Counterfactual rule generation for fuzzy rule-based classification systems

Te Zhang
Lab for Uncertainty in Data and Decision Making (LUCID)
School of Computer Science
University of Nottingham
Nottingham, UK
te.zhang@nottingham.ac.uk

Christian Wagner
LUCID
School of Computer Science
University of Nottingham
Nottingham, UK
christian.wagner@nottingham.ac.uk

Jonathan.M.Garibaldi
LUCID
School of Computer Science
University of Nottingham
Nottingham, UK
Jon.Garibaldi@nottingham.ac.uk

***Abstract*—EXplainable Artificial Intelligence (XAI) is of increasing importance as researchers and practitioners seek better transparency and verifiability of AI systems. Mamdani fuzzy systems can provide explanations based on their linguistic rules, and thus a potential pathway to XAI. A factual rule based explanation generally refers to the given set of rules executed, or fired, for a given input. However, research has shown that human explanations are often counterfactual (CF), i.e. rather than explaining why a given output was reached, they show why other potential outputs were not. Although several machine learning-based CF explanation generation methods have been proposed in recent years, quasi none of them focus on fuzzy systems. Also, where they do, they focus on correlation, which limits the interpretive value of any CF explanations obtained as humans expect a causal relationship in rules, i.e. we are cause-effect thinkers. In this paper, we propose a new rule generation framework for Mamdani fuzzy classification systems, which we refer to as CF-MABLAR, building on the MARkov BLAnket Rules (MABLAR) framework. CF-MABLAR approximates the causal links between inputs and output(s) of fuzzy systems and generates CF rules by leveraging them. Uniquely, the CF rules obtained not only provide a basic CF explanation, but can also articulate how the given inputs would need to be changed to generate a different output, crucial for lay-user insight, verification and sensitivity-evaluation of XAI systems, for example in decision support around credit risk, cyber security and medical assistance.**

***Index Terms*—Fuzzy, rule, counterfactual, explanation, causal**

I. Introduction

Nowadays, AI models are used widely in many areas including risk sensitive areas, such as healthcare [1], where mistakes by AI models can result in serious consequences. More broadly, with AI becoming increasingly pervasive in consumer-facing products, its 'trustworthiness' is essential for sustained consumer-acceptance. It is important for users of AI to understand why an AI model makes the given prediction and decision – supporting AI accountability and trustworthyness. Consequently, eXplainable AI (XAI) has attracted an increasing interest in recent years [2].

The work was supported in part through the University of Nottingham, School of Computer Science PhD Scholarship Program, and UK EPSRC grant EP/P011918/1

In general, the purpose of XAI is to provide explanations for the behaviour of AI models. It is interesting to note that the purpose or utility of such explanations may differ substantially depending on the given AI application, as already alluded to. One may be interested in understanding the rationale of an AI to check whether it is making decisions as-expected (c.f. trustworthiness, verifiability), or in order to understand how one needs to change its inputs to achieve a desirable output (e.g. credit assessment).

For example, consider the case discussed in [3]. Here, the military trained a prediction model called 'Tank Prediction Model' (TPM), to discriminate between photos of enemy and friendly tanks. Later, the military found that the TPM classified the tanks on the basis of the weather shown in the photos, because all photos of friendly tanks were taken on sunny days, while all photos of enemy tanks were taken on overcast days. In this case, an XAI explanation could for example highlight to a user that: '*The tank is classified as an enemy tank, because the weather in the photo is overcast*'. As a result, the military in this case may decide that this behaviour of the AI is unreasonable, leading them to update/change their model altogether. At the same time, the example shows that *if* the military was happy with this behaviour, the XAI explanation would inform them on how to change an input (image) to provoke a change in classification—insight which may be of use from an adversarial point of view. In practice, the purpose of explanations is a key driver for the type and format of explanations sought.

Within the context of AI-driven decision support, Mamdani Fuzzy Systems (FSs) have been widely used in many areas [4]–[6] and have shown good ability to present knowledge in an interpretable way [7]. Mamdani FSs can provide explanations for their output(s) through their linguistic rules. For example, 'This flower is Iris Setosa, because its sepal length, sepal width, petal length and petal width are all low' is an explanation as can be derived from a rule of a Mamdani Iris classification FS. Broadly speaking, such explanations are called *factual explanations*, because they describe which features contributed to the model's output based on the

given input [8]. A substantial number of popular Mamdani rule generation algorithms exist, including the Wang-Mendel algorithm and its variants [9]–[12], evolution computation-based rule generation algorithms [13]–[15] and other data-driven rule generation algorithms, such as [16], [17] etc. These approaches are strong at generating 'factual' rules from data directly.

Beyond the 'factual' rules used in AI systems, human explanations are often counterfactual (CF) [18]. When a model gives a particular output, users benefit from not only knowing why a particular output is obtained, but also why other outputs are not obtained. For example, in the above Iris case, why the flower is classified to Iris Setosa, and not Iris Versicolour? To explain this, explanations need to contain CF elements. In the above case, a potential CF explanation would be 'This flower would be classified as Iris Versicolour if its sepal width and sepal length were medium'. Here, 'sepal width and sepal length were medium' is a CF element, because the given inputs of sepal width and sepal length are low, and not medium. Such explanations are called CF explanations [8].

