Counterfactual linguistic rule-based explanations based on locally relevant causal mechanisms

Te Zhang * *Student Member, IEEE,* Christian Wagner * *Senior Member, IEEE,*
**Lab for Uncertainty in Data and Decision Making (LUCID),*
School of Computer Science, University of Nottingham, Nottingham, UK Email: {te.zhang, christian.wagner}@nottingham.ac.uk

***Abstract*—Counterfactual (CF) explanations provide a potentially powerful mechanism to deliver meaningful explanations of AI decisions. CF explanations are convincing when they reflect causal relationships between variables, because humans are cause-effect thinkers. Prior work has established a rule generation framework called CF-MABLAR, which is designed to generate causal rules that provide CF explanations. However, in the real-world, an effect is often the result of multiple causal mechanisms, and rules obtained by CF-MABLAR may not capture the actual causal mechanism that leads to the effect, which we called the locally relevant causal mechanism. Consequently, CF explanations generated by CF-MABLAR have the risk of containing redundant components, which reduces the explainability of the obtained CF explanations. To address this issue, in this paper, we provide a detailed discussion about two key aspects of generating CF explanations from a causal perspective: 1) *which variables* require intervention and 2) *what magnitude* of an intervention is needed. We propose CF-MABLAR-local which allows users to generate CF explanations based on locally relevant causal mechanisms. We conduct experiments on several real-world data sets to compare CF explanations generated through different methods, and analyse the impact of different parameterizations in CF-MABLAR-local.**

***Index Terms*—Fuzzy, Causality, rules, counterfactual, XAI**

I. Introduction

Counterfactual (CF) explanations provide a potentially powerful mechanism to deliver meaningful explanations of AI decisions [1]–[3]. A *CF explanation* aims to answer how to change the input of an AI model to obtain an output of the model which is different from the current one [4], [5]. CF explanations can provide users with additional information about a model's operation and offer guidance based on CF information [1], [6]. In addition, human explanations are often CF [2], [7]. Thus, explanations which contain CF information are in general in line with a human way of explaining, and can provide a more effective explanation.

A key issue in generating CF explanations is to determine which variables should be 'intervened' on (i.e., have their values changed). As discussed in [6], the variables to be changed should be causally related to the target variable, as human are cause-effect thinkers and we expect explanations should reflect causal relationships [6], [8], [9].

Within the context of XAI, rule-based systems have been widely used as they can offer factual explanations using linguistic, human-accessible rules [10]. Aiming at generating CF explanations for rule-based models, Stepin et al. [11] proposed a novel framework for rule-based models based on correlations between variables. We refer to this as the Cor-CF framework in this paper. To generate *causal* CF explanations using fuzzy rule-based systems, Zhang et al. [6] proposed the CF Markov blanket rule generation framework (CF-MABLAR), which achieves causal CF rules generation by leveraging the Markov blanket information obtained from a causal graph of a given data set.

However, in the real-world, an effect often arises via multiple causal mechanisms [12], [13]. Here, the term 'causal mechanism' refers to how a set of variables influences the target variable through a specific process. Thus, for different samples, the specific causal mechanism leading to their respective effects, referred to as the *locally relevant* causal mechanism, may be different [12]. For example, both COVID-19 and flu can lead to a person's fever. If a patient with a fever has COVID-19 but does not have flu, the locally relevant causal mechanism leading to their fever is most likely the first causal mechanism. In contrast, for another patient with a fever who only has the flu and not COVID-19, the locally relevant casual mechanism leading to their fever is the second causal mechanism.

In some real-world applications, CF explanations are expected to specifically focus on locally relevant causal mechanisms. In this paper, we refer to such explanations as *locally relevant Causal and Counterfactual Explanations*, or simply '*local CCF explanations*'. For example, suppose an AI model designed to predict the causes of fever is applied in the above mentioned scenario, and there is a fever patient who has the flu but not COVID-19. A CF explanation provided by the AI model based on causal relationships could be: "To cure your fever, you should: take medication A to treat the flu, and not take medication B to treat COVID-19." This CF explanation can help a doctor understand that the model considered both flu and COVID-19 when diagnosing a patient's fever. However, for the patient, the 'not take medication B to treat COVID-19' is redundant and potentially confusing, because COVID-19 is *not relevant* to their situation.

As discussed above, generating local CCF explanations can avoid redundant information which is unrelated to a user's situation. However, how to generate such CF explanations for rule-based systems is still an open problem. To address this issue, in this paper, we discuss the key issues in generating such CF explanations, and propose CF-MABLAR-local. CF-MABLAR-local is designed for scenarios that require local CCF explanations. The main contributions of this paper are as follows:

1) We analyse and discuss two issues in generating CF explanations: 1) *which variables* require intervention, and 2) *what magnitude* of an intervention is needed.
2) We propose CF-MABLAR-local designed to generate local CCF explanations. In addition, to address situations where the interventions suggested by the CF explanations obtained by CF-MABLAR may be redundant or unrealistic in the real world, CF-MABLAR-local provides an alternative method for computing interventions based on existing samples.
3) We conduct experiments to evaluate CF-MABLAR-local on several real-world data sets to compare the CF explanations obtained by CF-MABLAR-local based on different settings, and analyse the results applicable to each setting.

