Information Sciences 618 (2022) 379–399

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

An empirical study on how humans appreciate automated counterfactual explanations which embrace imprecise information

Ilia Stepin [a,b,*], Jose M. Alonso-Moral [a,b], Alejandro Catala [a,b], Martín Pereira-Fariña [c]

[a] Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela, Rúa de Jenaro de la Fuente Domínguez, s/n, 15782 Santiago de Compostela, A Coruña, Spain
[b] Departamento de Electrónica e Computación, Universidade de Santiago de Compostela, Rúa Lope Gómez de Marzoa, s/n, 15782 Santiago de Compostela, A Coruña, Spain
[c] Departamento de Filosofía e Antropoloxía, Universidade de Santiago de Compostela, Plaza de Mazarelos, s/n, 15705 Santiago de Compostela, A Coruña, Spain

ARTICLE INFO

Article history:
Received 15 February 2022
Received in revised form 19 October 2022
Accepted 21 October 2022
Available online 4 November 2022

2020 MSC:
94D05
03B52
03E72
68T30
68T37

Keywords:
Explainable artificial intelligence
Interpretable fuzzy modeling
Fuzzy rule-based classification
Counterfactual explanation
Human evaluation

ABSTRACT

The explanatory capacity of interpretable fuzzy rule-based classifiers is usually limited to offering explanations for the predicted class only. A lack of potentially useful explanations for non-predicted alternatives can be overcome by designing methods for the so-called counterfactual reasoning. Nevertheless, state-of-the-art methods for counterfactual explanation generation require special attention to human evaluation aspects, as the final decision upon the classification under consideration is left for the end user. In this paper, we first introduce novel methods for qualitative and quantitative counterfactual explanation generation. Then, we carry out a comparative analysis of qualitative explanation generation methods operating on (combinations of) linguistic terms as well as a quantitative method suggesting precise changes in feature values. Then, we propose a new metric for assessing the perceived complexity of the generated explanations. Further, we design and carry out two human evaluation experiments to assess the explanatory power of the aforementioned methods. As a major result, we show that the estimated explanation complexity correlates well with the informativeness, relevance, and readability of explanations perceived by the targeted study participants. This fact opens the door to using the new automatic complexity metric for guiding multi-objective evolutionary explainable fuzzy modeling in the near future.

1. Introduction

Artificial intelligence (AI)-based algorithms show striking accuracy in a wide range of domains and applications [1]. However, the most accurate models are known to produce scarcely explainable decisions [2]. This lack of explainability may damage the overall trust in AI [36]. In the light of possible negative consequences of following such automatic decisions without

* Corresponding author at: Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela, Rúa de Jenaro de la Fuente Domínguez, s/n, 15782 Santiago de Compostela, A Coruña, Spain.
E-mail addresses: ilia.stepin@usc.es (I. Stepin), josemaria.alonso.moral@usc.es (J.M. Alonso-Moral), alejandro.catala@usc.es (A. Catala), martin.pereira@usc.es (M. Pereira-Fariña).

https://doi.org/10.1016/j.ins.2022.10.098
0020-0255/

having them explained, legal regulations concerning data processing are becoming widely adopted, e.g. the General Data Protection Regulation (GDPR) in the European Union [33]. Moreover, a new European regulation on AI is in progress and highlights the importance of preserving the European values by promoting trustworthy and responsible human-centric AI [9,34].

The gap between obscurity of automatic decisions and their explainability can be overcome by using interpretable models [37]. Among all AI tools, such soft computing techniques as fuzzy sets and systems have been shown to be not only interpretable but also explainable [3]. Thus, two key advantages are distinguished when relating the properties of interpretability and explainability of fuzzy systems. First, their transparent (i.e., interpretable) structure allows for making unambiguous inferences of why the given output was produced. Second, the use of linguistic variables and rules enables such systems to be explainable, i.e., to produce comprehensible explanations in natural language.

Nevertheless, the ability to demonstrate evidence on why specific output is produced (i.e., explain the factual output) may not be sufficient to display the underlying reasoning to the end user. Therefore, a factual explanation may need to be complemented with an explanation of why some other output was not produced. Opposed to factual explanations justifying the given prediction, counterfactual (CF) explanations (or counterfactuals) inform the end user about minimally different alterations to the input features for the outcome to change [41]. In the context of classification problems, CF explanations are typically designed as answers to the template question "Why was P predicted rather than Q?" where P is the output (factual) class and Q is a non-predicted hypothesized alternative CF class [29].

CF explanation generation is often regarded as an optimization problem in search of the data point of another class which represents the closest data point alternative to the test instance in an n-dimensional Euclidean space [46]. In the context of fuzzy sets and systems, however, such minimal changes may be described not only by means of a continuous variable representing numerical feature values (which we call "quantitative CFs" in this paper) but also by a discrete linguistic variable whose values are linguistic terms (which we refer to as "qualitative CFs" in this paper). In the former case, distinctive (numerical) features point to specific values, which are minimally different from those the test instance has, that should be set for the outcome to change. In the latter case, linguistic terms represent sets of suitable CF feature values in form of text and conceal the underlying numerical intervals.

The difference in end user's perception of these types of CF explanations remains unclear [45]. On the one hand, it may be affected by peculiarities of the structure of explanation, such as the number of explanatory features or explanation length. On the other hand, user's perception may be influenced by a degree of precision of the explanation content. Thus, qualitative CFs may be regarded as pieces of imprecise information which can facilitate understanding of the communicated explanation but may, however, be underinformative or even misleading to the end user. Conversely, quantitative CFs specify fine-grained changes to values of features. Last but not least, existing metrics for measuring quality of CF explanations (e.g., validity, proximity, diversity, among others) are strongly related to the data used for explanation generation [31]. However, those metrics ignore perceptual skills of the explanation's recipient and may not be sufficient for assessing the overall explanation effectiveness. In order to make another step towards human-centric AI, it therefore appears necessary to propose novel means of capturing and assessing human perception of explanations.

As part of previous work [41], we introduced a method for generating qualitative CF explanations applied to decision trees (DT). Then, we generalized this method to fuzzy information granules [43]. In this paper, our contribution is fourfold. First, we extend our previous work with a generalized Euclidean distance-based metric for CF explanation generation which better grasps membership function values. Second, we propose a novel genetic-based quantitative CF explanation generation method. Third, we define a new metric for assessing the complexity of automated explanations. Fourth, we carefully validate both qualitative and quantitative CF explanations via human evaluation in agreement with the best known practices for fair and sound evaluation of Natural Language Generation (NLG) and analyze the findings in terms of explanation complexity as expected to be perceived by the end user.

The rest of the manuscript is structured as follows. Section 2 presents a brief overview of existing methods for quantitative and qualitative CF explanation generation. Section 3 introduces our methods for generating CF explanations associated to fuzzy rule-based classification systems (FRBCS). Section 4 describes the key characteristics of the experimental design for subsequent human evaluation studies. Section 5 goes in detail with the analysis of the data collected in two evaluation surveys. Section 6 discusses the findings and offers suggestions on how they can be exploited. Finally, we outline directions for future work and conclude in Section 7.

2. Related work

CF explanation generation has in recent years attracted increasing attention from researchers in the AI field. As CFs oppose actual and potential outcomes, they are most widely used to explain the output of various classifiers, from linear machine learning models to deep neural networks [42]. Further, they are extensively found across different application domains. For example, CFs are found applicable in healthcare where they, e.g., serve to provide a patient with a bigger picture of the risk of developing diabetic retinopathy [26] or in banking where CFs suggest recommendations on necessary changes to have a loan application approved if previously rejected [16]. In addition, CF explanations are as well extensively used in robotics (e.g., in planning – to justify the choice of a robot over other feasible but unfavored possible solutions [44]). Despite numerous potential application domains, the use of CFs is advised to be controled due to possible malicious implications. As

such, they have been misused or misinterpreted (what may lead to data breaches) in cases of, e.g. password masking or e-voting [20]. Other privacy concerns include inferring sensitive patterns of the training data or manipulations with the revealed internals of the model [40].

In the context of qualitative CFs, a number of generation methods output CF sets to support diversity. For example, Sokol and Flash inspect the internal structure of DTs in their "Glass-Box" framework for generating CF sets [40]. Thus, the authors retrieve CF sets from the decision paths ranking them by their leaf-to-leaf distance to the actual prediction. On a similar note, Stepin et al. generate set-based (i.e., qualitative) CFs from either crisp or fuzzy DTs [41] but also regarding fuzzy information granules [43] while introducing an extra-linguistic layer to approximate numerical intervals or membership function values, respectively, using predefined linguistic terms.

Whereas the aforementioned methods are model-specific, i.e., they only allow for explaining counterfactually the given output of the DT itself, DT-based approaches are also used for model-agnostic methods. In their LOcal Rule-based Explanation (LORE) method, Guidotti et al. employ a genetic algorithm to first synthesize a local neighborhood around the test instance which is subsequently used to train a DT and generate CF sets [17]. The collection of CF sets is then reconstructed from the decision paths. Then, the minimally different CF set is selected on the basis of the (minimal) number of Boolean split conditions of the DT that the given CF path does not satisfy. Maaroof et al. extend LORE to fuzzy logic-based applications by proposing Contextualised LORE for Fuzzy attributes (C-LORE-F) [26]. Alternatively to LORE, the researchers formulate a local neighborhood generation approach for solving the uniform cost search problem. Potential neighbors are generated by applying iterative changes over a single feature taking into account intersections between two corresponding fuzzy sets. Further, the authors propose to induce the rules instead of building up a DT using the Dominance-based Rough Set Approach (DRSA) where the decision rules take into consideration the preference directions of the input variables. In addition, Fernández et al. extract CF sets from a random forest classifier by partly fusing individual tree predictors [12]. Further, their Random Forest Optimal Counterfactual Set Extractor (RF-OCSE) prunes the search space of candidate CFs using the minimum observable approach to filter out CFs whose distance to the test instance exceeds the best up-to-now distance.

On the other hand, quantitative (i.e., single-point-output) CF explanation generation methods address the optimization problem searching for an individual data point found to be minimally different from the test point under consideration in accordance with the selected distance function, e.g., Manhattan distance weighted by the inverse median absolute deviation [46]. Similarly, Moore et al. use a differentiable model on the basis of a gradient-based method over the cross entropy loss function to identify a single minimally distant CF data point [30].

Alternatively, genetic algorithms are also frequently used to generate CFs [39]. Model-agnostic genetic algorithms are used not only to generate a local neighborhood but also to identify a specific optimal CF data point. In addition to the standard genetic algorithm, Lash et al. apply local search to non-mutated children so that the best solution is preserved for the next generation [24]. Sharma et al. propose another approach called Counterfactual Explanations for Robustness, Transparency, Interpretability, and Fairness of Artificial Intelligence (CERTIFAI) where a genetic algorithm based on natural selection, mutation, and crossover appeals to user feedback (regarding feature mutation, feature range specification, and enquiries for a specific number of explanations) [39]. Whereas these user constraints allow for generating actionable human-centric explanations, imposing too severe restrictions may overreduce the search space resulting in generating null explanations. In addition, Schleich et al. make use of a complete search space in their GeCo framework [38]. Thus, the authors present a customizable genetic algorithm enhanced with two optimization techniques to reduce memory costs and running time. The compressed δ-representation of the input features reduces the memory storage required for mutation-related calculations whereas the so-called partial evaluation optimizes the evaluation of the classifier, as static components of the classifier can be pre-evaluated using an equivalent sub-model of the same classifier [38].

Finally, both qualitative and quantitative generation methods are primarily evaluated with automatically computable metrics (e.g., fidelity, validity, proximity, or diversity) [12,17,31]. Unfortunately, empirical studies involving human evaluation for assessing the goodness of automated CFs are scarcely found in the literature. Baaj and Poli show that explanations based on the use of linguistic terms appear rather satisfactory and convincing despite being overly repetitive for a general audience [5]. Wang and Yin state that CFs increase understanding for users who have sufficient domain knowledge but fail to calibrate trust in the model [47]. Further, Lucic et al. demonstrate that CFs help users understand why a model makes large errors [25]. Olson et al. show that CFs can be also effective for non-expert users in the identification of flawed agents [32]. In addition, Woodcock et al. stress that lay users trust CFs only if the information gap in the existing domain knowledge between them and expert users is not significant, specifically in the healthcare domain [48]. Nevertheless, unlike our work, none of the aforementioned studies contrasts the output of single-point-output quantitative generation methods and set-based qualitative ones.

