Generating locally relevant explanations using causal rule discovery

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*—In the real-world an effect often arises via multiple causal mechanisms. Conversely, the behaviour of AI systems is commonly driven by correlations which may–or may not–be themselves linked to causal mechanisms in the associated real-world system they are modelling. From an AI and XAI point of view, it is desirable for AI systems to model and communicate primarily, if not exclusively, causal mechanisms between variables, affording strong generalisation performance and effective explanations. Indeed, as we discuss in this paper, it is critical for explanations for a given effect not only to reflect possible causal mechanisms, but to highlight the specific causal mechanisms which led to the effect in the given instance. In this light, we proceed to propose a rule generation framework which generates rules for fuzzy systems that capture possible causal mechanisms between the input variables and the target variable as discovered by data-driven causal discovery algorithms for the given data set. For a given sample, i.e., a specific set of inputs, the obtained fuzzy system provides local explanations which distinguish the locally relevant causal mechanism(s) of its effect from other possible, but not applicable causal mechanisms, and thus avoids both overly simplistic single-cause and exhaustive–potentially misleading explanations. Experiments show that the fuzzy systems obtained by the proposed framework achieve comparable performance compared to classical correlation-based approaches, and provide local explanations which indicate the specific causal mechanism for different effects.**

***Index Terms*—Fuzzy, Causal graph, Causal weights, Rules**

I. Introduction

In the real-world, a specific effect is often the combined result of multiple causal mechanisms and associated set of inputs [1]. For example, infection with COVID-19 or flu virus can lead to COVID-19 or flu respectively, and both COVID-19 and the flu can cause a fever. In this scenario, the *fever* is the effect and there are two possible causal mechanisms of fever: 1) 'infection with the COVID-19 virus' → 'COVID-19' → 'fever' and 2) 'infection with the flu virus' → 'flu' → 'fever'. If a patient with a fever has COVID-19 but does not have flu, the locally relevant causal mechanism of 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 of their fever is the second causal mechanism.

It is important to distinguish the *locally relevant* causal mechanism of a given effect from other *possible* but situationally unrelated causal mechanisms. Returning to the above fever example, when a doctor explains to the patient with flu why they have a fever, the doctor usually explains, 'Because you were infected with the flu virus, it resulted in you contracting flu, and flu subsequently led to your fever.' rather than saying, 'Because you were infected with the flu virus and not with the COVID virus, you contracted flu and not COVID-19. Flu caused your fever, not COVID-19.' In other words, explanations generally focus on the locally relevant causal mechanism for the given specific, i.e. 'local'–in XAI terms–case. In addition, only by identifying the locally relevant causal mechanism can doctors implement appropriate treatment measures to cure the patient. In other words, identifying the locally relevant causal mechanism of a given effect enables suitable treatment and is critical to enabling meaningful explanations in the context of AI.

In recent years, with the increasing demand for explainable artificial intelligence (XAI), fuzzy systems have attracted interest due to their potential for strong explainability and high prediction accuracy. The linguistic rules of fuzzy systems reflect relationships in a human understandable structure and thus can deliver explanations of the operation of the fuzzy system model [2].

At the same time, the automated generation of rules from a given dataset through data-driven approaches such as the Wang-Mendel(WM) algorithm [3], FURIA [4] etc., is quasi the norm in many areas of application. Here, to generate fuzzy systems with high performance, such as high classification accuracy in a classification problem, data-driven approaches exploit statistical correlation between variables within the given data set–rather than necessarily capturing causal mechanisms. As a result, rules obtained by such approaches often reflect correlations between variables. However, the 'IF-THEN' structure of rules implicitly conveys to users that there *is* a causal relationship between the variables in the antecedent and the consequent. Humans are cause-effect thinkers [5] and rules in rule-based systems are expected to reflect causal mechanisms between variables. In other words, resulting systems risk sacrificing the potential for strong explainability of fuzzy systems in exchange for strong performance given the data available - potentially limiting the utility of using fuzzy systems compared to other modelling techniques.

Generating fuzzy systems that capture possible causal mechanisms of the target variable from a given data set is a challenging area [6]. Beyond generating working rule bases and explanations, an area which has seen comparatively little attention is the generation of fuzzy systems with the capacity to provide local explanations that distinguish the locally relevant causal mechanism for each specific effect from the set of possible captured causal mechanisms–as alluded to above. To address this problem, in this paper we propose a rule generation framework which we refer to as MArkov BLanket Rule generation-Causal Weight (MABLAR-CW).

The main contributions of this paper are as follows:

- We introduce the concept of locally relevant explanations and discuss their relevance to generating meaningful local explanations for fuzzy systems.
- We propose MABLAR-CW: a rule generation framework to generate fuzzy system rules which capture the set of possible causal mechanisms between the input variables and a given target variable, given a data set.
- We conduct experiments to evaluate the proposed framework on several real-world data sets in terms on performance and explainability, exploring in particular its potential for delivering improved local interpretability.

This paper is organised as follows: Section II provides the background. Section III introduces MABLAR-CW. Section IV presents the experiment results and the analysis of the obtained results. Section V presents the conclusions.

