0.9,7.54,0.38)	*gaussL* (x;0.9,7.64,0.24)	*gaussL* (x;0.9,7.54,0.38)	*gaussL* (x;0.9,8.11,0.38)
1	*gaussU* (x;9.36,0.85)	*gaussU* (x;9.39,0.53)	*gaussU* (x;9.5,0.85)	*gaussU* (x;9.01,0.85)
1	*gaussL* (x;0.9,9.36,0.38)	*gaussL* (x;0.9,9.39,0.24)	*gaussL* (x;0.9,9.5,0.38)	*gaussL* (x;0.9,9.01,0.38)

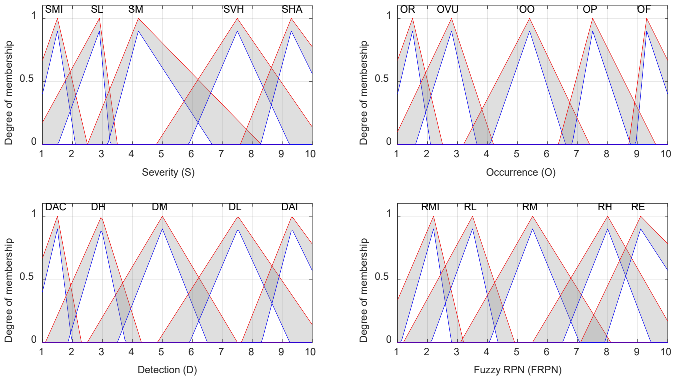

Fig. 1. Triangular Type-II fuzzy membership functions considered for the fuzzy-based FMECA.

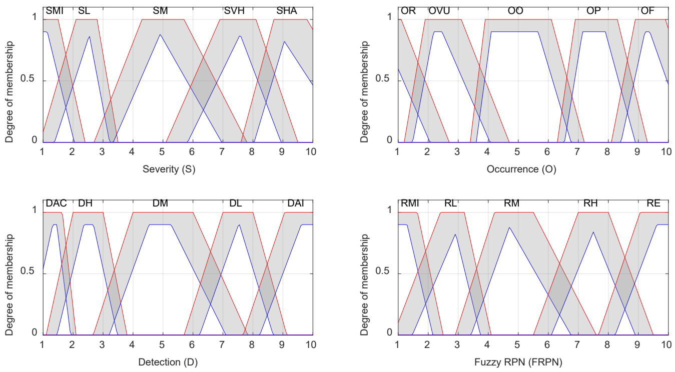

Fig. 2. Trapezoidal Type-II fuzzy membership functions considered for the fuzzy-based FMECA.

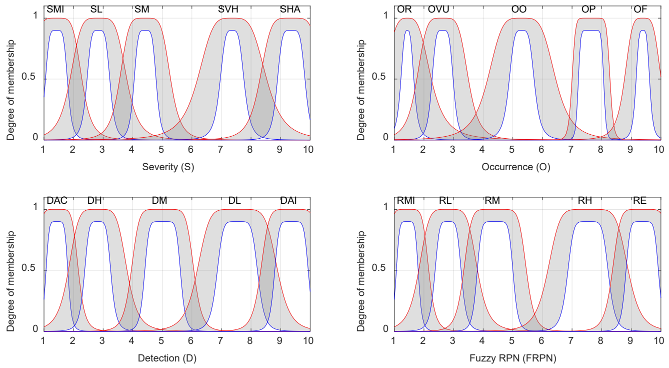

Fig. 3. G-bell Type-II fuzzy membership functions considered for the fuzzy-based FMECA.

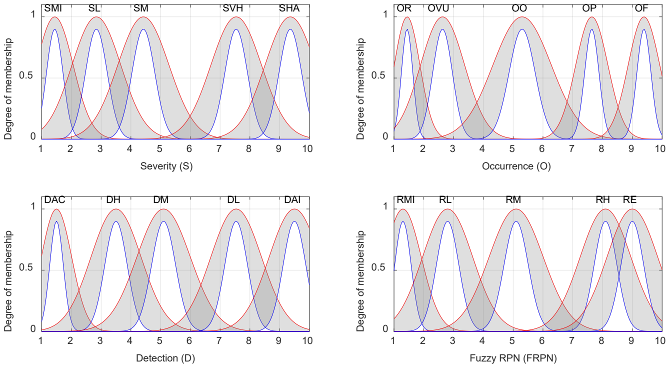

Fig. 4. Gaussian Type-II fuzzy membership functions considered for the fuzzy-based FMECA.

