Critical Thinking About Explainable AI (XAI) for Rule-Based Fuzzy Systems

Jerry M. Mendel, *Life Fellow, IEEE*, and Piero P. Bonissone, *Life Fellow, IEEE*

***Abstract*—This article is about explainable artificial intelligence (XAI) for rule-based fuzzy systems [that can be expressed generically, as $y(\mathrm{x}) = f(\mathrm{x})$]. It explains why it is *not valid* to explain the output of Mamdani or Takagi–Sugeno–Kang rule-based fuzzy systems using IF-THEN rules, and why it *is valid* to explain the output of such rule-based fuzzy systems as an *association* of the compound antecedents of a small subset of the original larger set of rules, using a phrase such as "these linguistic antecedents are *symptomatic* of this output." Importantly, it provides a novel multi-step approach to obtain such a small subset of rules for three kinds of fuzzy systems, and illustrates it by means of a very comprehensive example. It also explains why the choice for antecedent membership function shapes may be more critical for XAI than before XAI, why linguistic approximation and similarity are essential for XAI, and, it provides a way to estimate the quality of the explanations.**

***Index Terms*—Explainable artificial intelligence (XAI), fuzzy system, linguistic approximation (LA), Mamdani fuzzy system, quality of explanations, rule-based fuzzy system, similarity, Takagi–Sugeno–Kang (TSK) fuzzy system.**

I. Introduction

EXPLAINABLE artificial intelligence (XAI) refers to the transparency and inherent interpretability of AI models used to derive conclusions. XAI using rule-based fuzzy systems (*fuzzy system*, for short) is and has been for some time a very hot research topic, e.g., [1]–[18]. Hagras [19] summarizes three approaches to realize XAI:

1) Deep explanation: Modified deep learning techniques to learn explainable features (e.g., deep learning and neural networks).
2) Interpretable models: Techniques to learn more structured, interpretable, causal models.
3) Model induction: Techniques to infer explainable model from any model as a black box (e.g., if-then rules).
This article focuses on the latter approach for fuzzy systems.

Transparency is needed in domains regulated by government agencies, such as insurance, loans, mortgages, etc., in which it is essential to show that the model did not use or create any bias. Other domains include high-stakes (e.g., portfolio re-balancing), or mission-critical applications (e.g., power plant set-point selection), in which the final decisions must be approved by a board or accepted by plant operators who are then accountable for such decisions.

A transparency/interpretability requirement usually comes at the expense of performance, because black-box models can be optimized for performance in ways that would be difficult to apply to transparent-box models that must maintain interpretability.

There is a sentiment in the fuzzy community that fuzzy rules would be of great value in XAI because such rules use words (which are modeled as fuzzy sets) and so they lend themselves naturally to XAI. This article challenges that sentiment, in a constructive way, and is *motivated* by [20], especially the underlined sentence (the underlining is ours), in:

> Interpretability[1] is one of the core arguments often put forward by fuzzy scholars in favor of fuzzy models—usually in a very uncritical way. In fact, many authors seem to take it for granted that fuzzy models or, more specifically, fuzzy rule-based models, can easily be understood and interpreted by a human user or data analyst. Many of these authors apparently equate "fuzzy" with "linguistic" and "linguistic" with "interpretable," which, of course, is far too simple. <u>A real "proof" of interpretability would require the presentation and careful inspection of a fuzzy model learned from data, which is almost never done.</u> At best, a discussion of that kind is replaced by the computation of certain *interpretability measures*, which, however, are disputable and pretend to a level of objectivity that is arguably not warranted for this criterion.

The overarching question to be addressed in this article is: *Is it valid to say that the output of a type-1 (T1) fuzzy system*[2] *can be described by the rules that led to it?* This may seem like an absurd question to practitioners of T1 fuzzy systems, since their starting point is a set of IF-THEN rules, but, as will be seen below, it actually is a very profound question.

In order to answer this question one must first examine the nature and design of a fuzzy system, because doing this will expose the decisions that have been made, so that those decisions can be re-examined, under the constraint of XAI, by means of critical thinking.

The main contributions of this article are as follows.

Manuscript received January 27, 2021; revised April 7, 2021; accepted May 5, 2021. Date of publication May 13, 2021; date of current version December 1, 2021. *(Corresponding author: Jerry M. Mendel.)*

Jerry M. Mendel is with the Department of Electrical Engineering, University of Southern California, Los Angeles, CA 90007 USA (e-mail: jmmprof@me.com).

Piero P. Bonissone is with the Piero P Bonissone Analytics LLC, San Diego, CA 92104 USA (e-mail: bonissone@gmail.com).

This article has supplementary material provided by the authors and color versions of one or more figures available at https://doi.org/10.1109/TFUZZ.2021.3079503.

Digital Object Identifier 10.1109/TFUZZ.2021.3079503

[1] "Interpretable" and "explainable" are viewed as synonyms in this article, in the sense that they both imply *understandable*.

[2] Due to page limitations, this investigation is limited to T1 fuzzy systems.

1) Explanation of why it is *not valid* to explain the output of Mamdani or Takagi–Sugeno–Kang (TSK) fuzzy systems using IF-THEN rules.
2) Explanation of why it *is valid* to explain the output of such fuzzy systems as an *association* of the antecedents of a small subset of the original larger set of rules, using a phrase such as "These linguistic antecedents are *symptomatic* of this output."
3) Novel multistep approach to obtain such a small subset of rules for three kinds of fuzzy systems, illustrated by a very comprehensive example.
4) Explanation of how linguistic approximation (LA) can be used to express the antecedent membership functions (MFs) (the symptoms) linguistically.
5) Method for estimating the quality of linguistic explanations.

II. Background

This section provides background about some popular fuzzy systems that is necessary for examining XAI for fuzzy systems. Its results are adapted from [21].

A. Introduction

Suppose that a fuzzy system has p inputs $x_1 \in X_1, \ldots, x_p \in X_p$, and one output $y \in Y$, where x_1 is described by Q_1 linguistic terms ($T_{x_1} = \{X_{1j}\}_{j=1}^{Q_1}$), …, x_p is described by Q_p linguistic terms ($T_{x_p} = \{X_{pj}\}_{j=1}^{Q_p}$), and y is either described by Q_y linguistic terms ($T_y = \{Y_j\}_{j=1}^{Q_y}$) or by a function $g(x_1, \ldots, x_p) = g(\mathbf{x})$, where $\mathbf{x} = col(x_1, \ldots, x_p)$.

Definition 1. The structure of the l^{th} generic *Zadeh rule* [22] for a fuzzy system is ($l = 1, \ldots, M$)

$$R_Z^l: \text{ IF } x_1 \text{ is } F_1^l \text{ and } \cdots \text{ and } x_p \text{ is } F_p^l, \text{ THEN } y \text{ is } \mathrm{G}^l. \quad (1)$$

The structure of the l^{th}generic *TSK rule* [23], [24] for a fuzzy system is

$$R_{TSK}^l: \text{ IF } x_1 \text{ is } F_1^l \text{ and } \cdots \text{ and } x_p \text{ is } F_p^l, \text{ THEN } y \text{ is } g^l(\mathbf{x}). \quad (2)$$

In both (1) and (2), $F_1^l \in T_{x_1}$, … and $F_p^l \in T_{x_p}$, and in (1), because $G^l \in T_y$ is a T1 fuzzy set, it is described by its MF $\mu_{G^l}(y)$. The most common choice for $g^l(\mathbf{x})$ is $c_0^l + \sum_{j=1}^p c_j^l x_j$.

Definition 2: When a fuzzy system uses Zadeh (TSK) rules and a Mamdani implication operator (minimum or product) it is called a *Mamdani (TSK) fuzzy system* [25], [26].

Zadeh and TSK rules have the same antecedent structure. When $\mathbf{x} = \mathbf{x}'$, the antecedents are "fired", leading to a *firing level* in the l^{th} rule for each antecedent, $f^l(x_i')$, so that the *overall firing level* for each rule, $f^l(\mathbf{x}')$, is[3]

$$f^l(\mathbf{x}') = \bigstar_{i=1}^p f^l(x_i') = \bigstar_{i=1}^p \mu_{F_i^l}(x_i'). \quad (3)$$

In (3), $\bigstar$ denotes a t-norm, either the minimum or product.[4]

During the design of a Mamdani or TSK fuzzy system, each antecedent is granulated ahead of time into Q_i linguistic terms, which establishes that the total number of rules $M = Q_1 Q_2 \cdots Q_p$. In the early days of the design of Mamdani fuzzy systems, the consequent variable y was also granulated ahead of time into Q_y linguistic terms, and then the M rules were provided by one or more application-experts. Those days are arguably long gone, and what now occurs is that each consequent variable is modeled either by a single number, a MF or a mathematical function.

B. Mamdani Fuzzy System: Centroid Defuzzifier

The centroid defuzzifier combines the fired rule output fuzzy sets $f^l(\mathbf{x}')\bigstar\mu_{G^l}(y)$ using the union (i.e., a t-conorm, usually the maximum), and then finds the centroid, $y_c(\mathbf{x}')$, of this set, as

$$y_c(\mathbf{x}') = \frac{\sum_{i=1}^N y_i \max\limits_{l=1,\ldots,M} f^l(\mathbf{x}')\bigstar\mu_{G^l}(y_i)}{\sum_{i=1}^N \max\limits_{l=1,\ldots,M} f^l(\mathbf{x}')\bigstar\mu_{G^l}(y_i)}. \quad (4)$$

In (4), each $f^l(\mathbf{x}')\bigstar\mu_{G^l}(y)$ has been discretized at the N points, $y_1, \ldots, y_N$, and $y_c(\mathbf{x}')$ is shown as an explicit function of $\mathbf{x}'$ because $f^l(\mathbf{x}')\bigstar\mu_{G^l}(y)$ is a function of $\mathbf{x}'$, so, for each $\mathbf{x}'$ a different value is obtained for y_c.

C. Mamdani Fuzzy System: Center of Sets (COSs) Defuzzifier

In COS defuzzification [26] each rule consequent set is replaced by a singleton, c^l, with amplitude equal to the firing level, after which the centroid of these singletons is found, i.e.,

$$y_{\text{COS}}(\mathbf{x}') = \frac{\sum_{l=1}^M c^l f^l(\mathbf{x}')}{\sum_{l=1}^M f^l(\mathbf{x}')}. \quad (5)$$

D. Comments

During the designs of (4) and (5), the parameters of the antecedent and consequent MFs are usually tuned using training and testing data.

Observe from (4) and (5) [as well as (SM-1); TSK fuzzy systems are covered in the supplementary material (SM)], that, regardless of which fuzzy system one uses, its output is just a number, and $f^l(\mathbf{x}')$, which depends only on antecedent MFs, is common to all of them. So, one should be able to back-engineer the compound linguistic antecedent $''x_1$ is F_1^l, and $\cdots$ and x_p is $F_p^{l''}$ from $f^l(\mathbf{x}')$, which would then be a starting point for a linguistic way to explain $y(\mathbf{x}')$.

III. Hidden Consequence of Using a Fuzzy System

An IF-THEN rule in the framework of crisp logic is a material *implication* ($p \rightarrow q$) and there can be many different MFs for such an implication (e.g., Lukasiewicz, Kleene–Dienes, etc.), each of which obeys the classical truth table for it, that is given in Table I.

[3] Equation (3) is for singleton fuzzification; more complicated computations are needed for nonsingleton fuzzification. Due to page limitations, non-singleton fuzzification is not examined in detail in this article; however, the conclusions that are drawn at the end of this article are also valid for it.

[4] These are the simplest mathematical operations for "and," which is the word in (1) and (2) that connects the p antecedent components.

TABLE I
TRUTH TABLE: THIRD COLUMN IS FOR MATERIAL IMPLICATION, AND LAST TWO COLUMNS ARE FOR MAMDANI IMPLICATIONS

p	q	$p \rightarrow q$	$\min(p,q)$	$p \cdot q$
1	1	1	1	1
1	0	0	0	0
0	1	1	0	0
0	0	1	0	0

[a] 1 ≡ True and 0 ≡ False.

Fuzzy logic begins with any of the MFs for crisp material implication (whose values are either 0 or 1) and then replaces them with the MFs of fuzzy sets (whose grades are between or equal to 0 and 1). Table 11.1 in [27] lists 14 such MFs all of which satisfy the truth table for crisp sets.