The rest of this paper is organised as follows: Section II provides the background of this paper. Section III provides a detailed analysis and discussion of the key issues in generating CF explanations, and introduces the details of CF-MABLAR-local. Section IV presents and analyses the experiment results. Section V provides the conclusions.

II. Background

In this section, we provide background on causal graphs and existing causal fuzzy rule-generation frameworks which focus on different facets of causal relationships.

A. Causal graphs

To intuitively represent causal relationships between variables, Pearl [14] proposed the concept of causal graph. Fig. 1 shows an example of a causal graph.

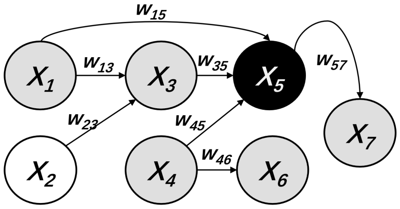


Fig. 1. An example of a causal graph

As shown in Fig. 1, a causal graph is directed and acyclic. Each node within the graph represents a variable. If an edge exists between two nodes, it indicates a causal relationship between the corresponding variables, pointing from the cause to the effect. The weights of edges in a causal graph describe a notion of the strength of causal link between two variables. The concept of the latter is non-trivial: it can be interpreted as the degree of influence of the cause on the effect [15], [16] or the probability that there is a causal relationship between the variables [17], depending on the way it is obtained.

For a variable within a causal graph, if all the edges on a path leading to this variable point toward it, then the path is a *causal path* for this variable. If the starting node of a causal path has no parents in the causal graph, this path is a complete causal path. Each complete causal path shows the relationships between the target variable and one set of its direct and/or indirect causes. Thus, each complete causal path represents one possible causal mechanism of the target variable–as captured by the causal graph. The latter is critical to keep in mind, i.e. in practice, when causal graphs are generated from data, there is no guarantee that the given graph is accurate. Of course, the aim is to derive accurate and complete causal graphs, but where the latter is not the case, causal paths may be missed and established causal paths may not actually be real.

B. MABLAR - Markov blanket rule generation framework

To generate rules which reflect causal relationships between variables, Zhang et al. [18] established the Markov blanket rule generation framework (MABLAR). The MABLAR framework has different variants designed to capture different facets of causal relationships. The standard variant of the MABLAR framework, i.e., MABLAR-ST, was originally proposed in [19].

The process of MABLAR-ST contains four steps. MABLAR-ST identifies a causal graph of the given data set and identifies the MB of the target variable in Step 1 and 2, respectively. Then, MABLAR-ST constructs a subset which only contains variables within the MB of the target variable in Step 3. Finally, in Step 4, MABLAR-ST generates rules from the constructed subset using data-driven algorithms (e.g. the WM algorithm [20]). By removing variables which are not causally related to the target variable, MABLAR-ST reduces the risk of generating rules which reflect correlation between variables and achieves causal rule generation.

C. Markov blanket rule generation using causal weights

To generate rules which can provide explanations that reveal the locally relevant causal mechanism of a given sample, Zhang et al. [13] proposed Markov blanket rule generation using causal weights (MABLAR-CW) as a variant of the MABLAR framework.

The process of MABLAR-CW contains four steps: Step 1 is to generate a causal weighted graph from a given data set, extracting causal information from the data set using established causal discovery algorithms such as ICA-LiNGAM [21]. In Step 2, MABLAR-CW identifies possible causal mechanisms by identifying all complete causal paths of the target variable, because each complete causal path represents a possible causal mechanism of the target variable as explained in Section II-A. After identifying all complete causal paths of the target variable within the obtained causal weighted graph, MABLAR-CW constructs a subset of the original data set for each completed causal path which contains only the variables on that path. In Step 3, MABLAR-CW uses the subset corresponding to each completed causal path to generate a fuzzy sub-system (FSS) for each completed causal path using data-driven approaches. Finally, in Step 4, MABLAR-CW assigns a causal score to each FSS using the causal weights from the obtained causal weighted graph.

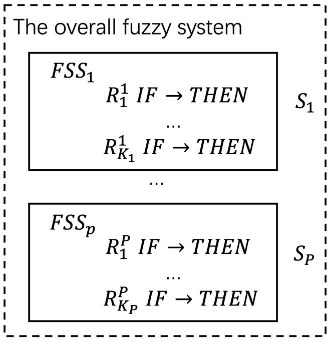


Fig. 2. The structure of a fuzzy system obtained by MABLAR-CW

The overall fuzzy system obtained by MABLAR-CW is constructed as a set of fuzzy sub-systems akin to an ensemble, as shown in Fig. 2. In Fig. 2, FSS_i represents the FSS corresponding to the i-th completed causal path. P is the number of completed causal paths identified by MABLAR-CW. R_j^i represent the j-th rule of the ith FSS. K_i represents the number of rules in FSS_i. S_i represents the causal score of the ith FSS. By modelling complete causal paths using a set of fuzzy sub-systems, the fuzzy system obtained by MABLAR-CW distinctly captures different possible causal mechanisms reflected in the causal weighted graph.