B. Rule base for the Type-II Fuzzy Inference System

In the present approach, the fuzzy rules are defined for all the possible combinations between the three risk factors O, S, and D in the rule's antecedent and RPN in the rule's consequent; the risk priority number will be named Fuzzy Risk Priority Number FRPN. Because each of the three risk criteria has five categories, 125 possible combinations (rules) of these categories exist. Each rule will have associated with its respective FRPN category.

The following set of rules is an example of the proposed fuzzy rules:

- *Fuzzy Rule 26*: If (**SEVERITY** is **SL**) and (**OCCURRENCE** is **OR**) and (**DETECTION** is **DAC**) then (**FRPN** is **MI**)
- *Fuzzy Rule 59*: If (**SEVERITY** is **EM**) and (**OCCURRENCE** is **OO**) and (**DETECTION** is **DL**) then (**FRPN** is **RM**)

C. Properties of the Type-II Fuzzy Inference System

The fuzzy inference system was implemented considering the following properties: FIS type Type-II Mandani; And method: Min; Or method: Max; Implication Method: Min; Aggregation Method: Max; Type reduction method: Karnik-Mendel; Defuzzification method: Centroid.

D. Type-II fuzzy FMECA cases

We defined four Type-II fuzzy FMECA cases, combining the membership functions triangular and gaussian shown in section VII.A., and considering the fuzzy rules described in section VII.B., and FIS properties described in VII.C.; Table XIV shows the combinations of membership functions for the four Type-II fuzzy FMECA

TABLE XIV
FMECA BASED ON TYPE-II FUZZY INFERENCE SYSTEM

Config	MFSEV	MFOCC	MFDET	MFRPN
T2-FIS 01	gaussian	triangle	triangle	triangle
T2-FIS 02	g-bell	triangle	trapezoidal	gaussian
T2-FIS 03	g-bell	triangle	trapezoidal	g-bell
T2-FIS 04	g-bell	trapezoidal	trapezoidal	gaussian
T2-FIS 05	trapezoidal	triangle	g-bell	g-bell
T2-FIS 06	trapezoidal	gaussian	g-bell	g-bell

VII. Results

Results are organized in two sections: Section VIII. A. shown the results for the linear weighted kappa and Section VIII. B. contains results for the quadratic weighted kappa. Each section contains the results of agreement between the reference ranking RPI(SC_4) and the following methods: 1) Fuzzy VIKOR, 2) ITHWD; 3) RPI(SC5); 4) Type-I Fuzzy Inference System, and 5) Type-II Fuzzy Inference System.

The end of the section compares the reference ranking and the ranking obtained by 5 methods with the highest concordance coefficient.

A. Results considering the linear weighted kappa

Table XV shows the linear weighted agreement coefficient κ_{w-lin}, the value of the test statistics z, the strength of agreement, and the hypothesis test result for the FMECA methods RPI(SC5), Fuzzy VIKOR, and ITHWD, when compared with the reference ranking RPI(SC4).

The computed κ_{w-lin} takes values from 0.55 to 0.65, and the scenario RPI(SC_5) achieves the better agreement with κ_{w-lin} equal to 0.65, which can be considered a substantial agreement according to the strength of agreement suggested in Table III

Regarding the hypothesis test, the critical value for the test statistics is $z_{0.05} = 1.645$. The null hypothesis H_0 *raters' agreement is no better than the agreement expected by chance* is rejected for all cases.

TABLE XV
LINEAR WEIGHTED KAPPA κ_{w-lin} FOR RPI(SC5), VIKOR, AND ITHWD

	RPI(SC_5)	Fuzzy VIKOR	ITHWD
κ_{w-lin}	0.65	0.55	0.6
Strength of agreement	Substantial	Moderate	Moderate
z	3.346	2.832	3.226
H_0 test	Reject	Reject	Reject

Table XVI shows the rankings and agreement results between RPI(SC_4) and the FMECA based on the six Type-II

FIS cases detailed in Table XIV. The best agreement coefficient value is 0.70 and corresponds to configurations T2-FIS-01 composed of membership functions type gaussian to represent severity and triangular membership functions to represent the occurrence, detection, and the fuzzy RPN. There is perfect agreement between T2-FIS and RPI(SC_4) in failure modes FM3, FM5, FM8, and FM9.