3. Explanation generation methods

3.1. Notation

The methods proposed in this study address a multi-class classification problem, i.e., learning a mapping function $h: X \longrightarrow Y$ from a dataset $X = \{x_i\}|_{i=1}^{n}$ containing n labeled instances to a discrete output variable (class) $Y = \{y_j\}|_{j=1}^{m}$ where m is the number of classes. The dataset is characterized by the set of p numerical[1] features $F = \{f_k\}|_{k=1}^{p}$, which are mapped to the corresponding linguistic variables. By definition [49], each feature is a tuple $f_k = \left\langle L^{f_k}, T_L^{f_k}, U^{f_k}, G^{f_k}, M^{f_k} \right\rangle, \forall f_k \in F$ where L^{f_k} is the name of the feature f_k, $T_L^{f_k} = \left\{ t_l^{f_k} \right\}|_{l=1}^{s}$ is the set of linguistic terms defined in the universe of discourse U^{f_k}, G^{f_k} and M^{f_k} being syntactic and semantic rules, respectively. Let $V_T = \bigcup T_L^{f_k}, \forall f_k \in F$ denote the set of all linguistic terms.

In our experiments (see Sections 4 and 5), we aim to explain (both factually and counterfactually) the output of an FRBCS [23] which is defined by the following components:

- a knowledge base containing a set of input and output variables and a rule base which represents a set $R = \{r_i(w_i)\}|_{i=1}^{|R|}$ of weighted fuzzy rules of the form $r_i(w_i)$: IF L^{f_1} is $t_1^{f_1}$ $\left[\text{AND} \ldots L^{f_k} \text{ is } t_k^{f_k} \ldots \text{AND} \ldots\right]$ THEN y IS y_i, where $r_i \in R, w_i \in [0,1]$ is the rule weight (i.e., the higher w_i the more relevant r_i), $t_k^{f_k} \in T_L^{f_k}, f_k \in F, y_i \in Y$;
- a fuzzy processing structure containing fuzzification and defuzzification interfaces as well as a fuzzy reasoning mechanism. Given an input vector $\mathbf{x} = [x_1, \ldots, x_p]$ and a rule $r_i \in R$, its activation degree a_i is computed as $a_i(\mathbf{x}) = \mu_{t_1^{f_1}}(x_1) \otimes \ldots \otimes \mu_{t_k^{f_k}}(x_k) \otimes \ldots \otimes \mu_{t_p^{f_p}}(x_p)$, being $\mu_{t_k^{f_k}}(x_k)$ the membership degree of the value x_k for the linguistic term t_k associated to feature f_k, and $\otimes$ is a t-norm such as minimum or product.

Any rule r_i can be denoted as a tuple $r_i(w_i) = \langle AC_i, cq_i \rangle$ where AC_i is an antecedent (i.e., a non-empty set of feature-value pairs) and cq_i is a consequent (i.e., a class label).

The output class $y_{FAC} \in Y$ predicted by an FRBCS is said to be the factual explanation class. All the rules from the rule base that lead to the predicted outcome form a set of factual explanation rules $R_{FAC} = \bigcup_{r_j \in R} \{r_j | cq_j = y_{FAC}\}$, being $R_{FAC} \subseteq R$. Similarly, all the non-predicted classes form a set of CF classes, with a collection of the corresponding rules mapped to each of them: $R_{CF} = \bigcup_{r_j \in R} \left\{ r_j | cq_{r_j} = y_{CF} \right\}, Y_{CF} = \{y_{CF} | y_{CF} \in Y \setminus y_{FAC}\}$.

Given an FRBCS s, a data instance $\mathbf{x} \in X$, and the classification output y_{FAC} predicted by s, each class $y_j \in Y$ is associated with a single explanation of why $\mathbf{x}$ is classified in the given way. Hence, there exists only one factual explanation $E_{FAC}(s, \mathbf{x}, y_{FAC})$. In addition, there is a non-empty set of CF explanations $E_{CF}(s, \mathbf{x}, Y_{CF}) = \bigcup_{y_{CF} \in Y_{CF}} E_{CF}(s, \mathbf{x}, y_{CF})$ for each non-predicted class $y_{CF} \in Y_{CF}$.

Throughout the manuscript, we assume that the output is explained in its entirety if the corresponding explanation contains a factual explanation specifying why the given decision is made as well as $|Y| - 1$ CF explanations indicating why all the alternative classification options are discarded. Therefore, a (full) explanation for a data instance $\mathbf{x} \in X$ is assumed to contain one factual explanation and a non-empty set of CF explanations: $E(s, \mathbf{x}, Y) = E_{FAC}(s, \mathbf{x}, y_{FAC}) \cup E_{CF}(s, \mathbf{x}, Y_{CF})$. Accordingly, explanation generation methods aim to produce (1) a factual explanation for the test instance and (2) the most relevant CF explanations for all the CF classes.

3.2. Factual explanation generation

We design the process of explanation generation to include three main stages (text planning, sentence planning, and surface text realization) as in the NLG pipeline proposed by Reiter and Dale [35]. We selected this NLG pipeline because it is by far the most commonly used in the scientific community [14]. It is worth noting that we apply the same NLG pipeline no matter if we consider either factual or CF explanations:

- **Text planning**, where the information to be conveyed in the text is identified (content determination), as well as some order and general structure of the text is planned. In the case of CF explanations, content determination relies on relevance estimation (as described in the next section).
- **Sentence planning**, which includes grouping of messages when needed (sentence aggregation) and decisions about the words/expressions to be used (referring expression generation and/or lexicalization). This stage is crucial to avoid repetitions and make the output text more natural.

[1] The use of categorical features is out of the scope of this work.

- **Surface text realization**, which consists of generating a syntactically, morphologically, and orthographically correct text. This last stage is implemented using a pool of templates dynamically instantiated, populated and mixed with a Python wrapper of the SimpleNLG library [6].

Specifically, the factual explanation generation process presupposes the following steps: factual explanation rule selection, linguistic approximation of the feature values used in the antecedent (optionally), and linguistic realization. First, the factual explanation rule is selected from all the rules whose consequent is the predicted class. To do so, we calculate the product of the activation degree a_j of each rule $r_j \in R$ and its associated rule weight w_j, s.t. $argmax\ w_j \cdot a_j$, i.e., the factual explanation rule has the maximum product of the activation degree a_j and rule weight w_j. Second, if the rules are semantically grounded, i.e., if they use meaningful strong fuzzy partitions (SFP), the feature values in the factual explanation are readily available and mapped to the corresponding linguistic terms (e.g., "IF Color IS *Pale* AND Strength IS *Standard* THEN Beer style IS Blanche" where *Pale* and *Standard* are expert-defined linguistic terms). Otherwise, i.e., if only local semantics are available (e.g., "IF Color IS *MF0* AND Strength IS *MF1* THEN Beer-style IS Blanche" where *MF0* and *MF1* are two membership functions with local semantics), linguistic approximation is necessary to generate a meaningful explanation. Notice that the mechanism of linguistic approximation is also used for qualitative CF explanation generation and will be described in detail in the next section. Finally, once the relevant pieces of information are identified, linguistic realization is performed.

3.3. Qualitative counterfactual explanation generation

In this section, we introduce a new method for generating qualitative CF explanations (hereinafter denoted as *EUC*). This method can be regarded as an extension of our previously proposed method (hereinafter denoted as *XOR*) [43]. The *EUC* method aims to be more sensitive than *XOR* to variations in membership functions. Despite certain methodological differences, both methods form a pipeline containing the following steps to be described in detail below (see Fig. 1): CF rule representation, relevance estimation, linguistic approximation (optional in terms of the local/global semantics attached to the FRBCS), and textual explanation generation.

CF rule representation. First of all, the test instance (as well as all the CF candidates) must be represented in a compatible form. Both *EUC* and *XOR* methods reason over the information retrieved from the rule base. Multiple candidates form CF sets which are labeled in accordance with the selected linguistic terms for the given features. Thus, we regard CF sets as collections of data instances covered by the rules leading to the desired CF class. In this sense, there exist as many potential CFs as there are rules that lead to the desired CF class.

For a given FRBCS, a test instance $\mathbf{x} \in X$ can be represented as a vector $\mathbf{x} = \overline{x}_{1\times|V_T|} = \left[\mu_x(t_i)\right]_{i=1}^{|V_T|}$ of membership function values of each linguistic variable. Similarly, each CF rule can be regarded in terms of the membership function values that the linguistic variables take on. Therefore, each CF rule $r_{CF} \in R_{CF}$ is vectorized over V_T for compatibility purposes so that the collection of such vectorized rules makes up a rule-term matrix $M_{|R_{CF}|\times|V_T|}$ where the i-th row corresponds to a CF rule and the j-th column corresponds to the given linguistic term $t_j \in V_T$. Hence, the rule-term matrix is populated with such membership values as functions of a given linguistic term $M_{ij} = \mu_x(t_{ij})$.

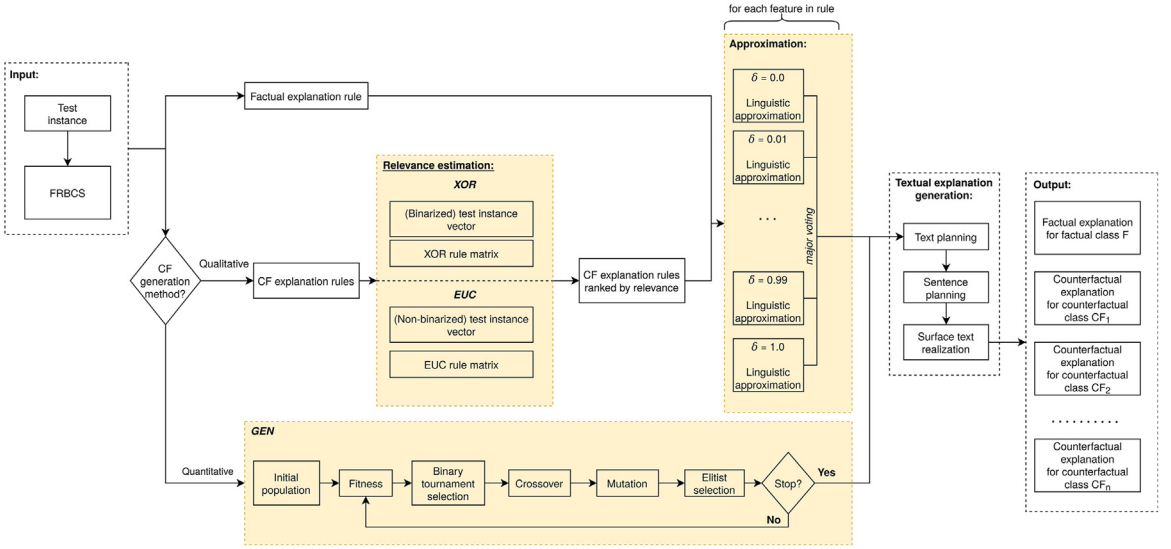

Fig. 1. CF explanation generation pipeline. The shadowed building blocks influence the surface realization of the output explanation.

It is worth noting that the *XOR* method additionally binarizes both the test instance vector and the rule-term matrix, at the cost of information loss because of the test instance and rule vectors being approximated. Instead, the *EUC* method represents the original information without further approximation. This is claimed to better capture fuzzy variable ambiguity and avoid potential information loss.

Relevance estimation. Given vector representations of the candidate CF rules, it becomes essential to identify the CF set that is minimally different from (and therefore most relevant to) the test instance. Whereas *XOR* calculates relevance by minimizing the number of different bits, *EUC* relates each vectorized CF rule to the test instance vector in a $|V_T|$-dimensional space and measures CF relevance as the Euclidean distance d between pairs of vectors $\langle \overline{x}, \overline{r_{CF_i}} \rangle, 1 \leqslant i \leqslant |R_{CF}|$; being $\overline{r_{CF_i}} = \overline{M_{i,*}}$ the vector associated to row i in matrix M, i.e., the vector which corresponds to CF rule i.