II. Background

In the sections below, we provide essential background on causal graphs, algorithms designed to discover causal structure from data–specifically linear non-Gaussian acyclic models (LiNGAM), as well as existing causal fuzzy rule-generation frameworks.

A. Causal graphs

Causal relationships between variables in the real world are complicated. To intuitively visualise causal relationships, in [7], the author proposed the concept of causal graphs. A causal graph is directed and acyclic. Within the graph, each node represents a variable. If two variables have a causal relationship, this is reflected by an edge between them, pointing from the cause to the effect.

Fig. 1 shows an example of a causal graph. The figure provides examples of regularly used concepts in causal graphs, specifically: the set of variables within the data (circular nodes); the target variable or output of the system (black node); causal paths (edges); causal direction (arrows, i.e. directional edges); causal weights w_{ij} specifying the weight of the edge between nodes x_i and x_j; the Markov blanket (MB) of the target variable, which is the set of variables that consists of the target variable's parents, children and spouses in the causal graph (shaded in grey); and *direct cause variables*, i.e. the variables which have a directional causal path pointing *towards* the target variable (hatched).

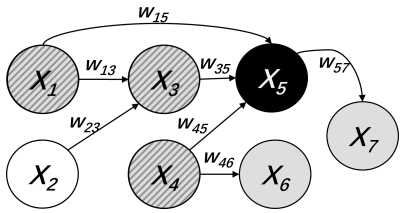


Fig. 1. An example of a causal graph

It is worth noting that the concept of causal weights mentioned above is complex. For example, in [8], the causal weights between two variables are the probability that there is a causal relationship between the variables. In [9] and [10], the causal weights between two variables represent the degree of influence of the cause on the effect. In [11], the authors suppose that the value of a variable is a linear summation of its corresponding cause variables, and the causal weights between two variables are the coefficient of their corresponding linear function. Although there are various definitions of causal weights, they share a common element: they describe a notion of the strength of causal link between two variables.

In this paper, we adopt the definition of causal weight as introduced in [11], where the causal weight of the edge between two variables is the coefficient of their corresponding causal linear function. We adopt this definition because it intuitively demonstrates the causal strength between variables. Furthermore, there are already several classical algorithms designed to directly estimate this weight from the given observed data set without requiring prior knowledge–a capability which we seek to leverage–which will be introduced in the next subsection.

B. Linear non-Gaussian acyclic models

LiNGAM-based algorithms are popular algorithms to estimate a causal weighted graph for a given data set. LiNGAM assumes that causal relationships between variables follow a set of conditions [11]: 1) The value of variable X_i is the linear sum of its cause variables, along with a noise variable n_i and a constant c_i. 2) n_i follows a non-Gaussian distribution with non-zero variance and n_i are independent of each other. We stress that when the given data set fails to meet these conditions, LiNGAM-based algorithms may find incorrect causal relationships between variables. In this paper, this means that the fuzzy system obtained by MABLAR-CW may contain fuzzy sub-systems that capture incorrect causal mechanisms, resulting decreased performance and flawed local explanations. Causal discovery is an active area of research, and alternative, including non-linear approaches may provide improved capability in future [12], [13].

The LiNGAM model can be represented as follows:

$$x_i = \sum_{o(j)<o(i)} w_{ji} x_j + n_i + c_i, \tag{1}$$

where x_i is the value of variable X_i, w_{ji} is the weight of the edge between X_i and X_j, c_i is a constant, and $o(i)$ is the rank of variable X_i in the causal order. In a causal order, if $o(j) < o(i)$, then X_j must be a parent or an ancestor of X_i [11]. Dropping c_i, (1) can be transformed into the following matrix form:

$$\boldsymbol{X} = \boldsymbol{W}\boldsymbol{X} + \boldsymbol{N}, \tag{2}$$

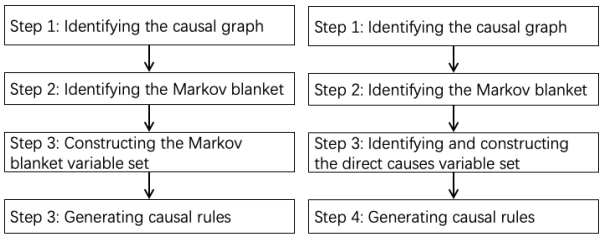
Solving for W in (2) enables obtaining the causal relationships between the variables.

In this paper, we adopt the DirectLiNGAM algorithm [14] to solve (2), as it offers good robustness in respect to initialisation conditions [15]. In other words, given the same data set, the results obtained by the DirectLiNGAM algorithm remain stable across different runs. The stability of the DirectLiNGAM algorithm benefits MABLAR-CW by resulting in the generation of the same local causal explanations for the same effect—if MABLAR-CW were to be executed multiple times on the same data set, i.e., improving the stability of local explanations. This stability of local explanation improves the trustworthiness of the generated local explanations, conditional on the correctness of the explanation itself—we acknowledge that trustworthiness in this space is a broad and complex question in of itself– beyond the remit of this paper.