Cases T2-FIS-02, T2-FIS-03, and T2-FIS-04 achieve a agreement coefficient of 0.7, 0.65, and 0.50, respectively. Nevertheless, the three cases have a perfect agreement with respect to RPI(SC_4) in the same failure modes, FM3, FM5, and FM8.

Cases T2-FIS 05 and 06 achieve the same agreement coefficient of 0.35, which can be considered “fair,” as Table III suggests. When comparing T2-FIS 05 and RPI(SC_4), there is a perfect agreement in two rankings (FM3 and FM5), and when comparing T2-FIS 06 and RPI(SC_4), there is a perfect agreement in failure modes FM2 and FM3. Notably, FM3 was ranked as priority 3 for all the six Type-II methods, and FM8 was ranked as priority 1 for five Type-II methods.

TABLE XVI

Linear Weighted Kappa κ_{w-lin} between Reference Ranking RPI(SC4) and Type-II Fuzzy Inference System

	RPI(SC_4)	T2-FIS 01	T2-FIS 02	T2-FIS 03	T2-FIS 04	T2-FIS 05	T2-FIS 06
FM1	4	5	5	8	8	10	11
FM2	5	4	4	4	7	4	5
FM3	2	2	2	2	2	2	2
FM4	7	6	6	9	9	9	10
FM5	11	11	11	,11	11	11	9
FM6	6	3	3	3	3	1	1
FM7	9	10	10	10	10	8	8
FM8	1	1	1	1	1	3	3
FM9	8	8	7	5	4	5	6
FM10	3	7	8	6	6	6	4
FM11	10	9	9	7	5	7	7
κ_{w-lin}	Reference	0.70	0.65	0.50	0.40	0.35	0.35
Strength of agreement	-	Substantial	Substantial	Moderate	Fair	Fair	Fair
z	-	3.604	3.346	2.574	2.059	1.802	1.802
H_0 test	-	Reject	Reject	Reject	Reject	Reject	Reject

B. *Results for quadratic weighted kappa*

Table XVII shows the quadratic weighted agreement coefficient κ_{w-quad}, the test statistics z, the strength of agreement, and the hypothesis test result for the FMECA methods RPI(SC5), Fuzzy VIKOR, and ITHWD when compared with the reference ranking RPI(SC4). As shown, quadratic weighted kappa takes values between 0.727 and 0.855, revealing the scenario RPI(SC5) as those one achieving better agreement of 0.855, which can be considered an almost perfect agreement.

Using now coefficient κ_{w-quad}, Table XVIII shows the results for the six Type-II FMECA configurations. Results show that the best agreement coefficient value equals 0.8, corresponding to T2-FIS-01.

TABLE XVII

Quadratic Weighted Kappa κ_{w-quad} for RPI(SC5), VIKOR, and ITHWD

	RPI(SC_5)	Fuzzy VIKOR	ITHWD
κ_{w-quad}	0.855	0.727	0.809
Strength of agreement	Perfect	Substantial	Substantial
z	2.834	2.412	2.683
H_0 test	Reject	Reject	Reject

TABLE XVIII

Quadratic Weighted Kappa κ_{w-quad} between Reference Ranking RPI(SC4) and Type-II Fuzzy Inference System

	T2-FIS 01	T2-FIS 02	T2-FIS 03	T2-FIS 04	T2-FIS 05	T2-FIS 06
κ_{w-quad}	0.864	0.818	0.736	0.618	0.555	0.518
Strength of agreement	Perfect	Perfect	Substantial	Substantial	Moderate	Moderate
z	2.864	2.714	2.442	2.05	1.839	1.719
H_0 test	Reject	Reject	Reject	Reject	Reject	Reject

Two Type-II FIS cases (T2-FIS 03 and T2-FIS 04) achieved the worst agreement value of 0.555 and 0.518, respectively. Unlike the result for the linear coefficient, where these two cases achieve the same kappa, the quadratic weighted kappa is slightly different for both cases.

These results show that the weighting scheme affects the magnitude of the agreement coefficient concerning the linear weight scheme and is also sensitive to the rankings whose agreement it measures.