Beyond factual explanations, CF explanations can complement a meaningful explanation [8], [18]. Although CF explanations have a long history in some areas such as philosophy and social sciences [19], they are a fairly recent topic in the XAI area, attracting much attention only in recent years [20], [21]. From an XAI perspective, CF explanations can be viewed as providing two different types of contributions:

- They can help users–commonly expert-users–to assess whether an AI is making decisions as expected. An example here is tank-classification example above, where a CF explanation can help highlight that the discrimination strategy of the AI is not reasonable.
- Where validating the correct functionality is not the focus, CF explanations can guide users to understand how they can change inputs to provoke a change in output. For example, CF explanations can help a credit-applicant understand that their high-spending behaviour is the reason their loan application was rejected and inform them to what degree they need to change this behaviour in order to be considered for a loan.

Recently, many Machine Learning (ML)-based CF explanation generation methods were proposed, such as [22], [23]. However, only few of them focus on Mamdani fuzzy systems. Addressing this gap, most recently, [24] proposed a CF explanation generation framework for decision trees and fuzzy rule-base systems, which can be extended to Mamdani fuzzy systems. However, the above mentioned CF explanation generation frameworks mainly focus on correlation. Humans are cause-effect thinkers [25], which means they expect explanations to reflect causal relationships. Thus, correlation-based CF explanations generation methods, which may not capture actual causal relationships, risk misleading users.

As discussed above, ideally Mamdani fuzzy systems can provide CF explanations which reflect causal relationships. Furthermore, we expect these explanations to answer the *intervention question*, i.e. that they indicate which variable would have to be changed–and how–in order to change the system output. In this paper, we summarise the intervention question as follows:

- How should be modify a given input $\boldsymbol{X}$ whose original corresponding output is 'A' in order for the input to be associated with a different output 'B'?

We will give a detailed analysis of the intervention question in Section III-A. In order to achieve this type of CF explanations, we develop a CounterFactual-MARkov BLAnket Rules (CF-MABLAR) generation framework which builds on [26] and [24]. The main contributions are as follows:

1) We review the nature of CF explanations, highlighting differences and flagging outstanding areas of research.
2) We develop a new framework which can generate CF rules for Mamdani fuzzy systems. The obtained rules can provide CF explanations which generally reflect causal relationships and answer the intervention question.
3) We conduct experiments to demonstrate and evaluate the developed framework.
4) We discuss the developed framework and next steps.

The rest of this paper is organised as follows: Section II provides relevant background. Section III presents the developed framework. Section IV demonstrates the developed framework and discusses the counterfactual explanations provided by the developed framework. Section V shows the experiment results. Finally, Section VI provides the conclusions and highlights further research directions.

II. Background

A. Counterfactual XAI

Nowadays, many XAI methods are being proposed (see [2], [27], [28] for some recent overviews) and are used in many real-world applications, such as debugging of machine learning models [29], healthcare [30] and explaining predictions of classifiers [31]. By providing explanations for the outputs of AI models, XAI can make the operation of AI models be understood by users [18], which in turn can support users' trust in AI models.

As the demand for human-like explanations is increasing, AI researchers are paying more attention to particular properties of explanations and their sub-types [18]. Researchers in social science have found that human explanations are often contrastive [8], [32]. Consequently, the research of CF explanations has attracted more and more AI researchers in the XAI community [33], [34], because CF explanations build on this contrastive and intuitive mechanism [18].

Consider that M is a classification model and $\boldsymbol{x}_{ft}$ is an input vector of M. If the prediction of M for $\boldsymbol{x}_{ft}$ is y_{ft}, then, from an XAI perspective, a CF explanation needs to explain why the output of the model is y_{ft} and not y_{ct} [35]. Here, y_{ct} is another output of $\boldsymbol{x}_{ft}$. Using CF explanations can alleviate the challenge of explaining the internal workings of

a complex AI models, because CF explanations don't need to explain all causes of 'why y_{ft} is obtained' [8]. Instead, CF explanations only need to address those related to 'why y_{ct} is not obtained'.

CF explanations can also help users to identify components of an AI system which require change, e.g., the outputs associated with a given input, which is helpful for users to validate AI models. For example, sometimes an AI model may not give users the desired output, in that cases, users will challenge the output given by the AI model, because they think the AI model is not working correctly. In that case, they need CF explanations to see if the output makes sense. If it does not - users are able to decide that the AI model is not working well for the given problem and take mitigating measures as appropriate.

A popular way to generate CF explanations is to find a vector which is the most similar with the input vector needed to be explained [19], such as the LIME-Counterfactual [31] and the framework proposed in [22]. Aiming at the problem of generating CF explanation for rule-based models, [24] proposed a novel framework for decision tree models and fuzzy rule-based models. In [36], the authors proposed a CF explanation generation framework for the FSs obtained by the fuzzy unordered rule induction algorithm [37]. However, these approaches do not focus on the challenge that where models are correlation-based, explanations, including CF explanations, reflect correlations, rather than causal relationships. How to design an integrated framework to construct FSs based on a set of causal rules, which in turn can provide causal CF explanations is still an open problem.