For a given sample, MABLAR-CW calculates a *causal index* of each FSS for the sample by leveraging the rule firing strength in each FSS and the corresponding causal score (i.e., S_i). The causal mechanism corresponding to the FSS with the highest causal index is identified as the locally relevant causal mechanism for the given sample. More details of MABLAR-CW can be found in [13], including the calculation of causal scores and the calculation of causal index.

III. Locally relevant Causal and Counterfactual explanation generation – CF-MABLAR-local

A. Critical discussion of CF-MABLAR

A CF explanation should answer the following two questions [6]: 1) *which variables* require intervention (the 'which variables' question), and 2) *what magnitude* of an intervention is needed (the 'what magnitude' question).

For the 'which variables' question, the variables requiring an intervention should have causal relationships with the target variable, because only the intervention with these variables can affect the target variable [6]. Thus, Zhang et al. [6] proposed CF-MABLAR, designed to generate *causal* CF explanations, as shown in Fig. 3 (a). CF-MABLAR adopts MABLAR-ST (see Section II-B) to capture causal relationships between variables. Consequently, as shown in Fig. 3(a), CF-MABLAR generates CF explanations which focus on the causal relationships between the target variable and the variables within its Markov blanket. We refer to such CF explanations as *causal CF Markov blanket explanations*.

However, as discussed in Section I, in some real-world scenarios, users may expect a CF explanation to only focus on the locally relevant causal mechanism. In that scenarios, local CCF explanations are more suitable than causal CF Markov blanket causal explanations, because local CCF explanations avoid redundant information which is unrelated to the users' context.

For the 'what magnitude' question, according to the Occam's razor, CF-MABLAR seeks to find the minimal intervention on the inputs. To achieve this, CF-MABLAR adopts the Rule Similarity (RS) index, which measures the similarity between two rules which have identical variables in their antecedents [6], [22]. To facilitate discussion, in this paper, we refer to this method as the RS based method. The RS index is calculated as follows:

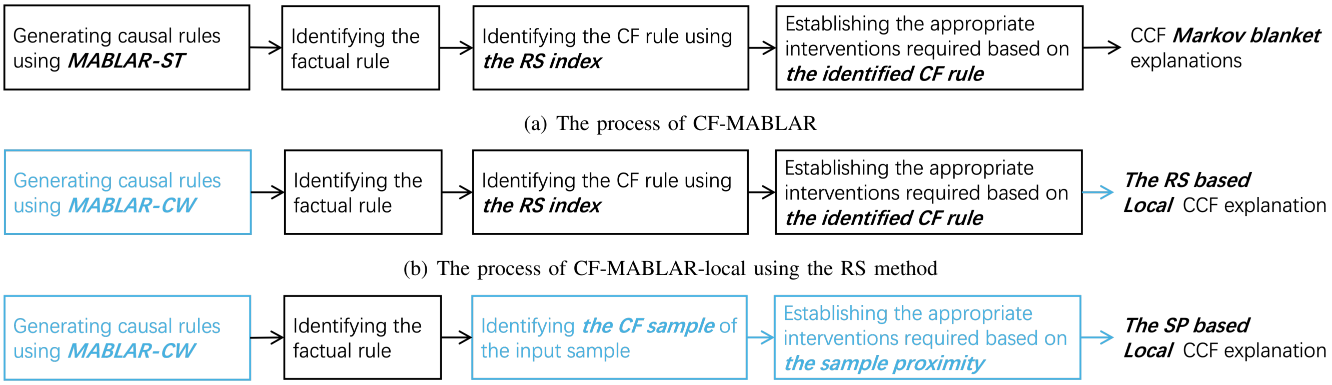
$$RS(k_1, k_2) = \sum_{i=1}^{D} S(A_i^{k_1}, A_i^{k_2}), \tag{1}$$

where $RS(k_1, k_2)$ represents the similarity between rule k_1 and k_2. D is the number of inputs. $A_i^{k_1}$ and $A_i^{k_2}$ are the antecedent fuzzy sets of the i-th input for rule k_1 and k_2, respectively. $S(A_i^{k_1}, A_i^{k_2})$ is the similarity between $A_i^{k_1}$ and $A_i^{k_2}$. In CF-MABLAR, the Jaccard ratio [23] is adopted to calculate the similarity between two fuzzy sets. Thus,

$$S(A_i^{k_1}, A_i^{k_2}) = \frac{\int_{x\in X} \min(\mu_{A_i^{k_1}}(x), \mu_{A_i^{k_2}}(x))}{\int_{x\in X} \max(\mu_{A_i^{k_1}}(x), \mu_{A_i^{k_2}}(x))}. \tag{2}$$

To ensure that the sample post-intervention achieves the desired CF output, the RS based method expects that the firing strength of the rule with the highest firing strength for the sample post-intervention reaches one [6]. To achieve this, CF-MABLAR ensures the sample post-intervention has a membership degree of one for every fuzzy set in the antecedent of the CF rule. However, in real-world applications, the firing strength of this rule is not always one. Consequently, the RS based method may result in redundant interventions being computed. Also, the RS based method may generate post-intervention samples which do not exist in the real world (e.g., a person who is one year old and two meters tall). In other words, the obtained interventions may not be actionable.