The null hypothesis H_0 was rejected in all the simulations. This means that the results achieved can be considered statistically significant.

VIII. DISCUSSION

Results confirm that the quadratic weighting scheme produces concordance values greater than the linear weighting scheme, as documented in [47]. In the FMECA context, the relationship between categories is not always linear and difficult to establish; this relationship should determine the weighting scheme used to calculate κ_w. However, we used the linear and quadratic weighting schemes in this paper.

The results show that FMECA methods achieve a higher linear weighted kappa and a higher quadratic weighted kappa. In practical terms, it can be stated that the main difference between the obtained results of kappa using the two weighting schemes can be determined by the strength of agreement labels detailed in Table III. For example, the method T2-FIS 01 can be considered "substantial," and its respective obtained using the same approach can be considered "almost perfect."

A more in-depth study is needed to quantify the influence of the weighting scheme on Cohen's kappa.

Table XIX shows the ranking for the reference FMECA RPI(Sc4), the T2-FIS 01, the RPI(Sc5), T2-FIS 02, and ITHWD, and their corresponding κ_{w-lin} and κ_{w-quad}; the rankings were ordered from highest to lowest kappa.

Fig. 5 shows a radar chart for the three FMECA methods shown in Table XIX and the reference one RPI(Sc_4). This graphic greatly simplifies the comparison between the different rankings assigned to each failure mode. The blue line in Fig. 5 represents the reference FMECA ranking RPI(SC_4), and the red line represents the ranking for the method with the highest κ (T2-FIS 01).

Because the FMECA case study has only a few failure modes, it is possible to identify the differences between the five FMECA methods. The ranking for failure modes FM1, FM2, FM5, FM10, and FM11 is the same for the reference RPI(SC_4) and T2-FIS 01; both models agree 5 times and disagree 6 times. Comparing the base case with RPI(SC_5), the rankings agree 4 times (FM5, FM8, FM9, and FM11) and disagree 7 times. For T2-FIS 02, the rankings agree 3 times (FM3, FM5, and FM8) and disagree 8 times. For ITHWD, the rankings agree 3 times (FM1, FM5, and FM7) and disagree 8 times. All the methods listed in Table XIX agree to classify FM5 as the lowest priority failure mode.

Notice that the number of agreements and disagreements can indicate only the level of concordance between two raters. However, it does not provide an effective metric to measure it; Cohen's coefficient deals with this issue and gives a concordance level based on the coincidences between ratings and the agreement that can occur by chance.

TABLE XIX
DIFFERENT RANKINGS FOR FMECA IMPROVEMENT METHODS

Failure mode	RPI (SC_4)	T2-FIS 01	RPI (SC_5)	T2-FIS 02	ITHWD
FM1	4	4	5	5	4
FM2	5	5	7	4	6
FM3	2	1	4	2	1
FM4	7	6	9	6	10
FM5	11	11	11	11	11
FM6	6	7	3	3	3
FM7	9	8	6	10	9
FM8	1	2	1	1	5
FM9	8	9	8	7	7
FM10	3	3	2	8	2
FM11	10	10	10	9	8
κ_{w-lin}	Ref.	0.70	0.650	0.65	0.60
κ_{w-quad}	Ref.	0.864	0.855	0.818	0.809

IX. CONCLUSION AND FUTURE WORK

This paper introduces an approach based on Cohen's kappa agreement coefficient to compare different methods used in the FMECA context. A simple and further analyzed FMECA case study was selected to conduct the comparisons, including rankings obtained through four methods reported in the literature. In addition, FMECA based on Type-I Fuzzy and Type-II Fuzzy Inference systems were developed and conducted to rank the failure modes. From our results and its previous discussion, one pulls out four critical conclusions:

1) The comparison between different FMECA methods is commonly based on the qualitative comparison between rankings and sometimes considering a balance between the three risk factors; nevertheless, this approach can be impractical for more extensive problems;
2) The proposed approach aims to contribute to the quantitative comparison between methods used to improve the failure modes prioritization regarding a reference ranking;
3) The results show that Cohen's κ coefficient provides a quantitative level for the agreement between two different rankings in the FMECA analysis context;
4) For this application, the ranking based on Type-II Fuzzy Inference System achieves the best agreement regarding the method selected as the reference;
5) The selection of the weighting scheme is another essential aspect to take into account in the proposed approach; in this particular application, the quadratic weighting scheme allows obtaining a better strength of agreement;
6) The reference FMECA's ranking identification is a critical aspect of the success of the proposed approach; nevertheless, this approach is practical when trying to

test the effectiveness of new FMECA methods. Other methodologies already applied to a given problem can be used as references.