C. Markov blanket rule generation frameworks

Using fuzzy sets for causal discovery has achieved significant progress, such as the fuzzy PC/FCI SCI approach [6], fuzzy cognitive maps [16]–[18], causal fuzzy neural networks [19], and have been widely used in many real-world applications [20], [21]. In this subsection, we highlight the Markov blanket rule generation framework (MABLAR) [22] and the MABLAR-causal direction framework (MABLAR-CD) [23] as they are specifically designed to improve the interpretability of fuzzy systems using causal discovery.

MABLAR and MABLAR-CD are two data-driven rule generation frameworks which are initially formulated in [22] and [23] for generating rules which capture the causal relationships between variables. Fig. 2(a) and Fig. 2(b) present the process of MABLAR and MABLAR-CD, respectively.



(a) MABLAR (b) MABLAR-CD

Fig. 2. MABLAR and MABLAR-CD

There are four steps in MABLAR and MABLAR-CD. As shown in Fig. 2, Step 1 and Step 2 of MABLAR and MABLAR-CD are identical. Both frameworks identify the causal graph of the given data set and identify the MB of the target variable in their first two steps. However, in contrast to MABLAR, which constructs a subset, marked as D_{MB}, which only contains variables within the MB of the target variable in Step 3, MABLAR-CD further identifies the direct cause(s) of the target variable and constructs a subset, marked as D_{CD}, which only contains all direct cause variables. Finally, in Step 4, using data-driven algorithms (e.g. the WM algorithm), MABLAR generates rules from D_{MB}, while MABLAR-CD generates rules from D_{CD}

The key difference between MABLAR and MABLAR-CD is the causal mechanism they use for rule generation. MABLAR uses the causal mechanism between the target variables and variables within D_{MB} to generate rules, with the aim of avoiding generating rules based solely on correlations. MABLAR-CD uses the causal mechanism between the target variables and variables within D_{CD} to generate rules which capture relationships between the target variable and its *direct* cause variables. While this is effective, as discussed in Section I, many direct causal mechanisms may possibly lead to a given effect, in practice only one or a small set of these may be at play in a given instance. MABLAR or MABLAR-CD only focus on one causal mechanism, resulting in the obtained fuzzy systems using single causal mechanism to generate local explanations for different effects. This limits the value of their explanations in the specific, local case in comparison to more focused explanations targeting the specific causes of the given instance–which is what we address in this paper.

III. MABLAR-CW – IN A CLASSIFICATION CONTEXT

In this section, focusing on the classification problem, we provide the details of MABLAR-CW, including its rationale, and its processes for rule generation, prediction and local explanation generation.

A. The rationale of MABLAR-CW

MABLAR-CW uses a given data set and its corresponding causal graph to generate fuzzy systems. The obtained fuzzy system is constructed as a set of fuzzy sub-systems akin to an ensemble. Each fuzzy sub-system models one complete causal path in the causal graph of the given data set. Here, a *complete* causal path of a target variable is a path of which all edges points to the target variable, and the starting node of this path has no parents in the causal graph. For example, in Fig. 1, $X_1 \rightarrow X_3 \rightarrow X_5$ is a complete causal path of X_5.

In the causal graph of a given data set, 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. By modelling complete causal paths using a set of fuzzy sub-systems, the fuzzy system obtained by MABLAR-CW captures possible causal mechanisms shown in the causal graph. Then, for a given sample, i.e., a specific set of inputs,

MABLAR-CW uses the causal weights from the causal graph to distinguish the locally relevant causal mechanism from the possible causal mechanisms captured by the fuzzy system.

In the next subsection, we first introduce the fuzzy system generation process of MABLAR-CW.

B. The fuzzy system generation process of MABLAR-CW

Fig. 3 shows the fuzzy system generation process of MABLAR-CW.

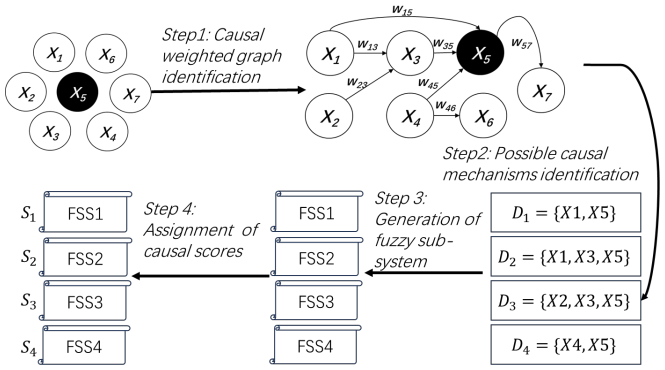


Fig. 3. The training process of MABLAR-CW

To understand the strategy of MABLAR-CW, consider a casual graph such as in Fig. 1, with a target variable X_5, representing the class of a sample, with values being either C_1 or C_2. The fuzzy system generation process of MABLAR-CW proceeds in the following steps:

Step 1: Causal weighted graph identification. The inputs of MABLAR-CW are the given data set and its corresponding causal weighted graph. In this paper, we adopt the DirectLiNGAM algorithm to achieve this goal as explained in Section II-B.

