Explain the world – Using causality to facilitate better rules for fuzzy systems

Te Zhang, *Student Member, IEEE,* Christian Wagner, *Senior Member, IEEE,*
Jonathan M. Garibaldi, *Fellow, IEEE,*

***Abstract*—The rules of a rule-based system provide explanations for its behaviour by revealing the relationships between the variables captured. However, ideally, we have AI systems which go beyond explainable AI (XAI), that is, systems which not only explain their behaviour, but also communicate their 'insights' in respect to the real world. This requires rules to capture causal relationships between variables. In this paper, we argue that those systems where the rules reflect causal relationships between variables represent an important class of fuzzy rule-based systems with unique benefits. Specifically, such systems benefit from improved performance and robustness; facilitate global explainability and thus cater to a core ambition for AI: the ability to communicate important relationships amongst a system's real-world variables to the human users of AI. We establish two causal-rule focused approaches to designing fuzzy systems, and show the distinctions in their respective application scenarios for the explanations of the rules obtained by these two methods. The results show that rules which reflect causal relationships are more suitable for XAI than rules which 'only' reflect correlations, while also confirming that they offer robustness to over-fitting, in turn supporting strong performance.**

***Index Terms*—Rule-based systems, causality, XAI, causal explanations.**

I. Introduction

As attention to eXplainable AI (XAI) increases, fuzzy systems have attracted increasing interest due to their potential for high explainability and high prediction accuracy, as derived from their linguistic rules [1]. In real-world applications, users may not only seek to understand the behaviour of an AI model, but may also want to leverage the rules of AI systems to gain insights into the real-world processes. The rules of an AI system usually tell users what the AI system is doing. However, they do not necessarily tell us what is happening in the real-world processes modelled by the AI system. Nowadays, most AI systems are adopting data-driven approaches to generate rules. To achieve good performance, these approaches exploit statistical correlations between the variables within the data set—rather than necessarily capturing causal relationships.

However, humans are cause–effect thinkers. We expect *rules* to reflect causation: causation is at the heart of what a rule is. Similarly, when considering real-world processes, we generally seek to determine the causal relationships underlying those processes so as to understand them, be they disease or the motions of stars and planets. Thus, *rules* which reflect correlations between variables can mislead users, and cannot in general provide meaningful explanations for real-world processes. Indeed, they risk undermining the capacity for rule-based AI systems to be explainable.

The concept of causality is a complex and ongoing topic for discussion across disciplines [2]–[4], with no universally accepted definition. For semantic purposes in this paper, we adopt an informal working definition of causality as *a relation where one event or state (a cause) contributes, and is at least partially responsible for the realisation of another event or state (an effect)*—adapted from [5], [6].

In the context of AI, Galles and Pearl defined the concept of a *causal model* which describes the world by random variables and uses structural equations to represent causal mechanisms in the world [7]. Based on the concept of a causal model, Halpern and Pearl proposed the 'Halpern-Pearl definition of causality' [8], designed to eliminate imprecision or conceptual fuzziness that may arise when describing causality using natural language semantics by instead providing a precise mathematical description.

In this paper, we propose a framework which uses data-driven algorithms to identify the causal relationships between variables. Thus, the framework and its outputs inherit the definition of causality underpinning the given causal discovery algorithms used. We summarise the definitions and assumptions of the specific causal discovery algorithms adopted in the experimental part of the paper in Section II-B.

When computationally identifying causal relationships, it is generally with the aspiration that these relationships do not only hold within the given data set but also apply in the real world. In other words, causal discovery aims to identify true causal relationships–but there is no guarantee that the relationships found for a given data set are indeed true causal relationships, or even that the set of causal relationships found is complete. Indeed in practice, such as due to technical or ethical reasons, data sets may exhibit bias or contain incomplete variables–or miss key variables altogether, which can contribute to causal relationships being missed, and

T. Zhang, C. Wagner and J. Garibaldi are with the Laboratory for Uncertainty in Data and Decision Making (LUCID) and the Intelligent Modelling and Analysis Research Group, School of Computer Science, the University of Nottingham, Nottingham, NG8 1BB, U.K. e-mail: (te.zhang, christian.wagner, jon.garibaldi)@nottingham.ac.uk

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

causal relationships discovered not corresponding to actual causal relationships in the real world.

This paper builds on the assumption that where data sets exist, it is at least in principle possible and worthwhile to design AI systems which seek to identify and model relationships between the data sets' variables which are of a causal nature. Where such identification of causal relationships is successful, and where these AI systems are explainable, the explanations of their underlying models will map to explanations of the real-world processes which they capture [9]. Through explanations of such systems, users can gain insights into the real-world processes. For example, Ziatdinov et al. report a case where a causality-based machine learning algorithm uncovers a real-world process in the domain of ferroelectric materials from the data which had not been discovered by domain experts before [10].

In prior work, we established two frameworks designed to automatically generate rules reflecting causal relationships between variables of a given data set. Both approaches leverage the causal structure within data as modelled by causal graphs—further discussed in Section III. We initially focused on an approach which targets variable selection based on the concept of the Markov blanket, which we refer to as the Markov blanket rule generation framework (MABLAR) [11]. We further developed on this, with an approach which targets variable selection based on its *direct* cause(s), which we refer to as the Markov blanket rule generation-causal direction framework (MABLAR-CD) [12].

In this paper, we go beyond both of these initial conference papers. We provide a detailed comparison of the two frameworks, establish, and show the scenarios in which each framework is suitable. In doing so, the main contributions of this paper are as follows:

1) We show how the rules generated by frameworks such as MABLAR and MABLAR-CD capture different aspects of causal relationships between variables and discuss the respective benefits this brings to the obtained fuzzy system, including reduced complexity and improved performance.
2) We explain why MABLAR captures local causal relationships of the target variable, while MABLAR-CD captures its direct causal relationships, and discuss how the adoption of either is dependent on the objectives sought.
3) We present in-depth experiments on a series of synthetic and real-world data sets, comparing the explainability and performance of rules obtained from classical data-driven rule generation approaches and those obtained from MABLAR and MABLAR-CD.
4) We show how an imperfect Markov blanket can have a negative impact on the performance of the obtained 'causal' fuzzy systems, but how even this can still be helpful for generating rules which capture causal relationships between variables.

The rest of this paper is organised as follows: Section II provides the background. Section III introduces MABLAR and MABLAR-CD and highlights the difference between their implementations. Section III explains why their rules capture different types of causal relationships between the variables. Section IV presents the experiment results. Section V and Section VI present a discussion and the conclusions, respectively.

II. Background

A. Fuzzy system explainability

Fuzzy systems are said to be interpretable because their linguistic rules provide model-based explanations to users. The explanations are called 'model-based' because they explain the operation of the (fuzzy) system, which in turn *may* directly mimic a process in the real world. Mamdani rules [13] are suitable for applications where the emphasis lies on model explainability, because both antecedents and consequents of their rules are composed of fuzzy sets which can be represented by linguistic terms, thus providing model-based explanations to users in a human understandable way [14]–[16].

The interpretation of the 'THEN' part of fuzzy system rules is directly related to the semantic interpretation of fuzzy systems. According to the choice of implication and aggregation functions, it can be interpreted in a conjunctive or an implicative way, corresponding respectively to conjunctive rules and implicative rules [17], [18]. The semantic meaning of implicative rules can be viewed as more precise compared to conjunctive rules [17], making them particularly suitable to represent causal relationships between variables [3]. However, most works focusing on the generation of fuzzy rules from data, such as [19], [20], focus on the conjunctive interpretation of rules.

Apart from the semantic meaning of rules, there is already a substantial body of literature focusing on the explainability of fuzzy systems specifically in terms of the complexity of their rule bases – irrespective of the causal nature of the rules [21]. In general, less complex fuzzy systems are considered to have a higher degree of explainability. However, the complexity of a fuzzy system is affected by many factors. For example, at the individual rule level, a rule with a lower number of input variables is less complex, i.e. a rule with a single type of logical connective in their antecedents, e.g.. only using 'AND' or only using 'OR', has less complexity than one with both types. At the rule base level, a rule base with fewer rules is less complex than one with many. Other factors affect the complexity of a fuzzy system, including the structure of the rules, the number of fuzzy partitions of each variable, the type of membership functions, etc. See [14], [15], [22], [23] for a comprehensive discussion.

To improve the explainability of a fuzzy system, many data-driven rule generation approaches, designed to reduce the complexity of the obtained fuzzy systems, have been proposed. For example, the ALM algorithm [24], the SR algorithm (SR) [25] and the FARC-HD algorithm [26] are designed to remove redundant rules from an initial rule base. The CHI algorithm [27] and the HILK++ framework [28] are designed to generate compact rule bases directly from the data or from experts. The EasIeR algorithm [29], the Higgin & Goodman algorithm [30] and the fuzzy unordered rule induction algorithm (FURIA) [20] are designed to generate a compact rule base by iteratively updating using heuristic strategies. Here we also note that FURIA determines membership functions at the specific level of each rule, resulting in rules generated by FURIA not offering the same level of semantic interpretability *across* rules as compared to approaches which establish fixed membership functions for linguistic terms a priory - such as the Wang-Mendel (WM) algorithm.