One branch of fuzzy logic research and applications uses these fuzzy implication MFs to perform approximate reasoning (e.g., [28]–[31]), but this is not the branch followed by most engineering and computer science advocates of fuzzy systems, due to the inherent bias in those MFs, which is due to rows three and four of the truth table. Loosely speaking, in engineering and computer science applications, if data are false one does not want any system output other than zero.

Mamdani fuzzy systems [25] do not begin with any of these fuzzy implication MFs, but instead use either minimum or product implication, both of which no longer satisfy the truth table for implication back in the crisp domain (see last two columns of Table I) [32]. Whereas the first two rows in Table I are the same for material implication and Mamdani implications, the last two rows are different. So, when a Mamdani implication is used, *the crisp logical meaning of the starting IF-THEN rules*[5] *is lost*. This means that *linguistic interpretations for the output of a Mamdani fuzzy system should not be stated using an IF-THEN structure*, because to do so carries the connotation of a crisp material implication, which is no longer the case.

A TSK IF-THEN rule has no connection at all to an implication, because its consequent is a mathematical function, and so there is no truth table for it.

Because logic has either vanished in a Mamdani fuzzy system, or does not exist at all in a TSK fuzzy system, it is not valid to explain their outputs using an IF-THEN rule structure. Instead, for XAI one can *associate* rule antecedents with $y(\mathbf{x}')$, in much the same way that one can associate symptoms with a disease, using linguistic summaries, such as

$$\begin{aligned}&\text{Given } \mathbf{x}=\mathbf{x}',\ \text{associated with } y(\mathbf{x}') \text{ are } x_1' \text{ is (use a}\\ &\qquad\text{linguistic term)}, \ldots,\ \text{and } x_p'\\ &\qquad\text{is (use a linguistic term)}\end{aligned} \tag{6a}$$

or

$$\begin{aligned}&\text{Given } \mathbf{x}=\mathbf{x}',\ \text{a linguistic term for } x_1',\ \ldots,\ \text{and a}\\ &\qquad\text{linguistic term for } x_p' \text{ are symptomatic of } y(\mathbf{x}').\end{aligned} \tag{6b}$$

How to determine those linguistic terms is explained in the sequel.

IV. FORWARD PROBLEM

At a very high level, the input-output formula for any of the Section II T1 fuzzy systems is

$$y(\mathbf{x}) = f(\mathbf{x}). \tag{7}$$

Until XAI became important, the designer of (7) was only interested in using it to solve a *forward problem*, i.e., given $\mathbf{x} = \mathbf{x}'$ compute $y(\mathbf{x}') = f(\mathbf{x}')$.

It is during the design process that the exact nature of $f(\mathbf{x})$ is established, and this requires making the following *many* decisions [21, ch. 4], [35].

1) Kind of fuzzy system: Mamdani or TSK.
2) Kind of fuzzifier: singleton or non-singleton.
3) Kind of t-norm: minimum or product.
4) How fired rule output sets are combined: union, addition, no combining.
5) Kind of defuzzifier (centroid, center-of-sets, etc.).
6) Number of input variables p.
7) Number and labels of terms per input: $\{X_{1j}\}_{j=1}^{Q_1}$, ..., $\{X_{pj}\}_{j=1}^{Q_p}$.
8) Number and labels of terms for, or structure of output (Mamdani fuzzy systems use $\{Y_j\}_{j=1}^{M}$ or $\{c^j\}_{j=1}^{M}$; TSK fuzzy systems use $\{g^l(\mathbf{x})\}_{l=1}^{M}$, where $g^l(\mathbf{x})$ must be specified).
9) Number of rules M.
10) Kind of MFs, e.g., triangle, trapezoid, Gaussian, etc.
11) Kind of MF parameters, e.g., pre-specified or optimized (tuned).

When one is only interested in using (7) to solve a forward problem, then it is now well known that:

1) TSK fuzzy systems often can achieve better performance than Mamdani fuzzy systems because of their functional rule consequents.
2) Nonsingleton fuzzification is more complicated than singleton fuzzification, but it can lead to improved performance when measured variables are noisy.
3) Both minimum and product t-norms lead to acceptable system performance, but the product may be easier to use since it does not require a test, and is inclusive because it does not throw away most of the information, whereas the minimum is very exclusive, since it chooses only the smallest value.
4) When computation time is important, taking the union of fired rule output sets is not advocated.
5) The number of input variables no longer seems to be an issue because today competitive neural networks are

[5] This is well-known, and has been stated in different ways, e.g., Mamdani [33] states: "The terms 'logic' in logic circuits and "fuzzy logic" in fuzzy logic control are purely incidental, and a matter of historical evolution." Some authors (e.g., Mendel [34] included) have often in the past referred to a Mamdani fuzzy system as a *fuzzy logic system*, but, since such a system breaks the MF of a material implication, the use of "logic" in this phrase is misleading, which is why the word "logic" was dropped from the title in [21].

designed with thousands of layers and billions of parameters.

6) Greater resolution is accomplished by using more terms for each variable.
7) The linguistic names that are given to each term's fuzzy sets are unimportant, because (7) does not involve the names of the fuzzy sets, but instead only involves MF formulas.
8) For p inputs there will be $M = Q_1 Q_2 \cdots Q_p$ rules. Increasing p and/or Q_j causes *rule explosion*.
9) Fuzzy systems seem to be robust to the choices of the MFs, so some prefer Gaussians or bell-shaped rather than triangles or trapezoids, whereas others prefer triangles or trapezoids.
10) Tuning of MF parameters using a supervised training method is better than specifying those parameters manually, because supervised training uses data (*the data speaks*).
11) *Universal approximation* holds (i.e., has been proved) for many fuzzy systems, which leaves one with confidence that (7) is able to uniformly approximate any real continuous function on a compact domain to arbitrary degree of accuracy.

V. Number of Antecedents

The number of antecedents p for the design of a fuzzy system (7) and its subsequent use in the forward problem is not a limiting factor, although in most real-world applications p has never been too large. For example, in fuzzy logic control, p equals the number of states and is often very small, e.g., in fuzzy PID control [36], [37] the states are error and error rate, so $p = 2$; or, for forecasting of time series (e.g., [38], [39]) p equals the length of a window of past sampled values, which is usually p = 2–8 past values. On the other hand, in XAI of a rule-based fuzzy system humans will be provided with explanations, and they must be able to understand them, so a small choice for p is now a limiting factor.

When p rule antecedents are connected by using the word "and," one is asking a human to mentally perform a p-fold correlation. Psychologists have told the first author that humans can barely correlate two things, and to ask them to correlate three or more things is beyond their capability. Of course, if the explanations are provided to an application-area expert (e.g., a cardiologist, stock trader, etc.), then, because of their intense training and prior knowledge, they should be able to correlate more than two antecedents, but surely not 10, 20, etc. of them.

So, for XAI of fuzzy systems, choosing a relatively small value of p seems quite important, which strongly suggests that the architecture used for such a fuzzy system is very important. The architectures of the Section II fuzzy systems are *flat,* i.e., all p antecedents occur in the rules. This architecture may be inappropriate for XAI of fuzzy systems, unless p can be kept very small. A hierarchical architecture may be more appropriate for XAI of such systems, but because there can be many such architectures (e.g., [40]–[44]), determining which one or more of these is (are) suitable for XAI of fuzzy systems, is at present unknown, and, how to explain the output from any of them remains to be studied.

The following approach for reducing the number of input variables, adapted (with some modifications) from [26], and analogous to forward step-wise regression, intuitively seems quite appropriate for XAI: round-1 determines which of p single antecedent fuzzy systems (each of which could be any one of the Section II systems) achieves best performance; round-2 determines which of $p - 1$ two antecedent fuzzy systems (in which one of the antecedents is the winner from round-1, and each of which could be any one of the Section II systems), achieves best performance; round-3 determines which of $p - 2$ three antecedent fuzzy systems (in which two of the antecedents are the winners from rounds 1 and 2, and each of which could be any one of the Section II systems), achieves best performance; etc. By limiting such a design to a small number of rounds, one would have rules with one to a small number of antecedents. The final system still has a flat structure, albeit with fewer than the starting p input variables, and so the methodology of this article can also be applied to its rules.

VI. Linguistic Approximation

XAI for fuzzy systems, as interpreted in this article, means: given $\mathbf{x} = \mathbf{x}'$ and $y(\mathbf{x}')$, establish the rule antecedent associations that can be used to explain $y(\mathbf{x}')$. One can say that such an XAI problem is akin to a *system identification problem* [45].

Our concern in the rest of this article is to study for which of the Section IV design decisions a fuzzy system is explainable, and what changes may be needed about those design decisions to make such a system explainable.

An explanation is done using words (W), but as has been seen above, *words no longer play any role in obtaining (7)*. So, before one is able to interpret $y(\mathbf{x}')$ one must first introduce words for the antecedents and consequents. This can be done by means of LA.

LA [26] is a phrase that has often been used in the past for obtaining a linguistic description of a fuzzy set. There are many methods for accomplishing it. Some obtain T1 fuzzy sets (e.g., see in [46, Appendix 3A] for a review of polling, direct rating, reverse rating and two kinds of interval estimation methods, and their references), some obtain interval type-2 fuzzy sets (e.g., see [47] for a review and comparison of the Interval Approach, Enhanced Interval Approach and the Hao–Mendel Approach; see, also, [48]), and others obtain general type-2 fuzzy sets (e.g., [49]).

All methods begin with an agreed upon application-dependent vocabulary of terms (words). Data are provided or collected for each term, and then those data are mapped into the fuzzy set for the term. One may argue that, because words mean different things to different people, linguistic uncertainties are present, so that interval type-2 or general type-2 fuzzy sets should be used, because their additional parameters are able to model such uncertainties, whereas T1 fuzzy sets cannot do this. One may also argue that, unless one is already an expert about fuzzy sets, the collected term data should not be a MF, because the concept of a MF will be unfamiliar to the non-expert, and will therefore introduce methodological uncertainties into LA. Unfortunately, it is not possible to separate methodological and linguistic uncertainties.

It is not the purpose of this article to get into details of which kind of fuzzy set to use, or how to collect word data either from a group of subjects or an individual, or how to then map that data into a fuzzy set model. For the purposes of this article, one only needs to understand that, as a result of LA, a *codebook* can be obtained for each antecedent and consequent variable, whose elements are pairs (W, MF_W).

For simplicity, it is assumed herein that each variable is explained linguistically by using the same number of V words (this keeps the notation light; otherwise, more subscripts are needed), so that

$$\text{Codebook}(x_i) = \{(W_j(x_i), MF_{W_j}(x_i))\}_{j=1}^{V} \; i = 1, \ldots, p \quad (8)$$

$$\text{Codebook}(y) = \{(W_{j_y}(y), MF_{j_y}(y))\}_{j_y=1}^{V}. \quad (9)$$

Example 1: Humans usually have little difficulty in understanding the words *low*, *moderate* and *high*. A codebook for these three words, taken from [50], by using the upper MFs of its interval type-2 fuzzy set word models, when $x \in X = [0, 10]$, is

$$\{(low, (0, 0, 1.22, 4.76), (moderate, (2.04, 4.02, 6.26, 8.51), (high, (5.92, 9.02, 10, 10)\}. \quad (10)$$

In (10), each MF is a trapezoid (a, b, c, d), *low* (*high*) is a left (right) shoulder and *moderate* is an interior MF.

How many words are in each codebook depends on how finely one wants to granulate a variable, which depends on how precisely one wants to explain $y(\mathbf{x}')$. The level of granulation has to be decided upon ahead of time, because it directly affects the MFs that are synthesized from word data, e.g., if it is decided ahead of time that the variable blood pressure (p) will be granulated coarsely into three terms, $\underline{l}$ow, $\underline{m}$oderate and $\underline{h}$igh, then word data for these three terms will lead to $\text{MF}_L^3(p)$, $\text{MF}_M^3(p)$ and $\text{MF}_H^3(p)$. If later it is decided that blood pressure should be granulated more finely into the five terms, $\underline{v}$ery $\underline{l}$ow, $\underline{l}$ow, $\underline{m}$oderate, $\underline{h}$igh and $\underline{v}$ery $\underline{h}$igh, then word data for these five terms will lead to $\text{MF}_{VL}^5(p)$, $\text{MF}_L^5(p)$, $\text{MF}_M^5(p)$, $\text{MF}_H^5(p)$ and $\text{MF}_{VH}^5(p)$. $\text{MF}_L^5(p) \neq \text{MF}_L^3(p)$, $\text{MF}_M^5(p) \neq \text{MF}_M^3(p)$ and $\text{MF}_H^5(p) \neq \text{MF}_H^3(p)$, because word data will be different when one knows the granulation is five versus three.