Step 2: Possible causal mechanisms identification. In this step, MABLAR-CW first identifies complete causal paths of the target variable in the causal graph of the given data set. Then, MABLAR-CW constructs a variable subset D_j for the jth complete causal path of the target variable. D_j contains all the variables within the jth path. For the example shown in Fig. 3, four complete causal paths, i.e., four possible causal mechanisms of X_5 are identified: 1) $X_1 \rightarrow X_5$, 2) $X_1 \rightarrow X_3 \rightarrow X_5$, 3) $X_2 \rightarrow X_3 \rightarrow X_5$, and 4) $X_4 \rightarrow X_5$. Thus, 4 subsets, i.e., D_1-D_4, are constructed.

Step 3: Generation of fuzzy sub-systems. In this step, MABLAR-CW generates one fuzzy sub-system from each subset obtained in Step 2 using data-driven approaches, such as the WM algorithm and FURIA, which makes sure each identified possible causal mechanism of the target variable is modelled by one fuzzy sub-system. In this paper, we adopt the WM algorithm, because it is simple but effective and provides reasonably good performance [24]. For this example, four fuzzy sub-systems are generated, They are 'FSS1' - 'FSS4' as shown in Fig. 3.

Step 4: Assignment of Causal Scores. In this step, each fuzzy sub-system is assigned a causal score based on the weights of its corresponding complete causal path of the target variable. The causal score of each fuzzy sub-system measures the causal impact of its corresponding causal mechanism on the target variable and is calculated using (3).

$$S_j = \sum_{i=1}^{E} \frac{W_i}{i}, \tag{3}$$

where S_j is the causal score of the jth fuzzy sub-system, E is the number of edges in its causal path, i is the index of the edge within the path from the path's endpoint to its starting point, and the W_i is the weight of the ith edge. For example, Fig. 4 shows a causal path with four variables, namely V_1 to V_4. Below each edge are their corresponding index numbers, while above each edge are their corresponding weights. In this case, the target variable is V_4 and the causal score of this path is $\frac{W_{12}}{3} + \frac{W_{23}}{2} + W_{34}$. We will further explain the rationale of (3) in the next subsection. In the example shown in Fig. 3, four causal scores are obtained, denoted as S_1, S_2, S_3, and S_4.

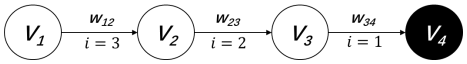


Fig. 4. The causal path from V_1 to V_4

We now have the final fuzzy system represented by the set of fuzzy sub-systems. The following rule is the kth rule of the jth fuzzy sub-system:

$$\begin{aligned} &\text{If } x_1 \text{ is } A_{1j}^k \text{ and } x_2 \text{ is } A_{2j}^k \text{ and ... and } x_d \text{ is } A_{dj}^k \\ &\text{Then } Class \text{ is } C_j^k \text{ with causal impact } S_j, \end{aligned} \tag{4}$$

where $\boldsymbol{x} = [x_1, \ldots, x_d]$ is a training sample with d input variables, A_{dj}^k is the antecedent fuzzy set of the k-th rule for the d-th input variable in the jth fuzzy system, C_j^k is the consequent class of k-th rule in the jth fuzzy system, and S_j is the causal score obtained by (3). For this example, $C_j^k \in \{C_1, C_2\}$ and the second rule of FSS1 in Fig. 3 can be 'If x_1 is A_{11}^2, Then Class is C_1 with causal impact S_1'

In the next subsection, we explain the rationale of (3).

C. The causal score calculation in MABLAR-CW

MABLAR-CW uses (3) to assign a causal score for each fuzzy sub-system. Equation (3) is motivated by the consideration that causal impacts of indirect causal variables on the target variable often occur through the sequential transmission of causal effects onto the target variable via intermediate variables. Here, intermediate variables refer to the variables along a causal path between the target variable and its indirect causal variables. For example, in Fig. 4, V_3 is the intermediate variable between V_4 and V_2. Considering that the further indirect causal variables are from the target variable, the more intermediate variables are in play to transmit their causal impact. This leads to a more pronounced attenuation of causal effects. As shown in Fig. 4, V_1 requires two intermediate variables while V_2 only requires one. Thus,

we assign a smaller weight for the variable which is further away from the target variable in the causal path.

In addition, in this paper, we adopt the LiNGAM model, which assumes an effect is a linear sum of its direct/indirect cause variables [11], for causal discovery. Thus, in (3), we sum the weighted causal impact of all variables in the corresponding causal path.

In the next subsection, we introduce the prediction process of the fuzzy system obtained by MABLAR-CW.

D. The prediction process of MABLAR-CW

Using a specific input sample $\boldsymbol{x_{test}} = [x_1^{te}, x_2^{te}, x_3^{te}, x_4^{te}, x_6^{te}, x_7^{te}]$, Fig. 5 illustrates the prediction process of MABLAR-CW following the same example as Fig. 3. Additionally, to provide a better description, Fig. 5 also shows each step of FSS1.

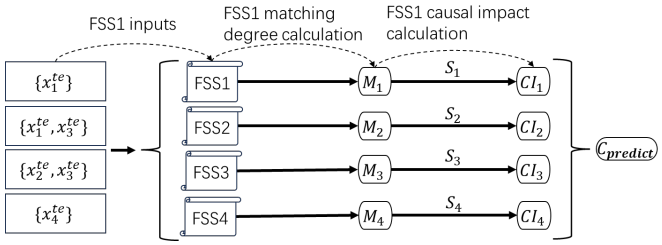


Fig. 5. The prediction process of MABLAR-CW

Given an input sample, MABLAR-CW generates a prediction based on the locally relevant causal mechanism as shown in the steps below.