None of these however are designed specifically to reduce the complexity of the rule base by finding the causal subset of the rules. We proceed to discuss causal discovery as an initial stepping stone to such approaches.

B. *Causal discovery based on observational data*

To describe causal relationships between variables, Pearl proposed the causal graph model [31]. Causal graphs are directed acyclic graphs (DAGs) where 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, and the causal relationships between a variable and its parent variables are direct causal relationships. Fig. 1 presents an example of a causal graph.

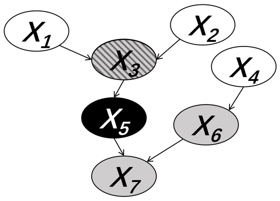


Fig. 1. An example of a causal graph

Randomised controlled trials are the gold standard for distinguishing between causal relationships and correlations [2], [3], [32]. However, in most cases, we can only obtain causal relationships between variables through existing observational data rather than active intervention, because of the limitations of realistically possible experimental techniques. Addressing the challenge of achieving this, many data-driven causal discovery algorithms have been proposed–they can be classified into the following three categories based on their purpose:

1) Causal structure learning algorithms: Given a set of variables $\{X_1, X_2, \cdots, X_m\}$, the goal of causal structure learning is to determine the structure of the causal graph for this set of variables. Classical causal structure learning algorithms include the PC algorithm [32], MMHC [33], MMPC [34], and RESIT [35].
2) Causal direction inference algorithms: Suppose there is a causal relationship between two variables X_1 and X_2. The goal of causal direction inference is to determine the direction of the edge between X_1 and X_2 in a causal graph, i.e. to distinguish between $X_1 \leftarrow X_2$ and $X_1 \rightarrow X_2$. Classical direction inference learning algorithms include ANM [36], IC algorithm [37], ICA-LiNGAM [38], and DirectLiNGAM [39].
3) Confounding variable detection algorithms: Given a set of variables $\{X_1, X_2, \cdots, X_m\}$, the goal of confounding variable detection is to determine whether there is an unobserved variable $v_c \notin \{X_1, X_2, \cdots, X_m\}$ which has causal relationships with at least one variable in $\{X_1, X_2, \cdots, X_m\}$. Classical confounding variable detection algorithms include FCI [32], and RFCI [40].

Apart from the above causal discovery algorithms, recent efforts have also targeted the use of fuzzy sets for fuzzy causal discovery [41]. Further, the concept of causal structure representation has been explored in areas including fuzzy cognitive maps [42]–[44], causal fuzzy neural networks [45], and leveraged in applications such as [46], [47].

These algorithms provide a variety of ways to discover causal relationships between variables within data sets in different contexts and are each subject to a number of assumptions and specific interpretation or definition of causality.

For example, the PC algorithm proposed by Spirtes and Glymour assumes that the value of a variable is independent of all other variables conditional on all of its direct causes, and these causal dependencies generate statistical dependencies [32]. This assumption is also known as causal Markov condition (CMC) [2], [3], [32]. Pairs of variables that satisfy these conditions are viewed as causal structures in the PC algorithm. Based on this assumption, the PC algorithm determines causal relationships between variables using the statistical dependencies obtained from the given data set.

The PC algorithm is generally considered a reliable method because it starts with a complete undirected graph, thus searching the entire search space [48]. However, it cannot determine the causal directions between variables in a Markov Equivalence class[1] [50]. To solve this problem, researchers have proposed causal function-based methods, such as the linear Non-Gaussian acyclic model (LiNGAM) [38] and additive noise models (ANM) [36]. These methods build on the definition of causality used in the PC algorithm, and further assume that an effect can be represented by a function of its direct causes and some noise terms [51]. For example, ANM assumes that the effect can be represented as a function of its direct causes, plus additive noise that is independent of the direct causes [36]. Causal function-based methods are effective at determining the causal relationships between two variables. However, these methods are usually

[1] A Markov Equivalence class is a set of causal graphs that encode the same set of conditional independencies [49].

more time consuming than the PC algorithm. See [48], [52], [53] for some recent overviews.

C. Markov blanket of a target variable

The concept of the Markov blanket (MB) was originally established by Pearl in 1988 [54]—the MB of a variable T is the minimal set of variables given which all other variables are independent of T [54], [55].

Throughout this paper, we leverage this concept of the MB as used within *causal discovery* as discussed in Section II-B. As discussed above and explained in [3], the causal graph of the data set articulates the dependencies between its variables. Therefore, the Markov blanket of a variable in a given data set can be obtained from the causal graph of the given data set [55]–[57]. Within this context, Aliferis et al. defined the MB of T in a causal graph as follows [56]: the MB of T in a causal graph is the set of parents, children, and spouses of T. For example, the grey nodes in Fig. 1, make up the MB of the variable X_5, consisting of: X_3 (a parent of X_5), X_6 (a spouse of X_5) and X_7 (a child of X_5).

The variables within the MB provide the local causal structure of a target variable [56]. Thus, the MB can be used to determine whether a variable is redundant in a causal sense, or indeed, whether it is causally linked to the target variable [55].

III. Towards causal explanations using causal graphs

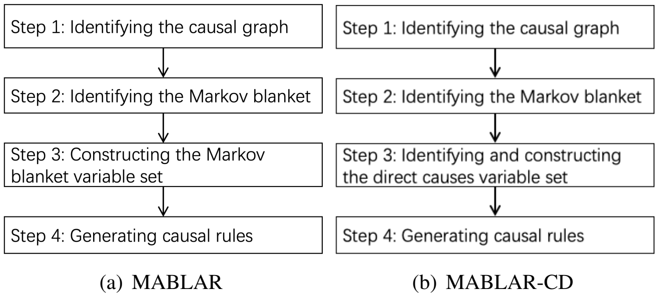
Following that brief general overview of causal learning, we now cover MABLAR and MABLAR-CD, initially introduced in short form in [11] and [12]. We introduce both approaches and show how the rules obtained by MABLAR capture local causal relationships, while the rules obtained by MABLAR-CD capture direct causal relationships. For facilitating the description, Table I presents the notation used in the rest of this paper.

TABLE I
Notation

D	The original data set
D_{MB}	The subset of D which only contains the MB of the target variable
D_{CD}	The subset of D_{MB} which only contains the direct causes within the MB of the target variable
$\lvert D\rvert$	The number of variables in the data set D

A. The processes of MABLAR and MABLAR-CD

MABLAR and MABLAR-CD are two data-driven rule generation frameworks initially formulated in [11] and [12] for generating rules which capture the causal relationships between variables shown in the causal graph of the given data set. Fig. 2(a) and Fig. 2(b) present the processes of MABLAR and MABLAR-CD, respectively. The algorithmic processes of MABLAR and MABLAR-CD are shown in the supplementary materials.


Step 1: Identifying the causal graph
Step 2: Identifying the Markov blanket
Step 3: Constructing the Markov blanket variable set
Step 4: Generating causal rules
Step 1: Identifying the causal graph
Step 2: Identifying the Markov blanket
Step 3: Identifying and constructing the direct causes variable set
Step 4: Generating causal rules

(a) MABLAR (b) MABLAR-CD

Fig. 2. MABLAR and MABLAR-CD

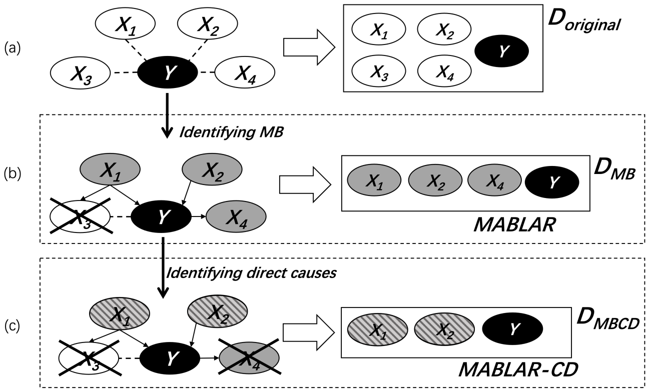


Fig. 3. The rationale of MABLAR-based frameworks

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 Markov blanket of the target variable in their first two steps. However, in contrast to MABLAR, which constructs D_{MB} in Step 3, MABLAR-CD further identifies the direct cause(s) of the target variable and constructs D_{CD} in Step 3. Finally, in Step 4, using data-driven algorithms (e.g. the WM algorithm), MABLAR automatically generates rules from D_{MB}, while MABLAR-CD generates rules from D_{CD}

In the next subsection, we introduce the rationale of MABLAR and MABLAR-CD.

B. Rationale of MABLAR and MABLAR-CD

MABLAR and MABLAR-CD are designed to focus on causal rule generation by removing variables which are not causally related to the target variable.