B. Mamdani fuzzy systems

Mamdani type fuzzy systems were proposed in [38]. The *k*-th rule of a Mamdani fuzzy system can be expressed as follows:

Rule k:

$$\begin{aligned} &\text{If } x_1 \text{ is } A_1^k \text{ and } x_2 \text{ is } A_2^k \text{ and } ... \text{ and } x_m \text{ is } A_m^k \\ &\text{Then } y \text{ is } B^k \end{aligned} \tag{1}$$

where each rule has a corresponding input vector $\boldsymbol{x} = [x_1, x_2, ..., x_d]$ and maps the fuzzy set $A_i^k \subset R$ of the input space into the fuzzy set $B^k \subset R$ of the output space. A_i^k is the corresponding fuzzy subset of the i-th input (i.e., x_i) in rule k. The linguistic term 'and' is represented by '$\wedge$', representing fuzzy conjunction.

C. Overview of rule generation algorithms

One advantage of fuzzy systems is their good interpretability, which is mainly derived from their linguistic rule bases. While rules of fuzzy systems were initially generated by experts, with the development of ML techniques, many ML-based rule generation algorithms were proposed, such as the Wang-Mendel algorithm [9], the Wang-Mendel completed algorithm [39], the reduced weighted Wang-Mendel algorithm [11] and improving Wang–Mendel method [40] etc. Often, fuzzy systems of which rules generated by ML-based algorithms have better performance (e.g., classification accuracy in a classification task) than fuzzy systems of which rules are obtained by experts [41]. However, rules obtained by ML-based methods usually have a lower interpretability compared with rules obtained from experts, because rules obtained by ML-based methods usually are very complicated to achieve high performance [42].

Nowadays, as XAI has attracted more and more attention, many ML-based rule generation algorithms have been proposed to improve the interpretability of rules by reducing the complexity of rules. One type of them treats the rule generation problem as a multi-objective optimization problem [13]–[15] . This type of algorithms treats the number of inputs, the number of rules, the performance of corresponding fuzzy system and so on as different optimization objectives. Then, using a classical evolutionary algorithm, e.g., MOEA/D [43]. They solve the optimization problem and obtain the rule base. Another type of ML rule generation decreases the complexity of rules by changing the structure of rules. As discussed in [44], when there are a large number of inputs, hierarchical FSs have less complexity and better interpretability than 'flat' FSs.

Another factor affecting the interpretability of rules is the relationship reflected by rules. The above mentioned rule generation methods can reduce the complexity of rules, however, all of them are focusing on correlation. In real-world applications, people expect rules to reflect causality [25]. To address this issue, a causal rule generation framework, MABLAR, was proposed in [26]. MABLAR is adopted in this paper to capture (or at least approximate) the causal relationships between variables, and the details of MABLAR are shown in the next subsection.

D. The MABLAR framework

The MABLAR framework is a two-step rule generation framework which can be adopted for Mamdani fuzzy systems. Rules generated by MABLAR can reflect, or at least approximate, the causal relationships between the inputs and output(s). Fig. 1 shows the MABLAR framework.

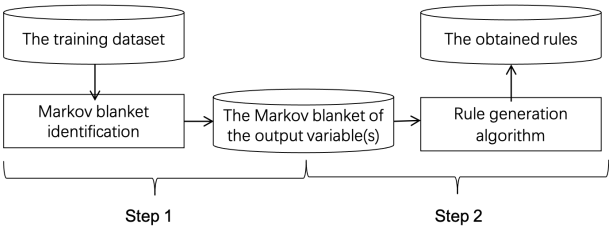


Fig. 1. The MABLAR framework [26]

There are two steps of the MABLAR framework:

- Step 1: Find the Markov blanket of the output variable(s).
- Step 2: Generate fuzzy rules using rule generation algorithms, e.g., the Wang-Mendel algorithm [9].

As we mentioned in Section II.B, most existing ML-based rule generation algorithms mainly focus on correlations. This results in some obtained rules reflecting spurious correlations between inputs and output(s). In this paper, such rules are called spurious rules. In the MABLAR framework, only the variables in the Markov blanket of the output variable(s) will be used to generate rules. Thus, the MABLAR framework can capture, or at least approximate, the causal relationship between inputs and output(s) at the rule level.

III. Counterfactual rule generation

In order to generate rules of Mamdani fuzzy systems which capture causal relationships, and then generate Cf explanations leveraging these causal relationships which can be used to answer the intervention question, we develop a new framework, called CF-MABLAR, which builds on [26] and [24]. The details of the CF-MABLAR will be presented in this section.

A. Analysis of the intervention question

As we mention in Section I, we expect to generate rules which can provide causal CF explanations that can answer the 'intervention question'. Here, we give a more detailed description of the intervention question focused in this paper:

- Intervention Question for fuzzy systems: Suppose we have already obtained a Mamdani FS *F*. Given a test sample $\boldsymbol{x}_{test}$ with d inputs, i.e., $\boldsymbol{x}_{test} = [x_1, x_2, ..., x_d]$, the output of $\boldsymbol{x}_{test}$ obtained from *F* is A. What $\boldsymbol{X}_{test}$ would be, if the Mamdani FS gave another output B?