To address the aforementioned issues arising in CF-MABLAR, we propose CF-MABLAR-local in the following subsection.


Generating causal rules using *MABLAR-ST*
Identifying the factual rule
Identifying the CF rule using *the RS index*
Establishing the appropriate interventions required based on *the identified CF rule*
CCF *Markov blanket* explanations
Generating causal rules using *MABLAR-CW*
Identifying the factual rule
Identifying the CF rule using *the RS index*
Establishing the appropriate interventions required based on *the identified CF rule*
The RS based Local CCF explanation
Generating causal rules using *MABLAR-CW*
Identifying the factual rule
Identifying *the CF sample* of the input sample
Establishing the appropriate interventions required based on *the sample proximity*
The SP based Local CCF explanation

(a) The process of CF-MABLAR

(b) The process of CF-MABLAR-local using the RS method

(c) The process of CF-MABLAR-local using the Sample-proximity-based method

Fig. 3. Comparison between CF-MABLAR and CF-MABLAR-local (the differences between them are marked in blue)

B. Overview of CF-MABLAR-local

CF-MABLAR-local is designed to generate local CCF explanations for fuzzy systems.Fig. 3 (b) and (c) show the process of CF-MABLAR-local using different methods for establishing interventions, where the components in the blue rectangle are the key different parts compared to CF-MABLAR. In Fig. 3, the step 'establishing the interventions required' includes establishing which variables require intervention and to what degree.

CF-MABLAR-local inherits the CF explanation generation mechanism from CF-MABLAR, which combines factual rules, interventions, and CF rules to generate the final explanation for the sample (Section III-D shows an illustrative example). As shown in Fig. 3, compared to CF-MABLAR which uses MABLAR-ST for causal rule generation, CF-MABLAR-local adopts MABLAR-CW to generate causal rules, because, as introduced in Section II-C, a fuzzy system obtained by MABLAR-CW can reveal the locally relevant causal mechanism of a given sample.

CF-MABLAR-local requires a single fuzzy system as the basis for generating CF explanations. However, as shown in Fig. 2, a fuzzy system generated by MABLAR-CW is an ensemble of a set of fuzzy sub-systems (FSSs). To address this issue, for a given sample, CF-MABLAR-local selects the FSS with the highest causal index as the basis for generating CF explanations, because, as further detailed in Section II-C, the FSS with the highest causal index reveals the locally relevant causal mechanism for the sample.

As shown in Fig. 3 (b), CF-MABLAR-local can adopt the RS based method to calculate the interventions. However, as discussed in Section III-A, the RS based method may result in redundant interventions being computed, and/or unactionable interventions. To address this issue, as shown in Fig.3 (c), CF-MABLAR-local provides an alternative method for establishing interventions based on existing samples. The following subsection details the alternative method and analyses the scenarios in which it and the RS based method are suitable.

C. Establishing interventions based on sample proximity

The principle of the sample-proximity-based intervention method (the SP based method) is to identify the CF sample closest to the target sample among all CF samples of the target–to minimize the intervention. The method relies on close CF samples being present within the data, in which case it provides a tangible CF explanation. When no such samples are in the data, the resulting CF explanations may be larger (further away) than needed in practice.

Given an input sample $\boldsymbol{x}_{input} = [x_1, x_2, ..., x_d]$ which contains d input variables, the CF sample $\boldsymbol{x}^{cf} = [x_1^{cf}, x_2^{cf}, ..., x_d^{cf}]$ of $\boldsymbol{x}_{input}$ is obtained as follows:

$$\begin{aligned} \boldsymbol{x}^{cf} &= \arg\min_{\boldsymbol{x}} \ dist(\boldsymbol{x}, \boldsymbol{x}_{input}) \\ \text{s.t.} \quad & f(\boldsymbol{x}) = y_{\boldsymbol{x}}, \\ & f(\boldsymbol{x}_{input}) \neq y_{\boldsymbol{x}} \end{aligned} \tag{3}$$

And the intervention $diff_i$ for x_i is calculated as: $diff_i = x_i^{cf} - x_i$. In (3), $f(\boldsymbol{x})$ represents the system prediction of $\boldsymbol{x}$. $y_{\boldsymbol{x}}$ represents the label of $\boldsymbol{x}$. $dist(\cdot, \cdot)$ represents the distance between two samples. In this paper, we adopt the Euclidean distance, as it has been widely used. One can adopt another distance depending on the actual application scenario.