Section IX explains how to obtain a small subset, $S_{ss}(\mathbf{x}')$, of $M_{ss}(\mathbf{x}')$ rule indexes, whose associated rule compound-antecedents are used to explain $y(\mathbf{x}')$. Similarity, involving the p codebooks in (8), is used to express these compound antecedents linguistically, as follows: [rule# :$l \in S_{ss}(\mathbf{x}')$; input# :$i = 1, \ldots, p$; and codebook MF #:$j = 1, \ldots, V$].

1) For each rule #, l, find the *maximum Jaccard similarity*,[6] s_J, between each of its rule's p antecedents, $F_i^l(x_i)$, and the words in its respective codebook, $\{W_j(x_i)\}_{j=1}^{V}$, i.e., compute

$$s_J(F_i^l, W_j(x_i)) = \frac{\int_{X_i} \min[F_i^l(x_i), W_j(x_i)]dx_i}{\int_{X_i} \max[F_i^l(x_i), W_j(x_i)]dx_i} \equiv s_J^l(i, j) \quad (11)$$

$$\arg\max_j \{s_J^l(i, j)\}_{j=1}^{V} \equiv j^*(l). \quad (12)$$

For each antecedent (i), there are V similarities, one of which is the largest, for which $j^*(l) \in \{1, \ldots, V\}$.

2) Use $j^*(l)$ and the codebook in (8) to describe $F_i^l(x_i)$ linguistically as $W_{j^*(l)}(x_i)$.

How to use the codebook in (9) is deferred to Section IX; it depends on which fuzzy system is used to compute $y(\mathbf{x}')$.

VII. Choices for MFs

It is well known (e.g., [54]) that the MFs of a rule's antecedent variables partition the fuzzy system's *state space,* $X_1 \times \cdots \times X_p$, into hyper-regions in which only a subset of $M_s(\mathbf{x}')$ rules contribute to $y(\mathbf{x}')$, and if that subset is small then this lends to its interpretability. This suggests that the choice for the shape of the antecedent MFs may be important for XAI. Two popular choices for MFs are examined next.

A. Gaussian MFs

Gaussian MFs are easy to use during the design of a fuzzy system because their derivatives do not require tests. This makes them very attractive for tuning procedures that use derivatives.

In theory, Gaussian MFs are never exactly zero; hence, they do not partition the state space into smaller hyperregions, which means that, in theory, all M rules must be used to explain $y(\mathbf{x}')$. Of course, in practice, when $\mathbf{x} = \mathbf{x}'$ some of the Gaussian MFs may be of such small numerical value that they are effectively zero, in which case the state space is effectively partitioned into smaller hyper-regions. All of this is rather vague, i.e., how small is small?

Truncating Gaussians at grade ε is one approach to resolving this; but, different choices for ε lead to different truncations, which in turn lead to different hyper-regions in $X_1 \times \cdots \times X_p$, and consequently to different subsets of the M rules. If, however, there is agreement on what ε should be for a real-world application, then truncation is no longer an issue, and they are suitable for XAI; otherwise, their use is problematic for XAI.

B. Triangle and Trapezoidal MFs

Triangle and trapezoidal MFs are more challenging to use during the design of a fuzzy system because their derivatives require tests. This may make them unattractive for tuning procedures that use derivatives. However, when these MFs are zero, they are exactly zero; hence, they partition the state space into smaller unambiguous hyper-regions, which is very useful for XAI, but only if these smaller hyper-regions contain a relatively small number of rules, which they do, as is explained in Section VIII.

So, triangle and trapezoidal MFs may be more suitable for XAI than Gaussian MFs. More effort will be required during the design phase, but no ambiguity will occur for XAI.

[6] Although there are many kinds of fuzzy similarity (compatibility) measures (e.g., [51], [52], [46, Appendix 4A.1]), Jaccard similarity [53] is the most popular and is relatively easy to compute.

VIII. Finding the Subset of Rules that Act as a Starting Point for Explaining $y(\mathbf{x}')$

This section explains how to locate exactly which region (called a *first-order rule partition*) of $X_1 \times \cdots \times X_p$ $\mathbf{x} = \mathbf{x}'$ resides in, and subsequently which smaller subset of rule indexes, $S_s(\mathbf{x}')$, containing $M_s(\mathbf{x}') < M$ elements, led to $y(\mathbf{x}')$. The compound antecedents of the rules associated with $S_s(\mathbf{x}')$ will be used as a starting point to explain $y(\mathbf{x}')$.

A. First-Order Rule Partitions

Definition 3 (see [54]): A firing level *contributes* to the output of a T1 fuzzy system only if it is non-zero. This occurs in $X_1 \times \cdots \times X_p$ where the MFs of *all* of a rule's antecedents are simultaneously nonzero.

Definition 4 (see [54]): A *T1 first-order rule partition* of $X_1 \times \cdots \times X_p$ is a collection of non-overlapping hyper-rectangles in each of which the *same* number of *same* rules is fired whose firing levels contribute to the output of a T1 fuzzy system.[7]

First-order rule partitions provide a *coarse sculpting* of the state space; their numbers can be increased by granulating x_i more finely into more fuzzy sets.

As is stated in [55]: One of the important things learned from [54, Section III] is that first-order rule partitions of $X_1 \times \cdots \times X_p$ are completely determined by the respective rule partitions of each $(i = 1, \ldots, p)$ X_i separately, because, when minimum or product t-norms are used in (3), if even one component of a rule's firing level is zero then that rule does not contribute to the system's output.

Consequently, first-order rule partitions are determined initially for each X_i, and then for $X_1 \times \cdots \times X_p$.

It is known [55] that the total number of first-order rule partitions of $X_1 \times \cdots \times X_p$, $N^1(X_1, \ldots, X_p)$, is the product of the number of those partitions for each X_i, $N^1(X_i)$; and, that the same number of same rules that are fired in each first-order rule partition of $X_1 \times \cdots \times X_p$, $N_R(k_1, \ldots, k_p)$, is the product of the same number of same rules fired in each variable's first-order rule partition, $N_R(k_i)$, i.e.,

$$N^1(X_1, \ldots, X_p) = \prod_{i=1}^{p} N^1(X_i) \tag{13}$$

$$N_R(k_1, \ldots, k_p) = \prod_{i=1}^{p} N_R(k_i). \tag{14}$$

In (14), k_i is an index of the T1 first-order rule partitions of X_i, i.e., $k_i = 1, \ldots, N^1(X_i)$. In order to use these two equations, one must first determine $N^1(X_i)$ and $N_R(k_i)$ for each X_i; this is easy to do by following the procedure given in Table II.

Example 2: Fig. 1 depicts five T1 fuzzy sets and their first-order rule partitions,[8] obtained by using the Table II construction procedure. These five MFs have nine first-order rule partitions, whose names are 1, 2, …, and 9.

TABLE II
Two-Step Procedure to Establish T1 First-Order Rule Partition Quantities for x_i on a Plot of Its MFs

Step	Description
1	Scan the axis of x_i with an imaginary dashed vertical line from left to right. Count the number of intersections of this line with the MFs of x_i; this gives $N_R(k_i)$. Where this number, or the nature of the same rules, changes draw a dashed (vertical) *first-order rule partition line*; it represents the boundary of a T1 first-order rule partition. Insert a dashed vertical line at both the start and end of X_i. For each x_i, the interval of real numbers between adjacent dashed vertical lines is its T1 first-order rule partition. For triangle and trapezoidal MFs, first-order rule partition lines occur along the x_i axis when the MF first becomes non-zero or finally becomes zero. For a Gaussian MF, truncated at grade ε, these lines occur at the points at which the MF intersects a horizontal line at level ε.
2	Count the number of T1 first-order rule partitions of X_i, the total being $N^1(X_i)$; then, $k_i = 1,\ldots,N^1(X_i)$.

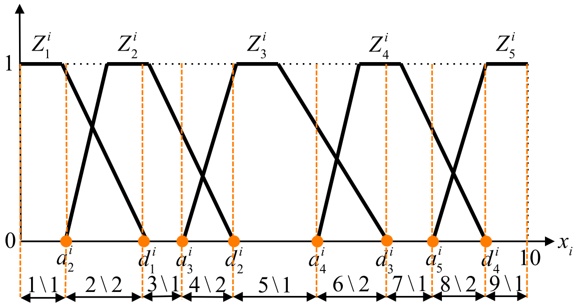


Fig. 1. First-order rule partitions for five T1 MFs and the number of activated MFs in each partition. In $\alpha\backslash\beta, \alpha = k_i$ is the index of the partition and $\beta = N_R(k_i)$ is the number of same rules fired in that partition.

Observe that only one or two rules are active (fire) in any of the partitions. This is due to MFs overlapping just with their nearest neighbors. If these MFs were used in a one-variable T1 fuzzy system, that system would have five rules, and even in such a simple system the number of rules that are active (1 or 2) ranges from 20% to 40% of the maximum number of five rules, which is a substantial reduction in the number of rules that explain what's happening in different regions of X_i.

Suppose that a two-variable system has each variable described by the same MFs as in Fig. 1. Then, this system would have 25 rules, and [according to (13)] there would be $9 \times 9 = 81$ first-order rule partitions each containing [according to (14)] either 1, 2 or 4 rules, each of which is a very substantial reduction[9] (84%–96%) from the maximum number of 25 rules. The percentage reduction in the number of active rules in a first-order rule partition becomes larger and larger as the number of input variables increases, which is very encouraging for XAI.

[7] Bonissone and Chang [56] call these *cells*. Their results, which include a formula for computing the total number of cells $[(2t-1)^n$, in which t is the number of terms (which is our Q), and n is the number of states (which is our p)] require MFs to overlap only with their nearest neighbors; this is not a requirement for first-order rule partitions.

[8] Many other examples of first-order rule partitions for T1 fuzzy systems are in [54] and its Supplementary Material.

[9] If the MFs overlap with more than their nearest neighbors, then these reduction percentages decrease.

TABLE III
FIRST-ORDER RULE PARTITION INFORMATION FOR FIG. 1

Partition index	Partition end-points		Active Rule Information ($j_i=1,\ldots,N_R(k_i)$)	
$k_i(x_i')\equiv k_i'$	$a(k_i\,\vert\,x_i)$	$b(k_i\,\vert\,x_i)$	$N_R(k_i)$	$Z^i_{j_i}$
1	0	a^i_2	1	Z^i_1
2	a^i_2	d^i_1	2	Z^i_1,Z^i_2
3	d^i_1	a^i_3	1	Z^i_2
4	a^i_3	d^i_2	2	Z^i_2,Z^i_3
5	d^i_2	a^i_4	1	Z^i_3
6	a^i_4	d^i_3	2	Z^i_3,Z^i_4
7	d^i_3	a^i_5	1	Z^i_4
8	a^i_5	d^i_4	2	Z^i_4,Z^i_5
9	d^i_4	10	1	Z^i_5

B. First-Order Rule Partition Information Table

By carrying out the Table II procedure for each of the p antecedent variables, one can construct a *first-order rule partition information table* for each variable.

Example 3: Table III is a first-order rule partition information table for $p=2$ and $Q=5$, that uses the MFs in Fig. 1. Its entries are self-explanatory.

C. Indexing Rules

Unfortunately, [54], [55], [57], and [58] only show how to construct first-order rule partitions, and how to determine the number of rules that are active in them, but do not explain how to determine what those rules actually are. While this was okay for forward problems, it is not okay for XAI, because XAI requires the exact rules so that antecedent associations can be linguistically stated for $y(\mathbf{x}')$.

In order to determine which rules are in a specific first-order rule partition, one must first establish how M rules, each with p antecedents, are stored. Tables (arrays) that do this work for $p=1$ and 2, and maybe for $p=3$, but they do not work for $p\geq 4$. Instead of multidimensional arrays, one can use a *lexicographical ordering* that maps p antecedent indexes into a single rule index.