- $d_{XOR}(\overline{x}, \overline{r_{CF_i}}) = \frac{\sum_j |x^j - r^j_{CF_i}|}{|V_T|} \in [0, 1]$;
- $d_{EUC}(\overline{x}, \overline{r_{CF_i}}) = \sqrt{\sum_j \left(x^j - r^j_{CF_i}\right)^2} \in [0, \infty)$.

where x^j and $r^j_{CF_i}$ are the j-th elements in vectors $\overline{x}$ and $\overline{r_{CF_i}}$, respectively.

The candidate CF rules are then ranked in accordance with the given distance metric. Subsequently, we include the minimally distant (or most relevant) CF rules for each CF class in the pool E_{CF} of the resulting CF explanations for the given test instance **x**. If multiple CF rules are equally minimally distant from **x**, such rules are deemed equally explanatory. In this case, the most relevant CF is selected randomly. Representing the test instance and CF rules in a Euclidean $|V_T|$-dimensional space is hypothesized to better capture fuzzy-specific properties of an FRBCS. For example, the Euclidean distance appears more sensitive to changes in membership function values. The number of unique values that the *XOR*-based distance can take on is limited by $|V_T|$. In consequence, several CF rules may result in having the same relevance score while being distinct in the number of features or their labeling. On the contrary, *EUC* provides a more flexible and diverse measure of relevance of different CF rules and therefore gives a better insight into the fuzzy system's behavior.

Linguistic approximation. If the linguistic terms are not based on a SFP and therefore not semantically grounded, the selected CF rule must be enhanced with an additional linguistic layer so that the output explanation is meaningful to the end user. Once the CF rules are ranked by relevance and the most relevant CF is identified, it must therefore be linguistically approximated. To do so, each fuzzy set corresponding to the linguistic term of the selected CF rule is mapped to the gold standard annotations. Note that this mapping is actionable if the α-cut is applied to such a fuzzy set given some threshold value δ. To illustrate the process of linguistic approximation, consider a fuzzy set *FS* characterized by a trapezoidal membership function and three linguistic terms ($T = \{t_1, t_2, t_3\}$) which are candidates to be associated with *FS* (see Fig. 2 for details). Given some cut-off threshold value δ_1, the fuzzy set *FS* can be projected to an interval of numerical values $L = [v_{\delta_1}, v_{\delta_2}]$. In addition, each linguistic term $t_i \in T$ can be projected to an interval $t_{i\delta 1} (1 \leqslant i \leqslant |T|)$. Then, the interval L can be compared with the intervals $t_{i\delta 1}$ using the Jaccard Similarity Index [13]:

$$\forall L \approx t^f_a \in V_T : S(t_{i\delta 1}, L) = \frac{t_{i\delta 1} \cap L}{t_{i\delta 1} \cup L} \in [0, 1], \tag{1}$$

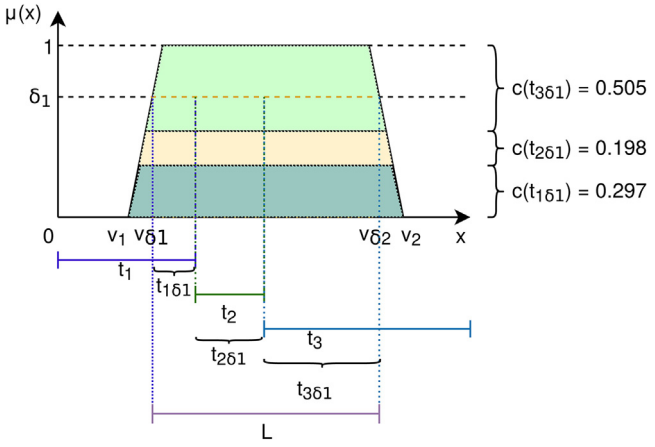

Fig. 2. Illustrative example of the linguistic approximation mechanism.

Table 1
Approximation confidence score calculation.

Term	δ	Approximation confidence
t_1	[0.0, 0.3)	30/101 = 0.297
t_2	[0.3, 0.5)	20/101 = 0.198
t_3	[0.5, 1.0]	51/101 = 0.505

where $t_{i\delta 1}$ is the numerical interval closer to the linguistic term t_i, and L is the numerical interval associated to the selected α-cut. As follows from Fig. 2, $S(t_{3\delta 1}, L) > S(t_{2\delta 1}, L) > S(t_{1\delta 1}, L)$. Hence, the feature f_j characterized by fuzzy set *FS* is verbalized as "f_j is t_3" in this case.

Note that the threshold value δ for the α-cut serves as a hyperparameter. The previously proposed *XOR* method uses heuristics to specify δ manually. Instead, both qualitative CF generation methods now use major voting in order to reduce possible approximation error. Thus, given some small enough *step*, we inspect all the approximated linguistic terms over the cut-off interval $[0, 1]$ for each term in the given CF rule and assign a confidence score to each term t_i as follows: $c(t_i) = \frac{\#t_i}{1+\frac{1}{step}}$, being $\#t_i$ the number of times t_i is the winner.

For each feature f_j involved in the classification and considered in the output explanation, we apply major voting to identify which linguistic term is covered by the widest range of the inspected approximations using the approximation confidence score $c(t_i)$ as a reference, so that the selected linguistic term is $t_j \in V_T | argmax\ c(t_j)$. Considering the example in Fig. 2, let *step* be 0.01. We therefore perform $n = 1 + 1/0.01 = 101$ linguistic approximations. Suppose that the term under consideration is mapped to the set of linguistic terms as indicated in Table 1.

Approximation confidence scores are calculated for all the competing linguistic terms. Since we aim to use the most frequently found term among all the considered threshold values, the linguistic term that has the highest score (in this case, t_3) is selected for the output explanation. It is worth noting that in this illustrative example, the selected linguistic term is the same as the one selected when considering only δ_1. However, in the general case they may be different. Therefore, it is recommended to follow the major voting approach instead of relying only on a single δ value selected heuristically.

As only two building blocks (relevance estimation and linguistic approximation) influence the output explanation (see the shadowed blocks in Fig. 1), *XOR* and *EUC* generate CFs following one of the three scenarios below:

- the two methods select the same rule to be the most relevant, the approximation algorithm gets the same semantically grounded linguistic terms;
- the two methods select two different CF rules (e.g., "IF f_1 IS MF_0 and f_2 IS MF_0 THEN y_{cf}" and "IF f_1 IS MF_1 and f_2 IS MF_1 THEN y_{cf}") which nevertheless generate identical CF explanations due to a large enough overlap between the corresponding fuzzy sets. This scenario is possible when all the features used in both rules are identical and their non-semantically grounded values overlap to a large enough extent;
- the two methods select two different CF rules (e.g., "IF f_1 IS MF_0 and f_2 IS MF_0 THEN y_{cf}" and "IF f_1 IS MF_2 and f_3 IS MF_4 THEN y_{cf}") where feature values are approximated to different linguistic terms.

Textual explanation realization. At the last stage, the selected factual and CF pieces of information are converted to explanations in natural language while applying the NLG pipeline introduced in the previous section. It is worth noting that the text and sentence planning along with text realization for a factual explanation follow the structure of the corresponding winner rule from the rule base. Thus, a factual explanation is assumed to include a subordinate clause of cause (e.g., "The data instance x is of class y_f because f_1 is v_1 and f_2 is v_2"), which lists the features and the corresponding values or linguistic terms that influenced the actual decision. On the other hand, a CF explanation is verbalized in natural language as a complex conditional sentence that adopts the structure of the rule, e.g., "x would be of class y_{cf} if f_1 were v_2 and f_3 were v_4" for the given CF class y_{cf}.

Implementation details. The *XOR* and *EUC* methods are implemented as open source software in Python and are made publicly available at a Gitlab repository[2].

3.4. Quantitative counterfactual explanation generation

In this section, we present a new method for CF explanation generation which is grounded in evolutionary and bio-inspired computation algorithms for explainable AI [11]. More precisely, we have implemented a Genetic Algorithm (hereafter denoted as *GEN*) which takes as the starting point the genetic fuzzy tuning approach previously proposed by Alonso et al. [4]. Indeed, the original algorithm was first introduced by Cordon and Herrera [7] and later adapted to explainable SFP tuning in [4].

GEN manages a population P with N individuals which evolve in g generations. The given test instance $\mathbf{x}$ is used for building the first individual of the population. Each individual is associated to a real-coded chromosome which is made up of p

[2] https://gitlab.citius.usc.es/ilia.stepin/fcfexpgen (branch "xor_euc_gen")

genes, with each gene representing one of the features in F. Since all the features are numerical, gene $i \in [1, p]$ encodes the double value associated to feature i. The rest of the population is generated randomly. Thus, a random value is assigned to each gene i within its variation interval which is determined by the numerical range associated to feature i. The pseudocode of the developed algorithm is as follows (see the *GEN* shadowed block in Fig. 1):

1. **Initialize** the generation counter, $g = 0$, and evaluate the initial population, $P^{(0)}$. Evaluating a population means computing *Fitness* for each individual in the population. Here, *Fitness* is computed as the Euclidean distance between the data instance $\hat{x}$ associated to the current chromosome and the original test instance $\mathbf{x}$, if the inferred output is in agreement with the target CF class. Otherwise, *Fitness* equals the maximum distance which comes out from the Euclidean distance between the two vectors representing the extreme values (min/max) for the variation intervals associated to each feature. Hence, the smaller *Fitness*, the better.
2. **while** $g <$ *MaxGener* **and** *Fitness* $\geqslant$ *StopThres* **and** *Nbest* $\leqslant$ *NrepThres*

   $g := g + 1$
   Select $P^{(g)}$ *from* $P^{(g-1)}$
   Crossover $P^{(g)}$
   Mutate $P^{(g)}$
   Elitist selection $P^{(g-1)}$
   Evaluate $P^{(g)}$

   end while

The procedure ends either when the maximum number of generations (*MaxGener*) is reached, or *Fitness* is under the predefined threshold (*StopThres*), or the number of consecutive generations for which the best fitness value remains the same (*Nbest*) is greater than the predefined threshold (*NrepThres*). On the one hand, *MaxGener* should be defined empirically in terms of the complexity of the dataset under consideration. It must be large enough to guarantee that *GEN* converges to a good enough solution. On the other hand, *StopThres* and *NrepThres* are threshold values to speed up the procedure, so that the algorithm stops before *MaxGener* is reached in case *Fitness* is small enough or becomes constant for a large enough number of generations. For each generation, the following steps are repeated:

- The selection of $P^{(g)}$ from $P^{(g-1)}$ is made as a deterministic tournament selection procedure. Each individual in the new population, $P^{(g)}$, is chosen from the previous one, $P^{(g-1)}$, after making a tournament that involves *TS* individuals randomly selected from $P^{(g-1)}$. The best individual is selected in any tournament. The selection pressure can be adjusted by changing $TS \leqslant N$. The larger *TS*, the smaller the chance of weak individuals to be selected. For example, if $TS = N$, then all the individuals in $P^{(g)}$ are equal to the best one in $P^{(g-1)}$, what is unsatisfactory from the point of view of diversity in the population.
- The $BLX - \alpha$ crossover operator [10] is applied to $P^{(g)}$. The parents, i.e., the selected chromosomes in the current population, are crossed over in pairs. Each pair of parents, $dad = (d_1, \cdots, d_p)$ and $mom = (m_1, \cdots, m_p)$, is replaced in the new population by two offsprings, $O_d = (o_{d1}, \cdots, o_{dp})$ and $O_m = (o_{m1}, \cdots, o_{mp})$, where o_{dj} and o_{mj} are random values from the intervals $[min_{dj}, max_{dj}]$ and $[min_{mj}, max_{mj}]$, respectively. I_j=$[I_j^l, I_j^u]$ is the variation interval of gene j. According to the taxonomy for the crossover operator presented by [21], $\alpha = 0.3$ is a suitable value for letting $BLX - \alpha$ exploit the nature of real coding as follows:

  $$min_{dj} = maximum\left(I_j^l, d_j - \alpha \cdot |d_j - m_j|\right)$$
  $$max_{dj} = minimum\left(d_j + \alpha \cdot |d_j - m_j|, I_j^u\right)$$
  $$min_{mj} = maximum\left(I_j^l, m_j - \alpha \cdot |m_j - d_j|\right)$$
  $$max_{mj} = minimum\left(m_j + \alpha \cdot |m_j - d_j|, I_j^u\right)$$

- A uniform mutation operator is considered. The value of the selected gene is changed by another one generated randomly within its variation interval.
- The elitist selection ensures perpetuating the best individual from the given generation to the next one. If the best individual, B_i in $P^{(g-1)}$, is not included in $P^{(g)}$, then the worst individual in $P^{(g)}$ is replaced by B_i.