Fig. 3 presents the rationale of MABLAR and MABLAR-CD. In Fig. 3, dashed lines represent correlations between variables and solid arrows between variables represent causal relationships. As shown in Fig. 3 (a), all the variables within D are correlated with the output variable, i.e., Y. However, In real-world applications, D often contains variables which exhibit strong correlation with the target variable but are not actually causally linked with it.

The relationships between such variables and the output (i.e., target) variable are routinely captured as rules by classical data-driven rule generation algorithms (see Section II-A). Therefore, if a classical data-driven algorithm is directly applied to D, the obtained rules will conflate correlation and causal relationships between the input and output variables. We refer to rules which are derived purely based on correlation as *spurious rules*.

MABLAR is designed to ignore relationships based solely on correlation, leveraging the actual or approximate[2] causal graph of the data set. Fig. 3 (b) and (c) present an example of a causal graph of the original data set. As shown in Fig. 3 (b) and (c), X_3 is removed by both MABLAR and MABLAR-CD and all remaining variables are causally related with the output variable. We highlight that although X_4 is causally related with the output variable, it is still removed by MABLAR-CD, because MABLAR-CD is designed to generate rules which capture the*direct causes* of the target variable.

To identify the variables which are causally related with the target variable, both MABLAR and MABLAR-CD need to find the causal graph of D, i.e., implement causal discovery on D, in their Step 1 as shown in Fig. 2. In this paper, this is achieved by using data-driven causal discovery algorithms, specifically the PC algorithm.

We highlight that most data-driven causal discovery algorithms perform causal discovery based on strong assumptions of the given data set [48]. For example, the PC algorithm assumes that causal relationships between variables should adhere to certain constraints of conditional independence among variables and performs causal discovery on the based on the results of tests for conditional independence among variables [32].

However, in real-world applications, the data set may not necessarily satisfy the assumptions of the adopted algorithm, which may result a sub-optimal causal graph. In this paper, this means that the fuzzy systems obtained by MABLAR or MABLAR-CD have the risk of generating rules which approximate—but may not reflect perfectly—causal relationships between variables, which may reduce model performance. The design and selection of causal discovery algorithms is in of itself an active research area and out of range of this paper. We refer the reader to [52], [53] for recent overviews, including on the selection of causal discovery algorithms.

In the next subsection, we discuss the benefits of the rules obtained by MABLAR-based frameworks for fuzzy systems.

C. Benefits of rules obtained through MABLAR frameworks

Rules obtained through MABLAR or similar frameworks can capture (or at least approximate) the causal relationships between the variables, which benefits the obtained fuzzy systems in the following three ways [11], [12]:

1) *Improving the robustness of fuzzy systems*: Models which focus on exploring the correlations between variables run the risk of degrading the performance (e.g. classification accuracy) when the distribution of the test data set differs from the distribution of the training data set. However, as discussed in [58], models based on causal relationships are more robust than models based on correlations when facing the situation that the distributions of the training and testing data are different. MABLAR mainly focus on exploiting causal relationships between the variables. Thus, fuzzy systems derived from MABLAR are more robust in distribution shift scenarios.
2) *Reducing the complexity of fuzzy systems*: The complexity of a fuzzy system based on rules obtained by MABLAR or similar frameworks is usually less than or equal to that of those obtained by classical data-driven rule generation algorithms, as they contain fewer rules, because spurious rules are largely absent from the resulting rule bases. Also, in general, $|D_{CD}| \leq |D_{MB}| \leq |D_{original}|$. Thus, the rules obtained by MABLAR-based frameworks often contain fewer variables in their antecedents compared to rules obtained using classical data-driven rule generation algorithms, which also reduces the complexity of the fuzzy system constructed.
3) *Providing causal explanations*: Fuzzy systems constructed by rules obtained by classical data-driven rule generation algorithms run the risk of containing spurious rules which can mislead users. In contrast, rules obtained by MABLAR-based frameworks capture (or at least approximate) the causal relationships between the input and output variables, minimising the risk of misleading users and increasing users' insights into the real-world.

D. Causal relationships captured by MABLAR and MABLAR-CD

As noted in the previous section, the causal relationships captured by rules obtained by MABLAR and MABLAR-CD are different, because MABLAR and MABLAR-CD use different subsets of the variables for the rule generation. MABLAR uses D_{MB} for the rule generation. The variables within D_{MB} provide the local causal structure of the target variable, with D_{MB} containing both the direct cause variables and the direct effect variables of the output variable. In contrast, MABLAR-CD is designed to use D_{CD} for rule generation, focusing only on the direct causes of the output variable. Thus, the rules obtained by MABLAR

[2] In the sense of identified from data but not necessarily being the true real-world causal structure

capture local causal relationships between the target variable and all variables within its MB, while the rules obtained by MABLAR-CD capture the relationships between the target variable and its direct cause(s).

In real-world applications, whether one selects MABLAR or MABLAR-CD should depend on the problem. The rules obtained by MABLAR map the variables within the input variables which have causal relationships with the output variable. Moreover, the fuzzy systems which consist of rules obtained by MABLAR are likely to exhibit better performance than those consisting of rules obtained by MABLAR-CD for most data sets, because MABLAR-CD uses less information from the data set. This reduced utilisation of the data may lead to reduced performance, especially when the overall information content of the data set is crucial for achieving optimal results. In other words, while MABLAR-CD is designed to focus on the direct causes only, completing, for example, a classification task within this setting can be much more challenging than when also considering variables representing causally linked direct effects. For example, it is often harder to diagnose the presence of a disease based on the presence of a given virus and the degree of infection of the patient than it is by considering the symptoms the patient is exhibiting. Conversely, establishing what the disease is exactly may not substantially benefit from considering the symptoms.

While potentially offering better performance, the rules obtained by MABLAR run the risk of containing antecedents which conflate the direct cause and the direct effect variables of the output variable. Therefore, when users are seeking to understand the causal relationships between the output variable and its direct cause variables within a given data set, they should select MABLAR-CD.

In the next section, we explore this brief discussion experimentally, leveraging first a synthetic data set to enable a clear illustration of the underlying mechanisms, followed by several real-world data sets to evaluate the MABLAR-based frameworks in terms of robustness and explainability.

IV. Experiments

In this section, we present the results of experiments on both synthetic and real-world datasets, comparing the explainability and the performance of rules obtained from classical data-driven rule generation approaches with those obtained from MABLAR-based frameworks.

A. Experimental settings

Two established data-driven rule generation approaches are used as benchmarks, and compared with the rule generation approaches of both MABLAR frameworks discussed in Section III. The first approach is the WM algorithm, which, as arguably the most traditional data-driven approach, provides a consistent basis for comparison [12], [59]. The other one is FURIA, representing a popular, high-performance technique, which already by design narrows the original variable sets to improve the interpretability of the fuzzy system—albeit without focusing on causal aspects, as is done in this paper. Specifically, unlike the MABLAR-based frameworks which use causal information, FURIA narrows the original variable sets by focusing on the correlations between the variables. Thus, the comparison between MABLAR, MABLAR-CD and FURIA highlights some of the differences between rules derived when focusing specifically on causal relationships vs rules derived from correlations. In addition, we also provide the performance of the random forest model [60] in the supplementary materials as a baseline of classical black box classifiers for reference. While doing so provides some indication of the performance of black-box models, we note that of course other black-box models can have stronger performance. Equally, causal rule-based systems with a larger number of fuzzy set partitions will deliver higher performance—while generally negatively impacting explainability. In other words, the measurement of performance is useful in context, but not in absolute terms.

We purposely adopt a traditional fuzzy set design, using trapezoidal and triangular MFs for the WM algorithm. The parameters of the MFs of the WM algorithm are estimated using the Fuzzy C-means algorithm (FCM) [61]. The ideal number of fuzzy partitions of each variable is generally problem dependent. However, in this paper, we use the same number of fuzzy partitions (i.e., three fuzzy partitions) for all variables to facilitate comparisons.

We adopt the python-weka-wrapper3 package, which is available in [62], to implement FURIA. All parameters are set to their default settings. We adopt trapezoidal MFs for FURIA in-line with the algorithm. In FURIA, the parameters of the MFs and the fuzzy partitions of each variables are automatically determined by FURIA itself.

In terms of causality, we adopt the PC algorithm [32] for all data sets for learning the causal structure, and the ANM algorithm [36] to infer causal directions which cannot be determined by the PC algorithm. The two adopted algorithms are implemented using the causal learning Python library, which is published in [63] and is available in [64]. In practice, causal discovery algorithms suitable for the intended application should be chosen. However, in this paper, we focus on the above as the universal setup to facilitate experiments and comparisons.

We provide more details about the experiment setting including the parameters setting in the supplementary materials.