For simplicity, assume that each rule antecedent has the same number of MFs (i.e., $Q_i=Q$, for $i=1,\ldots,p$), namely $\{Z^i_{j_i}\}^Q_{j_i=1}$. Let $(j_1,\ldots,j_p)$, $\forall j_i=1,\ldots,Q$, be mapped into the single index $l=1,\ldots,M=Q^p$, i.e.,

$$(j_1,j_2,\ldots,j_p)^Q_{\forall j_i=1}\rightarrow\{l\}^M_{l=1}. \tag{15}$$

Equation (15) is not a unique mapping, and in this article the following *lexicographical ordering* is used[10]

$$l=1+\sum_{i=1}^{p}(j_i-1)Q^{p-i}. \tag{16}$$

Example 4: $p=2$ and $Q=5$, as in Fig. 1, so that

$$l=1+\sum_{i=1}^{2}(j_i-1)\cdot 5^{2-i}=1+(j_1-1)\cdot 5+(j_2-1). \tag{17}$$

Using this formula, one obtains the 25 mappings from $(j_1,j_2)^5_{j_1=1,j_2=1}\rightarrow\{l\}^{25}_{l=1}$ that are summarized in Table SM-2, which also contains the rule's consequent.

D. Determining Rules Associated With $\mathbf{x}=\mathbf{x}'$

Using a table like Table III and (16), it is now possible to determine exactly which rules, and how many of them, are associated with $\mathbf{x}=\mathbf{x}'$, by using the following procedure:

1) For each antecedent value, x_i' $(i=1,\ldots,p)$, use its Table III to locate and save its k_i', $N_R(k_i')$ and $Z^i_{j_i}(j_i=1,\ldots,N_R(k_i'))$.
2) Compute the number of active rules in $P^1(k_1',k_2',\ldots,k_p')$, $N_R(k_1',k_2',\ldots,k_p')$, using (14).
3) Create the $N_R(k_1',k_2',\ldots,k_p')$ p-antecedent *active MF combinations* from the step-1 MFs.
4) Using (16), determine the rule number and rule consequent G^l for each of the step-3 combinations. This establishes the $M_s(\mathbf{x}')$-element set $S_s(\mathbf{x}')$.

Clearly

$$M_s(\mathbf{x}')=N_R(k_1',k_2',\ldots,k_p'). \tag{18}$$

Example 5: This is a continuation of Example 4 and illustrates this four-step procedure for two values of $\mathbf{x}'$.

First, suppose that $x_1'\in[a^1_3,d^1_2]$ and $x_2'\in[a^2_4,d^2_3]$. Using Table III for both x_1' and x_2', it follows that: $x_1'\in[a^1_3,d^1_2]\rightarrow k_1'=4\rightarrow\{Z^1_2,Z^1_3\}$ are active, and $x_2'\in[a^2_4,d^2_3]\rightarrow k_2'=6\rightarrow\{Z^2_3,Z^2_4\}$ are active; hence, there will be $N_R(4,6)=2\times 2=4$ two-antecedent rules with the antecedent pairings $\{Z^1_2,Z^2_3\}$, $\{Z^1_2,Z^2_4\}$, $\{Z^1_3,Z^2_3\}$, $\{Z^1_3,Z^2_4\}$. It then follows that[11] [the arrows in (19) and (20) imply the use of (17)]

$$\begin{cases}\{Z^1_2,Z^2_3\}\rightarrow j_1=2,\ j_2=3\rightarrow l=8\rightarrow G^8\\ \{Z^1_2,Z^2_4\}\rightarrow j_1=2,\ j_2=4\rightarrow l=9\rightarrow G^9\\ \{Z^1_3,Z^2_3\}\rightarrow j_1=3,\ j_2=3\rightarrow l=13\rightarrow G^{13}\\ \{Z^1_3,Z^2_4\}\rightarrow j_1=3,\ j_2=4\rightarrow l=14\rightarrow G^{14}.\end{cases} \tag{19}$$

Equation (19) shows that this $\mathbf{x}'$has the compound antecedents of rules R^8,R^9,R^{13} and R^{14} associated with it, so that $M_s(\mathbf{x}')=4$ and $S_s(\mathbf{x}')=\{8,9,13,14\}$.

Suppose, next that $x_1'\in[0,a^1_2]$ and $x_2'\in[a^2_2,d^2_1]$. Again, using Table III for both x_1' and x_2' , it follows that: $x_1'\in[0,a^1_2]\rightarrow k_1'=1\rightarrow\{Z^1_1\}$ is active; and $x_2'\in[a^2_2,d^2_1]\rightarrow k_2'=2\rightarrow\{Z^2_1,Z^2_2\}$ are active; hence, there will be $N_R(1,2)=1\times 2=2$ two-antecedent rules with the antecedent pairings $\{Z^1_1,Z^2_1\}$, $\{Z^1_1,Z^2_2\}$. It then follows that

$$\begin{cases}\{Z^1_1,Z^2_1\}\rightarrow j_1=1,\ j_2=1\rightarrow l=1\rightarrow G^1\\ \{Z^1_1,Z^2_2\}\rightarrow j_1=1,\ j_2=2\rightarrow l=2\rightarrow G^2\end{cases}. \tag{20}$$

Equation (20) shows that this $\mathbf{x}'$ has the compound antecedents of rules R^1 and R^2 associated with it, so that $M_s(\mathbf{x}')=2$ and $S_s(\mathbf{x}')=\{1,2\}$.

[10] This formula was provided to the first author by Prof. D. Wu, and is a representation of p nested DO loops in which the outermost loop is for j_1,followed by a loop for j_2, etc., until the final innermost loop is for j_p.

[11] If the fuzzy system is TSK then in (19) and (20) $G^l=G^l(\mathbf{x}')$.

IX. Proposed Methods for Finding a Smaller Subset of Rule Antecedents That Explain $y(\mathbf{x}')$

At this point, for $\mathbf{x} = \mathbf{x}'$ a subset $S_s(\mathbf{x}')$ with $M_s(\mathbf{x}')$ rule indexes has been determined, but no use has yet been made of the consequents of the associated rules, or of $y(\mathbf{x}')$. This section shows how to use both of these in order to determine an even smaller subset $S_{ss}(\mathbf{x}')$ of $M_{ss}(\mathbf{x}') \leq M_s(\mathbf{x}')$ rule-indexes, whose rule compound-antecedents best explain $y(\mathbf{x}')$. Those antecedents must then be expressed linguistically, using codebooks and similarity (see Section VI). Because some of those linguistic compound-antecedents may be the same, the $M_{ss}(\mathbf{x}')$ explanations may be further reduced to $M_{sss}(\mathbf{x}')$ explanations, which are then finally used to explain $y(\mathbf{x}')$.

Finding such smaller subsets depends on the kind of fuzzy system[12] that was used to obtain $y(\mathbf{x}')$.

A. Mamdani Fuzzy System: COS Defuzzification

This fuzzy system is $y(\mathbf{x}') = y_{\mathrm{COS}}(\mathbf{x}')$ [see (5)] in which $l = 1, \ldots, M$ is replaced by $l \in S_s(\mathbf{x}')$. One must now ask: How does one find a smaller subset of rules from only a single number, $y_{\mathrm{COS}}(\mathbf{x}')$, when the $M_s(\mathbf{x}')$ rule consequents are the numbers $\{c^l\}_{l\in S_s(\mathbf{x}')}$? There is no unique answer to this question. Two plausible answers are given next.

1) Answer 1: Use Nearest Consequent Neighbors: This answer finds the consequent values that are close to $y_{COS}(\mathbf{x}')$, and then associates the antecedents of those consequents with $y_{COS}(\mathbf{x}')$, as follows:

a) Compute $\min_{l\in S_s(\mathbf{x}')}[y_{\mathrm{COS}}(\mathbf{x}') - c^l] \equiv m_1(\mathbf{x}')$, where $\operatorname{argmin}_{l\in S_s(\mathbf{x}')}[y_{COS}(\mathbf{x}') - c^l] \equiv l_{m_1(\mathbf{x}')}$.

b) Specify a positive margin parameter ε_1 and then compute the interval I_1 where

$$I_1 \equiv [y_{COS}(\mathbf{x}') - (m_1(\mathbf{x}') + \varepsilon_1), y_{COS}(\mathbf{x}') + (m_1(\mathbf{x}') + \varepsilon_1)]. \tag{21}$$

c) Determine $l \ni c^l \in I_1$, where $l \in S_s(\mathbf{x}')$. There is at least one such l-value, namely $l_{m_1(\mathbf{x}')}$.

d) Collect these values of l into the $M_{ss}(\mathbf{x}')$-element set $S_{ss}(\mathbf{x}')$ where $M_{ss}(\mathbf{x}') \leq M_s(\mathbf{x}')$. This is the subset of rule indexes whose rule's compound-antecedents are used to explain $y_{\mathrm{COS}}(\mathbf{x}')$.

e) Use similarity and codebooks, as explained in Section VI, to express these compound antecedents linguistically.

f) Eliminate duplicate explanations, so that only $M_{sss}(\mathbf{x}') \leq M_{ss}(\mathbf{x}')$ of them remain.

This method needs the margin parameter ε_1, and different results may be obtained for different values of it.

2) Answer 2: Use Most Significant Contributing Terms: This answer finds the components of (5) that have contributed most significantly to $y_{\mathrm{COS}}(\mathbf{x}')$, as follows [$l \in S_s(\mathbf{x}')$]:

a) Express (5) as

$$y_{\mathrm{COS}}(\mathbf{x}') = \frac{\sum\limits_{l\in S_s(\mathbf{x}')} c^l f^l(\mathbf{x}')}{\sum\limits_{l\in S_s(\mathbf{x}')} f^l(\mathbf{x}')} = \sum_{l\in S_s(\mathbf{x}')} c^l w^l(\mathbf{x}') \tag{22}$$

$$w^l(\mathbf{x}') \equiv \frac{f^l(\mathbf{x}')}{\sum\limits_{l\in S_s(\mathbf{x}')} f^l(\mathbf{x}')}. \tag{23}$$

b) Compute the following *percentage contribution*:

$$\mathrm{Per}^l_{\mathrm{COS}}(\mathbf{x}') = \frac{c^l w^l(\mathbf{x}')}{y_{\mathrm{COS}}(\mathbf{x}')} \times 100. \tag{24}$$

c) Choose a threshold t_1 ($50\% \leq t_1 \leq 100\%$) and then (*Criterion*) save only the smallest number of $\mathrm{Per}^l_{\mathrm{COS}}(\mathbf{x}')$ such that their sum[13] is greater than or equal to t_1.

d) Collect the values of l, for which this Criterion is satisfied, into the $M_{ss}(\mathbf{x}')$-element set $S_{ss}(\mathbf{x}')$, where $M_{ss}(\mathbf{x}') \leq M_s(\mathbf{x}')$. This is the subset of rule indexes whose rule's compound-antecedents are used to explain $y_{\mathrm{COS}}(\mathbf{x}')$.

e) Use similarity and codebooks, as explained in Section VI, to express these compound antecedents linguistically.

f) Eliminate duplicate explanations, so that only $M_{sss}(\mathbf{x}') \leq M_{ss}(\mathbf{x}')$ of them remain.

g) Compute the minority decision (MD*)* as the sum of the $M_s(\mathbf{x}') - M_{ss}(\mathbf{x}')$ percentages not used in the *Criterion*. *MD* is a *measure of disagreement* supporting the $M_{ss}(\mathbf{x}')$ compound antecedents used to explain $y_{COS}(\mathbf{x}')$.

This method needs threshold parameter t_1 , and different results may be obtained for different values of it.

B. Mamdani Fuzzy System: Centroid Defuzzification

This fuzzy system is $y(\mathbf{x}') = y_c(\mathbf{x}')$ [see (4)] in which $l = 1, \ldots, M$ is replaced by $l \in S_s(\mathbf{x}')$. One must now ask: How does one find a smaller subset of rules from only a single number, $y_c(\mathbf{x}')$, when the $M_s(\mathbf{x}')$ rule consequents are the MFs, $\{\mu_{G^l}(y)\}_{l\in S_s(\mathbf{x}')}$?

One may again argue that there is no unique answer to this question, but, because MFs are now available for the consequent of each rule, a very strong case can be made for using similarity; so, only one answer to this question is given next.