Here, B is another output of *F*, and the term 'intervening' means changing values. For example, the intervention question in the Iris case can be 'What should the input be for the given flower to be classified as Iris Versicolour?'. As alluded to in Section I, there are at least two obvious use cases for CF explanations, i.e. AI-validation and input-intervention. Answering the intervention question is helpful to achieve these goals.

We divided the intervention question for fuzzy systems into the following two sub-questions:

- Sub-Question 1 (SQ1): *Which variables should we change?* In real world applications, datasets for ML tasks usually contain many variables. Most existing CF explanation methods, such as [22], [24], [45], tend to select all variables for change. However, variables in datasets for ML tasks usually have high correlation with the output variable, but not all of them have a causal relationship with the output variable. As we mention in Section I, if two variables only have high correlation but no causal relationship with each other, then the intervention in one variable will have no effect on the other (see the 'tank and weather' example in Section I). In that case, the CF explanations generated by existing CF methods contain redundant variables, which increases the complexity of the CF explanations. Thus, to solve SQ1, we should only change the variables which have a *causal relationship* with the output variable.
- Sub-Question 2 (SQ2): *How much should we change?* Usually, there are many possible interventions for an input in real world applications. For example, in the 'tank and weather' case shown in Section I, the weather shown in the photo could be 'changed from overcast to sunny using image-editing software' or 'overcast to rainy' etc. Thus, we need to decide which one should be selected. In this paper, according to the Occam's razor, we seek to obtain the expected output with the minimal intervention/change on the inputs.

B. The similarity measurement of rules

Within the framework, we need to measure the similarity between two rules. In this paper, the SR metric proposed in [46] is adopted to calculate the similarity between two rules due to its simpleness. The SR index is calculated as follows:

$$SR(k_1, k_2) = \sum_{i=1}^{D} S(A_i^{k_1}, A_i^{k_2}), \tag{2}$$

where $SR(k_1, k_2)$ 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 this paper, the Jaccard ratio [47] is adopted to calculate the similarity between two fuzzy sets. Thus,

$$S(A_i^{k_1}, A_i^{k_2}) = \frac{\int_{x \in X} \min(\mu_{A_i^{k_1}}(x), \mu_{A_i^{k_2}}(x))}{\int_{x \in X} \max(\mu_{A_i^{k_1}}(x), \mu_{A_i^{k_2}}(x))}. \tag{3}$$

C. The process of CF-MABLAR

The developed CF-MABLAR framework contains five steps. Fig. 2 shows the pipeline of the CF-MABLAR. And a detailed description is shown in Table I.

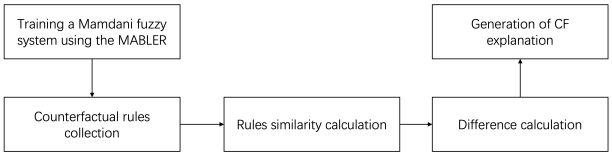


Fig. 2. The pipeline of CF-MABLAR

TABLE I
THE PROCESS OF THE IMPROVED FRAMEWORK

Step 1:	Train a Mamdani fuzzy system using MABLAR framework.
Step 2:	Find the factual rule and collect all CF rules of the test sample.
Step 3:	Calculate the similarity between the factual rule and each counterfactual rule, and find the counterfactual rule k with the highest similarity with the factual rule.
Step 4:	For each input of the test sample x_i, calculate the difference d_i between x_i and x_i^{cf}, where x_i^{cf} is obtained by (4)
Step 5:	Generate the CF explanation based on the factual rule and the CF rule.

Step 1 is to solve SQ1. In step 1, all rules are obtained by the MABLAR framework. As we mentioned in Section II.D, all input variables in rules obtained by MABLAR should have a causal relationship with the output(s). Thus, the intervention on the inputs of rules obtained by MABLAR should affect the output(s).

Step 2 is to find potential CF rules. The terms 'factual rule and 'counterfactual rule' are defined as follows:

- *Factual rule: The rule has the highest firing strength of the test sample.*
- *CF rules: Rules have a different consequent part with the factual rule.*

Usually, there is more than one rule fired for a test sample in a Mamdani fuzzy system. However, to keep the final CF explanation concise, we only consider the rule with the highest firing strength as the factual rule. Thus, each sample has only one corresponding factual rule.

Step 3 and Step 4 are used to solve SQ2. Usually, a number of different CF variants (with different consequents) can commonly be generated for a given rule. in a Mamdani rule base. As discussed in Section III.A, we expect to obtain the expected output with the minimal intervention. Thus, we need to find the CF rule which is most similar with the factual rule in Step 3, i.e., find the CF rule which have the highest SR index with the factual rule.