B. Evaluation using synthetic data

We first generated a synthetic data set to simulate a scenario where the ground truth causal relationships between different variables are known. The data set was generated by the method shown in [65]. The details of this generation are as follows:

$$
\begin{aligned}
&X_1 = sin(X_4^2) + X_2^2 + cos(X_7) + E_1, && E_1 \sim U(-0.1, 0.1)\\
&X_2 = X_3 + E_2, && E_2 \sim U(-0.5, 0.5)\\
&X_3 = E_3, && E_3 \sim U(-1.0, 1.0)\\
&X_4 = sin(X_2) + sin(2X_3) + E_4, && E_4 \sim U(-0.5, 0.5)\\
&X_5 = tanh(X_6 + X_7 + X_2) + E_5, && E_5 \sim U(-0.3, 0.2)\\
&X_6 = sin(X_2) + cos(2X_4) + E_6, && E_6 \sim U(-0.5, 0.5)\\
&X_7 = cos(X_6 + X_3) + E_7, && E_7 \sim U(-0.3, 0.3)\\
&Y = \mathbb{I}(X_3 + X_1^2 - 1.5 + \epsilon > 0), && \epsilon \sim N(0, 1)
\end{aligned}
$$

The target variable is Y. The generated data set contains 5000 samples and is available at HERE (and will be published with this paper). Table II shows the D_{MB} and D_{MBCD} obtained by the causal discovery algorithms adopted in this paper and the ground truth, respectively. The causal graph obtained by causal discovery algorithms and the ground truth causal graph are shown in the supplementary materials. We conducted three experiments on the synthetic data set to evaluate the performance, explainability, and robustness of the fuzzy systems obtained by the different approaches. As the variables in the synthetic data set do not have a physical meaning, here, we only evaluate the model-based explanations of the fuzzy systems obtained by the different approaches. The potential real-world causal explanations provided by the fuzzy systems will be evaluated in the next subsection using a real-world data set where each variable has a physical meaning.

TABLE II
D_{MB} AND D_{MBCD} OF THE TARGET VARIABLE ON THE SYNTHETIC DATA SET

	PC-ANM	Ground truth
D_{MB}	$\{X1,X2,X3,X4,X6,X7\}$	$\{X1, X3\}$
D_{MBCD}	$\{X1, X3\}$	$\{X1, X3\}$

1) Experiment 1: Evaluation on performance and model-based explanations: We first evaluate the performance and the model-based explanations of the fuzzy systems obtained by these different approaches. As the performance index, we adopt the average of the classification accuracy over 5-fold cross-validation. Both indices are better when higher.

Stratified sampling is adopted to keep the original class proportions in the training data set of each fold the same.

We adopt the number of rules as the evaluation index of the model-based explanations, because in the setting of this synthetic experiment, the only factor that differs between the fuzzy systems obtained by the different approaches and has an impact on the model-based explanations of the system is the number of rules.

Table III shows the results of each fuzzy system during the 5-fold cross-validation. '# Rules' in Table III represents 'number of rules'. An algorithm with a prefix of 'MB' (resp., 'MBCD') indicates that the fuzzy system is obtained using the MABLAR (resp., MABLAR-CD) framework.

When interpreting the results, we caution that these are illustrative only, as they are based on one synthetic model and its associated data set. From Table III, we can observe the following: MB-WM, MBCD-WM have a higher performance and smaller number of rules than WM. Similarly, MB-FURIA and MBCD-FURIA have a higher performance and smaller number of rules than FURIA. This observation indicates that MABLAR-based frameworks can improve the model-based explanations of the obtained fuzzy systems while improving the performance.

2) Experiment 2: Increasing robustness in distribution shift scenarios: To evaluate the robustness of the fuzzy systems obtained by the different approaches when the distributions of the training and testing data are different, we first use the whole data set generated in Experiment 1 as the training set. Then, we generate two separate testing sets to simulate scenarios where the distribution of the testing data is the same or different from the training data. The first testing set contains 5000 samples, which are generated the same way as the training set, but with a different Python random seed. The second testing set also contains 5000 samples, but the samples are generated by changing 'sin' to 'cos', 'cos' to 'sin', and '$tanh$' to 'tan' in the functions. Finally, for each rule generation approach, we trained a fuzzy system using the training set, but test the obtained fuzzy system on both testing sets. The results are shown in Fig. 4.

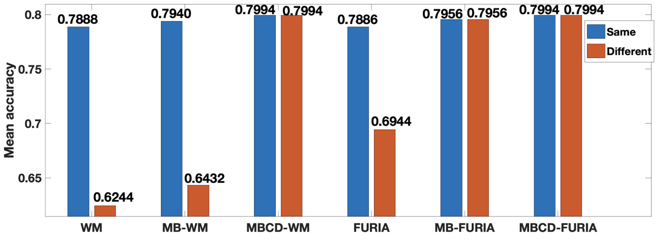


Fig. 4. Experimental results with Experimental Setting 2

In Fig. 4, the blue bars represent the testing results obtained on the testing set generated in the same way as the training set, while the red bars represent the testing results obtained on the testing set generated using the modified generation model.

From Fig. 4, we can make the following observations: 1) The performance of the fuzzy systems obtained by WM and FURIA is significantly worse when the distribution of the testing set is different from that of the training set. In contrast, the performance of the fuzzy systems obtained by MBCD-WM and MBCD-FURIA does not worsen. 2) The performance of the fuzzy systems obtained by MB-WM also worsens significantly when the distribution of the testing set is different from that of the training set. However, compared to the results of WM, MB-WM shows less performance degradation.

TABLE III
PERFORMANCE AND # RULES OF DIFFERENT FUZZY SYSTEMS

	WM	MB-WM	MBCD-WM	FURIA	MB-FURIA	MBCD-FURIA
Accuracy(Std)	0.7972(0.0180)	0.7998(0.0187)	**0.8140(0.0085)**	0.7998(0.0011)	0.8015(0.0062)	0.8057(0.0063)
# Rules	323.8(3.3)	188.8(1.1)	**9(0)**	7.6(1.4)	5.4(2.0)	**4.6(1.1)**

The above observations indicate that fuzzy systems whose rules only capture the causal relationships between the variables are more robust than those which also capture correlations between the variables when facing a distribution shift between the training set and the testing set. Another way of interpreting this is that the focus on the causal relationships makes the systems robust to overfitting.

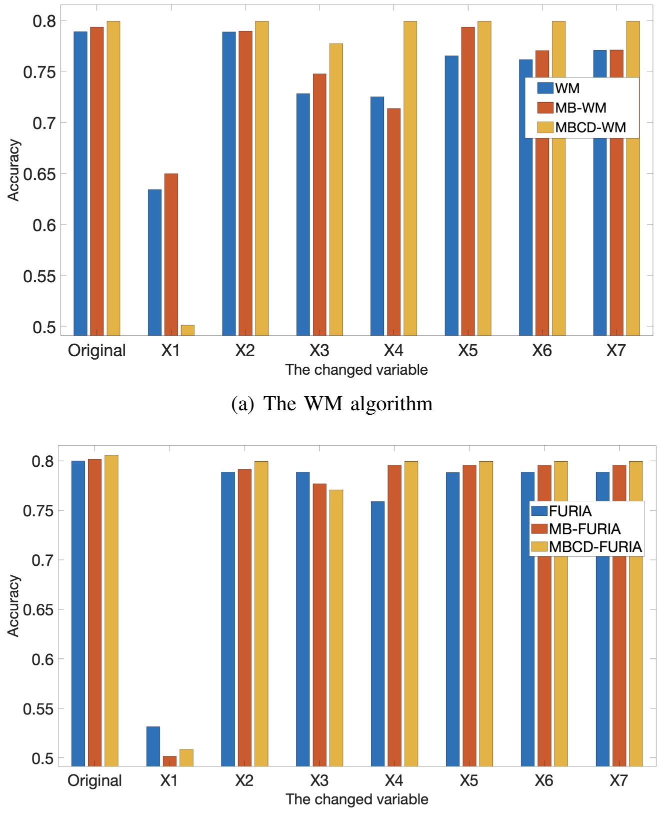
3) Experiment 3: Evaluating the importance of the causal variables for the performance of the model: To evaluate the importance of the variables which have causal relationships with the target variable, we first trained a fuzzy system using the whole data set obtained in Experiment 1. Then, we used the testing set which has the same distribution as the training set in Experiment 2. Finally, we replaced the variables $X1$–$X7$ one by one in the testing set in turn with random noise (generated from a standard normal distribution, i.e. $\mu(0,1)$) and tested the fuzzy system on each of these modified testing sets. In Fig. 5, $X1$–$X7$ on the x-axis represent the variable which is replaced by the random noise. The 'original' on the x-axis represents the performance of the model on the unmodified test set (i.e. no variable is replaced by random noise).

From Fig. 5 we can observe the following: 1) The performance of all obtained fuzzy systems decreases the most when '$X1$' is replaced, and the performance of most obtained fuzzy systems decreases the most when $X3$ is replaced. 2) The performance of MB-WM and MB-FURIA show the most significant performance decrease when $X1$ and $X4$ are replaced, while those of MBCD-WM and MBCD-FURIA decrease the most when $X1$ and $X3$ are replaced with noise. This indicates that the variables which have causal relationships with the target variable have the most significant impact on the model's performance.