Answer: Use Similarity

This answer finds words for $y(\mathbf{x}')$ and $\{\mu_{G^l}(y)\}_{l\in S_s(\mathbf{x}')}$, and then uses similarity to associate a subset of these consequent words with the word for $y(\mathbf{x}')$, as follows:

a) After the design of (4) is completed, map each $\{\mu_{G^l}(y)\}_{l=1}^M$ into its most similar word in the already-available codebook (9), i.e., (rule# : $l = 1, \ldots, M$; consequent Codebook MF#: $j_y = 1, \ldots, V$):[14]

- Find the maximum Jaccard similarity, s_J, between $\mu_{G^l}(y)$ and $\{MF_{j_y}(y)\}_{j_y=1}^V$, i.e., compute

$$s_J(\mu_{G^l}(y), MF_{j_y}(y)) = \frac{\int_Y \min[\mu_{G^l}(y), MF_{j_y}(y)]dy}{\int_Y \max[\mu_{G^l}(y), MF_{j_y}(y)]dy} \equiv s^l_J(j_y) \tag{25}$$

$$\arg\max_{j_y}\{s^l_J(j_y)\}_{j_y=1}^V \equiv j^*_y(l). \tag{26}$$

[12] See Section VI in the Supplementary Material for how to do this for a normalized TSK fuzzy system.

[13] If exactly one term satisfies this criterion, then it corresponds to the maximum value of $\mathrm{Per}^l_{\mathrm{COS}}(\mathbf{x}')$ in (24).

[14] This only has to be done once and stored, because the M consequent MFs do not depend on the input to the fuzzy system.

For each l, there are V similarities, one of which is the largest, for which $j_y^*(l) \in \{1, \ldots, V\}$.

•) Use $j_y^*(l)$ and the codebook in (9) to describe $\mu_{G^l}(y)$ linguistically as $W_{j_y^*(l)}(y)$.

b) Map output $y_c(\mathbf{x}')$ into a word using similarity, i.e.

- Treat $y_c(\mathbf{x}')$ as a fuzzy singleton, where

$$\mu_{y_c(\mathbf{x}')}(y) \equiv \begin{cases} 1 \text{ when } y = y_c(\mathbf{x}') \\ 0 \quad \text{otherwise} \end{cases}. \tag{27}$$

- Compute the Jaccard similarity between $y_c(\mathbf{x}')$ and $\{MF_{j_y}(y)\}_{j_y=1}^{V}$, where $(j_y = 1, \ldots, V)$.[15]

$$\begin{aligned} s_J(y_c(\mathbf{x}'), MF_{j_y}(y)) &= \frac{\int_Y \min[\mu_{y_c(\mathbf{x}')}(y), MF_{j_y}(y)]dy}{\int_Y \max[\mu_{y_c(\mathbf{x}')}(y), MF_{j_y}(y)]dy} \\ &= \frac{MF_{j_y}(y_c(\mathbf{x}'))}{1} = MF_{j_y}(y_c(\mathbf{x}')) \end{aligned} \tag{28}$$

$$\arg\max_{j_y}\{MF_{j_y}(y_c(\mathbf{x}'))\}_{j_y=1}^{V} \equiv j_y^*. \tag{29}$$

- Describe $y_c(\mathbf{x}')$ linguistically as $W_{j_y^*}(y)$.

c) Compute the similarity between $W_{j_y^*}(y)$ and the consequent words of the rules associated with the elements of $S_s(\mathbf{x}')$, $\{W_{j_y^*(l)}(y)\}_{l \in S_s(\mathbf{x}')}$, i.e., compute $\{s_J(W_{j_y^*}(y), W_{j_y^*(l)}(y))\}_{l \in S_s(\mathbf{x}')}$.

d) Let $S_{ss}(\mathbf{x}')$ be the $M_{ss}(\mathbf{x}')$-element subset of rule indexes such that $[l \in S_s(\mathbf{x}')]$ $s_J(W_{j_y^*}(y), W_{j_y^*(l)}(y)) \geq t_2 > 0.5$.[16] This is the subset of rule indexes whose linguistic rule compound-antecedents are used to explain $y_c(\mathbf{x}')$.

e) Eliminate duplicate explanations, so that only $M_{sss}(\mathbf{x}') \leq M_{ss}(\mathbf{x}')$ of them remain.

This method needs similarity threshold parameter t_2 , and different results will be obtained for different values of it.

X. Summary and Discussion

A. Summary

To see the forest from the trees, a summary of the steps one follows, when using our approach to XAI for fuzzy systems, is presented as follows:

1) Design your fuzzy system (Mamdani, TSK, etc.) in (7) to achieve quantitative performance goals. This will provide the number of antecedent variables (p), MFs for them [number (Q) and shape], rules [number (M) and their exact structure], and rule consequents (number, MF or function).

2) Perform LA for each input variable and output variable. To do this, you have to decide on the granularity for LA, collect data from a group of subjects about each word (or provide it yourself) [48], and then map that data into a fuzzy set for each word. The results are a codebook for each variable and output, each of whose entries is a two-element pair, (Word, MF).

3) Compute the Jaccard similarity measure between each antecedent variable's MFs (from Step 1) and the MF in the codebook for that variable (from Step 2), and then map each antecedent variable's MF into the codebook word with the largest similarity value.

4) For a given value $\mathbf{x}'$of the input $\mathbf{x}$.
   a) Compute $y(\mathbf{x}')$ using the results from Step 1.
   b) Determine how many and exactly which rules are active, using rule partitions (see Section VIII).
   c) Use one of the approaches that are described in Section IX (they depend on which kind of fuzzy system has been designed in Step 1) to determine the subset of uniquely different compound antecedent linguistic expressions (making use of Steps 2 and 3) that are used to explain $y(\mathbf{x}')$.[17]
   d) Express each uniquely different linguistic explanation using either (6a) or (6b).

B. Discussion

1) Our XAI approach is applicable to any set of rules, even rules that come after rule-reduction (e.g., [59]), or from Pareto-optimal designs that tradeoff performance (e.g., RMSE) with complexity (e.g., number of rules, [5], [7]).

2) We advocate *achieve acceptable performance first and then explain it*.

3) We do not advocate using the linguistic antecedent combinations obtained from our approach followed by a redesign of the original fuzzy system to obtain acceptable performance, because explanation is qualitative, whereas performance is quantitative.

4) Limitations of our XAI approach are: LA requires data collection about the words that are used in the explanations, and this requires more effort; type-1 fuzzy sets are unable to capture both the intra and interuncertainties that will be in the collected word data, but type-2 fuzzy sets can do this; so, our multi-stage approach needs to be extended to T2 fuzzy sets; and if different semantics are used, then LA has to be done for each of them.

XI. Comprehensive Example

A. Preliminaries

This section provides a comprehensive example that illustrates our approaches to XAI. So that it is a non-trivial example, rules have three input variables, each of which is described by five MFs (see Fig. 2), so that there can be 125 rules. MF parameters for all of these MFs are provided in Table SM-3 of the SM. Equations for their left and right legs are in Table SM-4.

For LA, to keep things simple, the three words and MFs that are in Example 1, and are depicted in Fig. 3, were used. Equations for the left and right legs of these MFs are in Table SM-5.

Jaccard similarities (s_J) were computed for each of the 15 MFs in Fig. 2 and three MFs in Fig. 3, and are given in Table

[15] Equations(28) and (29) have a graphical interpretation. On a plot of the MFs for the output variable y, locate $y_c(\mathbf{x}')$ on the y-axis and draw a vertical line. This line intersects one or more of the MFs, as given by (28). Equation (29) chooses the "winner" as the MF with the largest grade. What may be new is the recognition that (28), which in the past has been done graphically, is actually a similarity computation.

[16] The constraint that $t_2 > 0.5$ seems intuitively correct for similarity.

[17] A result of the comprehensive example that is in Section XI is a two-stage approach that modifies Step c.

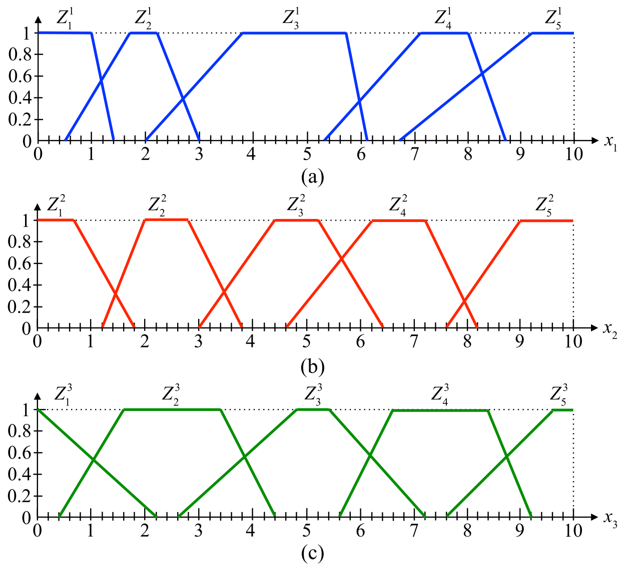


Fig. 2. Five MFs for each of the three inputs. (a) x_1. (b) x_2. (c) x_3.

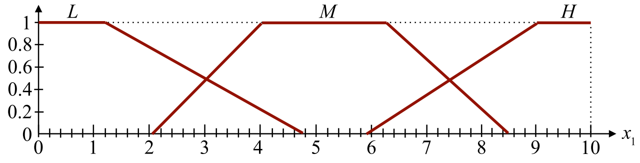


Fig. 3. LA MFs for *low* (*L*), *moderate* (*M*), and *high* (*H*).

SM-6. The largest s_J for each MF provides its LA, and is

$$\begin{cases} x_1 : Z_1^1 \to L,\ Z_2^1 \to L,\ Z_3^1 \to M,\ Z_4^1 \to H,\ Z_5^1 \to H \\ x_2 : Z_1^2 \to L,\ Z_2^2 \to L,\ Z_3^2 \to M,\ Z_4^2 \to M,\ Z_5^2 \to H \\ x_3 : Z_1^3 \to L,\ Z_2^3 \to L,\ Z_3^3 \to M,\ Z_4^3 \to H,\ Z_5^3 \to H \end{cases} \tag{30}$$

$\mathbf{x}'$ was chosen so that two MFs would be active for each of its three elements, meaning that eight rules are active [$M_s(\mathbf{x}') = 8$]. The specific $\mathbf{x}'$chosen is

$$\begin{cases} x_1' = 2.4\ (Z_2^1, Z_3^1 \text{ are active}) \\ x_2' = 5.4\ (Z_3^2, Z_4^2 \text{ are active}) \\ x_3' = 9\ (Z_4^3, Z_5^3 \text{ are active}) \end{cases} \tag{31}$$

Which MFs are *active* in (31) is determined by locating each component of $\mathbf{x}'$ on its respective plot in Fig. 2, and projecting a vertical line at that point upwards to see which MFs it intersects.

To determine exactly which eight rules are fired, first-order rule partition figures were created for the MFs of x_1, x_2 and x_3, beginning with Figs. 2(a)–(c), respectively. Each resulting figure closely resembles Fig. 1, and is therefore not included here. From those figures, it follows that:

1) $x_1' = 2.4$ is in partition[18] 4\2, so $k_1' = 4$ and Z_2^1, Z_3^1 are active;
2) $x_2' = 5.4$ is in partition 6\2, so $k_2' = 6$ and Z_3^2, Z_4^2 are active; and
3) $x_3' = 9$ is in partition 8\2, so $k_3' = 8$ and Z_4^3, Z_5^3 are active.

Knowing the indexes on these three pairs of active MFs, one can compute the rule index for each fired rule, by using (16), which for three variables reduces to

$$l = 25(j_1 - 1) + 5(j_2 - 1)5 + j_3. \tag{32}$$

Table IV summarizes the computations; the resulting l-values are collected into the following set [$M_s(\mathbf{x}') = 8$].