After we find the most similar CF rule, we need to know how much should we adjust to make sure that after the intervention, the CF rule should have the highest firing strength to obtain the desired output. Thus, in Step 4, we first find the value x_i^{cf} which can make the i-th input have the highest membership degree in rule k on A_i^k:

$$x_i^{cf} = \arg\max_{x_i} \mu_{A_i^k}(x_i) \tag{4}$$

where $\mu_{A_i^k}(x_i)$ is the membership degree of x_i to A_i^k. k is the rule index of the counterfactual rule whose consequent part is the expected output and has the highest similarity with the factual rule. Then, we calculate the difference between x_i and x_i^{cf} in Step 4:

$$diff_i = x_i^{cf} - x_i \tag{5}$$

Here, $diff_i$ is the value that we should adjust for x_i.

Finally, in Step 5, we will generate the final linguistic CF explanation by combining the factual rule, '$diff_i$' and the CF factual rule. An example will be shown in Section IV.B.

D. Properties of CF explanations obtained by CF-MABLAR

Compared with existing ML-based CF explanation generation methods, e.g., [24], the CF-MABLAR can help to improve the interpretability of CF explanations in the following two aspects:

1) Capture of causal relationships between inputs and output(s): Rules of CF-MABLAR are first obtained from MABLAR. As mentioned in Section II.D, MABLAR can capture (or at least approximate) the causal relationship between inputs and output(s). Thus, rules of CF-MABLAR generally capture causal relationships between inputs and output(s). As discussed in [48], human are cause-effect thinkers and rules reflecting causal relationships are more in line with the human way of thinking. Also, we discussed in Section I and Section III.A, interventions are effective only if there is a causal relationship between two variables. Existing CF explanation generation methods for fuzzy rule-based models, such as [24], focus on correlations. Thus, their CF explanations may contain inputs that 'only' correlate with the output. As discussed before, intervention on such inputs will have no effect on the output(s). In this case, the CF explanation will give wrong knowledge and mislead users. For example, 'changing weather from overcast to sunny can lead to enemy tanks become fridently tank'.

2) Reduce the complexity of CF explanations: Usually, not all variables in a dataset have causal relationships with the output variable. Existing ML-based CF explanation generation methods will use all variables to generate CF explanations. However, CF-MABLAR only uses the variables in the Markov blanket of the output variable(s). Usually, the Markov blanket of the output variable(s) contains a smaller (or equal) number of variable than the whole dataset [26]. Thus, CF explanations obtained by CF-MABLAR contain less elements than the CF explanation obtained by correlation-based methods.

IV. A DEMONSTRATIVE EXAMPLE

A. Settings

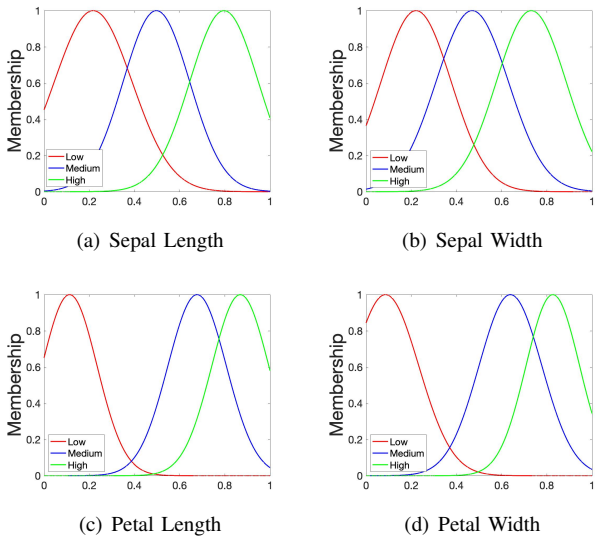
Use the Iris dataset [49] for a simple and accessible demonstration. The dataset contains four features, i.e., inputs of the fuzzy system. The four features are 'Sepal Length'(SL), 'Sepal Width'(SW), 'Petal Length'(PL) and 'Petal Width'(PW), respectively. To keep rules concise, each feature is divided into three fuzzy partitions. Fig. 3 shows the fuzzy partitions of each feature and the corresponding linguistic term. The Gaussian membership function is adopted in this paper, because Gaussian membership functions only require two parameters to be adjusted, which helps the optimization process [7]. All feature values have already been normalized into the range $[0, 1]$.

The output of the Iris dataset is the classifications of iris flowers. There three classifications in the dataset: 'Iris Setosa', 'Iris Versicolour' and 'Iris Virginica'. For simplicity, we select all the samples of 'Iris Setosa' and 'Iris Versicolour'. Thus, the problem becomes a binary classification problem and the dataset contains 100 samples.

B. Illustrative example of the CF-MABLAR framework

We demonstrate the example following the steps shown in Table I.

Step 1: Training a Mamdani fuzzy system using the MABLAR framework. In this step, the PC-algorithm [50] and the Wang-Mendel algorithm [9] are used to find the Markov blanket of the output and generate rules, respectively.



(a) Sepal Length (b) Sepal Width

(c) Petal Length (d) Petal Width

Fig. 3. Fuzzy partitions of each features

One can also use other algorithms according to specific applications. *The obtained Markov blanket* contains sepal length, petal length and petal width. Note that, the obtained causal graph may still not be perfect. According to [51], biologists actually only use 'Petal length' and 'Petal width' to classify Iris. Part of rules of the obtained Mamdani fuzzy system are shown in Table II.