Once *GEN* ends, we have identified a new data instance $\hat{x}$ that is assumed to minimally change the original test instance **x** while making the FRBCS infer the desired CF output[3]. Then, it is time for generating the related CF explanation in natural language. To do so, we once again apply the NLG pipeline described previously. First of all, we compute the percentage of modification $D_j = 100 * \frac{\hat{x}_j - x_j}{I_j}$ associated to each feature j to go from **x** to $\hat{x}$. The text which describes D_j is as follows: x_j is [*slightly*] *increased* | *decreased*; where *increased* appears if $D_j > 0$. On the contrary, *decreased* is used if $D_j < 0$. In addition, the linguistic modifier *slightly* appears only in case of small modifications, i.e., only if $0.9 \leqslant D_j \leqslant 5$, which means the percentage of modification is smaller or equal than 5%. Notice that nothing is said about feature j if $D_j < 0.9$. In this case, we consider the feature j to remain the same assuming that such a small change (less than 0.9%) does not have sufficient explanatory power for the recipient of the explanation. This assumption is made heuristically in accordance with our previous experience with designing NLG systems while keeping in mind the limited processing capability of human beings [28]. As a result, the generated textual explanations are shorter and easier to process while referring only to relevant changes.

Afterwards, at the sentence planning stage, for the sake of simplicity and naturalness, we aggregate those pieces of information associated to different features which are affected by the same type of modification (e.g., "f_1 and f_2 are *slightly increased*" replaces to "f_1 is *slightly increased* and f_2 is *slightly increased*"). We also apply lexicalization for each feature to be described in a fully meaningful way. Therefore, *increased* and *decreased* are replaced by more meaningful terms (e.g., *strength is bigger* or *color is darker*).

Finally, text realization is done again using the following template and the SimpleNLG library with the aim of ensuring syntactically, morphologically and orthographically correct final text: "[Output Class Name] *would be* [CF Class Name] *if* [Name of the most Relevant Feature$_j$] *were* [linguistic description of D_j] (new data value) [AND...]". Notice that the new values for the features associated with the most relevant changes are given in brackets.

Implementation details. The *GEN* method is implemented as a piece of open source software in Python and is made publicly available at a Gitlab repository[4]. It is also integrated with the open source software GUAJE[5] which is devoted to facilitating the design of explainable fuzzy systems [3]. The following *GEN* parameters are considered when generating the quantitative CF explanations under evaluation in the rest of the paper: population length ($N = 30$), tournament size ($TS = 2$), mutation probability ($mprob = 0.1$), crossover probability ($cprob = 0.8$), α-crossover ($\alpha = 0.3$), *MaxGener* = 1000, *StopThres* = 0, *NrepThres* = 30. The interested reader is kindly referred to Appendix A for further details about how such parameters were selected.

4. Evaluation design

In this section, we specify some of the key features that subsequent human evaluation studies rely upon. Section 4.1 introduces the dataset and FRBCS whose classifications are explained. Then, Section 4.2 presents a novel metric for measuring the complexity of automated explanations.

4.1. Dataset and fuzzy inference system

The experiments have been carried out using the BEER dataset[6]. It contains characteristics of 400 instances of beer each of which belongs to one of 8 classes (Blanche, Lager, Pilsner, IPA, Stout, Barleywine, Porter, or Belgian Strong Ale). All data instances are described in terms of three features: color, strength, and bitterness. The corresponding linguistic terms and their ranges of values are displayed in Table 2. It is worth noting that all linguistic terms are commonsense and fully meaningful because they were provided by expert brewers.

In our experiments, we generate explanations for an FRBCS associated with the Fuzzy Unordered Rule Induction Algorithm (FURIA) [22]. The min–max inference mechanism [27] is applied so that both conjunction (AND) and implication (THEN) are implemented by the t-norm minimum, and the output accumulation is done by the t-conorm maximum. All membership functions are trapezoidal. All rule weights are set to the default value of 1. In addition, it is necessary to apply linguistic approximation as part of the explanation generation pipeline because FURIA rules are endowed only with local semantics. It is worth noting that such a linguistic approximation makes use of meaningful SFP-based linguistic terms as well as their combinations. Thus, explanations may contain combinations of adjacent terms (e.g., "*Feature*$_1$ is *Term*$_1$ or *Term*$_2$") with the aim of enhancing further their explanatory capacity. Fig. 3 illustrates the SFP associated to color.

In this work, we use the same FRBCS that was previously designed and evaluated in [43] with 10-fold cross-validation, achieving 95.5% of correctly classified instances and F1-score equals 0.954 (see the confusion matrix in Table 3 for further details). Notice that, with the aim of avoiding generation of misleading explanations and mainly because the present work focuses on the intended human evaluation, the misclassified test instances are excluded from further analysis in the rest of this manuscript. Whereas explaining misclassification is a challenging problem, it falls outside the scope of this work.

[3] Due to the well-known random heuristic nature of genetic algorithms, they avoid stacking in a local minimum but they can not always guarantee the convergence to the global minimum. Anyway, as shown in Appendix A, GEN succeeds to be effective in the search of "sub-optimal" solutions which are expected to be close enough to the optimal one.

[4] https://gitlab.citius.usc.es/ilia.stepin/fcfexpgen (branch "xor_euc_gen")

[5] https://gitlab.citius.usc.es/jose.alonso/guaje/

[6] The BEER dataset is publicly available athttps://dx.doi.org/10.13140/RG.2.2.20313.67680

Table 2
Numerical intervals associated to each SFP-based linguistic term.

Feature	Linguistic term	Range of values
Color	Pale	[0.0, 3.0]
	Straw	[3.0, 7.5]
	Amber	[7.5, 19.0]
	Brown	[19.0, 29.0]
	Black	[29.0, 45.0]
Bitterness	Low	[7.0, 21.0]
	Low-medium	[21.0, 32.5]
	Medium–high	[32.5, 47.5]
	High	[47.5, 250.0]
Strength	Session	[0.035, 0.052]
	Standard	[0.052, 0.067]
	High	[0.067, 0.090]
	Very high	[0.090, 0.136]

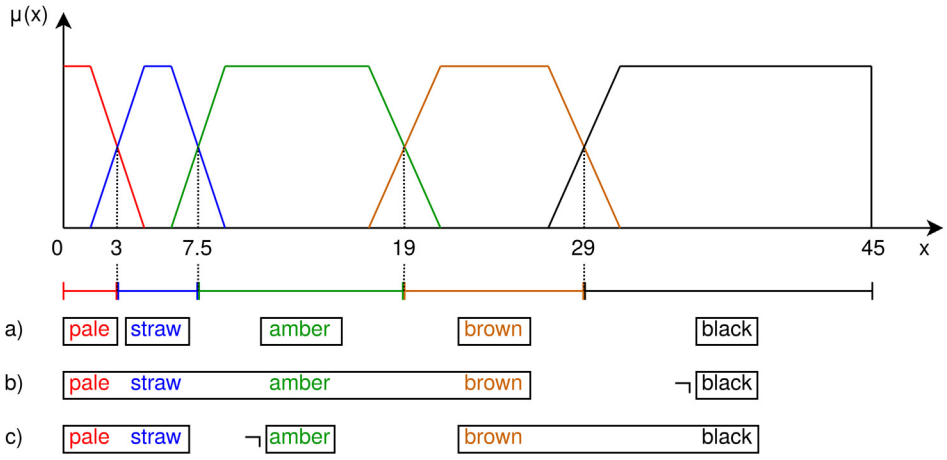

Fig. 3. Interpretation of SFP-based linguistic terms associated to Color.

Table 3
FURIA confusion matrix. UC stands for Unclassified instances.

	Predicted class								
Observed class	**BLA**	**LAG**	**PIL**	**IPA**	**STO**	**BAR**	**POR**	**BSA**	**UC**
Blanche (BLA)	50								
Lager (LAG)		48	1				1		
Pilsner (PIL)		1	49						
IPA			1	43		5			1
Stout (STO)					50				
Barleywine (BAR)				5		43	1		1
Porter (POR)		1			1		47		1
Belgian Strong Ale (BSA)					1	1	1	47	

4.2. Perceived explanation complexity

The use of explanations in natural language poses the problem of adequate estimation of explanation complexity. For example, it remains unclear whether the use of adjacent linguistic terms in an explanation (e.g., "*...if color were pale or straw*") increases or decreases understandability (and therefore effectiveness and usability) of such an explanation.

As the starting inspiring point for our proposal of automatic calculation of explanation complexity, we refer to existing readability tests in linguistics, which estimate how easily a text can be read by the intended audience. More precisely, the well-known *Gunning Fog Index* [19] is the weighted average of the normalized sentence length and the percentage of complex words in the text. Similarly, an estimate of complexity of a feature-based linguistic explanation (as perceived by the end user) may rely on the explanation length as well as on the number of features and linguistic terms used in the explanation.

In light of the above, we formally define the perceived explanation complexity (*PEC*) of an automated explanation e as follows:

$$PEC(e) = \lambda * \frac{min(l(e), \sigma)}{\sigma} + (1 - \lambda) * \frac{1}{|F|} \sum_{i=1}^{F_e} \frac{t^{f_i}}{|T_L^{f_i}|} \quad (2)$$

where $\lambda \in [0, 1]$ is the weight regularizing the impact of the explanation length and number of features and terms used in the explanation, $l(e)$ is the explanation length in characters, σ is a normalization hyperparameter over the explanation length, $|F|$ is the total number of features in the dataset, F_e is the number of unique features used in the given explanation, t^{f_i} is the number of terms associated with the i-th feature used in the explanation, $|T_L^{f_i}|$ is the power of the set of linguistic terms of the i-th feature.

In the case of the qualitative methods *XOR* and *EUC*, the basic linguistic terms to take into account are those already described in Table 2. However, in order to guarantee a fair comparison between quantitative and qualitative CF explanations, it is necessary to linguistically represent numerical feature value changes suggested by the quantitative method *GEN*. The sets of linguistic terms associated to each feature by the *GEN* method are the following:

$T_L(Color) = \{darker, slightly\ darker, lighter, slightly\ lighter\}$.

$T_L(Bitterness) = \{smaller, slightly\ smaller, bigger, slightly\ bigger\}$.

$T_L(Strength) = \{smaller, slightly\ smaller, bigger, slightly\ bigger\}$.

To illustrate computation of *PEC(e)*, let us consider the following example: given a data instance, $\lambda = 0.5$ and $\sigma = 150$, we have three alternative CF explanations with their corresponding complexity scores.

- *XOR*: "Beer style would be Stout if color were black."
  $PEC\ (e) = 0.5\ *\ \frac{46}{150} + 0.5\ *\ \frac{1}{3}\ *\ \frac{1}{5} = 0.153 + 0.033 = 0.186$
- *EUC*: "Beer style would be Stout if bitterness were low or low-medium, color were black, and strength were standard or high or very high."

  $PEC(e) = 0.5\ *\ \frac{130}{150} + 0.5\ *\ \frac{1}{3}\ *\ \left(\frac{2}{4} + \frac{1}{5} + \frac{3}{4}\right) = 0.433 + 0.242 = 0.675$

- *GEN*: "Beer style would be Stout if color were bigger (30.501) and strength were smaller (0.078)."