Step 1: Matching degree calculation. In this step, the matching degree for each fuzzy sub-system is calculated in respect to the relevant subset of inputs relevant to it, as illustrated Fig. 5. The matching degree of each fuzzy sub-system is calculated using (5):

$$M_j = \max_k \prod_{d=1}^{D_j} \mu_{dj}^k(x_d), \tag{5}$$

where M_j is the matching degree of the jth fuzzy sub-system to its corresponding inputs, x_d is the dth input of the jth fuzzy sub-system, $\mu_{dj}^k(x_d)$ is the membership degree of x_d to A_{dj}^k, K is the number of rules in the jth fuzzy sub-system. Note that we are using the product t-Norm to compute the given rule's firing strength.

In other words, the matching degree is the maximum firing strength of the given sample for each sub-system. We mark the rule with the maximum firing strength of the given sample for the jth sub-system as R_j^{max}. The rationale for this is that the firing strengths indicate how well a set of inputs matches the given rule. Thus, as shown in (5), in this paper, for each fuzzy sub-system, MABLAR-CW calculates the firing strength of each rule for its corresponding inputs. Considering that the locally relevant causal mechanism should highly match the sample and facilitate the comparison between different causal mechanisms, in (5), only the highest firing strength is selected to represent the matching degree of each causal mechanism modelled by the corresponding fuzzy sub-system. As shown in Fig.5, in this case, 4 matching degrees are obtained, i.e., M_1-M_4, corresponding to R_1^{max}-R_4^{max}.

Step 2: Locally relevant causal impact calculation. As discussed in Section I, a locally relevant causal mechanism is one which matches the input sample, i.e., it is possible given the inputs, and it should have a strong causal impact on the effect. Thus, in this step, MABLAR-CW calculates the locally relevant causal impact CI_j of each jth fuzzy sub-system using $CI_j = M_j \cdot S_j$, i.e. the product of the causal score arising from the causal graph and the firing strength of the given input set for R_j^{max}, i.e., the matching degree M_j.

Step 3: Final prediction via voting. The final prediction for the given sample is selected as the consequent of R_j^{max} which results the highest CI_j. MABLAR-CW uses the 'winner takes all' voting strategy, to reflect the intention to focus on the most locally relevant causal mechanism. For this example, suppose the consequent of R_j^{max} with the highest CI_j is C_1. The the final prediction for $\boldsymbol{x_{test}}$, i.e., $C_{predict}$ in Fig. 5, would be C_1. Of course, alternative approaches could explore combining a set of best-fitting rules and causal mechanism reflecting that in the the real world frequently multiple causal mechanisms come together to underpin an effect.

Having established how MABLAR-CW generates an output for a given input sample, we proceed to explain how MABLAR-CW generates an associated local explanation.

E. Local explanation generation of MABLAR-CW

MABLAR-CW is designed not only to generate a prediction, but to also provide the locally relevant explanation for the effect predicted for a given input sample, thus providing the most relevant causal explanation for it. As explained above, CI_j measures the locally relevant causal impact of the jth fuzzy sub-system on the predicted effect. Thus, given a sample, the local explanation of its effect predicted by the fuzzy system obtained by MABLAR-CW is R_j^{max} which results the highest CI_j. Having established both the prediction and explanation stages of MABLAR-CW, we proceed to experimental evaluation and analysis in the next section.

IV. Experiments

In this section, we evaluate the performance and explainability of the fuzzy systems generated using MABLAR-CW.

A. Experiment settings

We compare MABLAR-CW with 3 different rule generation approaches:MABLAR, MABLAR-CD and the WM algorithm, respectively. MABLAR and MABLAR-CD provide local explanations for different effects based on the same causal mechanism, while MABLAR-CW provides local explanations for different effects by identifying their locally relevant causal mechanism. Thus, the comparison between MABLAR, MABLAR-CD and MABLAR-CW highlights the

difference between different ways of using causal mechanisms for generating local explanations. The WM algorithm, as arguably the most classical data-driven rule generation approach, is adopted as the baseline correlation-based rule generation approach, because it provides a consistent basis to compare all approaches evaluated [25]. The WM algorithm is also adopted in the corresponding rule generation step of MABLAR, MABLAR-CD and MABLAR-CW to maintain consistence between different frameworks.

We select 5 real-world data sets, which are widely used as benchmarks for rule generation, from the UCI data repository [26] and the Kaggle website [27]. Table I shows the details of the selected data sets.

TABLE I
DATA SETS USED IN THIS PAPER

Name	Samples	Class	$\|D\|$	$\|D_{MB}\|$	$\|D_{CD}\|$
Breast [26]	699	2	9	6	6
Iris [26]	150	3	4	4	1
Mammographic mass (MAM) [27]	830	2	6	4	4
Pima Indian Diabetes (PID) [27]	768	2	8	8	6
Wine [26]	178	3	13	13	5

Trapezoidal and triangle membership functions are adopted as they facilitate explainability [28]. One can adopt different membership functions for specific problems. Here, each variable is divided into 3 fuzzy partitions for the convenience of facilitating the assignment of linguistic labels. The parameters of membership functions are estimated by the K-means clustering method. More details can be found in [23] which follows the same approach. All variables are normalized to [0,1].