TABLE IV
COMPUTATION OF RULE INDEXES, L, WHEN $\mathbf{x}' = \text{col}(2.4, 5.4, 9)$

j_1	j_2	j_3	$25(j_1-1)$	$5(j_2-1)$	l
2	3	4	25	10	39
2	3	5	25	10	40
2	4	4	25	15	44
2	4	5	25	15	45
3	3	4	50	10	64
3	3	5	50	10	65
3	4	4	50	15	69
3	4	5	50	15	70

TABLE V
CALCULATIONS OF $f^l(\mathbf{x}')$ WHEN $\mathbf{x}' = col(2.4, 5.4, 9)$, AND VALUE OF RULE CONSEQUENT, c^l

l	$f^l(\mathbf{x}')$ formula	$f^l(\mathbf{x}')$	c^l
39	$Z_{\underline{2}}^1(2.4)\cdot Z_{\underline{3}}^2(5.4)\cdot Z_{\underline{4}}^3(9)$	0.156	6
40	$Z_{\underline{2}}^1(2.4)\cdot Z_{\underline{3}}^2(5.4)\cdot Z_{\underline{5}}^3(9)$	0.437	9
44	$Z_{\underline{2}}^1(2.4)\cdot Z_{\underline{4}}^2(5.4)\cdot Z_{\underline{4}}^3(9)$	0.094	8.3
45	$Z_{\underline{2}}^1(2.4)\cdot Z_{\underline{4}}^2(5.4)\cdot Z_{\underline{5}}^3(9)$	0.263	8
64	$Z_{\underline{3}}^1(2.4)\cdot Z_{\underline{3}}^2(5.4)\cdot Z_{\underline{4}}^3(9)$	0.046	6.7
65	$Z_{\underline{3}}^1(2.4)\cdot Z_{\underline{3}}^2(5.4)\cdot Z_{\underline{5}}^3(9)$	0.129	7
69	$Z_{\underline{3}}^1(2.4)\cdot Z_{\underline{4}}^2(5.4)\cdot Z_{\underline{4}}^3(9)$	0.028	8.4
70	$Z_{\underline{3}}^1(2.4)\cdot Z_{\underline{4}}^2(5.4)\cdot Z_{\underline{5}}^3(9)$	0.078	7.4

$$S_s(\mathbf{x}') = \{39, 40, 44, 45, 64, 65, 69, 70\} \tag{33}$$

In this example, the product t-norm was used, so that the firing level for each rule is the product of three MFs, i.e., [$l \in S_s(\mathbf{x}')$]

$$f^l(\mathbf{x}') = \prod\nolimits_{i=1}^{3} \mu_{F_i^l}(x'_i) = \prod\nolimits_{i=1}^{3} Z_{j_i(l)}^i(x_i'). \tag{34}$$

In (34), $j_i(l)$ was found by using the l-value in the last column of Table IV, and then reading off its j_1, j_2 and $j_3$values, that are in the first three columns of that table.

$Z_{j_i(l)}^i(x_i')$ was then found by locating x_i' on its respective part of Fig. 2, determining whether it was a left or right leg of Z_j^i, and then using the appropriate formula from Table SM-4. Results for (34) are given in Table V. A *rank ordering of the computed firing levels* [used below in (37) and (41)] is

$$\begin{cases} l = 40 : f^{40}(\mathbf{x}') = 0.437 : & l = 44 : f^{44}(\mathbf{x}') = 0.094 \\ l = 45 : f^{45}(\mathbf{x}') = 0.263 : & l = 70 : f^{70}(\mathbf{x}') = 0.078 \\ l = 39 : f^{39}(\mathbf{x}') = 0.156 : & l = 64 : f^{64}(\mathbf{x}') = 0.046 \\ l = 65 : f^{65}(\mathbf{x}') = 0.129 : & l = 69 : f^{69}(\mathbf{x}') = 0.028 \end{cases}. \tag{35}$$

[18] See the caption to Fig. 1 for an explanation of the backslash notation.

B. Mamdani With COS Defuzzification

In order to compute $y_{COS}(\mathbf{x}')$ using (22) and (23), numerical values are needed for the consequent numbers c^l. Our chosen[19] values are in the last column of Table V. It is then straightforward to show, that

$$\sum_{l \in S_s(\mathbf{x}')} f^l(\mathbf{x}') = 1.231 \tag{36}$$

$$\begin{cases} w^{40}(\mathbf{x}') = 0.437/1.231 = 0.355 \rightarrow & c^{40}w^{40}(\mathbf{x}') = 3.195 \\ w^{45}(\mathbf{x}') = 0.263/1.231 = 0.214 \rightarrow & c^{45}w^{45}(\mathbf{x}') = 1.712 \\ w^{39}(\mathbf{x}') = 0.156/1.231 = 0.126 \rightarrow & c^{39}w^{39}(\mathbf{x}') = 0.756 \\ w^{65}(\mathbf{x}') = 0.129/1.231 = 0.105 \rightarrow & c^{65}w^{65}(\mathbf{x}') = 0.735 \\ w^{44}(\mathbf{x}') = 0.094/1.231 = 0.076 \rightarrow & c^{44}w^{44}(\mathbf{x}') = 0.631 \\ w^{70}(\mathbf{x}') = 0.078/1.231 = 0.063 \rightarrow & c^{70}w^{70}(\mathbf{x}') = 0.466 \\ w^{64}(\mathbf{x}') = 0.046/1.231 = 0.037 \rightarrow & c^{64}w^{64}(\mathbf{x}') = 0.248 \\ w^{69}(\mathbf{x}') = 0.028/1.231 = 0.023 \rightarrow & c^{69}w^{69}(\mathbf{x}') = 0.193 \end{cases} \tag{37}$$

$$y_{\mathrm{COS}}(\mathbf{x}') = \sum_{l \in S_s(\mathbf{x}')} c^l w^l(\mathbf{x}') = 7.936. \tag{38}$$

1) Nearest Consequent Neighbor Explanations: Comparing $y_{\mathrm{COS}}(\mathbf{x}') = 7.936$ with the eight values of c^l that are in Table V, observe that $m_1(\mathbf{x}') = 8 - 7.936 = 0.064$ and $l_{m_1(\mathbf{x}')} = 45$. Examining the numerical c^l values in Table V, the next c^l values closest to 7.936 are 8.3 ($l = 44$) and 8.4 ($l = 69$); to include them in I_1, ε_1 would have to be &ge; 0.4. If ε_1 is chosen to be larger and larger, more rule firing levels will be chosen. Here only three compound antecedents were used, and from Table V and (30), one finds [$S_{ss}(\mathbf{x}') = \{45, 44, 69\}$ and $M_{ss}(\mathbf{x}') = 3$]

$$\begin{cases} l = 45 : Z_2^1 \wedge Z_4^2 \wedge Z_5^3 & \rightarrow \ L \wedge M \wedge H \\ l = 44 : Z_2^1 \wedge Z_4^2 \wedge Z_4^3 & \rightarrow \ L \wedge M \wedge H \\ l = 69 : Z_3^1 \wedge Z_4^2 \wedge Z_4^3 & \rightarrow \ M \wedge M \wedge H \end{cases} \tag{39}$$

Observe that $l = 45$ and $l = 44$ give exactly the same linguistic explanation, namely

$$x_1 \text{ is } \textit{low}, \text{ and } x_2 \text{ is } \textit{moderate} \text{ and } x_3 \text{ is } \textit{high} \text{ are symptomatic of } y_{\mathrm{COS}}(\mathbf{x}') = 7.936. \tag{40a}$$

Additionally, $l = 69$ has the following linguistic explanation:

$$x_1 \text{ is } moderate, \text{ and } x_2 \text{ is } moderate \text{ and } x_3 \text{ is } \textit{high} \text{ are symptomatic of } y_{\mathrm{COS}}(\mathbf{x}') = 7.936. \tag{40b}$$

Equation (40a) and (40b) demonstrates that $M_{sss}(\mathbf{x}') = 2$.

2) Most Significant Contributing Terms Explanations: Using (37) and (38), one finds [see (24)]

$$\begin{cases} \mathrm{P}_{\mathrm{COS}}^{40}(\mathbf{x}') = \frac{3.195}{7.936}10^2 = 40.26\% \\ \mathrm{P}_{\mathrm{COS}}^{45}(\mathbf{x}') = \frac{1.712}{7.936}10^2 = 21.57\% \\ \mathrm{P}_{\mathrm{COS}}^{39}(\mathbf{x}') = \frac{0.756}{7.936}10^2 = 9.53\% \\ \mathrm{P}_{\mathrm{COS}}^{65}(\mathbf{x}') = \frac{0.735}{7.936}10^2 = 9.26\% \\ \mathrm{P}_{\mathrm{COS}}^{44}(\mathbf{x}') = \frac{0.631}{7.936}10^2 = 7.95\% \\ \mathrm{P}_{\mathrm{COS}}^{70}(\mathbf{x}') = \frac{0.466}{7.936}10^2 = 5.87\% \\ \mathrm{P}_{\mathrm{COS}}^{64}(\mathbf{x}') = \frac{0.248}{7.936}10^2 = 3.13\% \\ \mathrm{P}_{\mathrm{COS}}^{69}(\mathbf{x}') = \frac{0.193}{7.936}10^2 = 2.43\%. \end{cases} \tag{41}$$

Examining these rank-ordered $\mathrm{P}_{\mathrm{COS}}^l(\mathbf{x}')$ numbers, it is clear that the first two are dominant, with a total of 61.83%, which seems too low to convince others that only these two terms should be considered; hence, the first four terms, are used with a total of 80.62%; consequently, from Table V and (30), one finds [$S_{ss}(\mathbf{x}') = \{40, 45, 39, 65\}$ and $M_{ss}(\mathbf{x}') = 4$]

$$\begin{cases} l = 40 : Z_2^1 \wedge Z_3^2 \wedge Z_5^3 \rightarrow & L \wedge M \wedge H \\ l = 45 : Z_2^1 \wedge Z_4^2 \wedge Z_5^3 \rightarrow & L \wedge M \wedge H \\ l = 39 : Z_2^1 \wedge Z_3^2 \wedge Z_4^3 \rightarrow & L \wedge M \wedge H \\ l = 65 : Z_3^1 \wedge Z_3^2 \wedge Z_5^3 \rightarrow & M \wedge M \wedge H \end{cases} \tag{42}$$

Comparing (42) and (39), even though they only share one common value of l, $l = 45$, the same linguistic explanations have been obtained (40a,b) [$M_{sss}(\mathbf{x}') = 2$].

Examining $\mathrm{P}_{\mathrm{COS}}^l(\mathbf{x}')$ for $l = 44, 70, 64, 69$, one finds that MD = 19.38%. The antecedent combinations and their linguistic equivalents that are associated with MD are

$$\begin{cases} l = 44 : Z_2^1 \wedge Z_4^2 \wedge Z_4^3 \rightarrow & L \wedge M \wedge H \\ l = 70 : Z_3^1 \wedge Z_4^2 \wedge Z_5^3 \rightarrow & M \wedge M \wedge H \\ l = 64 : Z_3^1 \wedge Z_3^2 \wedge Z_4^3 \rightarrow & M \wedge M \wedge H \\ l = 69 : Z_3^1 \wedge Z_4^2 \wedge Z_4^3 \rightarrow & M \wedge M \wedge H \end{cases} \tag{43}$$

Observe that there are no new antecedent combinations! This means that, at least for this example, one could have used all eight active antecedent combinations, and did not have to use either of the above two approaches. We conjecture that this has occurred because of the low granularity of the three words used for our LA, and that if more words had been used this would not have occurred.

C. Mamdani With Centroid Defuzzification

In centroid defuzzification, each rule is assigned its own consequent MF, whose parameters are all tuned during training. Trapezoidal MFs were chosen for the consequents, each with their flat tops located (not symmetrically) about c^l (the consequent locations for the Mamdani fuzzy system with COS defuzzification). Our rationalization for doing this is that an optimized centroid defuzzified Mamdani fuzzy system should have its consequent MFs covering those of a COS defuzzified Mamdani fuzzy system, since both systems are optimizing the same objective function. The eight consequent MFs are in Fig. 4 (in red), and the formulas for each of their left and right legs are in Table SM-7.