C. Evaluation on real-world data sets

We proceed to evaluate the fuzzy systems obtained by the different approaches on 17 real-world data sets from the UCI data repository [66] in terms of both performance and explainability. We chose these 17 datasets because they are widely used as benchmarks for rule generation. Table IV shows the number of samples, and the number of classes of each data set, the number of variables in different subsets and the percentage of samples in each class.

We first evaluate the performance and explainability of the model-based explanations of the fuzzy systems obtained by the different approaches. The evaluation index of the model-based explanations is the same as for Experiment 1 on the synthetic data set. Considering some data sets have unbalance classes (e.g., the HTRU2 data set), we adopt both the average of the classification accuracy and the average of F1 scores over 5-fold cross-validation as the performance indices. The F1 score is calculated as follows [67]:



(a) The WM algorithm

(b) The FURIA algorithm

Fig. 5. Experimental results with Experimental Setting 4

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall}, \quad (1)$$

where

$$Precision = \frac{TP}{TP + FP}, \quad (2)$$

$$Recall = \frac{TP}{TP + FN}, \quad (3)$$

and

- TP represents the number of true positive cases
- TN denotes the number true negative cases;
- FP represents the number false positive cases;

TABLE IV
DATA SETS USED IN THIS PAPER[1]

Name	Samples	Class	$\|D\|$	$\|D_{MB}\|$	$\|D_{CD}\|$	Percentage of each class
Authorship	841	4	70	36	4	38%, 35%, 7%, 21%
Breast	699	2	9	6	3	66%, 34%
Dry bean	13611	7	16	2	0	10%, 4%, 12%, 26%, 14%, 15%, 19%
Ecoli	336	8	7	2	0	43%, 23%, 1%, 1%, 10%, 6%, 1%, 15%
Glass	214	6	9	7	1	33%, 36%, 8%, 6%, 4%, 14%
Haberman	306	2	3	1	1	74%, 26%
HTRU2	17989	2	8	7	0	91%, 9%
Iris	150	3	4	3	0	33%, 33%, 33%
Mammographic mas	830	2	6	4	2	51%, 49%
Page-blocks	5473	5	10	6	0	90%, 6%, 1%, 2%, 2%
Pendigits	10992	10	16	10	0	10%, 10%, 10%, 10%, 10%, 10%, 10%, 10%, 10%, 10%
Pima Indian Diabetes	768	2	8	6	3	65%, 35%
Sonar	208	2	60	1	0	53%, 47%
Vehicle	846	4	18	10	0	26%, 25%, 26%, 24%
Vowel	990	11	12	7	2	55%, 45%
Waveform-5000	5000	3	40	11	0	34%, 33%, 33%
Wine	178	3	13	8	2	33%, 40%, 27%

[1] The sum of percentage of each class for some data sets is not 100% because of rounding.

- FN is the number of false negative cases.

For multi-class data sets, the Macro F1 scores is adopted as it has been widely used as an extension of F1 scores in multi-classes problem. For the data set with N_c classes ($N_c > 2$), the Macro F1 scores is calculated as follows [68]:

$$Macro - F1 = \frac{\sum_{i=1}^{N_c} F1_i}{N_c}, \quad (4)$$

where $F1_i$ is the F1 score of the i the class. Both the accuracy and the (Macro) F1 scores are better when higher. The performance indices are shown in Table V and Table VI, with the best indices in bold.

Note that the rule generation methods based on MABLAR-CD generate no result for several data sets, because on these data sets the node of the target variable does not have any parent node in the resulting causal graph. Consequently, MABLAR-CD cannot be implemented for those data sets. Adopting a different causal discovery approach may result in a different causal graph as discussed above. We specifically include these data sets here to highlight this aspect of frameworks such as MABLAR-CD which are dependent on eliciting direct-cause variables.

The number of rules of each fuzzy system is shown in Table VIII and Table VII shows the execution time of different rule generation approaches on each data set.

From Table V to Table VIII we can make the following observations: 1) For most data sets, the number of rules of a fuzzy system obtained by a MABLAR-based framework is lower than that of one obtained by the original approach. 2) The fuzzy systems obtained by WM and FURIA achieve the highest accuracy on 10 data sets, while the fuzzy systems obtained by MABLAR-based frameworks achieve the highest accuracy on 7 data sets. The same trend is observed on the F1 scores results. 3) While not critical for the applications considered here, we note that the MB-WM and MBCD-WM have generally lower execution time. However, This trend is not observed in MB-FURIA and MBCD-FURIA.

The above observations indicate that MABLAR-based frameworks can improve the explainability of model-based explanations. However, as expected, MABLAR-based frameworks cannot always improve the performance for a given data set, i.e. based on the data available and parameters such as the causal discovery technique adopted, they may not be able to identify and/or leverage the set of causal variables to achieve performance levels which are higher than traditional rule generation approaches.

D. Evaluation of causal explanations

In this section, we select the Pima Indian Diabetes (PID) data set as an example to show how the rules obtained by MABLAR capture local causal relationships, and how the rules obtained by MABLAR-CD capture causal relationships. The goal here is to predict, using the variables included in the data set, whether or not a Pima Indian woman has diabetes [69].

In this case, note how the rules obtained by MABLAR can be used to guide users on how to diagnose whether a person has type-2 diabetes – based on the causal links to both direct causes and direct effects. The rules obtained by MABLAR-CD should indicate which variables are the actual cause(s) for type-2 diabetes – based solely on the direct cause variables.

Fig. 6 shows the causal graph of the PID data set obtained in this paper using the PC algorithm. In cases where the direction of an edge cannot be determined by the PC algorithm, the ANM algorithm has been used to determine its direction. The MB of the outcome consists of all grey nodes, and the striped grey nodes are the direct cause variables of the outcome. Table IX shows the variables used in each fuzzy system.

For this illustrative case, we adjusted the number of fuzzy partitions for the BMI and age variables from 3 to 4 to reflect the medical and World Health Organisation guidelines on

TABLE V
CLASSIFICATION ACCURACY OF DIFFERENT FUZZY SYSTEMS ON EACH DATA SET

	WM		MB-WM		MBCD-WM		FURIA		MB-FURIA		MBCD-FURIA	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
Authorship	0.2057	0.0031	0.0202	0.0099	0.7907	0.0335	**0.8971**	0.0063	0.8962	0.0139	0.7993	0.0162
Breast	0.9070	0.0195	0.9542	0.0148	0.9414	0.0233	0.9217	0.0134	**0.9245**	0.0129	0.8991	0.0099
Dry_bean	**0.9115**	0.0044	0.3812	0.0174			0.9084	0.0008	0.5852	0.0077		
Ecoli	**0.8215**	0.0433	0.5031	0.0266			0.7917	0.0354	0.5654	0.0370		
Glass	**0.6536**	0.1040	0.5560	0.0496	0.3643	0.0407	0.4230	0.0515	0.4299	0.0318	0.3878	0.0102
Haberman	0.6764	0.0321	0.7354	0.0756			0.7353	0.0018	**0.7377**	0.0038		
HTRU2	**0.9754**	0.0040	0.9744	0.0034			0.9741	0.0037	0.9743	0.0034		
Iris	0.9267	0.0365	0.9267	0.0435			0.9283	0.0256	**0.9417**	0.0129		
Mammographic	0.7518	0.0571	**0.8157**	0.0421	0.8000	0.0116	0.7813	0.0062	0.7822	0.0050	0.7861	0.0045
Page_blocks	**0.9379**	0.0034	0.9165	0.0133			0.9139	0.0032	0.9085	0.0040		
Penditgits	**0.9852**	0.0027	0.9597	0.0069			0.8995	0.0052	0.8192	0.0063		
Pima_diabetes	0.7135	0.0249	0.7122	0.0319	0.6887	0.0481	0.6976	0.0135	0.7103	0.0178	**0.7249**	0.0182
Sonar	0.1345	0.0495	0.4663	0.0101			**0.7151**	0.0156	0.6141	0.0300		
Vehicle	**0.6666**	0.0534	0.6524	0.0395			0.6356	0.0299	0.6235	0.0503		
Vowel	**0.7343**	0.3293	0.6354	0.1971			0.5212	0.0096	0.5220	0.0141		
Waveform_5000	0.3316	0.0146	**0.7570**	0.0208			0.7271	0.0055	0.7157	0.0046		
Wine	0.6962	0.0407	**0.8987**	0.0256	0.8365	0.0561	0.7374	0.0323	0.7879	0.0341	0.6291	0.0717