  $PEC(e) = 0.5\ *\ \frac{90}{150} + 0.5\ *\ \frac{1}{3}\ *\ \left(\frac{1}{4} + \frac{1}{4}\right) = 0.300 + 0.083 = 0.383$

Noteworthy, it always holds that $PEC(e) \in [0, 1]$. *PEC(e)* is null only if the explanation is empty and the associated weight $\lambda = 1$. On the contrary, the highest value of *PEC(e)* is obtained when the explanation length is equal to the normalization hyperparameter σ or all the dataset features and all the linguistic terms are included in the explanation. However, both of these special cases are of no interest, as the empty explanation has got null explanatory power whereas explanation including all the possible categories of features is clearly misleading.

5. Human evaluation

The human evaluation study consisted of two online questionnaires that allowed us to assess how the metric *PEC* is related to different explanation aspects. Section 5.1 presents the instruments and design of the first questionnaire (hereinafter referred to as *Survey GM* because the items to rate are associated to the so-called *Gricean Maxims* [15] as we will show below) as well as the analysis of collected data and the discussion of main results. In the light of lessons learned from this survey, we developed a subsequent one (hereinafter referred to as *Survey TS* because the focus is on assessing *Trustworthiness* and *Satisfaction* of the given explanations) whose experimental design and main discoveries are described in Section 5.2. In both surveys, all the subjects participated voluntarily and anonymously. This research obtained ethics approval from the University Ethics committee.

5.1. Survey GM: Evaluating CF explanations in terms of Gricean Maxims

5.1.1. Experimental settings

The first experiment was designed as a within-subject study. In order to perform a comparative analysis of qualitative and quantitative CF explanations, we considered only those test instances for which the qualitative methods (*XOR* and *EUC*) generated distinct explanations (thus avoiding misleading repetitions).

Since the BEER dataset has 8 classes, given a test instance we have 1 factual class and 7 alternative CF classes. Because the FURIA rules were trained and evaluated with 10-fold cross-validation, the 400 data instances in the BEER dataset were split 10 times into training set (90%) and test set (10%). As a result, we built 10 sets of FURIA rules. They were used to make predictions for all test instances in each fold (see details in Table 4). Then, we filtered out unclassified and misclassified test instances with the aim of avoiding the inclusion of void or misleading explanations to be evaluated in the survey. Notewor-

Table 4
Screening of test instances for defining the survey stimuli in terms of *XOR* and *EUC* CF explanations generated fold by fold. Unclassified instances are those for which no rule was activated. Misclassified instances are those where the FRBCS prediction does not match the ground-truth class label. Wrong factuals correspond to test instances for which wrong factual explanations were generated.

Fold	CV0	CV1	CV2	CV3	CV4	CV5	CV6	CV7	CV8	CV9
Test instances	40	40	40	40	40	40	40	40	40	40
Unclassified instances	–	–	2	–	–	1	–	–	–	–
Misclassified instances	2	2	3	1	3	1	2	–	3	3
Wrong factuals	1	–	1	–	1	2	–	1	–	2
Screened instances	37	38	34	39	36	36	38	39	37	35
CF explanations	259	266	238	273	252	252	266	273	259	245
Distinct CF pairs	25	28	29	24	20	36	32	74	28	38
Unique CF pairs	6	5	6	4	2	7	8	10	3	5

Table 5
Test instances and the corresponding CF explanations under study.

Task	Feature			Factual class	CF class	CF explanations		
	Color	Bitterness	Strength			XOR	EUC	*GEN*
1	17	87	0.096	Barleywine	IPA	if strength were high (*PEC=0.195*)	Beer style would be IPA if color were pale or straw or amber and strength were session or standard or high (*PEC=0.582*)	if strength were smaller (0.085) (*PEC=0.232*)
2	8	69	0.083	IPA	Barleywine	if strength were very high (*PEC=0.235*)	Beer style would be Barleywine if color were amber or brown or black and strength were very high (*PEC=0.465*)	if strength were bigger (0.096) (*PEC=0.252*)
3	5	34	0.068	Pilsner	Lager	if bitterness were low or low-medium or medium-high and color were amber (*PEC=0.488*)	Beer style would be Lager if bitterness were low or low-medium or medium-high and color were straw or amber (*PEC=0.552*)	if color were slightly darker (6.500) (*PEC=0.255*)
4	28	38	0.091	Belgian Strong Ale	Stout	if color were black (*PEC=0.186*)	Beer style would be Stout if bitterness were low or low-medium, color were black, and strength were standard or high or very high (*PEC=0.675*)	if color were darker (30.501) and strength were smaller (0.078) (*PEC=0.383*)
5	3	16	0.054	Blanche	Porter	if color were brown and strength were session or standard (*PEC=0.400*)	Beer style would be Porter if bitterness were low-medium or medium-high or high, color were brown, and strength were session or standard (*PEC=0.699*)	if color were darker (16.001) and strength were slightly smaller (0.052) (*PEC=0.416*)

thy, only 3 out of the 400 (0.75%) test instances (across all the folds) were unclassified, whereas 20 out of the 400 (5%) test instances were misclassified. Then, we generated CF explanations for each given prediction using the qualitative methods *XOR* and *EUC*. All in all, after careful screening, we identified all unique pairs of distinct CFs to exclude pieces of repeated explanations from the survey. Then, we picked 5 test instances representing illustrative cases (among the instances associated to all the previously identified unique CF pairs) that would be used as stimuli in the human evaluation study. Afterwards, we generated quantitative CF explanations for the selected stimuli using the *GEN* method.

Hence, *Survey GM* includes the following 5 tasks which were presented in a randomized order to each subject (see Table 5 for details). *Task 1* (predicted class: Barleywine, CF class: IPA) and *Task 2* (predicted class: IPA, CF class: Barleywine) represent pairs of classes where the classifier predicted the greatest number of incorrect results (see the confusion matrix from Table 3 in the previous section for details). Hereinafter we therefore refer to the first two tasks as "confusing" (CONF) while the rest of the tasks are deemed "non-confusing" (NON-CONF). In addition, the last three stimuli were selected by their relation to color. In *Task 3* (predicted class: Pilsner, CF class: Lager) both the predicted and CF classes are characterized by low values of color (i.e., from Pale to Amber in Table 2). In contrast, *Task 4* (predicted class: Belgian Strong Ale, CF class: Stout) the corresponding classes presume high values of color (i.e., Brown or Black in Table 2) for both factual and CF classes. Finally, the stimulus for *Task 5* (predicted class: Blanche, CF class: Porter) was selected to have contrastive values of color for the predicted and CF classes (i.e., Pale for Blanche versus Black for Porter).

Table 6
Explanation aspects under evaluation in *Survey GM*.

Related maxim of	Evaluation aspect	Description
Quantity	Informativeness	An estimate of how complete a CF explanation is perceived to be
Quality	Trustworthiness	An estimate of how credible a CF explanation is perceived to be
	Accuracy	An estimate of how precisely a CF explanation describes the CF class instances
Relevance	Relevance	An estimate of how pertinent the CF explanation details are in order to make a minimal change in feature values
Manner	Readability	An estimate of how grammatical a CF explanation is perceived to be

Table 7
Self-reported demographic data (*Survey GM*). The number of subjects comes along with the percentage in brackets for each category.

Demographic parameter	Number of participants
(a) Age	
18–25	3 (16.67%)
26–35	7 (38.89%)
36–45	5 (27.78%)
46–55	3 (16.67%)
(b) Gender	
Male	15 (83.33%)
Female	2 (11.12%)
Preferred not to say	1 (5.55%)
(c) Education	
Doctorate (Ph.D)	10 (55.56%)
Master's (M.A./M.Sc.)	7 (38.89%)
Bachelor's (B.A./B.Sc.)	1 (5.55%)
(d) English proficiency	
Native speaker	3 (16.67%)
Proficient (C2)	7 (38.89%)
Advanced (C1)	4 (22.22%)
Upper intermediate (B2)	4 (22.22%)
(e) Areas of expertise	
Explainable AI	12 (66.67%)
Fuzzy logic	9 (50.00%)
Mathematics	6 (33.33%)
Engineering	8 (44.44%)
Computer science	14 (77.78%)
Computational linguistics	4 (22.22%)
Social sciences	1 (5.56%)

The survey was implemented as an online questionnaire[7] which was developed in Python[8]. Each task screen included two panels. On the left panel, the factual explanation was given in the upper-left corner (for reference only) followed by three different CF explanations, each corresponding to one of the methods under study. The given test instance was depicted as a parallel coordinates plot below the explanations. On the right panel, the subjects were asked to rate each CF explanation on a 7-point Likert scale regarding several explanation aspects which are linked to the following Gricean Maxims [8,15]: *Maxim of quantity* (make your contribution as informative as is required without making it more informative than required); *Maxim of quality* (do not give information that is untruthful or lacks evidence); *Maxim of relevance* (present information pertinent to the discussion); and *Maxim of manner* (be clear and orderly, avoid ambiguity and obscurity).

It is worth noting that these four maxims were transformed into five explanation aspects (see Table 6). First, *informativeness* is related to the maxim of quantity and estimates whether the information present in the explanation sufficiently describes a necessary feature perturbation and whether it contains any unnecessary information. Then, the maxim of quality is represented by two explanation aspects. On the one hand, *trustworthiness* measures how credible the suggested changes are perceived (without them necessarily being accurate). On the other hand, *accuracy* indicates whether the suggestions found in the explanation are perceived to be correct and truly leading to the desired different outcome. In addition, the aspect of *relevance* aims to estimate how adequate the suggested changes are with respect to the test instance characteristics. Further, the aspect of *readability* estimates how grammatical and easy to read the given explanation is. Finally, in order

[7] https://tec.citius.usc.es/qxaisurvey1/
[8] https://gitlab.citius.usc.es/jose.alonso/surveygenerator

to estimate a degree of association between *PEC* and the estimated explanation aspects, Spearman's rank correlation coefficients (ρ) were calculated pairwise for the *PEC* scores and mean human evaluation scores of each explanation aspect. The threshold of $p = 0.05$ was used to confirm whether the correlation between *PEC* and the given explanation aspect exists.

5.1.2. Results

A total of 18 subjects participated in the *Survey GM*, each evaluating all the three explanation generation methods. All the demographic data collected from participants in *Survey GM* as well as their self-reported areas of expertise can be found in Table 7. To sum it up, 15 participants were males (83.33%), two were females (11.12%), and one person (5.55%) did not disclose its gender. In addition, all the participants held at least a Bachelor degree and had expertise in a wide range of sciences. Further, all the participants had at least the B2 level of English proficiency and represented various areas of expertise. Note that participants were allowed to select multiple areas.

Table 8 shows the mean and median human evaluation scores in *Survey GM* as well as the corresponding standard deviation (St.dev.). On average, the *EUC* explanations are perceived more informative than *GEN* or *XOR* explanations. However, the quantitative *GEN* method is perceived to generate more trustworthy explanations, *XOR* explanations being the second most credible, and the *EUC* method offering the least trustworthy explanations among the three methods. The *GEN* explanations are found more accurate than those generated by *XOR* and *EUC* methods. The *GEN* method also appears to generate more relevant explanations than *XOR* and *EUC*. Nevertheless, *XOR* explanations are perceived as grammatical as those offered by *GEN*, with *EUC* offering the least readable explanations, possibly due to their increased length.

As we consider all the sample explanations collectively, we observe important correlations between *PEC* and averaged scores for several explanation aspects. Thus, explanation complexity is found to moderately correlate with informativeness ($\rho = 0.545, p = 0.036$). In addition, strong negative correlations are observed between *PEC* and relevance ($\rho = -0.688, p = 0.005$) but also between *PEC* and readability ($\rho = -0.871, p < 0.001$). On the other hand, no conclusion can be made regarding the correlation either between *PEC* and trustworthiness ($\rho = -0.3, p = 0.278$) or between *PEC* and accuracy ($\rho = 0.07, p = 0.804$).