According to the principle of Occam's Razor, the intervention should be minimized as much as possible. Therefore, as shown in (3), the SP based method only considers the CF sample closest to $\boldsymbol{x}_{input}$. In addition, to ensure that $\boldsymbol{x}_{input}$ can obtain a different (or desired) output after intervention, as shown in (3), the SP method requires not only that the actual label of $\boldsymbol{x}^{cf}$ (i.e., $y_{\boldsymbol{x}^{cf}}$) differs from the model's prediction for $\boldsymbol{x}_{input}$ (i.e., $f(\boldsymbol{x}_{input})$), but also that the model's prediction for $\boldsymbol{x}^{cf}$ is correct. As the interventions obtained by the SP based method are based on samples that already *exist* in the real world, the SP based method reduces the risk of obtaining unrealistic interventions.

When adopting the SP based method, CF-MABLAR-local uses the rule which has the highest firing strength of the CF sample as the CF explanation part in the final explanation.

As noted, the SP based method requires a data set contains sufficient number of samples, because, when the data set has limited number of samples, the SP based method may fail to find the CF sample for a given sample. In such cases, we recommend using the RS method. In addition, we note that both the RS based method and the SP based method tend to intervene on all variables which have causal relationships (captured by MABLAR-ST/MABLAR-CW) with the target variable. However, in the real world, it may only be necessary to intervene on one or a few variables (rather than all) to achieve the desired counterfactual outcome [9]. For example, a person who is denied a loan application due to low education and low income may only need to improve either their education or income, rather than both, to successfully obtain the loan. This is worth exploring in future research to achieve even more precise and concise interventions.

D. Illustrative example

In this subsection, we use the Mammographic mass (MAM) data set [24] as an example to show the extended functionality in CF-MABLAR-local, which involves generating counterfactual explanations focused on locally relevant causal mechanisms using the SP based method. We choose the MAM dataset because its limited number of variables makes it feasible to show the counterfactual explanation generated by CF-MABLAR-local, as well as its corresponding causal graph, in this paper.

The MAM data set is used to predict the severity (benign or malignant) of a mammographic mass [24]. The data set contains six variables, which are ‘BI-RADS’, ‘Shape’, ‘Age’, ‘Margin’, ‘Density’ and ‘Severity’, respectively. The ‘Severity’ variable is the output variable. The ICA-LiNGAM algorithm [21] is adopted to generate a causal weighted graph from the data set. Fig. 4 shows the obtained causal graph. The target variable is marked as the black node in Fig. 4.

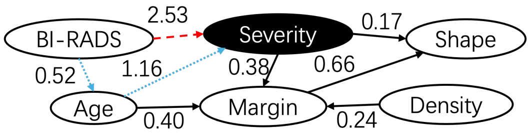


Fig. 4. The causal weighted graph of the MAM data set

As shown in Fig. 4, two possible causal mechanisms are identified, marked as red-dashed and blue-dotted lines in Fig. 4, respectively. Consequently, two FSSs are generated, modelling 1) the ‘BI-RADS → Severity’ and 2) the ‘BI-RADS → Age → Severity’ causal mechanisms, respectively.

TABLE I
THE SELECTED SAMPLE AND IT CORRESPONDING CF SAMPLE

	BI-RADS	Age	Shape	Margin	Density	Severity
Factual sample	0.0727	0.1282	0	0	0.6667	Benign
CF sample	0.0909	0.6154	0	0	0.6667	Malignant

The locally relevant causal mechanisms may differ across samples, further influencing the CF explanations generated by CF-MABLAR-local. Thus, to make sure our selection is unbiased, we randomly select a sample from the data set (marked as the factual sample). The value of the factual sample is shown in Table I (all values have been normalised to $[0, 1]$). The sample is classified into ‘Benign’ and the corresponding factual rule is ‘*BI-RADS is low and Age is low, then class is Benign*’. Thus, in this example, the FSS which models the ‘BI-RADS → Age → Severity’ causal mechanism is used as the basis model for the CF explanation generation.

In this example, the SP based method is adopted to generate CF explanations. There are 169 samples which have a different label compared to the factual sample and are correctly classified by the obtained fuzzy system. The value of the CF sample which is closest to the factual sample (marked as the CF sample) is also shown in Table I. Consequently, the intervention of ‘BI-RADS’ is 0.0182 and the intervention of ‘Age’ is 0.4872. In this case, there is no need to calculate the intervention of ‘Shape’, ‘Margin’ and ‘Density’, because these variables are not within the locally relevant mechanism of the factual sample.