We adopt the average classification accuracy over 5-fold cross-validation as the performance index. Stratified sampling is adopted to keep the original class proportions in the training data set of each fold the same. In order to keep the consistency of causal graphs and fuzzy partitions across different dataset partitions during the cross-validation process, the causal discovery and the determination of fuzzy partitions are implemented before cross-validation. More specifically, we first implement DirectLiNGAM on the original data set to obtain the causal weighted graph. Then we implement K-means on the original data set to determine the fuzzy partitions of each variable. Then we apply MABLAR, MABLAR-CD, WM, and MABLAR-CW using the initially determined causal graph and partitions as applicable.

B. Performance evaluation

Although MABLAR-CW is designed to maximise the local interpretability of fuzzy systems, we expect the obtained fuzzy systems also achieve comparable performance with those obtained by classical algorithms, e.g. the WM algorithm. Thus, we first compare the performance of the fuzzy systems obtained by the different frameworks.

Table II shows the results of each fuzzy system. We note that the fuzzy systems obtained by MABLAR-CW achieve higher or comparable performance with other frameworks in most data sets, however, the performance of MABLAR-CW has a significant decrease in the Iris data set, as well as MABLAR-CD–which warrants further investigation.

TABLE II
PERFORMANCE OF DIFFERENT FRAMEWORKS

	WM	MABLAR	MABLAR-CD	MABLAR-CW
Breast	0.9128	0.9571	0.9571	**0.9686**
Iris	**0.9667**	**0.9667**	0.6667	0.6667
MAM	0.8145	0.8277	0.8277	**0.8386**
PID	0.6850	**0.6889**	0.6276	0.6238
Wine	0.7521	0.7578	0.8711	**0.9160**

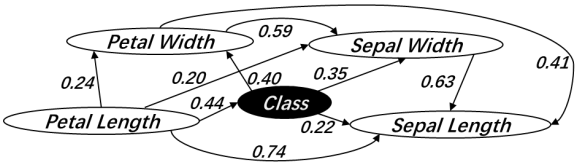


Fig. 6. The weighted causal graph of the Iris data set

Fig. 6 shows the causal graph of the Iris data set obtained by DirectLiNGAM in this paper. According to [29], both 'petal length' and 'petal width' should be used to classify an iris flower. However, as shown in Fig. 6, only the 'petal length' variable is identified by the algorithms as a direct cause variable for the target variable (i.e., the 'Class' node in Fig. 6), which indicates that DirectLiNGAM does not perform well on the iris data set. The sub-optimal causal weighted graph leads to insufficient information for classification in MABLAR-CD and MABLAR-CW, resulting in a decrease in the model performance. This may be a result of the Iris dataset not meeting the conditions for LiNGAM. In real-world applications, causal discovery algorithms suitable for a given data set should be chosen. See [12], [13] for recent overviews, including the selection of causal discovery algorithms. However, the example highlights that the quality of MABLAR frameworks are dependent on the quality of the causal graph.

Overall, we consider the performance satisfactory and sufficient to meaningfully consider the explanations generated by the approach. In the next subsection, focusing on the MAM data set, we compare local explanations provided by different frameworks. We choose the MAM data set because the causal graph complexity of the MAM data set is moderate, which is suitable for visualisation. We are hoping to present more extensive experimental results on more data sets in an upcoming journal paper.

C. Local explanations for the mammographic mass severity prediction

The MAM data set is used to predict the severity (benign or malignant) of a mammographic mass [30]. Fig. 7 shows the causal graph of the MAM data set obtained in this paper.

From the causal graph, 8 complete causal paths to the target variable *Severity* are identified:

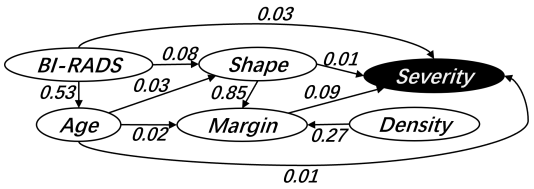


Fig. 7. The weighted causal graph of the MAM data set

1) BI-RADS $\xrightarrow{0.03}$ Severity
2) BI-RADS $\xrightarrow{0.53}$ Age $\xrightarrow{0.01}$ Severity
3) BI-RADS $\xrightarrow{0.53}$ Age $\xrightarrow{0.02}$ Margin $\xrightarrow{0.09}$ Severity
4) BI-RADS $\xrightarrow{0.53}$ Age $\xrightarrow{0.03}$ Shape $\xrightarrow{0.85}$ Margin $\xrightarrow{0.09}$ Severity
5) BI-RADS $\xrightarrow{0.53}$ Age $\xrightarrow{0.03}$ Shape $\xrightarrow{0.01}$ Severity
6) BI-RADS $\xrightarrow{0.08}$ Shape $\xrightarrow{0.85}$ Margin $\xrightarrow{0.09}$ Severity
7) BI-RADS $\xrightarrow{0.08}$ Shape $\xrightarrow{0.01}$ Severity
8) Density $\xrightarrow{0.27}$ Shape $\xrightarrow{0.01}$ Severity