Our proposed approach's main shortcoming is the reference FMECA ranking selection. In practical applications identifying a suitable FMECA reference ranking can become a demanding task. An acceptable procedure to conduct this kind of comparison could be to apply different FMECA approaches to a well-known problem whose failure modes' ranking can be considered optimal and then compute the concordance coefficient to identify the best FMECA method concerning this reference. Once the best FMECA method is identified, it can be applied to another case study with similar characteristics. The solution to this shortcoming is being addressed and included in future works.

Additional aspects are currently in development and will be included in forthcoming works:

1) The application of the proposed approach in the context of smart substations;
2) The definition of *tailor-made* scales for the FMECA risk factors in the context of smart substations;
3) The proposal of a new weighting scheme based on the risk mentioned above factors' scales;
4) The use of paradox-resistant concordance coefficients.

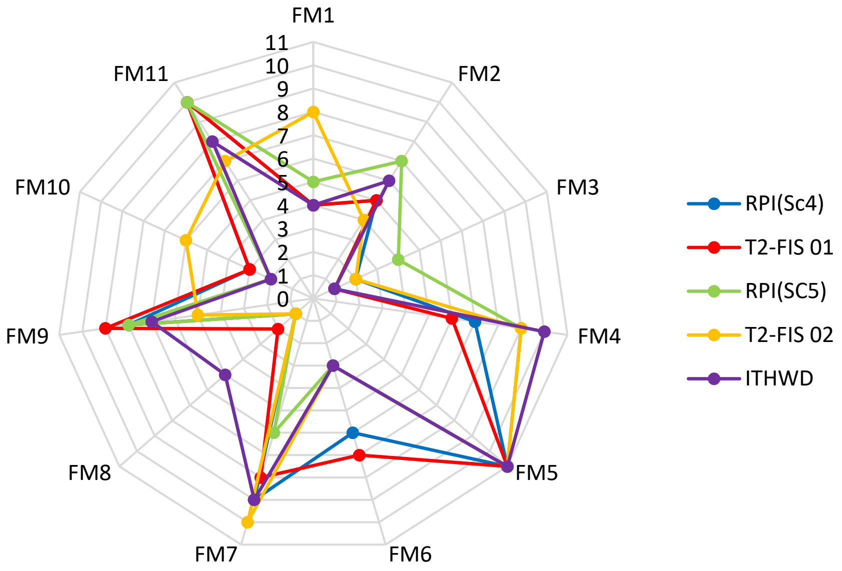

Fig. 5. Radar chart showing the reference FMECA ranking (blue line) and the three best approaches.

ACKNOWLEDGMENT

The authors express special thanks to Professor Blas Galván González for reading this paper's earlier version and offering many insightful comments and suggestions.

REFERENCES

[1] International Electrotechnical Comission, IEC 60812:2006 - Analysis techniques for system reliability – Procedure for failure mode and effects analysis (FMEA), 2006.

[2] X. Li, M. He, H. Wang, Application of failure mode and effect analysis in managing catheter-related blood stream infection in intensive care unit, Med. (United States). 96 (2017). https://doi.org/10.1097/MD.0000000000009339.

[3] J.B. Bowles, C.E. Peláez, Fuzzy logic prioritization of failures in a system failure mode, effects and criticality analysis, Reliab. Eng. Syst. Saf. 50 (1995) 203–213. https://doi.org/10.1016/0951-8320(95)00068-D.

[4] H.C. Liu, J.X. You, X.Y. You, Evaluating the risk of healthcare failure modes using interval 2-tuple hybrid weighted distance measure, Comput. Ind. Eng. 78 (2014) 249–258. https://doi.org/10.1016/j.cie.2014.07.018.

[5] J. Huang, J.X. You, H.C. Liu, M.S. Song, Failure mode and effect analysis improvement: A systematic literature review and future research agenda, Reliab. Eng. Syst. Saf. 199 (2020) 106885. https://doi.org/10.1016/j.ress.2020.106885.