As for the "confusing" tasks alone, a strong negative correlation is found only between *PEC* and trustworthiness ($\rho = -0.87, p = 0.024$). The human evaluation scores for the other explanation aspects do not allow us to draw any other significant conclusions on their association with *PEC*. As for the "non-confusing" tasks alone, the findings testify that more complex explanations are perceived less readable ($\rho = -0.983, p < 0.001$).

The main lessons learned from this survey are as follows: (1) most participants agreed that the online questionnaire was long because it involved many different evaluation aspects for the three different methods; and (2) *PEC* turned out to be a good estimate for some of the explanation aspects under study. Then, we may take profit from these facts when designing future surveys: provide subjects with short questionnaires that regard only those specific aspects which cannot be inferred from *PEC*.

5.2. Survey TS: Evaluating Trustworthiness and Satisfaction of explanations

5.2.1. Experimental settings

In the light of lessons learned from previous survey, we defined a subsequent one. *Survey TS* was designed to have a simplified structure and follow a between-subject design where each subject would assess only one given explanation generation method. We considered the same stimuli as in the previous survey but focused only on trustworthiness and satisfaction of explanations instead. In the new questionnaire[9], the subjects were asked to evaluate the given CF explanation only in terms of trustworthiness and satisfaction. In addition, we adhered to the DARPA[10] [18] guidelines for assessing these explanation aspects on a 5-point Likert scale.

As designed previously, the task screens were presented in a randomized order to each subject. Similarly to *Survey GM*, Spearman's rank correlation coefficients (ρ) were calculated to estimate the association between *PEC* scores and human evaluation scores for trustworthiness and satisfaction. The same threshold value of $p = 0.05$ was used to verify whether such correlations existed.

5.2.2. Results

Sixty subjects participated in *Survey TS*. Each method was assessed by 20 participants independently. All the demographic data collected from participants are detailed in Table 9. Out of all the participants, a total of 57 (95 %) disclosed their demographic data. Thus, 32 of all the participants reported to be males (56.14%), 21 participants were females (36.84%) whereas 4 people (7.02%) preferred not to indicate their gender. Similarly to *Survey GM*, all the participants self-assessed their English language proficiency to be of at least the B2 level, and 53 out of 57 subjects disclosed their area of expertise.

Table 10 summarizes the human evaluation scores in *Survey TS*. Regarding trustworthy, *XOR* and *GEN* explanations are on average perceived nearly the same, the *EUC* explanations slightly falling behind. A similar pattern is observed for satisfaction. The quantitative *GEN* explanations are, in general, found to be the most satisfying. Nevertheless, the qualitative *XOR* expla-

[9] https://tec.citius.usc.es/cfsurvey/

[10] The acronym DARPA stands for Defense Advanced Research Projects Agency, which is the research and development agency of the USA.

Table 8
Survey GM results. ALL corresponds to the average for the five tasks. CONF averages only confusing tasks (1 and 2). NON-CONF averages only non-confusing tasks (3, 4 and 5). The highest average values for each (group of) task(s) and explanation aspect are highlighted in bold. Notice that, *PEC* values for ALL, CONF, and NON-CONF are averaged scores for the corresponding groups of tasks.

Task	Method	*PEC*	Informativeness			Trustworthiness			Accuracy			Relevance			Readability		
			Mean	Median	St.dev.	Mean	Median	St.dev.	Mean	Median	St.dev.	Mean	Median	St.dev.	Mean	Median	St.dev.
1	*XOR*	0.195	4.667	4.500	1.152	4.889	5.000	1.451	4.667	5.000	1.609	4.722	5.000	1.447	**5.778**	7.000	1.592
	EUC	0.582	**5.333**	5.000	1.188	4.333	5.000	1.749	4.667	5.000	1.814	3.611	3.500	1.754	4.500	4.500	1.917
	GEN	0.232	5.222	6.000	1.517	**5.222**	5.000	1.166	**5.500**	6.000	1.581	**5.056**	6.000	1.697	5.556	6.000	1.580
2	*XOR*	0.235	4.611	4.500	1.614	**4.778**	5.000	1.396	**4.778**	5.000	1.592	5.111	5.000	1.323	**6.222**	7.000	1.003
	EUC	0.465	**5.000**	5.000	1.414	4.222	5.000	1.865	3.889	4.500	2.083	4.611	5.000	1.685	4.333	4.500	1.815
	GEN	0.252	4.722	4.500	1.742	**4.778**	5.000	1.353	4.667	5.000	1.715	**5.278**	5.500	1.487	5.611	6.000	1.501
3	*XOR*	0.488	4.722	5.000	1.526	4.111	4.000	1.676	**4.444**	4.000	1.688	4.389	5.000	1.335	4.556	4.500	1.617
	EUC	0.552	**4.833**	5.000	1.339	**4.333**	5.000	1.414	4.167	4.000	1.618	4.167	4.000	1.383	4.333	4.500	2.058
	GEN	0.255	4.333	4.500	1.572	4.167	4.500	1.791	4.278	4.000	1.447	**4.667**	5.000	1.572	**6.000**	6.000	1.237
4	*XOR*	0.186	4.500	4.000	1.581	4.389	4.500	1.819	3.889	3.500	1.676	4.667	4.500	1.879	**6.278**	6.000	0.826
	EUC	0.675	5.389	6.000	1.501	4.944	5.000	1.211	4.778	5.000	1.734	4.278	5.000	1.487	3.833	3.000	2.121
	GEN	0.383	**5.556**	6.000	1.338	**5.833**	6.000	1.200	**5.833**	6.000	1.339	**5.611**	5.500	1.290	5.778	6.000	1.166
5	*XOR*	0.400	4.722	5.000	1.638	**5.000**	5.000	1.495	5.000	5.000	1.283	**4.889**	5.000	1.451	**5.667**	6.000	1.609
	EUC	0.699	**4.889**	5.000	1.231	4.667	5.000	1.534	**5.056**	5.000	1.474	4.389	4.000	1.787	4.056	4.000	1.893
	GEN	0.416	**4.889**	5.000	1.323	4.333	5.000	1.749	4.833	5.000	1.791	**4.889**	5.500	1.676	5.389	6.000	1.539
ALL	*XOR*	0.301	4.644	5.000	1.553	4.633	5.000	1.575	4.556	5.000	1.187	4.756	5.000	1.486	**5.700**	6.000	1.480
	EUC	0.595	**5.089**	5.000	4.944	4.500	5.000	1.560	4.511	5.000	1.769	4.211	4.000	1.625	4.211	4.000	1.934
	GEN	0.308	4.944	5.000	1.531	**4.867**	5.000	1.567	**5.022**	5.000	1.649	**5.100**	5.000	1.551	5.667	6.000	1.398
CONF	*XOR*	0.215	4.639	4.500	1.570	4.833	5.000	1.404	4.722	5.000	1.579	4.917	5.000	1.381	**6.000**	7.000	1.331
	EUC	0.524	**5.167**	5.000	1.298	4.278	5.000	1.783	4.278	5.000	1.966	4.111	4.500	1.769	4.417	4.500	1.842
	GEN	0.242	4.972	5.000	1.630	**5.000**	5.000	1.265	**5.083**	5.500	1.680	**5.167**	6.000	1.577	5.583	6.000	1.519
NON-CONF	*XOR*	0.358	4.648	5.000	1.556	4.500	5.000	1.680	4.444	5.000	1.598	4.648	5.000	1.556	5.500	6.000	1.551
	EUC	0.642	**5.037**	5.000	1.359	4.648	5.000	1.389	4.667	5.000	1.625	4.278	4.000	1.535	4.167	4.000	1.979
	GEN	0.351	4.926	5.000	1.478	**4.778**	5.000	1.745	**4.981**	5.000	1.642	**5.056**	5.000	1.547	**5.722**	6.000	1.433

Table 9
Self-reported demographic data (*Survey TS*). The number of subjects comes along with the percentage in brackets for each category.

Demographic parameter	Number of participants
(a) Age	
18–25	9 (15.79%)
26–35	19 (33.33%)
36–45	10 (17.54%)
46–55	10 (17.54%)
56–65	7 (12.28%)
66+	2 (3.52%)
(b) Gender	
Male	32 (56.14%)
Female	21 (36.84%)
Preferred not to say	4 (7.02%)
(c) Education	
Doctorate (Ph.D)	33 (57.89%)
Master's (M.A./M.Sc.)	17 (29.82%)
Bachelor's (B.A./B.Sc.)	5 (8.77%)
Short-cycle terciary	1 (1.76%)
Post-secondary non-terciary	1 (1.76%)
(d) English proficiency	
Native speaker	9 (15.79%)
Proficient (C2)	20 (35.09%)
Advanced (C1)	21 (36.84%)
Upper intermediate (B2)	7 (12.28%)
(e) Areas of expertise	
Explainable AI	29 (54.72%)
Fuzzy logic	14 (26.42%)
Mathematics	6 (11.32%)
Engineering	11 (20.75%)
Computer science	35 (66.04%)
Computational linguistics	22 (41.51%)
Social sciences	5 (9.43%)

Table 10
Survey TS results. ALL corresponds to the average for the five tasks. CONF averages only confusing tasks (1 and 2). NON-CONF averages only non-confusing tasks (3, 4 and 5). The highest average values for each (group of) task(s) and explanation aspect are highlighted in bold. Notice that, *PEC* values for ALL, CONF, and NON-CONF are averaged for the corresponding groups of tasks.

Task	Method	*PEC*	Trustworthiness			Satisfaction		
			Mean	Median	St.dev.	Mean	Median	St.dev.
1	*XOR*	0.195	**3.100**	3.000	1.221	2.750	2.000	1.545
	EUC	0.582	3.000	3.000	1.183	2.550	2.000	1.161
	GEN	0.232	**3.100**	3.500	1.338	**3.100**	3.000	1.375
2	*XOR*	0.235	3.100	3.000	1.179	2.900	3.000	1.261
	EUC	0.465	2.850	3.000	1.108	2.800	2.000	1.288
	GEN	0.252	**3.300**	3.500	1.308	**2.950**	3.000	1.117
3	*XOR*	0.488	**3.700**	4.000	1.345	**3.150**	3.500	1.352
	EUC	0.552	2.850	3.000	1.108	2.650	2.000	1.108
	GEN	0.255	3.150	3.000	1.152	2.950	3.000	1.203
4	*XOR*	0.186	3.300	3.000	1.382	3.150	3.000	1.424
	EUC	0.675	3.300	3.500	1.145	2.900	3.000	1.044
	GEN	0.383	**3.800**	4.000	1.030	**3.900**	4.000	1.179
5	*XOR*	0.400	**3.700**	4.000	1.100	**3.550**	4.000	1.203
	EUC	0.699	3.600	4.000	1.020	3.400	4.000	1.158
	GEN	0.416	3.400	3.000	1.020	3.200	3.000	1.077
ALL	*XOR*	0.301	**3.380**	4.000	1.279	3.100	3.000	1.389
	EUC	0.595	3.120	3.000	1.151	2.860	2.500	1.192
	GEN	0.308	3.350	4.000	1.203	**3.220**	3.000	1.246
CONF	*XOR*	0.215	3.100	3.000	1.200	2.825	3.000	1.412
	EUC	0.524	2.925	3.000	1.149	2.675	2.000	1.233
	GEN	0.242	**3.200**	3.500	1.327	**3.025**	3.000	1.255
NON-CONF	*XOR*	0.358	**3.567**	4.000	1.296	3.283	3.500	1.343
	EUC	0.642	3.250	3.000	1.135	2.983	3.000	1.147
	GEN	0.351	3.450	4.000	1.102	**3.350**	3.000	1.222

nations turn out more satisfying for certain tasks (3 and 5) whereas the *EUC* explanations appear less favorable in 4 out of the 5 tasks as well as on average.