We select 3 samples from the MAM data set to help illustrate key difference between the local explanations provided by different frameworks for each sample. The input samples marked as ①, ②, ③, respectively, as well as the associated output class are shown in Table III. Note that all scalar values have been normalised to $[0, 1]$:

TABLE III
THE TESTED SAMPLES

	BI-RADS	Age	Shape	Margin	Density	Class
①	0.0909	0.5128	1	1	0.6667	Malignant
②	0.0909	0.7436	0	0.75	0.6667	Malignant
③	0.0909	0.6282	1	1	0.6667	Malignant

Table IV-VII show the local rule-based explanations provided by different frameworks for different samples in table form. Note that all rules use the AND logical connective exclusively, implemented as the product t-Norm as mentioned above.

TABLE IV
LOCAL EXPLANATIONS PROVIDED BY THE WM ALGORITHM

	BI-RADS	Age	Shape	Margin	Density	Class
①	Mid	Mid	High	High	High	Malignant
②	Mid	High	Low	Mid	High	Malignant
③	Mid	High	High	High	High	Malignant

TABLE V
LOCAL EXPLANATIONS PROVIDED BY MABLAR

	BI-RADS	Age	Shape	Margin	Class
①	Mid	Mid	High	High	Malignant
②	Mid	High	Low	Mid	Malignant
③	Mid	High	High	High	Malignant

In Table VII, the symbol $\times$ for ③ represents that the Age variable is removed in the antecedents of the local explanation for the effect of ③. From Table IV-VII we can observe the following: Local explanations provided by WM, MABLAR or MABLAR-CD have the same variables in their antecedents, while local explanations provided by MABLAR-CW have different antecedents. This observation supports the conclusion that the local explanations provided by MABLAR-CW can capture different causal mechanisms for the effect of different samples.

TABLE VI
LOCAL EXPLANATIONS PROVIDED BY MABLAR-CD

	BI-RADS	Age	Shape	Margin	Class
①	Mid	Mid	High	High	Malignant
②	Mid	High	Low	Mid	Malignant
③	Mid	High	High	High	Malignant

TABLE VII
LOCAL EXPLANATIONS PROVIDED BY MABLAR-CW

	BI-RADS	Age	Shape	Margin	Class
①	Mid	Mid	High	High	Malignant
②	Mid	High	Low	Mid	Malignant
③	Mid	$\times$	High	High	Malignant

We further highlight the comparison between the local explanations for ① and ③ provided by different frameworks. Note that, as shown in Table III, all values for ③ are the same as for input ①, except the age value, which is greater for ③. Having only one variable with differing values makes the comparison easier. As shown in Table IV-Table VI, for these two samples or input-effect pairs, the local explanations in each case use the same causal mechanism to explain why ① and ③ are malignant. For example, the local explanation provided by MABLAR-CD indicates that 'BI-RADS', 'Age', 'Shape' and 'Margin' are the causes for the 'Malignant' of both ① and ③. However, according to [31], the authors support the idea that a mammographic mass in middle-aged women is more likely to be malignant than in older women. Thus, 'Age is Mid' should be included in the local explanations for the effect of ①. In contrast, old age is not a risk factor for the malignant of a mammographic mass. Thus, it is intuitive for 'Age is High' to be removed as it is an unnecessary component in the local explanations for ③.

Noting this, reviewing Table IV - Table VII reveals that local explanations for ③ provided by WM, MABLAR and MABLAR-CD all include 'Age is high'. In contrast, 'Age is high' is not present from the local explanation provided by MABLAR-CW for ③. This observation supports that MABLAR-CW focuses explanations on the most locally relevant causal path. In addition, the local explanation for ③ indicates that the rule 'If BI-RADS is Mid AND Shape is High AND Margin is High, Then class is Malignant' achieves the highest CI_j. According to the prediction mechanism, this means that the 'Age' variable is not used by MABLAR-CW to make the finial prediction for ③. However, of course, this is only an illustrative example and broader evaluation will be necessary to underpin firmer conclusions. As the evaluation of the quality of explanations is significantly more challenging than the evaluation of system performance, the latter will not be a straightforward or quick process, but will required sustained research efforts in the future.

V. CONCLUSIONS

In this paper, we propose a rule generation framework called MABLAR-CW designed to focus on locally-

relevant causal mechanisms rather than correlations between variables–both for the generation of outputs, and for producing meaningful locally relevant explanations. Given a data set and its corresponding causal graph, MABLAR-CW uses causal direction information from the associated causal graph to generate a fuzzy system constructed as a set of fuzzy sub-systems. The obtained fuzzy system captures possible causal mechanisms of the target variable. Given an input sample, MABLAR-CW uses causal weight information from the causal graph to identify the most causally relevant mechanism and generate an output–and explanation. Initial experiments indicate that MABLAR-CW has the capacity to produce fuzzy systems which can provide locally relevant explanations by identifying the locally relevant causal mechanism for a given sample, while maintaining comparable performance compared to classical rule generation approaches. We have already conducted further experiments which will be published in an upcoming journal paper.