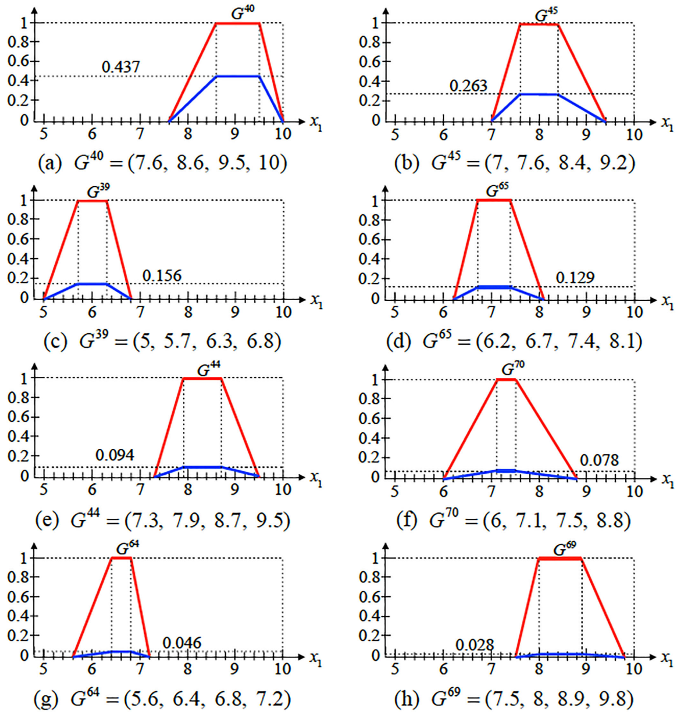
In order to compute $y_c(\mathbf{x}')$ in (4), one must first compute $f^l(\mathbf{x}') \cdot \mu_{G^l}(y)$ for $y \in Y$. This is done in Fig. 4 (in blue). Then one has to compute $\mu_B(y|\mathbf{x}') = \max_{l \in S_s(\mathbf{x}')} f^l(\mathbf{x}') \cdot \mu_{G^l}(y)$. This result is depicted in Fig. 5, from which it follows (by means of sampling and numerical integration), that

$$y_c(\mathbf{x}') = 8.7182. \tag{44}$$

In this section, the same resolution for the LA of output y was used as was assumed for the three inputs, namely L, M and H; and, the MFs for these terms are the ones in Fig. 3 with their associated equations.

Similarities between $G^l (l = 1, \ldots, 8)$ and L, M and H can be accomplished visually, by comparing each G^l MF in Fig. 4 with the three MFs in Fig. 3, and are [$S_s(\mathbf{x}')$ is given by (33) and

[19] During an actual design, where training data are available, these consequent numbers would have resulted from the tuning process. Here, they were chosen so that the most dominant rule ($l = 40$) would be most heavily weighted in (22), and that all other rules would be reasonably weighted.



(a) $G^{40} = (7.6,\ 8.6,\ 9.5,\ 10)$ (b) $G^{45} = (7,\ 7.6,\ 8.4,\ 9.2)$

(c) $G^{39} = (5,\ 5.7,\ 6.3,\ 6.8)$ (d) $G^{65} = (6.2,\ 6.7,\ 7.4,\ 8.1)$

(e) $G^{44} = (7.3,\ 7.9,\ 8.7,\ 9.5)$ (f) $G^{70} = (6,\ 7.1,\ 7.5,\ 8.8)$

(g) $G^{64} = (5.6,\ 6.4,\ 6.8,\ 7.2)$ (h) $G^{69} = (7.5,\ 8,\ 8.9,\ 9.8)$

Fig. 4. Consequent MFs (in red) for each of the eight fired-rules, and $f^l(\mathbf{x}')\cdot G^l$ (in blue) for $l = 40, 45, 39, 65, 44, 70, 64$ and 69.The firing level for each rule is also shown. (a) $G^{40} = (7.6, 8.6, 9.5, 10)$. (b) $G^{45} = (7, 7.6, 8.4, 9.2)$. (c) $G^{39} = (5, 5.7, 6.3, 6.8)$. (d) $G^{65} = (6.2, 6.7, 7.4, 8.1)$. (e) $G^{44} = (7.3, 7.9, 8.7, 9.5)$. (f) $G^{70} = (6, 7.1, 7.5, 8.8)$. (g) $G^{64} = (5.6, 6.4, 6.8, 7.2)$. (h) $G^{69} = (7.5, 8, 8.9, 9.8)$.

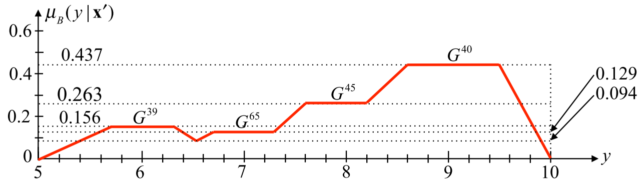


Fig. 5. Union of eight fired-rule output sets (only four contribute to it).

$M_s(\mathbf{x}') = 8]$

$$\begin{cases} G^{40} \to H,\ G^{45} \to H,\ G^{39} \to M,\ G^{65} \to M \\ G^{44} \to H,\ G^{70} \to M,\ G^{64} \to M,\ G^{69} \to H. \end{cases} \tag{45}$$

To map $y_c(\mathbf{x}') = 8.7182$ into a word using similarity, it was located on the Fig. 3 MF plots, from which it was easily established, that

$$y_c(\mathbf{x}') = 8.7182 \to H. \tag{46}$$

Observe, from (45), that four of the consequent MFs have been mapped into H. It is those four consequent words that are most similar to $y_c(\mathbf{x}') = 8.7182 \to H$ (the similarity value of H to H is unity). Consequently, the subset of most similar compound antecedents to[20] $y_c(\mathbf{x}')$ is: G^{40}, G^{45}, G^{44} and G^{69}. Using Table V and (30), one finds $[S_{ss}(\mathbf{x}') = \{40, 45, 44, 69\}$ and $M_{ss}(\mathbf{x}') = 4]$:

$$\begin{cases} l = 40 : Z_2^1 \wedge Z_3^2 \wedge Z_5^3 \to\ L \wedge M \wedge H \\ l = 45 : Z_2^1 \wedge Z_4^2 \wedge Z_5^3 \to\ L \wedge M \wedge H \\ l = 44 : Z_2^1 \wedge Z_4^2 \wedge Z_4^3 \to\ L \wedge M \wedge H \\ l = 69 : Z_3^1 \wedge Z_4^2 \wedge Z_4^3 \to\ M \wedge M \wedge H \end{cases} \tag{47}$$

These are the same linguistic results that were obtained above in (39) and (42), encapsulated by (40a) and (40b) $[M_{sss}(\mathbf{x}') = 2]$.

D. *Observations*

While it is dangerous to draw conclusions from results for only one value of $\mathbf{x}'$, it is worthwhile to observe that, for $\mathbf{x}' = \mathrm{col}(2.4, 5.4, 9)$, all approaches used to explain $y_{\mathrm{COS}}(\mathbf{x}')$ or $y_c(\mathbf{x}')$ led to only two linguistic explanations, the most dominant being $L \wedge M \wedge H$, followed by $M \wedge M \wedge H$. What is surprising is that these results could have been obtained directly simply by using the similarity results for the antecedents of the eight fired rules. The results don't seem to depend on the numerical value of the output.

Reducing eight compound symptoms to two is impressive. So, why is this occurring? We believe it is due to only using L, M and H as linguistic terms for each variable. Doing this force more than one of a variable's five MFs to be similar to the same linguistic term. If, e.g., we had used five such terms then this would not have occurred and then one would not get the reduction from eight to two symptoms.

Based on this observation, our example has revealed a new *two-stage method for XAI*

1) Given $\mathbf{x}'$, establish the exact number and form of each fired rule's compound antecedent, and convert them into linguistic expressions. If the number of such linguistic expressions is small (e.g., ≤ 3) then STOP; otherwise,
2) Use one of the Section IX's methods for finding a smaller subset of rule compound antecedents that explain $y_{\mathrm{COS}}(\mathbf{x}')$ or $y_c(\mathbf{x}')$.

This example can also be used to explain how to estimate the *quality of the explanations*. Each of the eight active rules given in Table V has non-zero possibility measures with the input. However, the MFs used to compute their firing levels, [see Fig. 2(a)–(c)], have finer granularity than the MFs used to describe them linguistically (see Fig. 3). While the explanations are driven by the active rules, the quality of the explanation is limited by the richness of the LA term set.

To select the most appropriate explanation from $L \wedge M \wedge H$ and $M \wedge M \wedge H$, explanation quality can be assessed by evaluating the possibility measures of both explanations with the input $\mathbf{x}' = \mathrm{col}(2.4, 5.4, 9)$ using the LA MFs in Fig. 3, i.e.,

$$\begin{cases} \mathrm{Quality}(L \wedge M \wedge H) = \min[L(2.4), M(5.4), H(9)] \\ \quad = \min(0.65, 1, 1) = 0.65 \\ \mathrm{Quality}(M \wedge M \wedge H) = \min[M(2.4), M(5.4), H(9)] \\ \quad = \min(0.18, 0.65, 1) = 0.18. \end{cases} \tag{48}$$

Equation (48) leads us to select $L \wedge M \wedge H$over $M \wedge M \wedge H$.

This example has also demonstrated that even a flat rule-based fuzzy system can have simple explanations.

[20] No threshold was needed.

XII. Conclusions and Suggestions for Future Research

This article has focused on answering the question: When $\mathbf{x} = \mathbf{x}'$, *is it valid to say that the output of a T1 rule-based fuzzy system*, $y(\mathbf{x}') = f(\mathbf{x}')$, *can be described by the rules that led to it?* It has done this by critically thinking about the many decisions that are made for some popular T1 fuzzy systems, each of which ultimately can be expressed generically, as $y(\mathbf{x}) = f(\mathbf{x})$. Those decisions have in the past only been associated with the design of the fuzzy system and its subsequent use to solve a forward problem, i.e., given $\mathbf{x} = \mathbf{x}'$ compute $y(\mathbf{x}') = f(\mathbf{x}')$.

This article has explained that when the constraint of explainability (XAI) is imposed on a rule-based fuzzy system then one is trying to solve the following *system identification problem*: given $\mathbf{x} = \mathbf{x}'$ and $y(\mathbf{x}')$, establish the rule antecedent associations that can be used to explain $y(\mathbf{x}')$.

Critical thinking has led to the following:

1) Explanation of why for XAI one should only *associate* rule antecedents (*symptoms*), and not an IF-THEN rule, with $y(\mathbf{x}')$, using a linguistic explanation, such as (6a) or (6b).
2) Recognition that, to explain $y(\mathbf{x}')$, words (W) must be introduced for rule antecedents and consequents using LA, which results in *codebooks* for each rule's antecedent and consequent variable, whose elements are pairs (W, MF_W).
3) Realization that similarity computations play a very important role in XAI; they connect codebook words and their MFs to rule antecedent MFs, consequents and even to $y(\mathbf{x}')$.
4) Recognition that triangle and trapezoidal MFs may be more suitable for XAI than Gaussian MFs, because the former partition the state space unambiguously (they are exactly zero), whereas the latter partition the state space ambiguously (they are never exactly zero).
5) Development of a novel multi-step procedure for $\mathbf{x} = \mathbf{x}'$ that determines exactly which rule antecedents occur, how many of them there are, and how to reduce that number to a smaller number of distinctly different linguistic explanations.
6) Proposal of a novel way to estimate the quality of the explanations.

Returning to the Section I question "When $\mathbf{x} = \mathbf{x}'$, *is it valid to say that the output of a T1 rule-based fuzzy system*, $y(\mathbf{x}') = f(\mathbf{x}')$, *can be described by the rules that led to it?*", our answer is NO. Instead, one can only say something like (6a) or (6b), which do not mention the word rule.