TABLE VI
THE F1 SCORES OF DIFFERENT FUZZY SYSTEMS ON EACH DATA SET

	WM		MB-WM		MBCD-WM		FURIA		MB-FURIA		MBCD-FURIA	
	F1-mean	Std	F1-mean	Std	F1-mean	Std	F1-mean	Std	F1-mean	Std	F1-mean	Std
Authorship	0.0853	0.0011	0.1022	0.0190	0.7615	0.0297	0.8392	0.0356	**0.8624**	0.0242	0.7401	0.0442
Breast	0.9545	0.0126	0.9558	0.0205	0.9279	0.0299	**0.9575**	0.0097	0.9518	0.0047	0.9495	0.0098
Dry-bean	**0.9254**	0.0031	0.2532	0.0258			0.9197	0.0018	0.4940	0.0116		
Ecoli	**0.4841**	0.0392	0.1945	0.0230			0.4521	0.0319	0.2642	0.0182		
Glass	**0.5868**	0.1167	0.5302	0.0830	0.1522	0.0636	0.4395	0.0639	0.4361	0.0609	0.2346	0.0082
Haberman	0.5527	0.0261	0.4552	0.0565			0.8215	0.0295	**0.8230**	0.0298		
HTRU2	0.9221	0.0087	0.9110	0.0172			0.9874	0.0008	**0.9879**	0.0003		
Iris	0.9254	0.0388	0.9389	0.0445			0.9257	0.0274	**0.9415**	0.0132		
mammographic	0.7541	0.0808	0.7907	0.1109	0.7598	0.0864	**0.8439**	0.0145	0.8416	0.0157	0.7779	0.0192
Page_blocks	**0.6502**	0.0091	0.5276	0.0516			0.1654	0.0208	0.1546	0.0347		
Penditgits	**0.9852**	0.0028	0.9583	0.0074			0.9493	0.0034	0.9169	0.0023		
Pima_diabetes	0.6750	0.0289	0.6610	0.0205	0.6917	0.0442	0.8064	0.0168	0.8057	0.0120	**0.8091**	0.0132
Sonar	0.4369	0.0280	0.3419	0.0529			**0.6792**	0.0231	0.5269	0.0932		
Vehicle	0.6614	0.0328	**0.6635**	0.0455			0.5908	0.0500	0.5777	0.0571		
Vowel	0.1336	0.0260	0.1233	0.0041			0.5402	0.0387	**0.5410**	0.0278		
Waveform_5000	0.1670	0.0063	0.7533	0.0195			**0.8027**	0.0061	0.7650	0.0046		
Wine	0.7624	0.0608	**0.9008**	0.0243	0.8248	0.0773	0.2168	0.0078	0.2425	0.0118	0.1992	0.0266

TABLE VII
THE EXECUTION TIME (SECONDS) OF FUZZY SYSTEM RULE GENERATION APPROACHES ON EACH DATA SET. NOTE THAT TIMES SHOWN FOR MB AND MBCD APPROACHES DO NOT INCLUDE THE TIME FOR CAUSAL DISCOVERY; THE LATTER IS SHOWN IN THE SUPPLEMENTARY MATERIALS.

	WM	MB-WM	MBCD-WM	FURIA	MB-FURIA	MBCD-FURIA
Authorship	16.26	10.95	0.60	0.29	0.24	0.13
Breast	1.60	0.95	0.24	0.04	0.03	0.04
Dry_bean	126.21	4.49		6.91	1.79	
Ecoli	2.54	0.10		0.02	0.02	
Glass	0.45	0.32	0.05	0.02	0.02	0.02
Haberman	0.15	0.10		0.02	0.01	
HTRU2	26.37	22.24		1.62	1.60	
Iris	0.16	0.11		0.01	0.01	
mammographic	0.88	1.27	0.17	0.02	0.02	0.06
Page_blocks	12.34	7.44		0.38	0.29	
Penditgits	301.82	108.18		5.18	6.91	
Pima_diabetes	2.87	2.36	0.23	0.04	0.04	0.06
Sonar	4.15	0.06		0.02	0.01	
Vehicle	6.64	4.11		0.12	0.09	
Vowel	8.82	3.39		0.19	0.14	
Waveform_5000	424.17	83.75		3.64	1.61	
Wine	0.74	0.56	0.10	0.01	0.01	0.02

TABLE VIII
THE NUMBER OF RULES OF DIFFERENT FUZZY SYSTEMS FOR EACH DATA SET

	WM		MB-WM		MBCD-WM		FURIA		MB-FURIA		MBCD-FURIA	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
Authorship	672.8	0.4	672.8	0.4	70.4	1.3	8.0	1.1	9.8	1.2	**7.4**	1.2
Breast	230.4	4.2	130.4	5.2	24.8	1.1	10.2	1.2	10.6	1.9	**9.4**	2.2
Dry_bean	1338.8	24.6	**9.0**	0.0			33.4	1.5	12.0	1.7		
Ecoli	81.2	5.2	9.0	0.0			6.8	1.0	**4.6**	0.5		
Glass	87	6.8	61.2	4.5	3.0	0.0	5.0	1.1	4.0	1.3	**1.4**	0.5
Haberman	25.4	0.5	3.0	0.0			3.2	2.1	**1.4**	0.5		
HTRU2	159.4	3.4	104.8	2.2			**12.6**	2.8	14.2	2.6		
Iris	19.6	1.8	10.2	0.4			3.4	0.5	**3.0**	0.0		
Mammographic	77.4	5.6	53	5.8	9	0	**4.2**	1.3	4.4	1.4	4.6	1.6
Page_blocks	187.4	5.5	72.6	7.1			17.0	2.3	**16.0**	1.9		
Penditgits	3960.8	37.7	2153.0	9.6			**204.0**	6.2	212.0	13.6		
Pima_diabetes	396.6	8.1	330.8	6.0	9.0	0.0	9.0	2.1	10.8	1.5	**7.4**	1.0
Sonar	166.2	0.4	3.0	0.0			4.2	1.0	**2.6**	0.8		
Vehicle	537.6	5.4	418.4	8.0			**10.2**	3.1	11.8	1.9		
Vowel	539.6	7.0	240.8	35.1			42.6	4.5	**40.4**	4.5		
Waveform_5000	4000	0.0	2920.2	14.9			158.6	5.2	**73.6**	13.2		
Wine	140.4	1.1	113.8	1.6	8.0	0.0	8.2	0.7	7.8	1.2	**6.6**	2.1

key categories [70], [71]. The linguistic terms for the BMI fuzzy partitions are underweight, healthy, overweight, and obese, respectively, whilst the linguistic terms for the age fuzzy partitions are ‘young adulthood’, ‘middle adulthood’, ‘young old’ and ‘old old’, which are taken verbatim from Lachman [70]. The complete rule base of each fuzzy system is shown in the supplementary materials.

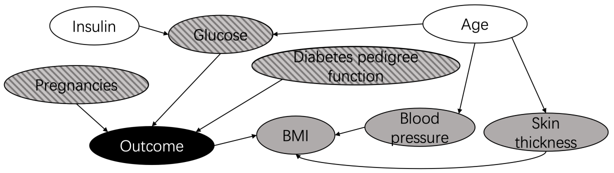


Fig. 6. The causal graph of the PID dataset

1) Local causal relationships captured by MABLAR: According to [72], the diagnosis of a disease depends on the symptoms that are causally related to the disease. Similarly, the rules obtained by MABLAR provide users with a guideline on how to diagnose type-2 diabetes. We describe it as a guideline – as the results are greatly influenced by the quality of the data set and the causal discovery algorithm. In other words, in practice, the resulting causal links can be viewed as suggestions which may guide experts – who may proceed to experimentally verify them.

As shown in Table IX, as expected, the WM approach indicates that all variables should be considered when diagnosing whether an individual has Type-2 diabetes. On the other hand, MB-WM suggests that insulin and age variables do not need to be considered. Similarly, MB-FURIA also removes the blood pressure and age variables. According to [73], [74], doctors usually only use the variable ‘Glucose’, which represents the plasma glucose concentration at two hours in an oral glucose tolerance test to diagnose whether a person has type-2 diabetes. Thus, compared to MB-WM and MB-FURIA, the rules obtained by WM and FURIA contain more variables which doctors consider unnecessary to consider when diagnosing diabetes. The results show that, although not perfect, the explanations provided by the rules obtained by MABLAR are closer to the diagnostic approach of experts. Thus, regarding the issue of diagnosing type-2 diabetes, the rules obtained by MABLAR provide more rational explanations compared to traditional methods.

2) Causal relationships captured by MABLAR-CD: In this case, the rules obtained by MABLAR-CD provide explanations for the etiology of type-2 diabetes, i.e. the factors that cause type-2 diabetes. As shown in Table IX, both MBCD-WM and MBCD-FURIA indicate that type-2 diabetes is caused by the variables pregnancies, glucose and Diabetes Pedigreee Function (DPF). In the real-world, the etiology of type 2 diabetes remains an unresolved issue. However, there are many risk factors being considered as the potential causes of type 2 diabetes. According to [75]–[77] (and some public sources, e.g. [78]), the recognized risk factors of type-2 diabetes, which are also included in the actual Pima Indian diabetes data set, are blood pressure, BMI, DPF and age. Among these four variables, DPF is the only variable considered both as a risk factor and included in the rules generated by MABLAR-CD. In this case, users can be more concerned with the DPF variable. The glucose variable is also worth highlighting. The rules in MBCD-WM and MBCD-FURIA both include the glucose variable; however, it is not recognized as a risk factor. In Fig. 6, the causal relationship between ‘BMI’, ‘Glucose’ and ‘Outcome’ is ‘Glucose → Outcome → BMI’. However, according to [76], the correct causal relationship should be ‘BMI → Outcome → Glucose’. A possible reason for this is that these two causal structures belong to the same Markov equivalence classes, which cannot be determined by the adopted PC algorithm (See Section II-B).