Finally, we note again that MABLAR-CW is dependent on the quality of the causal graph and thus generally, the quality and suitability of the causal discovery algorithm used. In addition, When the causal graph of the given data set shows a large number of possible causal mechanisms, MABLAR-CW has the risk of generating a fuzzy system which contain a large number of rules. Thus, MABLAR-CW is suggested to be used when the task is focused on the local explanations of a specific output. In future, we will explore predictive approaches which combines a set of best-fitting rules and causal mechanisms reflecting that in the real world frequently multiple causal mechanisms come together to underpin an effect. We also note that there is a potential to generate rules interactively in sequence to provide customised local explanations of the locally relevant causal mechanism, another topic within an exciting area for future study in fuzzy and rule based systems generally.

REFERENCES

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

[2] X. Zhu, D. Wang, W. Pedrycz, and Z. Li, "Fuzzy rule-based local surrogate models for black-box model explanation," *IEEE Transactions on Fuzzy Systems*, vol. 31, no. 6, pp. 2056–2064, 2023.

[3] 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.

[4] J. Hühn and E. Hüllermeier, "FURIA: an algorithm for unordered fuzzy rule induction," *Data Mining and Knowledge Discovery*, vol. 19, pp. 293–319, 2009.

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

[6] L. Kunitomo-Jacquin, A. Lomet, and J.-P. Poli, "Causal discovery for fuzzy rule learning," in *2022 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 2022, pp. 1–8.

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

[8] E. C. Garrido-Merchán, C. Puente, A. Sobrino, and J. A. Olivas, "Uncertainty Weighted Causal Graphs," *arXiv*, 2020, arXiv:2002.00429 [cs]. [Online]. Available: http://arxiv.org/abs/2002.00429

[9] 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 . . . , 2017, pp. 297–301.

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

[11] 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. 10, 2006.

[12] B. Schölkopf, "Causality for machine learning," in *Probabilistic and Causal Inference: The Works of Judea Pearl*, 2022, pp. 765–804.

[13] L. Yao, Z. Chu, S. Li, Y. Li, J. Gao, and A. Zhang, "A survey on causal inference," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 15, no. 5, pp. 1–46, 2021.

[14] S. Shimizu, T. Inazumi, Y. Sogawa, A. Hyvarinen, Y. Kawahara, T. Washio, P. O. Hoyer, K. Bollen, and P. Hoyer, "Directlingam: A direct method for learning a linear non-gaussian structural equation model," *Journal of Machine Learning Research-JMLR*, vol. 12, no. Apr, pp. 1225–1248, 2011.

[15] 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.

[16] B. Kosko, "Fuzzy cognitive maps," *International journal of man-machine studies*, vol. 24, no. 1, pp. 65–75, 1986.

[17] A. Sharma, A. Tselykh, E. Podoplelova, and A. Tselykh, "Knowledge-oriented methodologies for causal inference relations using fuzzy cognitive maps: A systematic review," *Computers & Industrial Engineering*, p. 108500, 2022.

[18] G. Felix, G. Nápoles, R. Falcon, W. Froelich, K. Vanhoof, and R. Bello, "A review on methods and software for fuzzy cognitive maps," *Artificial intelligence review*, vol. 52, pp. 1707–1737, 2019.

[19] W. Zhang, X. Zhang, and D. Chen, "Causal neural fuzzy inference modeling of missing data in implicit recommendation system," *Knowledge-Based Systems*, vol. 222, p. 106678, 2021.

[20] K. Zhao and B. Upadhyaya, "Adaptive fuzzy inference causal graph approach to fault detection and isolation of field devices in nuclear power plants," *Progress in Nuclear Energy*, vol. 46, no. 3-4, pp. 226–240, 2005.

[21] L. A. M. Rosales, S. E. P. Hernandez, and G. R. Gomez, "Coordination for synchronous cooperative systems based on fuzzy causal relations," *International Journal of Computer Science*, vol. 3, p. 4, 2008.

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

[23] T. Zhang, J. Ying, C. Wagner, and J. Garibaldi, "Towards causal fuzzy system rules using causal direction," in *2023 IEEE International Conference on Fuzzy Systems (FUZZ)*, 2023, pp. 1–6.

[24] D. Alvarez-Estevez and V. Moret-Bonillo, "Revisiting the w ang–m endel algorithm for fuzzy classification," *Expert Systems*, vol. 35, no. 4, p. e12268, 2018.

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

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

[28] 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.

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

[30] "Mammographic mass data set," https://www.kaggle.com/datasets/overratedgman/mammographic-mass-data-set, 2016, accessed: 2023-11-25.

[31] K. A. Bertrand, R. M. Tamimi, C. G. Scott, M. R. Jensen, V. S. Pankratz, D. Visscher, A. Norman, F. Couch, J. Shepherd, B. Fan *et al.*, "Mammographic density and risk of breast cancer by age and tumor characteristics," *Breast Cancer Research*, vol. 15, no. 6, pp. 1–13, 2013.