The rule with the highest firing strength for the factual sample after the intervention is ‘*The BI-RADS is high and the Age is high, then class is Malignant*’. So, we obtain the final explanation for the factual sample generated by CF-MABLAR-local. To make it clear, we divide the final explanation into three parts: the factual part, the CF part and the CF conclusion part:

- The factual part: The sample is *Benign*, because its *BI-RADS* is $\underline{\textbf{low}}$ and its *Age* is $\underline{\textbf{low}}$.
- The CF part: To be *Malignant*, its *BI-RADS* would to be $\underline{\textbf{higher by } 0.0182}$ and its *Age* would to be $\underline{\textbf{higher by } 0.4872\ }$.
- The CF conclusion part: In that case, its *BI-RADS* would be $\underline{\textbf{high}}$, and its *Age* would be $\underline{\textbf{high}}$, and this sample would be classified *Malignant*.

IV. Experiments

A. Experiment settings

CF-MABLAR-local is specifically designed for fuzzy systems. Therefore, in this section, it is compared to two other CF explanation generation frameworks also designed for fuzzy systems: CF-MABLAR and a correlation-based CF explanation generation framework, i.e., Cor-CF [11], discussed in Section I. As shown in Fig. 3, CF-MABLAR-local can use either the RS based method or the SP based method for the establishing the appropriate intervention. In this section, we compare both configurations of CF-MABLAR-local. For clarity, we mark CF-MABLAR-local using the RS based method and the SP based method as CF-MABLAR-local(RS) and CF-MABLAR-local(SP), respectively.

The Wang-Mendel algorithm [20] is used in all CF explanation generation frameworks for rule generation, as it provides a consistent basis for comparison [25]. Trapezoidal and triangle membership functions are adopted as they facilitate explainability [26]. They are designed using the data-driven way demonstrated in [18]. The ICA-LiNGAM algorithm [21] is adopted by both CF-MABLAR and CF-MABLAR-local to generate causal weighted graphs, because the ICA algorithm has low computational complexity, which is well-suited for repeated experiments [27], [28]. Different causal discovery algorithms and membership design methods can be adopted for specific problems.

Five real-world data sets are selected as these data sets are widely used as benchmarks for rule generation. All data sets except the Beer3 data set are from the UCI data repository [29] and the Kaggle website [24]. The Beer3 data set is available at [30]. Table II summaries the data sets used in this paper. $|D|$ in Table II represents the number of input variables. Considering that a data set which is too small might result the SP based method failing to find CF samples (as discussed in Section III-C), while an overly large data set could result in excessive computational overhead (especially when using the SP based method), this paper adopts a compromise by selecting data sets with a sample size between 100 and 1000.

TABLE II
DATA SETS USED IN THIS PAPER

Name	Samples	Class	$\lvert D \rvert$
Beer3 [30]	400	8	3
Breast [29]	699	2	9
Iris [29]	150	3	4
MAM [24]	830	2	5
Pima Indian Diabetes (PID) [24]	768	2	8

We adopt the average F-score over 5-fold cross validation as the performance index. For the evaluation of CF explanations, the following three indices are adopted[1]:

- Validity: Validity is the percentage of samples which obtain the desired output after intervention [31].
- Average Minimal Intervention (AMI): A better CF explanation should have a lower degree of intervention [31], which means a good CF explanation should change the inputs as little as possible. AMI measures the average amount of intervention for each sample in a data set.Thus, the AMI index is defined as follows [31]:

$$AMI = \frac{\sum_{j=1}^{n} \sum_{i=1}^{d} |x_i^j - \bar{x}_i^j|}{n}, \tag{4}$$

  where n is the number of samples requiring intervention and d is the number of inputs. x_i^j and $\bar{x}_i^j|$ are the actual value and the value-post-intervention, respectively, of the ith input variable of the jth input sample.
- Length of the best CF explanation (CFLength): The number of conditions in the rules used to provide the CF explanation, where lower is better [11].

B. Experiment results

All approaches achieve a validity of one in all data sets. Tables III - V show the remaining evaluation indices for the different approaches (the values in parentheses represent the standard deviation). The results of CF-MABLAR-local(RS) and CF-MABLAR-local(SP) are the same in Table III, because CF-MABLAR-local(RS) and CF-MABLAR-local(SP) adopts the same model for prediction. In this paper, the WM algorithm is adopted to generate rules and all rules in a fuzzy system obtained by the WM algorithm have the same length. Consequently, in Table IV, the standard deviations of Cor-CF and CF-MABLAR are zero in all data sets as they use a single fuzzy system obtained by the WM algorithm. From Table III - V, we can make the following observations:

TABLE III
THE F-SCORES ACHIEVED BY EACH APPROACH

	Cor-CF	CF-MABLAR	CF-MABLAR	CF-MABLAR
			-local(RS)	-local(SP)
Beer3	**0.76**(0.05)	0.75(0.02)	0.56(0.02)	0.56(0.02)
Breast	0.84(0.04)	0.86(0.05)	**0.92**(0.02)	**0.92**(0.02)
Iris	0.95(0.03)	0.94(0.05)	**0.95**(0.04)	**0.95**(0.04)
MAM	0.75(0.06)	**0.76**(0.04)	0.72(0.06)	0.72(0.08)
PID	0.61(0.02)	**0.66**(0.03)	0.57(0.02)	0.57(0.02)