[6] H.C. Liu, FMEA using uncertainty theories and MCDM methods, Springer Science+Business Media Singapore, Singapore, 2016. https://doi.org/10.1007/978-981-10-1466-6.

[7] A. Pillay, J. Wang, Modified failure mode and effects analysis using approximate reasoning, Reliab. Eng. Syst. Saf. 79 (2003) 69–85. https://doi.org/10.1016/S0951-8320(02)00179-5.

[8] H.C. Liu, Z. Li, W. Song, Q. Su, Failure mode and effect analysis using cloud model theory and PROMETHEE method, IEEE Trans. Reliab. 66 (2017) 1058–1072. https://doi.org/10.1109/TR.2017.2754642.

[9] J.-P. Brans, B. Mareschal, PROMETHEE methods, in: Mult. Criteria Decis. Anal. State Art Surv., Springer, New York, NY., 2005: pp. 163–195. https://doi.org/https://doi.org/10.1007/0-387-23081-5_5.

[10] F.E. Boran, S. Genç, M. Kurt, D. Akay, A multi-criteria intuitionistic fuzzy group decision making for supplier selection with TOPSIS method, Expert Syst. Appl. 36 (2009) 11363–11368. https://doi.org/10.1016/j.eswa.2009.03.039.

[11] S. Opricovic, Fuzzy VIKOR with an application to water resources planning, Expert Syst. Appl. 38 (2011) 12983–12990. https://doi.org/10.1016/j.eswa.2011.04.097.

[12] H. Zhao, J.X. You, H.C. Liu, Failure mode and effect analysis using MULTIMOORA method with continuous weighted entropy under interval-valued intuitionistic fuzzy environment, Soft Comput. 21 (2017) 5355–5367. https://doi.org/10.1007/s00500-016-2118-x.

[13] Z. Wang, J.M. Gao, R.X. Wang, K. Chen, Z.Y. Gao, W. Zheng, Failure Mode and Effects Analysis by Using the House of Reliability-Based Rough VIKOR Approach, IEEE Trans. Reliab. 67 (2018) 230–248. https://doi.org/10.1109/TR.2017.2778316.

[14] W. Song, X. Ming, Z. Wu, B. Zhu, A rough TOPSIS approach for

failure mode and effects analysis in uncertain environments, Qual. Reliab. Eng. Int. 30 (2014) 473–486. https://doi.org/10.1002/qre.1500.

[15] W.Z. Wang, X.W. Liu, S.L. Liu, Failure Mode and Effect Analysis for Machine Tool Risk Analysis Using Extended Gained and Lost Dominance Score Method, IEEE Trans. Reliab. 69 (2020) 954–967. https://doi.org/10.1109/TR.2019.2955500.

[16] G. Choquet, Theory of capacities, Ann. L'Institut Fourier. 5 (1954) 131–295. https://doi.org/10.5802/aif.53.

[17] H.C. Liu, X.Q. Chen, J.X. You, Z. Li, A New Integrated Approach for Risk Evaluation and Classification with Dynamic Expert Weights, IEEE Trans. Reliab. 70 (2021) 163–174. https://doi.org/10.1109/TR.2020.2973403.

[18] H. gang Peng, J. qiang Wang, Hesitant Uncertain Linguistic Z-Numbers and Their Application in Multi-criteria Group Decision-Making Problems, Int. J. Fuzzy Syst. 19 (2017) 1300–1316. https://doi.org/10.1007/s40815-016-0257-y.

[19] P. Liu, Y. Li, An improved failure mode and effect analysis method for multi-criteria group decision-making in green logistics risk assessment, Reliab. Eng. Syst. Saf. 215 (2021) 107826. https://doi.org/10.1016/j.ress.2021.107826.

[20] L. Morissette, S. Chartier, The k-means clustering technique: General considerations and implementation in Mathematica, Tutor. Quant. Methods Psychol. 9 (2013) 15–24. https://doi.org/10.20982/tqmp.09.1.p015.

[21] S. Hassan, J. Wang, C. Kontovas, M. Bashir, Modified FMEA hazard identification for cross-country petroleum pipeline using Fuzzy Rule Base and approximate reasoning, J. Loss Prev. Process Ind. 74 (2022) 104616. https://doi.org/10.1016/j.jlp.2021.104616.