If the algorithm could correctly identify this causal re-

TABLE IX
VARIABLES USED IN DIFFERENT FUZZY SYSTEMS

	Pregnancies	Glucose	***Blood Pressure***	Skin Thickness	Insulin	***BMI***	***Diabetes Pedigree Function***	***Age***
WM	✓	✓	✓	✓	✓	✓	✓	✓
MB-WM	✓	✓	✓	✓		✓	✓	
MBCD-WM	✓	✓					✓	
FURIA		✓	✓			✓	✓	✓
MB-FURIA		✓				✓	✓	
MBCD-FURIA	✓	✓					✓	

lationship from the given data set [3], the glucose variable in the rules obtained by MABLAR-CD would be replaced by the BMI variable. Consequently, the rules obtained by MABLAR-CD would provide better explanations for the etiology of type-2 diabetes, more consistent with the domain knowledge – and with the potential to guide medical experts in less researched contexts.

Overall, while not perfect, MABLAR-based frameworks provide important improvements in generating rules with the capacity to provide causal symptom explanations and direct cause explanations.

V. Discussion

MABLAR and MABLAR-CD are designed to generate rules which capture causal relationships between the input and output variables, which benefits fuzzy systems in terms of performance, robustness and explainability. Table III and Table VIII support the conclusion that both MABLAR and MABLAR-CD can reduce the complexity of the obtained fuzzy systems and improve the explainability of the obtained fuzzy systems in terms of model-based explanations. However, Table III and Table V show that MABLAR frameworks may not always improve the performance of the obtained fuzzy systems.

As we explained in Section III-D, the performance of fuzzy systems which are obtained by MABLAR should be the highest among all adopted approaches if the appropriate causal relationships are represented in the data set *and* are identified. In practice, as experiments on both the simulated data set and the real world data sets show, this is not necessarily the case, not least because the data sets may not fulfil the latter condition, and even when they do, the causal discovery algorithm may not identify the appropriate relationships correctly.

One possible reason for this is that the causal discovery algorithms used is not suitable for the data set, which results in the obtained Markov blanket not being perfect, i.e. the obtained Markov blanket is not the ground truth Markov blanket. When the obtained Markov blankets are imperfect, the rules obtained by a MABLAR-based framework run the risk of capturing non-causal relationships, which worsens the performance of the obtained fuzzy system and risks generating misleading rules/explanations. As shown in Table III, MBCD-WM and MBCD-FURIA perform better than WM and MB-WM. From Table II we can see that, on the simulated data set, the MB adopted by MABLAR-CD is equal to the ground truth MB. Thus, given the ground truth MB, it is clear that MABLAR-based frameworks can be helpful for improving the performance of the obtained fuzzy system. Also, the results shown in Table III and Table V show that even when MABLAR-based frameworks adopt an imperfect MB, they are still helpful for improving the performance of the obtained fuzzy system in several cases. The discussion about how to select causal discovery algorithms is outside the scope of this paper. However, it is a rich area for future work.

The results in Section IV-D1) and Section IV-D2) indicate that, although not perfect, the rules obtained by MABLAR and MABLAR-CD provide better causal explanations than do rules obtained by a classical data-driven rule generation algorithm. Also, even while not perfect, the rules obtained by MABLAR and MABLAR-CD are still helpful for increasing users' understanding of the real-world processes.

We would like to further discuss the explanations provided by the rules obtained by MABLAR-CD. In many real-world applications, the real cause(s) of a phenomenon are not known. In these cases, even where the rules obtained by MABLAR-CD capture imperfect causal relationships between the variables, they are still helpful for increasing the users' understanding of the real world. To confirm the direct cause of a real-world phenomenon, experts (e.g. in biology or medicine) traditionally engage in large-scale analysis and/or experiments. The rules obtained by MABLAR-CD can reduce the search scope for the direct cause(s). As we showed in Section IV-D2, doctors have narrowed the search scope for the causes of type-2 diabetes, which are also included in the PID data set, to the blood pressure variable, the BMI variable, the DPF variable and the age variable. The rules obtained by MABLAR-CD further narrow the search scope to the DPF variable, which highlights a mechanism enabling doctors to gather insight into the etiology of diseases such as type-2 diabetes more directly based on what is learnt by the fuzzy system obtained by MABLAR-CD.

In general, the experimental results support the conclusion that the rules obtained by MABLAR and MABLAR-CD benefit fuzzy systems in terms of both performance and explainability.

[3] We note that a data set itself may not allow this to be done, i.e. it may contain insufficient information.

VI. Conclusions

In this paper, we formulated MABLAR and MABLAR-CD, and showed that the rules obtained by MABLAR and MABLAR-CD benefit fuzzy systems in A) providing causal explanations which have the potential to expand users' insights into the real-world, B) improving the human explainability due to the reduction of both the number of rules and the complexity of the rules vis-a-vis classical data-driven rule generation methods, and C) delivering strong robustness when facing distribution shift. We have shown how MABLAR and MABLAR-CD capture different types of causal relationships, and have highlighted their respective suitability for different problems. The quality of the causal graph obtained from the given data set affects the quality of the rules generated by MABLAR and MABLAR-CD. Rules generated based on a perfect causal graph, i.e. a ground truth causal graph, can perfectly reflect the causal relationships between variables. However, the experimental results indicate that even imperfect causal graphs, which may be obtained by data-driven causal discovery algorithms, contribute to some extent in generating rules that reflect the causal relationships between the variables.

In the future, we will explore further approaches which focus on improving the quality of the rules when facing an imperfect causal graph. Also, MABLAR and MABLAR-CD respectively exploit the MB (Markov blanket) information and causal direction information from the causal graph to generate rules that reflect the causal relationships between variables. In the future, we will explore the use of additional causal information to generate improved 'truly causal' rule bases.

References

[1] 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.

[2] S. Kleinberg, *Causality, probability, and time*. Cambridge University Press, 2013.

[3] J. Pearl, *Causality: Models, Reasoning and Inference*, 2nd ed. USA: Cambridge University Press, 2009.

[4] J. Y. Halpern, *Actual Causality*, 1st ed. The MIT Press, 2016.

[5] T. Honderich, *The Oxford companion to philosophy*. OUP Oxford, 2005.

[6] "Causality in wikipedia," https://en.wikipedia.org/wiki/Causality, accessed: 2024-6-4.

[7] D. Galles and J. Pearl, "Axioms of causal relevance," *Artificial Intelligence*, vol. 97, no. 1-2, pp. 9–43, 1997.

[8] J. Y. Halpern and J. Pearl, "Causes and explanations: a structural-model approach part 1: causes," in *The conference on Uncertainty in Artificial Intelligence (UAI)*. AUAI, 2001, pp. 194–202.

[9] F. Oviedo, J. L. Ferres, T. Buonassisi, and K. T. Butler, "Interpretable and explainable machine learning for materials science and chemistry," *Accounts of Materials Research*, vol. 3, no. 6, pp. 597–607, 2022.

[10] M. Ziatdinov, C. T. Nelson, X. Zhang, R. K. Vasudevan, E. Eliseev, A. N. Morozovska, I. Takeuchi, and S. V. Kalinin, "Causal analysis of competing atomistic mechanisms in ferroelectric materials from high-resolution scanning transmission electron microscopy data," *npj Computational Materials*, vol. 6, no. 1, p. 127, 2020.

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

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

[13] E. Mamdani and S. Assilian, "An experiment in linguistic synthesis with a fuzzy logic controller," *International Journal of Man-Machine Studies*, vol. 7, no. 1, pp. 1–13, 1975.

[14] 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.

[15] J. M. Alonso, C. Castiello, and C. Mencar, *Interpretability of Fuzzy Systems: Current Research Trends and Prospects*. Springer Berlin Heidelberg, 2015, pp. 219–237.

[16] J. J. Jassbi, P. J. A. Serra, R. A. Ribeiro, and A. Donati, "A comparison of mandani and sugeno inference systems for a space fault detection application," in *2006 World Automation Congress*. IEEE, 2006, pp. 1–8.

[17] H. Jones, B. Charnomordic, D. Dubois, and S. Guillaume, "Practical inference with systems of gradual implicative rules," *IEEE Transactions on Fuzzy Systems*, vol. 17, no. 1, pp. 61–78, 2008.

[18] D. Dubois, H. Prade, and L. Ughetto, "A new perspective on reasoning with fuzzy rules," *International Journal of Intelligent Systems*, vol. 18, no. 5, pp. 541–567, 2003.

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

[20] 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.