TABLE IV
THE AVERAGE CFLENGTH ACHIEVED BY EACH APPROACH

	Cor-CF	CF-MABLAR	CF-MABLAR	CF-MABLAR
			-local(RS)	-local(SP)
Beer3	3(0)	3(0)	**2.985**(0.121)	**2.985**(0.121)
Breast	10(0)	9(0)	**2.415**(0.849)	**2.415**(0.849)
Iris	4(0)	3(0)	**2**(0)	**2**(0)
MAM	5(0)	5(0)	**2**(0)	**2**(0)
PID	8(0)	5(0)	**2.967**(0.177)	**2.967**(0.177)

TABLE V
THE AMI ACHIEVED BY EACH APPROACH

	Cor-CF	CF-MABLAR	CF-MABLAR	CF-MABLAR
			-local(RS)	-local(SP)
Beer3	0.32(0.13)	0.32(0.18)	0.57(0.11)	**0.16**(0.12)
Breast	1.46(0.24)	1.66(0.28)	0.85(0.21)	**0.81**(0.21)
Iris	0.68(0.30)	0.49(0.28)	0.48(0.27)	**0.44**(0.19)
MAM	0.72(0.19)	0.95(0.39)	0.22(0.16)	**0.21**(0.18)
PID	0.45(0.15)	0.37(0.17)	0.27(0.13)	**0.20**(0.13)

1) All approaches show comparable performance in most data sets. However, CF-MABLAR-local shows a signifiant decline on the Beer3 data set. This may be due to MABLAR-CW adopting the “winner-takes-all” principle for prediction, which results in much of the information beneficial for prediction being overlooked, thereby reducing the prediction performance. Overall, we consider the performance of different approaches satisfactory and sufficient to meaningfully consider the explanations generated by different approaches.
2) Both CF-MABLAR-local(RS) and CF-MABLAR-local(SP) achieve the shortest CFLength, which indicates that identifying the local causal mechanism of a given sample can effectively reduce the complexity of the generated CF explanations, thereby enhancing their explainability. In addition, both approaches achieve the lowest AMI, which indicates that identifying the local causal mechanism of a given sample has the potential to avoid redundant interventions as discussed in Section I and III-A.

[1] Evaluating evaluations are complex–as part of a forthcoming journal paper we expect to expand on the indices used here.

3) We highlight the comparison between CF-MABLAR-local(RS) and CF-MABLAR-local(SP). As shown in Table V, CF-MABLAR-local(SP) achieves lower AMI values in all data sets compared to CF-MABLAR-local(RS). As we discussed in Section III-A, the RS based method has a risk of generating redundant interventions. This observation supports that the SP based method reduces this risk by leveraging existing samples which already fired the expected rule.

V. Conclusions

To enable users to customize, based on their needs, the generation of CF explanations provided by a fuzzy system, we propose CF-MABLAR-local, an extended version of CF-MABLAR. It not only supports fuzzy system based counterfactual (CF) explanations based on variables that are causally related to the target variable, but also supports the generation of CF explanations which specifically focus on locally relevant causal mechanisms. Furthermore, CF-MABLAR-local also provides an alternative way of calculating interventions based on existing samples to reduce the risk of obtaining a redundant and/or infeasible interventions.

We note that the SP based method requires not only storing the model itself but also storing all existing (e.g. training) samples. When the number of these samples becomes excessively large, this significantly increases storage overhead. Furthermore, with increasing numbers of samples, computational overhead also increases. In future research, we will focus on ways to meeting the challenges which arise for larger data sets, such as via selecting representative samples to optimize storage and computation, thereby reducing resource requirements. Also, as discussed in Section III-C, a desired counterfactual outcome may only require an intervention on one or a subset of variables, which is worth further investigation in the future to generate more concise CF explanations.

References

[1] R. Guidotti, "Counterfactual explanations and how to find them: literature review and benchmarking," *Data Mining and Knowledge Discovery*, vol. 38, no. 5, pp. 2770–2824, 2024.

[2] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial intelligence*, vol. 267, pp. 1–38, 2019.

[3] R. Guidotti, A. Monreale, F. Giannotti, D. Pedreschi, S. Ruggieri, and F. Turini, "Factual and counterfactual explanations for black box decision making," *IEEE Intelligent Systems*, vol. 34, no. 6, pp. 14–23, 2019.

[4] 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*, ser. FAT* '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 90–98.

[5] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the gdpr," *Harv. JL & Tech.*, vol. 31, p. 841, 2017.

[6] T. Zhang, C. Wagner, and J. M. Garibaldi, "Counterfactual rule generation for fuzzy rule-based classification systems," in *IEEE International Conference on Fuzzy Systems*, 2022, pp. 1–8.