TABLE II
PART OF RULES OF THE OBTAINED FUZZY SYSTEM

	Sepal Length	Petal Length	Petal Width	Class
R1	Low	Low	Low	Setosa
R2	Low	Low	Medium	Setosa
R3	Medium	Medium	High	Versicolour

Step 2: Collect CF rules. To make sure our selection is unbiased, we randomly select a sample from the dataset. The feature value of the selected sample is $sepal\ length = 0.2963$, $petal\ length = 0.0976$ and $petal\ width = 0.0588$, respectively. The sample is classified into the Setosa class and the corresponding factual rule is '*The sepal length is low and the petal length is low and the petal width is low, then class is Setosa*'. The collection of counterfactual rules will contain all rules of which the consequent part is Versicolour, i.e., NOT Setosa. In total, there are six rules have Versicolour consequent part of the obtained Mamdani rule base, which is shown in Table III.

Step 3: Calculate the similarity. In our example, R6 in Table III has the highest SR index (shown in (2)) with the factual rule and the corresponding similarity is 0.9362.

Step 4: Calculate the difference. The x_i^{cf} calculated by (4) of R6 in Table III are 0.2172, 0.8683 and 0.7212, respectively. Thus, the difference between each input of the test sample and its corresponding x_i^{cf} are -0.0791, 0.0976 and 0.0588, respectively.

TABLE III
COUNTERFACTUAL RULES OF THE SELECTED SAMPLE

	Sepal Length	Petal Length	Petal Width	Class
R1	Medium	Medium	High	Versicolour
R2	High	Medium	High	Versicolour
R3	Low	Medium	High	Versicolour
R4	Medium	High	High	Versicolour
R5	High	High	High	Versicolour
R6	Low	High	High	Versicolour

Step 5: Generate CF explanation. So far, we can obtain the final explanation of the test sample. To make it clearer, we divide the CF explanation into three parts: the factual part, the CF part and the CF conclusion part:

- The factual part: The test sample is *Iris Setosa*, because its *sepal length* is **low** and its *petal length* is **low** and the *petal width* is **low**.
- CF part: To obtain *Iris Versicolour*, its *sepal length* should be **decreased 0.0791** and its *petal length* should be **increased 0.0151** and its *petal width* should be **increased 0.0029**.
- The CF conclusion part: In that case, its *sepal length* would be **low** and its *petal length* would be **high** and its *petal width* would be **high**, and it would be classified *Iris Versicolour*.

In this paper, we exclusively focus on conjunctive rules (i.e. using the *and* logical connective). Considering disjunctive rules employing *or*, or indeed a mix of both is possible, however increases complexity and is left to future work.

V. EXPERIMENTS AND RESULTS

We compare CF-MABLAR with the Correlation-based CF (Cor-CF) explanation framework for fuzzy rule-based classifiers proposed in [24]. The experiments are also conducted on the Iris data set to allow for concise comparison in this paper. The experiment settings are the same as Section IV.A.

A. Evaluation metric

The following three indices are adopted in this paper:

- Number of inputs (#inputs): The number of inputs of each rule. This index is used to measure the complexity of rules. A rule with more inputs can be viewed as having higher complexity, and thus lower interpretability [26], [52]. Although the complexity of rules is affected by multiple factors, such as the number of fuzzy partitions of each input, structure of rules etc. [52], the number of inputs is one intuitive index. All the remaining factors are the same for both frameworks.
- Average Minimal Intervention (AMI): A better CF explanation should have a lower degree of intervention [19], which means a good CF explanation should change the inputs as little as possible. Thus, we define an AMI index as follows:

$$AMI = \frac{\sum_{j=1}^{n}\sum_{i=1}^{d}(diff_i^j)^2}{n}, \tag{6}$$

where n is the number of samples to be intervened (in this paper, $n = 100$) and d is the number of inputs.

- Validity: Validity is the percentage of samples which obtain the desired output after intervention [19].

B. Results

We train two FSs using the settings shown in Section IV.A. Specifically, we train one Mamdani fuzzy system for the causality-focused CF-MABLAR, and another for the standard Cor-CF approach. Then, we generate a CF explanation from both frameworks for each sample in the data set. Thus, each sample has two CF explanations (one from the CF-MABLAR and one from the Cor-CF), and each framework generates 100 CF explanations, because there are 100 samples in the dataset as shown in Section IV.A. The results are shown in Table IV.