[21] 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 *IEEE Symposium Series on Computational Intelligence (IEEE-SSCI)*. IEEE, 2018, pp. 420–426.

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

[23] J. M. A. Moral, C. Castiello, L. Magdalena, and C. Mencar, *Explainable Fuzzy Systems: Paving the Way from Interpretable Fuzzy Systems to Explainable AI Systems*. Springer Nature, 2021.

[24] O. Cordón and F. Herrera, "A proposal for improving the accuracy of linguistic modeling," *IEEE Transactions on Fuzzy Systems*, vol. 8, no. 3, pp. 335–344, 2000.

[25] L.-C. Dutu, G. Mauris, and P. Bolon, "A linear-complexity rule base generation method for fuzzy systems," in *Proceedings of the 2015 Conference of the International Fuzzy Systems Association and the European Society for Fuzzy Logic and Technology*. Atlantis Press, pp. 520–527.

[26] J. Alcala-Fdez, R. Alcala, and F. Herrera, "A fuzzy association rule-based classification model for high-dimensional problems with genetic rule selection and lateral tuning," *IEEE Transactions on Fuzzy Systems*, vol. 19, no. 5, pp. 857–872, 2011.

[27] Z. Chi, J. Wu, and H. Yan, "Handwritten numeral recognition using self-organizing maps and fuzzy rules," *Pattern Recognition*, vol. 28, no. 1, pp. 59–66, 1995.

[28] J. M. Alonso and L. Magdalena, "Hilk++: an interpretability-guided fuzzy modeling methodology for learning readable and comprehensible fuzzy rule-based classifiers," *Soft Computing*, vol. 15, pp. 1959–1980, 2011.

[29] S. Porebski, "Evaluation of fuzzy membership functions for linguistic rule-based classifier focused on explainability, interpretability and reliability," *Expert Systems with Applications*, vol. 199, p. 117116, 2022.

[30] C. Higgins and R. Goodman, "Fuzzy rule-based networks for control," *IEEE Transactions on Fuzzy Systems*, vol. 2, no. 1, pp. 82–88, 1994.

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

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

[33] I. Tsamardinos, L. E. Brown, and C. F. Aliferis, "The max-min hill-climbing bayesian network structure learning algorithm," *Machine learning*, vol. 65, no. 1, pp. 31–78, 2006.

[34] I. Tsamardinos, C. F. Aliferis, and A. Statnikov, "Time and sample efficient discovery of Markov blankets and direct causal relations," in *Proceedings of the Ninth ACM SIGKDD International Conference on*

Knowledge Discovery and Data Mining. Association for Computing Machinery, 2003, pp. 673–678.

[35] J. Peters, J. M. Mooij, D. Janzing, and B. Schölkopf, "Causal discovery with continuous additive noise models," *Journal of Machine Learning Research*, vol. 15, no. 58, pp. 2009–2053, 2014.

[36] P. Hoyer, D. Janzing, J. M. Mooij, J. Peters, and B. Schölkopf, "Nonlinear causal discovery with additive noise models," in *Advances in Neural Information Processing Systems(NeurlPS)*. Curran Associates, Inc., 2008, pp. 689–696.

[37] T. S. Verma and J. Pearl, "Equivalence and synthesis of causal models," in *Probabilistic and Causal Inference: The Works of Judea Pearl*. Association for Computing Machinery, 2022, pp. 221–236.

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

[39] S. Shimizu, T. Inazumi, Y. Sogawa, A. Hyvärinen, Y. Kawahara, T. Washio, P. O. Hoyer, and K. Bollen, "DirectLiNGAM: A direct method for learning a linear non-gaussian structural equation model," *Journal of Machine Learning Research*, vol. 12, pp. 1225–1248, 2011.

[40] D. Colombo, M. H. Maathuis, M. Kalisch, and T. S. Richardson, "Learning high-dimensional directed acyclic graphs with latent and selection variables," *The Annals of Statistics*, vol. 40, no. 1, pp. 294–321, 2012.

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

[42] B. Kosko, "Fuzzy cognitive maps," *International Journal of Man-Machine Studies*, vol. 24, no. 1, pp. 65–75, 1986.

[43] 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.

[44] 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.

[45] 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.

[46] 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, pp. 226–240, 2005.

[47] 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, no. 4, pp. 270–276, 2008.

[48] 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.

[49] Y. He, J. Jia, and B. Yu, "Counting and exploring sizes of Markov equivalence classes of directed acyclic graphs," *Journal of Machine Learning Research*, vol. 16, no. 79, pp. 2589–2609, 2015.

[50] C. Glymour, K. Zhang, and P. Spirtes, "Review of causal discovery methods based on graphical models," *Frontiers in Genetics*, vol. 10, 2019.

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

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

[53] 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*, vol. 15, no. 5, pp. 1–46, 2021.

[54] J. Pearl, *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann Publishers, 1988.

[55] J.-P. Pellet and A. Elisseeff, "Using Markov blankets for causal structure learning," *Journal of Machine Learning Research*, vol. 9, no. 7, pp. 1295–1342, 2008.

[56] C. F. Aliferis, A. Statnikov, I. Tsamardinos, S. Mani, and X. D. Koutsoukos, "Local causal and Markov blanket induction for causal discovery and feature selection for classification part I: algorithms and empirical evaluation." *Journal of Machine Learning Research*, vol. 11, no. 7, pp. 171–234, 2010.

[57] K. Yu, L. Liu, and J. Li, "Learning Markov blankets from multiple interventional data sets," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 6, pp. 2005–2019, 2020.

[58] P. Cui and S. Athey, "Stable learning establishes some common ground between causal inference and machine learning," *Nature Machine Intelligence*, vol. 4, no. 2, pp. 110–115, 2022.

[59] 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.

[60] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[61] J. C. Bezdek, R. Ehrlich, and W. Full, "FCM: The fuzzy c-means clustering algorithm," *Computers & Geosciences*, vol. 10, no. 2, pp. 191–203, 1984.

[62] "Python weka wrapper3," https://fracpete.github.io/python-weka-wrapper3/, 2014, accessed: 2023-6-12.

[63] Y. Zheng, B. Huang, W. Chen, J. Ramsey, M. Gong, R. Cai, S. Shimizu, P. Spirtes, and K. Zhang, "Causal-learn: Causal discovery in python," *arXiv preprint arXiv:2307.16405*, 2023.

[64] "causal-learn: Causal discovery in python," https://github.com/py-why/causal-learn, 2021, accessed: 2023-6-12.

[65] T.-Z. Wang, S.-J. Huang, and Z.-H. Zhou, "Towards identifying causal relation between instances and labels," in *Proceedings of the SIAM International Conference on Data Mining*. SIAM, 2019, pp. 289–297.

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

[67] D. D. Lewis, "A sequential algorithm for training text classifiers: Corrigendum and additional data," *SIGIR Forum*, vol. 29, no. 2, pp. 13–19, 1995.

[68] D. D. Lewis, R. E. Schapire, J. P. Callan, and R. Papka, "Training algorithms for linear text classifiers," in *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery, 1996, p. 298–306.

[69] "Pima indians diabetes database," https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database, 2016, accessed: 2022-8-21.

[70] M. Lachman, "Adult development, psychology of," in *International Encyclopedia of the Social & Behavioral Sciences*, N. J. Smelser and P. B. Baltes, Eds. Oxford: Pergamon, 2001, pp. 135–139.

[71] W. E. Committee *et al.*, "Physical status: the use and interpretation of anthropometry," *World Health Organ Tech Rep Ser.*, vol. 854, pp. 312–344, 1995.

[72] J. G. Richens, C. M. Lee, and S. Johri, "Improving the accuracy of medical diagnosis with causal machine learning," *Nature Communications*, vol. 11, no. 1, p. 3923, 2020.

[73] W. C. Knowler, P. H. Bennett, R. F. Hamman, and M. Miller, "Diabetes incidence and prevalence in pima indians: a 19-fold greater incidence than in rochester, minnesota," *American Journal of Epidemiology*, vol. 108, no. 6, pp. 497–505, 1978.

[74] W. H. Organization, *Diabetes Mellitus: Report of a WHO Study Group*. World Health Organization, 1985, vol. 727.

[75] Y. Wu, Y. Ding, Y. Tanaka, and W. Zhang, "Risk factors contributing to type 2 diabetes and recent advances in the treatment and prevention," *International Journal of Medical Sciences*, vol. 11, no. 11, pp. 1185–1200, 2014.

[76] C. D. Society, "Guidelines for the prevention and control of type 2 diabetes in china (2017 edition)," *Chinese Journal of Practical Internal Medicine*, vol. 38, no. 4, pp. 292–344, 2018.

[77] B. Fletcher, M. Gulanick, and C. Lamendola, "Risk factors for type 2 diabetes mellitus," *Journal of Cardiovascular Nursing*, vol. 16, no. 2, pp. 17–23, 2002.

[78] "Diabetes risk factors," https://www.cdc.gov/diabetes/basics/risk-factors.html, 2022, accessed: 2022-9-18.