[22] J.-S.R. Jang, C.-T. Sun, E. Mizutani, Neuro-Fuzzy and Soft Computing: A Computational Approach to Leraning and Machine Intelligence, Prentice Hall, Inc., Upper Saddle River NJ, 1997.

[23] J.L. Deng, Introduction to Grey Systems Theory, J. Grey Syst. 1 (1989) 1–24. https://doi.org/10.1007/978-3-642-16158-2_1.

[24] E. Akyuz, E. Celik, A quantitative risk analysis by using interval type-2 fuzzy FMEA approach: the case of oil spill, Marit. Policy Manag. 45 (2018) 979–994. https://doi.org/10.1080/03088839.2018.1520401.

[25] E. Bozdag, U. Asan, A. Soyer, S. Serdarasan, Risk prioritization in failure mode and effects analysis using interval type-2 fuzzy sets, Expert Syst. Appl. 42 (2015) 4000–4015. https://doi.org/10.1016/j.eswa.2015.01.015.

[26] J. Qin, Y. Xi, W. Pedrycz, Failure mode and effects analysis (FMEA) for risk assessment based on interval type-2 fuzzy evidential reasoning method, Appl. Soft Comput. J. 89 (2020) 106134. https://doi.org/10.1016/j.asoc.2020.106134.

[27] V. Anes, E. Henriques, M. Freitas, L. Reis, A new risk prioritization model for failure mode and effects analysis, Qual. Reliab. Eng. Int. 34 (2018) 516–528. https://doi.org/10.1002/qre.2269.

[28] P.C. Okwesili Jr., Risk Assessmnt Using Paired Comparison Expert Elicitation for Ranking of Compounding Outsourcing Facilities, The George Washington University, 2016.

[29] H. Frewen, E. Brown, M. Jenkins, A. O'Donovan, Failure mode and effects analysis in a paperless radiotherapy department, J. Med. Imaging Radiat. Oncol. 62 (2018) 707–715. https://doi.org/10.1111/1754-9485.12762.

[30] R. Renu, D. Visotsky, S. Knackstedt, G. Mocko, J.D. Summers, J. Schulte, A Knowledge Based FMEA to Support Identification and Management of Vehicle Flexible Component Issues, Procedia CIRP. 44 (2016) 157–162. https://doi.org/10.1016/j.procir.2016.02.112.

[31] M.G. Kendall, B.B. Smith, The Problem of m Rankings, Ann. Math. Stat. 10 (1939) 275–287. https://doi.org/10.1214/aoms/1177732186.

[32] J. Liu, W. Tang, G. Chen, Y. Lu, C. Feng, X.M. Tu, Correlation and agreement: overview and clarification of competing concepts and measures, Shanghai Arch. Psychiatry. 28 (2016) 115–120. https://doi.org/10.11919/j.issn.1002-0829.216045.

[33] K.L. Gwet, Handbook of Inter-Rater Reliability, 4th ed., Advanced Analytics, LLC, Gaithersburg, 2014.

[34] J.R. Landis, G.G. Koch, The Measurement of Observer Agreement for Categorical Data, Biometrics. 33 (1977) 159. https://doi.org/10.2307/2529310.

[35] T.R. Vetter, P. Schober, Agreement analysis: What he said, she said versus you said, Anesth. Analg. 126 (2018) 2123–2128. https://doi.org/10.1213/ANE.0000000000002924.

[36] P. Ranganathan, C.S. Pramesh, R. Aggarwal, Common pitfalls in statistical analysis : Measures of agreement, Perspect. 8 (2017) 187–191. https://doi.org/10.4103/picr.PICR.

[37] J. Teles, Concordance coefficients to measure the agreement among several sets of ranks, J. Appl. Stat. 39 (2012) 1749–1764. https://doi.org/10.1080/02664763.2012.681460.

[38] J. Cohen, A Coefficient of Agreement for Nominal Scales, Educ. Psychol. Meas. 20 (1960) 37–46. https://doi.org/10.1177/001316446002000104.

[39] M.L. McHugh, Lessons in biostatistics interrater reliability: the kappa statistic, Biochem. Medica. 22 (2012) 276–282. https://hrcak.srce.hr/89395.

[40] P.W. Mielke, K.J. Berry, A note on Cohen's weighted kappa coefficient of agreement with linear weights, Stat. Methodol. 6 (2009) 439–446. https://doi.org/10.1016/j.stamet.2009.03.002.