TABLE IV
RESULTS

	#inputs	AMI	Validity
CF-MABLAR	3	0.6443	1
Cor-CF	4	0.7394	1

From Table IV we can make the following observations:

1) The validity indices of both frameworks are 1, which means both two frameworks can obtain the desired output by the interventions provided by their CF explanations. As this is a low-complexity, binary classification task, this is expected.
2) Compared with Cor-CF, CF-MABLAR has a smaller number of inputs in its rules. Thus, rules generated by CF-MABLAR have less complexity, which means a better interpretability as discussed in Sections III.C and V.A. This behaviour is in-line with results originally presented in [26].
3) The AMI of CF-MABLAR is smaller than Cor-CF. This is because for Cor-CF intervention affects a larger number of inputs–including those which do not share a causal relationship with the output. Specifically, as discussed in Section IV.B, the Markov blanket of the output does note contain the variable 'sepal width', indicating that 'sepal width' does not have a causal relationship with the output for this data set. Thus, the intervention on 'sepal width' is a redundant intervention, which in turn limits the interpretability of CF explanation obtained by Cor-CF, because redundant interventions increase the complexity of CF explanation. Also, redundant intervention results in potentially misleading explanations, indicating to the user that changes in the 'sepal width' will result in a different classification, even though this is not the case—for the causal system.

VI. CONCLUSIONS

In this paper, we develop an integrated rule-based reasoning and Counterfactual (CF) explanation generation framework, which we refer to as CF-MABLAR, for Mamdani fuzzy systems. The proposed framework can generate rules which capture (or at least approximate) the causal relationships between inputs and output(s) for Mamdani inference– and it enables the generation of CF explanations leveraging these causal relationships. The CF explanation provided by CF-MABLAR can tell users how to intervene, i.e. how to change given input features of a sample to obtain a desired output. Compared with existing correlation-based CF explanation generation frameworks for fuzzy rule-based systems, the CF explanation have the potential to be more concise and avoid potentially misleading (CF) explanations.

Experiments in this paper are limited to the proof-of-concept stage. In future work, we expect to consider larger and more complex datasets, while discussing the challenge of Markov blanket approximation and its impact on the causal quality of the rule set as a whole.

Finally, note that in this paper, we consider that all input variables can be changed. However, this may not happen in the real-world. For example, if an input variable is the height of an adult, it will be unchangeable. Also, it is assumed that input variables can be modified to any extent, which is also unusually in the real-world. As alluded to in the introduction, there are at least two obvious use cases for CF explanations, i.e. AI-validation and input-intervention. In future work, we will explore how to generate CF explanations for different purposes and under constraints such as the immutability of inputs.

REFERENCES

[1] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information Fusion*, vol. 58, pp. 82–115, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253519308103

[2] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, and G.-Z. Yang, "Xai—explainable artificial intelligence," *Science Robotics*, vol. 4, no. 37, 2019.

[3] A. A. Freitas, "Comprehensible classification models: a position paper," *ACM SIGKDD explorations newsletter*, vol. 15, no. 1, pp. 1–10, 2014.

[4] Y. Jiang, Z. Deng, and S.-T. Wang, "0-order-l2-norm-takagi-sugeno-kang type transfer learning fuzzy system," *Tien Tzu Hsueh Pao/Acta Electronica Sinica*, vol. 41, pp. 897–904, 05 2013.

[5] Y.-Y. Lin, J.-Y. Chang, and C.-T. Lin, "A tsk-type-based self-evolving compensatory interval type-2 fuzzy neural network (tscit2fnn) and its applications," *IEEE Transactions on Industrial Electronics*, vol. 61, no. 1, pp. 447–459, 2014.

[6] Z. Deng, K.-S. Choi, L. Cao, and S. Wang, "T2fela: Type-2 fuzzy extreme learning algorithm for fast training of interval type-2 tsk fuzzy logic system," *Neural Networks and Learning Systems, IEEE Transactions on*, vol. 25, pp. 664–676, 04 2014.

[7] M. Pota, M. Esposito, and G. De Pietro, "Interpretability indexes for fuzzy classification in cognitive systems," in *2016 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. IEEE, 2016, pp. 24–31.

[8] M. Riveiro and S. Thill, ""that's (not) the output i expected!" on the role of end user expectations in creating explanations of ai systems," *Artificial Intelligence*, vol. 298, p. 103507, 2021.

[9] 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.

[10] 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.

[11] Z. Fan, J. Gou, C. Wang, and H. Chi, "A reduced weighted wang-mendel algorithm using the clustering algorithm to build fuzzy system," in *2016 International Conference on Progress in Informatics and Computing (PIC)*, 2016, pp. 8–12.

[12] W. Chen and J. Gou, "Improved wang-mendel scheme based on cooperation among input variables," *International Review on Computers and Software*, vol. 7, no. 5, 2012.

[13] M. Galende-Hernández, G. I. Sainz-Palmero, and M. J. Fuente-Aparicio, "Complexity reduction and interpretability improvement for fuzzy rule systems based on simple interpretability measures and indices by bi-objective evolutionary rule selection," *Soft Computing*, vol. 16, no. 3, pp. 451–470, 2012.

[14] H. Ishibuchi, K. Nozaki, N. Yamamoto, and H. Tanaka, "Construction of fuzzy classification systems with rectangular fuzzy rules using genetic algorithms," *Fuzzy sets and systems*, vol. 65, no. 2-3, pp. 237–253, 1994.

[15] H. Ishibuchi and T. Yamamoto, "Fuzzy rule selection by multi-objective genetic local search algorithms and rule evaluation measures in data mining," *Fuzzy sets and systems*, vol. 141, no. 1, pp. 59–88, 2004.