Considering all the methods and tasks together, the findings from *Survey TS* do not allow us to make any conclusion regarding the correlation either between *PEC* and trustworthiness ($\rho = 0.07, p = 0.803$) or between *PEC* and satisfaction ($\rho = -0.081, p = 0.775$). The same situation is observed irrespective of the "confusing" nature of the tasks. On the one hand, there is a negative correlation but not enough statistical evidence for making distinctive conclusions in the case of "confusing" tasks: *PEC* versus trustworthiness ($\rho = -0.516, p = 0.295$); and *PEC* versus satisfaction ($\rho = -0.371, p = 0.468$). On the other hand, correlation coefficients are smaller but, once again, lack evidence in case of "non-confusing" tasks: *PEC* versus trustworthiness ($\rho = -0.025, p = 0.949$); *PEC* versus satisfaction ($\rho = -0.209, p = 0.589$).

6. Discussion

The explanation generation methods under study have a number of strengths and weaknesses. The qualitative methods favor two essential properties of CFs. First, the output CFs turn out to be diverse, as they can be mapped to a set of individual data points that are all equally minimally different on a categorical scale. Second, these methods are expected to maximize the validity of the generated CFs, as the corresponding explanations mimic the rules from the rule base. Therefore, following such explanations maximizes the probability of the corresponding CF rule to fire. On the other hand, the proposed qualitative methods may generate explanations that include a high number of features, some of them possibly being irrelevant or poorly explanatory.

The human evaluation study testifies that more complex explanations are perceived to be more informative, whereas increasing complexity jeopardizes readability and relevance. These findings specify the necessity of a careful design of automated explanations for specific tasks and/or application domains and/or intended audience. Thus, high-stakes decisions may require the corresponding explanations to be more informative and therefore encourage the use of methods that guarantee higher *PEC* scores of their output explanations (*EUC*). On the other hand, if the intended audience involved only lay users, more readable and therefore less complex explanations (*XOR* or *GEN*) may be preferred.

PEC scores allow us not only to quantify the perceived complexity of automatically computed CFs but also discern the most favorable of them. Lower *PEC* values appear to represent lower explanation complexity from user's point of view and therefore be more comprehensive. It can be seen that explanation length has a major impact on explanation complexity if the number of explanation features is low or if the linguistic terms used for such features are selected from a wider range of terms. Indeed, the terms covering narrower intervals appear more characteristic for the corresponding features and therefore more comprehensive. Further, the use of the proposed metric favors shorter but more informative (in terms of the number of features and/or linguistic terms used) explanations. Hence, driven by a complexity-oriented approach to evaluating CFs, a better understanding of a feature-based explanation can be reached by finding a balance between short enough explanation length and the number of unique features and/or linguistic terms used in the explanation.

Importantly, *PEC* can help to choose among alternative but semantically equivalent explanations. For example, the piece of explanation "if color were pale or straw or amber or brown" can be replaced by the shorter "if color were not black" (see Fig. 3). Then, it becomes essential to define how many linguistic terms are necessary to be properly understood to guarantee a consistent use of the metric. We thus suggest two strategies to calculate the number of terms associated to a feature if the term under consideration is negated. On the one hand, it may be sufficient to calculate the sum of the non-negated terms. In this case, the number of linguistic terms in the aforementioned explanation $t^{Color} = |\{pale, straw, amber, brown\}| = 4$ (see Fig. 3b). On the other hand, it may be argued that, to fully understand the meaning of the negated term, it is only necessary to understand the meaning of the negated term itself (*black*, in this case) as well as that of the collective linguistic terms covering all the contrasting linguistic terms (i.e., lighter/darker than black). Thus, if the negated linguistic variable takes on either of the extreme values (e.g., *pale* or *black*), the number of terms associated with the given explanation for feature t^{f_i} always equals 2. Moreover, if the fuzzy partition presumes that both lower and higher values can be captured by other linguistic terms with respect to the negated term (e.g., "*...if color were not amber*"), the number of the associated terms always includes the negated term as well as the values from the extended set of terms covering both smaller and higher corresponding intervals (see Fig. 3c).

7. Concluding remarks and future work

In this paper, we presented one quantitative method (*GEN*) of CF explanation generation and two methods (*XOR* and *EUC*) of qualitative CF explanation generation for FRBCSs. As all of them provide the end user with output of different kinds, they can be used solely or complementarily to offer explanations on demand and customized for different user profiles. In addition, we proposed the new metric *PEC* for estimating the complexity of a given explanation (as expected to be perceived by an end user).

To evaluate the proposed methods, we collected human evaluation scores in an empirical study which comprised two online questionnaires. In addition, we computed *PEC* scores for each of the explanations under consideration in the study. We observed that a more complexly structured within-subject questionnaire (*Survey GM*) appears to provide a better insight into the goodness of automated explanations given an equivalent number of participants. However, collecting data in such a

survey is costly as it requires higher cognitive load and more time from the participants. Therefore, calculating *PEC* automatically allows the survey designer to set up and deploy a shorter questionnaire and thus easier to fill (*Survey TS*). It is worth noting that *PEC* strongly correlates with several explanation aspects but does so in different directions, so an FRBCS designer is advised to carefully select the method of explanation generation based on the peculiarities of the application domain and/or intended audience.

All in all, the insights from this work are expected to advance methods of generation and evaluation for various explanation approaches. As such, they are expected to be helpful for designing future human evaluation surveys in the area of explainable AI. Moreover, as part of future work, we will go deeper with selecting and fusing CF explanations with the aim of customizing them for users having different profiles in different application scenarios. Further research is therefore necessary: (1) to extend the proposed CF explanation generation methods beyond numerical features; (2) to better assess the impact of the *PEC* hyperparameters (σ and λ); and (3) to better understand the connection between complexity and trustworthiness of automated explanations. Notice that, the conclusions derived from the current study are only applicable to the target population under consideration. As part of future work, for the sake of generalization, we intend to design and carry out other similar experiments with a larger and wider panel of respondents, including non-expert lay users. Finally, we plan to use *PEC* as one of the criteria to optimize when designing explainable multi-objective evolutionary fuzzy systems.

Funding

Ilia Stepin is an *FPI* researcher (grant PRE2019-090153). Jose M. Alonso-Moral is a *Ramon y Cajal* researcher (grant RYC-2016–19802). This work was supported by the Spanish Ministry of Science and Innovation (grants RTI2018-099646-B-I00, PID2021-123152OB-C21, and TED2021-130295B-C33) and the Galician Ministry of Culture, Education, Professional Training and University (grants ED431F2018/02, ED431G2019/04, and ED431C2022/19). All the grants were co-funded by the European Regional Development Fund (ERDF/FEDER program).

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A

In addition to the human evaluation study on the automatically generated CFs, we performed three independent experiments on the genetic algorithm hyperparameter fine-tuning. In particular, we estimated the impact of the following hyperparameters associated to the *GEN* method: (i) the size of the population, (ii) the crossover probability and the corresponding alpha value, and (iii) the mutation probability. All the experiments were run for the five survey stimuli where both the predicted classes and the CF classes were known. The experimental results were assessed in terms of the best achieved fitness scores.

Fig. 4 summarizes the impact of the population size (10, 20, 30, 40, 50). It can be observed that the default population size (30) provides good results, on average, for all the test instances under consideration.

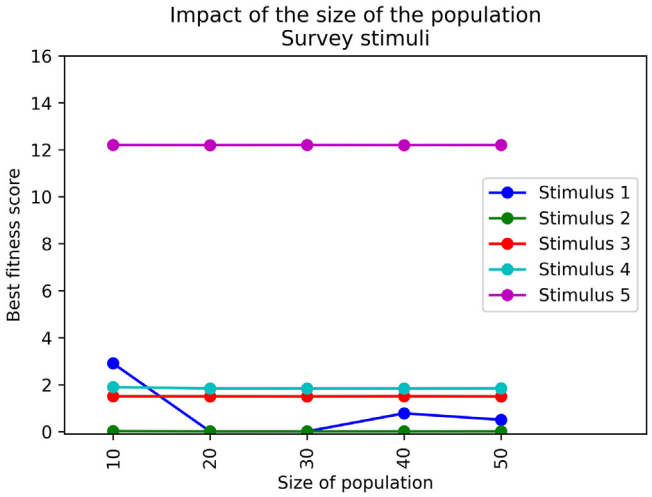

Fig. 4. An empirical assessment of the impact of the population size in the GEN method.

Fig. 5 shows the results of the experiment on the crossover probability values (0.7, 0.8, 0.9), considering different α values (0.2, 0.3, 0.4). In short, the combination of the crossover probability (0.8) and $\alpha = 0.3$ yields the best results for the considered CF data points.

Fig. 6 illustrates the impact of the selected mutation probability values (0.05, 0.1, 0.15, 0.2). It can be seen that doubling the default mutation probability value may result in worsened performance of the algorithm.

To sum it up, the analysis carried out allows us to conclude that the selected hyperparameter values do not only agree with the guidelines found in the literature (e.g., [21]) but also prove to be effective in the given experiments and can indeed be recommended for future use. All the detailed calculations as well as additional plots and the source code for replicating this experimental analysis can be found in our Gitlab repository:https://gitlab.citius.usc.es/ilia.stepin/fcfexpgen (branch "xor_euc_gen").

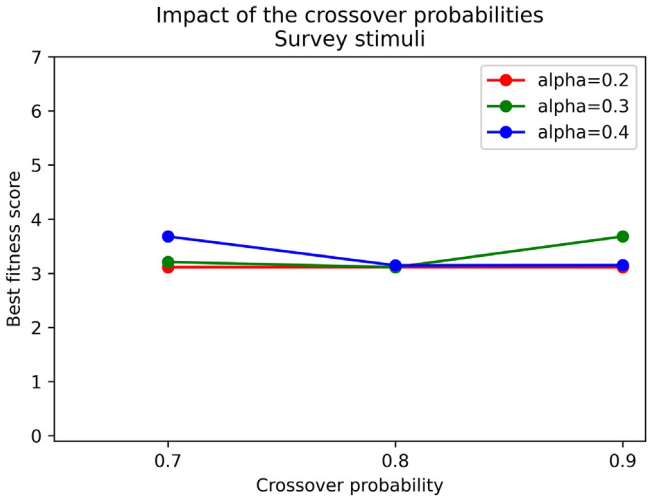

Fig. 5. An empirical assessment of the impact of the crossover hyperparameters (the crossover probability and the α crossover operator) in the GEN method.

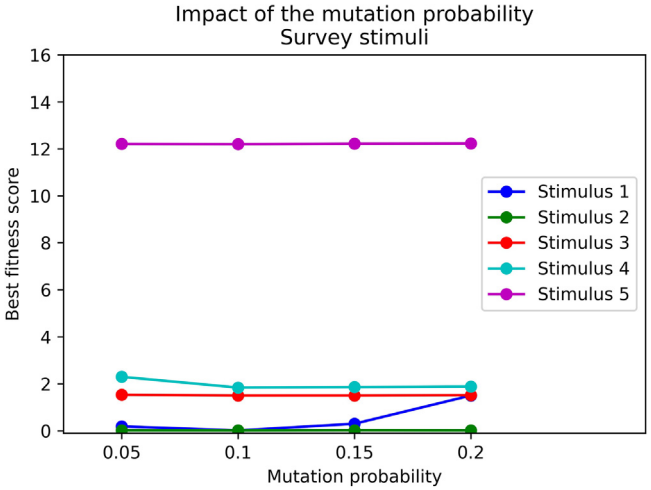

Fig. 6. An empirical assessment of the impact of the mutation probability in the GEN method.

References

[1] A. Abdul, J. Vermeulen, D. Wang, B.Y. Lim, and M. Kankanhalli. Trends and trajectories for explainable, accountable and intelligible systems: An HCI research agenda. In Proceedings of the Conference on Human Factors in Computing Systems (CHI), pages 1–18, Montreal QC, Canada, 2018. Association for Computing Machinery. https://doi.org/10.1145/3173574.3174156.

[2] A. Adadi, M. Berrada, Peeking inside the black-box: A survey on explainable artificial intelligence (XAI), IEEE Access 6 (2018) 52138–52160, https://doi.org/10.1109/ACCESS.2018.2870052.