[41] J. Sim, C.C. Wright, The kappa statistic in reliability studies: Use, interpretation, and sample size requirements, Phys. Ther. 85 (2005) 257–268. https://doi.org/10.1093/ptj/85.3.257.

[42] J. Cohen, Weighted kappa: nominal scale agreement with provision for scaled disagreement of partial credit, Psychol. Bull. 70 (1968) 213–220. https://doi.org/https://doi.org/10.1037/h0026256.

[43] M. Banerjee, M. Capozzoli, L. McSweeney, D. Sinha, Beyond kappa: A review of interrater agreement measures, Can. J. Stat. 27 (1999) 3–23. https://doi.org/10.2307/3315487.

[44] A. Ben-David, Comparison of classification accuracy using Cohen's Weighted Kappa, Expert Syst. Appl. 34 (2008) 825–832. https://doi.org/10.1016/j.eswa.2006.10.022.

[45] I. Guggenmoos-Holzmann, The meaning of kappa: Probabilistic concepts of reliability and validity revisited, J. Clin. Epidemiol. 49 (1996) 775–782. https://doi.org/10.1016/0895-4356(96)00011-X.

[46] M.J. Warrens, Cohen's kappa is a weighted average, Stat. Methodol. 8 (2011) 473–484. https://doi.org/10.1016/j.stamet.2011.06.002.

[47] M.J. Warrens, Cohen's quadratically weighted kappa is higher than linearly weighted kappa for tridiagonal agreement tables, Stat. Methodol. 9 (2012) 440–444. https://doi.org/10.1016/j.stamet.2011.08.006.

[48] J.L. Fleiss, J. Cohen, B.S. Everitt, Large sample standard errors of kappa and weighted kappa, Psychol. Bull. 72 (1969) 323–327. https://doi.org/10.1037/h0028106.

[49] S.M. Ross, Introduction to Probability and Statistics for Engineers and Scientists, Sixth Edit, Academic Press, London, 2021.

[50] S.M. Vieira, U. Kaymak, J.M.C. Sousa, Cohen's kappa coefficient as a performance measure for feature selection, 2010 IEEE World Congr. Comput. Intell. WCCI 2010. (2010). https://doi.org/10.1109/FUZZY.2010.5584447.

[51] P. Hammond, M. Suttie, V.T. Lewis, A.P. Smith, A.C. Singer, Detection of untreated sewage discharges to watercourses using machine learning, Npj Clean Water. 4 (2021) 1–10. https://doi.org/10.1038/s41545-021-00108-3.

[52] P. Agrawal, B. Trivedi, Evaluating Machine Learning Classifiers to detect Android Malware, 2020 IEEE Int. Conf. Innov. Technol. INOCON 2020. (2020) 8–13. https://doi.org/10.1109/INOCON50539.2020.9298290.

[53] A. Whata, C. Chimedza, Deep Learning for SARS COV-2 Genome Sequences, IEEE Access. 9 (2021) 59597–59611. https://doi.org/10.1109/ACCESS.2021.3073728.

[54] N. Cherrier, Interpretable Machine Learning for CLAS12 Data Analysis Thèse de doctorat, Université Paris-Saclay, 2021. https://tel.archives-ouvertes.fr/tel-03184811.

[55] Y. Lu, F. Teng, J. Zhou, A. Wen, Y. Bi, Failure mode and effect analysis in blood transfusion: A proactive tool to reduce risks, Transfusion. 53 (2013) 3080–3087. https://doi.org/10.1111/trf.12174.

[56] H.-C. Liu, Improved FMEA Methods for Proactive Healthcare Risk Analysis, Springer, Singapore, 2019. https://doi.org/10.1007/978-981-13-6366-5.

[57] H.C. Liu, L. Liu, N. Liu, L.X. Mao, Risk evaluation in failure mode and effects analysis with extended VIKOR method under fuzzy environment, Expert Syst. Appl. 39 (2012) 12926–12934. https://doi.org/10.1016/j.eswa.2012.05.031.

[58] J.M. Mendel, Uncertain Rule-Based Fuzzy Systems: Introduction and New Directions, 2nd ed., Springer International Publishing, Cham, 2017. https://doi.org/10.1007/978-3-319-51370-6.