[16] L.-C. Duţu, G. Mauris, and P. Bolon, "A fast and accurate rule-base generation method for mamdani fuzzy systems," *IEEE Transactions on Fuzzy Systems*, vol. 26, no. 2, pp. 715–733, 2018.

[17] C.-T. Lin, M. Prasad, and J.-Y. Chang, "Designing mamdani type fuzzy rule using a collaborative fcm scheme," in *2013 International Conference on Fuzzy Theory and Its Applications (iFUZZY)*, 2013, pp. 279–282.

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

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

[20] K. Sokol and P. A. Flach, "Counterfactual explanations of machine learning predictions: opportunities and challenges for ai safety," in *SafeAI@ AAAI*, 2019.

[21] 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.

[22] 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.

[23] D. Martens and F. Provost, "Explaining data-driven document classifications," *MIS quarterly*, vol. 38, no. 1, pp. 73–100, 2014.

[24] 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.

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

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

[27] R. R. Hoffman, G. Klein, and S. T. Mueller, "Explaining explanation for "explainable ai"," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 62, no. 1. SAGE Publications Sage CA: Los Angeles, CA, 2018, pp. 197–201.

[28] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins *et al.*, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information fusion*, vol. 58, pp. 82–115, 2020.

[29] T. Kulesza, M. Burnett, W.-K. Wong, and S. Stumpf, "Principles of explanatory debugging to personalize interactive machine learning," in *Proceedings of the 20th international conference on intelligent user interfaces*, 2015, pp. 126–137.

[30] J. Fox, D. Glasspool, D. Grecu, S. Modgil, M. South, and V. Patkar, "Argumentation-based inference and decision making–a medical perspective," *IEEE intelligent systems*, vol. 22, no. 6, pp. 34–41, 2007.

[31] M. T. Ribeiro, S. Singh, and C. Guestrin, """ why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[32] P. Lipton, "Contrastive explanation," *Royal Institute of Philosophy Supplements*, vol. 27, pp. 247–266, 1990.

[33] I. Stepin, J. M. Alonso, A. Catala, and M. Pereira-Fariña, "A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence," *IEEE Access*, vol. 9, pp. 11 974–12 001, 2021.

[34] T. Laugel, M.-J. Lesot, C. Marsala, X. Renard, and M. Detyniecki, "Unjustified classification regions and counterfactual explanations in machine learning," in *Joint European conference on machine learning and knowledge discovery in databases*. Springer, 2019, pp. 37–54.

[35] 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*, 2020, pp. 607–617.

[36] I. Stepin, A. Catala, M. Pereira-Fariña, and J. M. Alonso, "Factual and counterfactual explanation of fuzzy information granules," in *Interpretable Artificial Intelligence: A Perspective of Granular Computing*. Springer, 2021, pp. 153–185.

[37] J. Hühn and E. Hüllermeier, "Furia: an algorithm for unordered fuzzy rule induction," *Data Mining and Knowledge Discovery*, vol. 19, no. 3, pp. 293–319, 2009.

[38] E. H. Mamdani, "Application of fuzzy logic to approximate reasoning using linguistic synthesis," *IEEE transactions on computers*, vol. 26, no. 12, pp. 1182–1191, 1977.

[39] 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.

[40] J. Gou, F. Hou, W. Chen, C. Wang, and W. Luo, "Improving wang–mendel method performance in fuzzy rules generation using the fuzzy c-means clustering algorithm," *Neurocomputing*, vol. 151, pp. 1293–1304, 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231214014738

[41] S. Guillaume, "Designing fuzzy inference systems from data: An interpretability-oriented review," *IEEE Transactions on fuzzy systems*, vol. 9, no. 3, pp. 426–443, 2001.

[42] J. M. Alonso and L. Magdalena, "Special issue on interpretable fuzzy systems," pp. 4331–4339, 2011.

[43] Q. Zhang and H. Li, "Moea/d: A multiobjective evolutionary algorithm based on decomposition," *IEEE Transactions on evolutionary computation*, vol. 11, no. 6, pp. 712–731, 2007.

[44] T. R. Razak, J. M. Garibaldi, C. Wagner, A. Pourabdollah, and D. Soria, "Toward a framework for capturing interpretability of hierarchical fuzzy systems—a participatory design approach," *IEEE Transactions on Fuzzy Systems*, vol. 29, no. 5, pp. 1160–1172, 2020.

[45] B. Ustun, A. Spangher, and Y. Liu, "Actionable recourse in linear classification," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2019, pp. 10–19.

[46] 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.

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

[48] J. Pearl, *Causality*. Cambridge university press, 2009.

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

[50] P. Spirtes, C. N. Glymour, R. Scheines, and D. Heckerman, *Causation, prediction, and search*. MIT press, 2000.

[51] X. Qian, Y. Xu, and Z. Hu, *Flora of China: Irisaceae*. Beijing Science press, 1985.

[52] T. R. Razak, J. M. Garibaldi, C. Wagner, A. Pourabdollah, and D. Soria, "Interpretability and complexity of design in the creation of fuzzy logic systems — a user study," in *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*, 2018, pp. 420–426.