[3] J.M. Alonso, C. Castiello, L. Magdalena, C. Mencar, Explainable Fuzzy Systems - Paving the Way from Interpretable Fuzzy Systems to Explainable AI Systems, volume 970, Springer International Publishing (2021), https://doi.org/10.1007/978-3-030-71098-9.

[4] J.M. Alonso, O. Cordón, S. Guillaume, and L. Magdalena. Highly interpretable linguistic knowledge bases optimization: Genetic tuning versus solis-wetts. Looking for a good interpretability-accuracy trade-off. In Proceedings of the IEEE International Conference on Fuzzy Systems, pages 901–906, London, UK, 2007. https://doi.org/10.1109/FUZZY.2007.4295485.

[5] I. Baaj and J.-P. Poli. Natural language generation of explanations of fuzzy inference decisions. In Proceedings of the IEEE International Conference on Fuzzy Systems, pages 1–6, New Orleans, LA, USA, 2019. https://doi.org/10.1109/FUZZ-IEEE.2019.8858994.

[6] A. Cascallar-Fuentes, A. Ramos-Soto, A. Bugarín, Adapting SimpleNLG to Galician Language, in: In Proceedings of the International Conference on Natural Language Generation, Association for Computational Linguistics (ACL), 2018, https://doi.org/10.18653/v1/W18-6507.

[7] O. Cordón, F. Herrera, A Three-Stage Evolutionary Process for Learning Descriptive and Approximate Fuzzy Logic Controller Knowledge Bases from Examples, International Journal of Approximate Reasoning 17 (4) (1997) 369–407, https://doi.org/10.1016/S0888-613X(96)00133-8.

[8] R. Dale, E. Reiter, Computational Interpretations of the Gricean Maxims in the Generation of Referring Expressions, Cognitive science 19 (2) (1995) 233–263, https://doi.org/10.1016/0364-0213(95)90018-7.

[9] V. Dignum, Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way. Artificial Intelligence: Foundations, Theory, and Algorithms, Springer, Cham, 2019, https://doi.org/10.1007/978-3-030-30371-6.

[10] L.J. Eshelman and J.D. Schaffer. Real-Coded Genetic Algorithms and Interval-Schemata. In L. Darrell Whitley, editor, Foundations of Genetic Algorithms, volume 2 of Foundations of Genetic Algorithms, pages 187–202. Elsevier, 1993. https://doi.org/10.1016/B978-0-08-094832-4.50018-0.

[11] A. Fernandez, F. Herrera, O. Cordon, M. Jose del Jesus, F. Marcelloni, Evolutionary Fuzzy Systems for Explainable Artificial Intelligence: Why, When, What for, and Where to?, IEEE Computational Intelligence Magazine 14 (1) (2019) 69–81, https://doi.org/10.1109/MCI.2018.2881645.

[12] R.R. Fernández, I.M. de Diego, V. Aceña, A. Fernández-Isabel, J.M. Moguerza, Random forest explainability using counterfactual sets, Information Fusion 63 (2020) 196–207, https://doi.org/10.1016/j.inffus.2020.07.001.

[13] S. Fletcher, M.Z. Islam, Comparing sets of patterns with the Jaccard index, Australasian Journal of Information Systems 22 (2018), https://doi.org/10.3127/ajis.v22i0.1538.

[14] A. Gatt, E. Krahmer, Survey of the State of the Art in Natural Language Generation: Core tasks, applications and evaluation, Journal of Artificial Intelligence Research 61 (2018) 65–170, https://doi.org/10.1613/jair.5477.

[15] H.P. Grice, Logic and Conversation, in: P. Cole, J.L. Morgan (Eds.), Syntax and Semantics: Speech Acts, Academic Press, 1975, pp. 41–58, https://doi.org/10.1163/9789004368811_003.

[16] R. Guidotti, Counterfactual explanations and how to find them: literature review and benchmarking, Data Mining and Knowledge Discovery (2022) 1–55, https://doi.org/10.1007/s10618-022-00831-6.

[17] R. Guidotti, A. Monreale, F. Giannotti, D. Pedreschi, S. Ruggieri, F. Turini, Factual and Counterfactual Explanations for Black Box Decision Making, IEEE Intelligent Systems 34 (6) (2019) 14–23, https://doi.org/10.1109/MIS.2019.2957223.

[18] D. Gunning, E. Vorm, J.Y. Wang, M. Turek, DARPA's explainable AI (XAI) program: A retrospective, Applied AI Letters 2 (4) (2021), https://doi.org/10.1002/ail2.61 e61.

[19] R. Gunning, Technique of clear writing, McGraw-Hill, 1968.

[20] C. Herley and W. Pieters. If You Were Attacked, You'd Be Sorry: Counterfactuals as Security Arguments. In Proceedings of the 2015 New Security Paradigms Workshop, NSPW '15, pages 112–123, New York, NY, USA, 2015. Association for Computing Machinery. https://doi.org/10.1145/2841113.2841122.

[21] F. Herrera, M. Lozano, A.M. Sánchez, A Taxonomy for the Crossover Operator for Real-Coded Genetic algorithms: An Experimental Study, International Journal of Intelligent Systems 18 (3) (2003) 309–338, https://doi.org/10.1002/int.10091.

[22] J. Hühn, E. Hüllermeier, FURIA: an algorithm for unordered fuzzy rule induction, Data Mining and Knowledge Discovery 19 (3) (2009) 293–319, https://doi.org/10.1007/s10618-009-0131-8.

[23] H. Ishibuchi, T. Nakashima, M. Nii, Classification and modeling with linguistic information granules: Advanced approaches to linguistic Data Mining, Springer Science & Business Media (2004), https://doi.org/10.1007/b138232.

[24] M. Lash, Q. Lin, N. Street, J. Robinson, and J. Ohlmann. Generalized Inverse Classification. In Proceedings of the International Conference on Data Mining (SDM), pages 162–170. Society for Industrial and Applied Mathematics, 2017. https://doi.org/10.1137/1.9781611974973.19.

[25] A. Lucic, H. Haned, and M. de Rijke. Why Does My Model Fail? Contrastive Local Explanations for Retail Forecasting. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, FAT* '20, pages 90–98, Barcelona, Spain, 2020. Association for Computing Machinery. https://doi.org/10.1145/3351095.3372824.

[26] N. Maaroof, A. Moreno, A. Valls, M. Jabreel, M. Szelag, A Comparative Study of Two Rule-Based Explanation Methods for Diabetic Retinopathy Risk Assessment, Applied Sciences 12 (7) (2022) 1–18, https://doi.org/10.3390/app12073358.

[27] E.H. Mamdani, Application of Fuzzy Logic to Approximate Reasoning Using Linguistic Systems, IEEE Transactions on Computers 26 (12) (1977) 1182–1191, https://doi.org/10.1109/TC.1977.1674779.

[28] G.A. Miller, The magical number seven, plus or minus two: Some limits on our capacity for processing information, Psychological Review 63 (2) (1956) 81–97, https://doi.org/10.1037/0033-295x.101.2.343.

[29] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, Artificial Intelligence 267 (2019) 1–38, https://doi.org/10.1016/j.artint.2018.07.007.

[30] J. Moore, N. Hammerla, and C. Watkins. Explaining deep learning models with constrained adversarial examples. In Proceedings of the Pacific Rim International Conference on Artificial Intelligence (PRICAI), pages 43–56. Springer, 2019. https://doi.org/10.1007/978-3-030-29908-8_4.

[31] R.K. Mothilal, A. Sharma, and C. Tan. Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, FAT* '20, pages 607–617, Barcelona, Spain, 2020. Association for Computing Machinery. https://doi.org/10.1145/3351095.3372850.

[32] M.L. Olson, R. Khanna, L. Neal, F. Li, W.-K. Wong, Counterfactual state explanations for reinforcement learning agents via generative deep learning, Artificial Intelligence 295 (2021) 1–29, https://doi.org/10.1016/j.artint.2021.103455.

[33] Parliament and Council of the European Union. General Data Protection Regulation (GDPR), 2016. URL:http://data.europa.eu/eli/reg/2016/679/oj.

[34] Parliament and Council of the European Union. A European Approach to Artificial Intelligence, 2022. URL:https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence.

[35] E. Reiter, R. Dale, Building Natural Language Generation Systems, in: Studies in Natural Language Processing, Cambridge University Press, 2000, https://doi.org/10.1017/CBO9780511519857.

[36] M.T. Ribeiro, S. Singh, and C. Guestrin. "Why should I trust you?": Explaining the predictions of any classifier. In Proceedings of the International Conference on Knowledge Discovery and Data Mining (SIGKDD), pages 1135–1144, San Francisco, California, USA, 2016. Association for Computing Machinery. https://doi.org/10.1145/2939672.2939778.

[37] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nature Machine Intelligence 1 (5) (2019) 206–215, https://doi.org/10.1038/s42256-019-0048-x.
[38] M. Schleich, Z. Geng, Y. Zhang, and D. Suciu. GeCo: Quality Counterfactual Explanations in Real Time. In Proceedings of the Very Large Data Bases (VLDB) Endowment, volume 14(9), pages 1681–1693, 2021. https://doi.org/10.14778/3461535.3461555.
[39] S. Sharma, J. Henderson, J. Ghosh. CERTIFAI, A Common Framework to Provide Explanations and Analyse the Fairness and Robustness of Black-Box Models, Association for Computing Machinery, 2020, pp. 166–172, https://doi.org/10.1145/3375627.3375812.
[40] K. Sokol and P. Flach. One Explanation Does Not Fit All: The Promise of Interactive Explanations for Machine Learning Transparency. KI – Künstliche Intelligenz, 2020. https://doi.org/10.1007/s13218-020-00637-y.
[41] I. Stepin, J.M. Alonso, A. Catala, and M. Pereira-Fariña. Generation and evaluation of factual and counterfactual explanations for decision trees and fuzzy rule-based classifiers. In Proceedings of the IEEE World Congress on Computational Intelligence (WCCI), Glasgow, UK, 2020. https://doi.org/10.1109/FUZZ48607.2020.9177629.
[42] I. Stepin, J.M. Alonso, A. Catala, M. Pereira-Fariña, A Survey of Contrastive and Counterfactual Explanation Generation Methods for Explainable Artificial Intelligence, IEEE Access 9 (2021) 11974–12001, https://doi.org/10.1109/ACCESS.2021.3051315.
[43] I. Stepin, A. Catala, M. Pereira-Fariña, J.M. Alonso, Factual and Counterfactual Explanation of Fuzzy Information Granules, in: Interpretable Artificial Intelligence: A Perspective of Granular Computing, Springer International Publishing, 2021, pp. 153–185, https://doi.org/10.1007/978-3-030-64949-4_6.
[44] R. Sukkerd, R. Simmons, D. Garlan, Toward Explainable Multi-Objective Probabilistic Planning, in: IEEE/ACM 4th International Workshop on Software Engineering for Smart Cyber-Physical Systems (SEsCPS), Gothenburg, Sweden, 2018, pp. 19–25.
[45] S. Verma, J. Dickerson, and K. Hines. Counterfactual explanations for machine learning: A review. In Proceedings of the Machine Learning: Retrospectives, Surveys and meta-Analyses (ML-RSA) Workshop at the Conference on Neural Information Processing Systems (NeurIPS), 2020.
[46] S. Wachter, B. Mittelstadt, C. Russell, Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR, Harvard Journal of Law & Technology 31 (2) (2018) 841–887, https://doi.org/10.2139/ssrn.3063289.
[47] X. Wang, M. Yin, Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making, in: 26th International Conference on Intelligent User Interfaces, IUI '21, Association for Computing Machinery, 2021, pp. 318–328, https://doi.org/10.1145/3397481.3450650.
[48] C. Woodcock, B. Mittelstadt, D. Busbridge, G. Blank, et al, The Impact of Explanations on Layperson Trust in Artificial Intelligence-Driven Symptom Checker Apps: Experimental Study, Journal of Medical Internet Research 23 (11) (2021), https://doi.org/10.2196/29386 e29386.
[49] L.A. Zadeh, The concept of a linguistic variable and its application to approximate reasoning, Information Sciences 8 (3) (1975) 199–249, https://doi.org/10.1016/0020-0255(75)90036-5.