[7] R. M. Byrne, "Counterfactuals in explainable artificial intelligence (xai): Evidence from human reasoning." in *IJCAI*. California, CA, 2019, pp. 6276–6282.

[8] R. R. Hoffman and G. Klein, "Explaining explanation, part 1: theoretical foundations," *IEEE Intelligent Systems*, vol. 32, no. 3, pp. 68–73, 2017.

[9] L. E. Bynum, J. R. Loftus, and J. Stoyanovich, "A new paradigm for counterfactual reasoning in fairness and recourse," in *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, K. Larson, Ed. International Joint Conferences on Artificial Intelligence Organization, 8 2024, pp. 7092–7100, main Track. [Online]. Available: https://doi.org/10.24963/ijcai.2024/784

[10] Y. You, J. Sun, Y. Guo, Y. Tan, and J. Jiang, "Interpretability and accuracy trade-off in the modeling of belief rule-based systems," *Knowledge-Based Systems*, vol. 236, p. 107491, 2022.

[11] 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 *2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. IEEE, 2020, pp. 1–8.

[12] G. Klein, "Explaining explanation, part 3: The causal landscape," *IEEE Intelligent Systems*, vol. 33, no. 2, pp. 83–88, 2018.

[13] T. Zhang and C. Wagner, "Generating locally relevant explanations using causal rule discovery," in *IEEE International Conference on Fuzzy Systems*, 2024, pp. 1–8.

[14] J. Pearl, "Causal diagrams for empirical research," *Biometrika*, vol. 82, no. 4, pp. 669–688, 1995.

[15] C. Puente, M. López, J. Rodrigo, and J. Olivas, "Weighted graphs to model causality," in *Proceedings on the International Conference on Artificial Intelligence (ICAI)*. The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 2017, pp. 297–301.

[16] A. Sobrino, C. Puente, and J. A. Olivas, "Extracting answers from causal mechanisms in a medical document," *Neurocomputing*, vol. 135, pp. 53–60, 2014.

[17] E. C. Garrido-Merchán, C. Puente, A. Sobrino, and J. A. Olivas, "Uncertainty weighted causal graphs," *arXiv:2002.00429*, 2020.

[18] T. Zhang, C. Wagner, and J. M. Garibaldi, "Explain the world—using causality to facilitate better rules for fuzzy systems," *IEEE Transactions on Fuzzy Systems*, vol. 32, no. 12, pp. 6671–6683, 2024.

[19] T. Zhang and C. Wagner, "Learning causal fuzzy logic rules by leveraging Markov blankets," in *IEEE International Conference on Systems, Man, and Cybernetics*, 2021, pp. 2794–2799.

[20] L.-X. Wang and J. M. Mendel, "Generating fuzzy rules by learning from examples," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 22, no. 6, pp. 1414–1427, 1992.

[21] S. Shimizu, P. O. Hoyer, A. Hyvärinen, A. Kerminen, and M. Jordan, "A linear non-Gaussian acyclic model for causal discovery," *Journal of Machine Learning Research*, vol. 7, no. 72, pp. 2003–2030, 2006.

[22] A. Garcia-Garcia, M. Z. Reformat, and A. Mendez-Vazquez, "Similarity-based method for reduction of fuzzy rules," in *2016 Annual Conference of the North American Fuzzy Information Processing Society (NAFIPS)*, 2016, pp. 1–6.

[23] P. Jaccard, "Nouvelles recherches sur la distribution florale," *Bull. Soc. Vaud. Sci. Nat.*, vol. 44, pp. 223–270, 1908.

[24] "Kaggle data sets," https://www.kaggle.com/datasets, accessed: 2023-11-25.

[25] L.-X. Wang, "The wm method completed: a flexible fuzzy system approach to data mining," *IEEE Transactions on Fuzzy Systems*, vol. 11, no. 6, pp. 768–782, 2003.

[26] J. M. Mendel and P. P. Bonissone, "Critical thinking about explainable AI (XAI) for rule-based fuzzy systems," *IEEE Transactions on Fuzzy Systems*, vol. 29, no. 12, pp. 3579–3593, 2021.

[27] R. Guo, L. Cheng, J. Li, P. R. Hahn, and H. Liu, "A survey of learning causality with data: Problems and methods," *ACM Computing Surveys (CSUR)*, vol. 53, no. 4, pp. 1–37, 2020.

[28] C. Ruichu, C. Wei, Z. Kun, and H. Zhifeng, "A survey on non-temporal series observational data based causal discovery," *Chinese Journal of Computers*, vol. 40, no. 6, pp. 1470–1490, 2017.

[29] D. Dua and C. Graff, "UCI machine learning repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml

[30] "The beer data set," https://gitlab.citius.gal/jose.alonso/xai/-/blob/master/BEER3.txt.aux.arff?ref_type=heads, accessed: 2025-01-05.

[31] S. Verma, J. Dickerson, and K. Hines, "Counterfactual explanations for machine learning: A review," *arXiv preprint arXiv:2010.10596*, 2020.