Some future research suggestions are as follows:

1) Study how and when to choose a relatively small number of antecedent variables, so that antecedent associations are understandable by a human, e.g., should this be done at the beginning of a design by using a hierarchical architecture, or by some sort of antecedent reduction technique at the end of a design, or during the design?
2) Establish exactly how XAI can be accomplished when a hierarchical architecture is used. Usually a hierarchical architecture is driven by the desire to perform problem decomposition. Therefore, the output of the intermediate rule sets in the architecture can be used to provide a hierarchical explanation.
3) Extend the critical thinking of this article from T1 to interval type-2 (IT2) and general T2 (GT2) fuzzy sets, in which linguistic uncertainties are modeled by either IT2 or GT2 rule antecedent MFs, so that (6a) and (6b) would then include some measure of uncertainty about the explanation.
4) Fix all rule antecedent MFs via LA, and then only tune consequent MF parameters (see, also, [59]). Such a design would provide perfect correlation between its antecedent MFs and the LA used to explain an output.
5) Study the robustness (sensitivity) of XAI statements to truncation levels of Gaussian MFs, and the margin and threshold parameters for the three fuzzy systems, as described in Section IX, and in Section VI of the SM.
6) Extend the critical thinking of this article to rule-based fuzzy systems that are used for classification, to the two other kinds of XAI described by Hagras [19] (i.e., deep explanation and interpretable models), and to autoencoder kinds of models [15], when rule-based fuzzy systems are used for them.
7) Each first-order rule partition contains a set of active rules, whose compound antecedents map (by means of LA and similarity) into words. The result of doing this could be called an *explainability partition*. For our example in Section XI, there are $9^3 = 729$ first-order rule partitions, but only $3^3 = 27$ possible linguistic explanations, so many of the 729 explainability partitions are the same. Collecting all of the regions in $X_1 \times X_2 \times X_3$ that have the same explanation may provide new insights into XAI for a particular application. Extending this idea to larger dimension applications, as well as to more finely granulated LA are also worthy of study.

Acknowledgment

The authors would like to thank the reviewers of this article who made many useful comments and suggestions that have (hopefully) led to an improved version of this article.

References

[1] M. Setnes, and H. Roubos, "GA-fuzzy modeling and classification: Complexity and performance," *IEEE Trans, Fuzzy Syst.*, vol. 8, no. 5, pp. 509–522, Oct. 2000.

[2] H. Ishibuchi, and T. Yamamoto, "Interpretability issues in fuzzy genetic-based machine learning for linguistic modeling," in *Modeling With Words (Lecture Notes in Computer Sciences)*, vol. 2873. J. Lawry, J. Shanahan, and A. L. Ralescu, Eds., Berlin, Germany: Springer, 2003, pp. 209–228.

[3] J. Casillas, O. Cordon, F. Herrera and L. Magdalena, Eds., *Interpretability Issues in Fuzzy Modeling*, Berlin, Germany: Springer, 2003.

[4] H. Ishibuchi, T. Nakashima, and M. Nii, *Classification and Modeling With Linguistic Information Granules: Advanced Approaches to Linguistic Data Mining*, Berlin, Germany: Springer, 2004.

[5] H. Ishibuchi, "Multiobjective genetic fuzzy systems: Review and future research directions," in *Proc. IEEE Int. Conf. Fuzzy Syst.*, 2007, pp. 1–6.

[6] M. J. Gacto, R. Alcala, and F. Herrera, "Interpretability of linguistic fuzzy rule-based systems: An overview of interpretability measures," *Inf. Sci.*, vol. 181, pp. 4340–4360, Oct. 2011.

[7] O. Cordon, “A historical review of evolutionary learning methods for Mamdani-type fuzzy rule-based systems: Designing interpretable genetic fuzzy systems,” *Int. J. Approx. Reasoning*, vol. 52, pp. 894–913, Sep. 2011.

[8] P. Ducange and F. Marcelloni, “Multi-objective evolutionary fuzzy systems,” in *Proc. 9th Int. Workshop Fuzzy Log. Appl.*, 2011, pp. 83–90.

[9] M. Galende-Hernandez, G. I. Sainz-Palmero, and M. J. Fuente-Aparicio, “Complexity reduction and interpretability improvement for fuzzy rule systems based on simple interpretability measures and indices by bi-objective evolutionary rule selection,” *Soft Comput.*, vol. 16, pp. 451–470, Mar. 2012.

[10] M. Fazzolari, R. Alcala, Y. Nojima, H. Ishibuchi, and F. Herrera, “A review of the application of multi-objective evolutionary fuzzy systems: Current issues and further directions,” *IEEE Trans. Fuzzy Syst.*, vol. 21, no. 1, pp. 45–65, Feb. 2013.

[11] D. Garcia, J. C. Gamez, A. Gonzalez, and R. Perez, “An interpretability improvement for fuzzy rule bases obtained by the iterative learning approach,” *Int. J. Approx. Reasoning*, vol. 67, pp. 37–58, Dec. 2015.

[12] M. Antonelli, D. Bernardo, H. Hagras, and F. Marcelloni, “Multi-objective evolutionary optimization of type-2 fuzzy rule-based systems for financial data classification,” *IEEE Trans. Fuzzy Syst.*, vol. 25, no. 2, pp. 249–264, Apr. 2017.

[13] J. Adams and H. Hagras, “A type-2 fuzzy logic approach to explainable AI for regulatory compliance, fair customer outcomes and market stability in the global financial sector,” in *Proc. IEEE Int. Conf. Fuzzy Syst.*, 2020, pp. 1–8.

[14] K. Bolat and T. Kumbasar, “Interpreting variational autoencoders with fuzzy logic: A step towards interpretable deep learning based fuzzy classifiers,” in *Proc. IEEE Int. Conf. Fuzzy Syst.*, 2020, pp. 1–7.

[15] R. Chimatapu, H. Hagras, M. Kern, and G. Owusu, “Hybrid deep learning type-2 fuzzy logic systems for explainable AI,” in *Proc. IEEE Int. Conf. Fuzzy Syst.*, 2020, pp. 1–6.

[16] P. D’Alterio, J. Garibaldi, and R. John, “Constrained interval type-2 classification systems for explainable AI,” in *Proc. IEEE Int. Conf. Fuzzy Syst.*, 2020, pp. 1–8.

[17] J. Rozman, H. Hagras, J. Andreu-Perez, D. Clarke, B. Muller, and S. Fitz, “Privacy-preserving gesture recognition with explainable type-2 fuzzy logic reasoning,” in *Proc. IEEE Int. Conf. Fuzzy Syst.*, 2020, pp. 1–8.

[18] J. R. Trillo, A. Fernandez, and F. Herrera, “HFER: Promoting explainability in fuzzy systems via fuzzy exception rules,” in *Proc. IEEE Int. Conf. Fuzzy Syst.*, 2020, pp. 1–8.

[19] H. Hagras, “Toward human-understandable, explainable aI,” *Computer*, vol. 51, no. 9, pp. 32–40, Sep. 2018.

[20] E. Hüllermeier, “Does machine learning need fuzzy logic?,” *Fuzzy Sets Syst.,* vol. 281, pp. 292–299, Dec. 2015.

[21] J. M. Mendel, *Uncertain Rule-Based Fuzzy Systems: Introduction and New Directions*, 2nd ed., Cham, Switzerland: Springer, 2017.

[22] L. A. Zadeh, “Outline of a new approach to the analysis of complex systems and decision processes,” *IEEE Trans. Syst., Man, Cybern.,* vol. SMC-3, no. 1, pp. 28–44, Jan. 1973.

[23] T. Takagi, and M. Sugeno, “Fuzzy identification of systems and its application to modeling and control,” *IEEE Trans. Syst., Man Cybern.*, vol. 15, no. 1, pp. 116–132, Jan./Feb. 1985.

[24] M. Sugeno and G. T. Kang, “Structure identification of fuzzy model,” *Fuzzy Sets Syst.*, vol. 28, pp. 15–33, Oct. 1988.

[25] E. H. Mamdani, “Applications of fuzzy algorithms for simple dynamic plant,” *Proc. IEEE*, vol. 121, no. 12, pp. 1585–1588, Dec. 1974.

[26] M. Sugeno and T. Yasukawa, “A fuzzy-logic-based approach to qualitative modeling,” *IEEE Trans. Fuzzy Syst.*, vol. 1, no. 1, pp. 7–30, Feb. 1993.

[27] G. J. Klir and B. Yuan, *Fuzzy Sets and Fuzzy Logic: Theory and Applications*, Upper Saddle River, NJ, USA: Prentice-Hall, 1995.

[28] D. Dubois and H. Prade, “Fuzzy sets in approximate reasoning, part 1: Inference with possibility distributions,” *Fuzzy Sets Syst.*, vol. 40, pp. 143–202, Mar. 1991.

[29] H. Dubois, J. Lang, and H. Prade, “Fuzzy sets in approximate reasoning, part 2: Logical approaches,” *Fuzzy Sets Syst.*, vol. 40, pp. 203–244, Mar. 1991.

[30] R. Belohlavek, J. W. Dauben, and G. J. Klir, *Fuzzy Logic and Mathematics: A Historical Perspective*, New York, NY, USA: Oxford Univ. Press, 2017.

[31] L. A. Zadeh, “The concept of a linguistic variable and its application to approximate reasoning–I, II and III” *Inf. Sci.*, vol. 8, pp. 199–249, vol. 8, pp. 301–357 and vol. 9, pp. 43–80, 1975.

[32] J. M. Mendel, “Fuzzy logic systems for engineering: A tutorial,” *IEEE Proc.,* vol. 83, no. 3, pp. 345–377, Mar. 1995.

[33] E. H. Mamdani, “Fuzzy control: A misconception of theory and application,” *IEEE Expert*, vol. 4, pp. 17–18, Aug. 1994.

[34] J. M. Mendel, *Uncertain Rule-Based Fuzzy Logic Systems: Introduction and New Directions*, Upper Saddle River, NJ: Prentice-Hall, 2001.

[35] D. Wu and J. M. Mendel, “Recommendations on designing practical interval type-2 fuzzy systems,” *Eng. Appl. Artif. Intell.,* vol. 85, pp. 182–193, Oct. 2019.

[36] H. Ying, *Fuzzy Control and Modeling: Analytical Foundations and Applications*, Piscataway, NJ, USA: IEEE Press, 2000.

[37] L.-X. Wang, *A Course in Fuzzy Systems and Control*, Upper Saddle River, NJ, USA: Prentice-Hall, 1997.

[38] S.-M. Chen and Y.-C. Chang, “Multi-variable fuzzy forecasting based on fuzzy clustering and fuzzy rule interpolation techniques,” *Inf. Sci.*, vol. 180, pp. 4772–4783, Dec. 2010.

[39] O. Duru, “A fuzzy integrated logical forecasting model for dry bulked shipping index forecasting: An improved fuzzy time series approach,” *Expert Syst. Appl.*, vol. 37, no. 7, pp. 5372–5380, Jul. 2010.

[40] R. J. G. B. Campello and W. C. do Amaral, “Hierarchical fuzzy relational models: Linguistic interpretation and universal approximation,” *IEEE Trans. Fuzzy Syst.*, vol. 14, no. 3, pp. 446–453, Jun. 2006.

[41] L.-X. Wang, “Analysis and design of hierarchical fuzzy systems,” *IEEE Trans. Fuzzy Syst.*, vol. 7, no. 5, pp. 617–624, Oct. 1999.

[42] D. Wang, X.-J. Zeng, and J. A. Keane, “A survey of hierarchical fuzzy systems,” *Int. J. Comput. Cogn.*, vol. 4, no. 1, pp. 18–29, Mar. 2006.

[43] H. Hagras, “A hierarchical type-2 fuzzy logic controller architecture for autonomous mobile robots,” *IEEE Trans. Fuzzy Syst.*, vol. 12, no. 4, pp. 524–539, Aug. 2004.

[44] M.-L. Lee, H.-Y. Chung, and F.-M. Yu, “Modeling of hierarchical fuzzy systems,” *Fuzzy Sets Syst.*, vol. 138, pp. 343–361, Sep. 2003.

[45] L. Ljung, *System Identification—Theory For the User*, 2nd ed., Upper Saddle River, NJ, USA: Prentice-Hall, 1999.

[46] J. M. Mendel and D. Wu, *Perceptual Computing: Aiding People in Making Subjective Judgments*, Hoboken, NJ, USA: Wiley, 2010.

[47] J. M. Mendel, “A comparison of three approaches for estimating (synthesizing) an interval type-2 fuzzy set model of a linguistic term for computing with words,” *Granular Comput.*, vol. 1, pp. 59–69, Mar. 2016.

[48] H. Tahayori and A. Sadeghian, “Median interval approach to model words with interval type-2 fuzzy sets,” *Int. J. Adv. Intell. Paradigms*, vol. 4, no. 3, pp. 313–336, Feb. 2012.

[49] C. Wagner, S. Miller, J. M. Garibaldi, D. T. Anderson, and T. C. Havens, “From interval-valued data to general type-2 fuzzy sets,” *IEEE Trans. Fuzzy Syst.*, vol. 23, no. 2, pp. 248–269, Apr. 2015.

[50] D. Wu, J. M. Mendel, and S. Coupland, “Enhanced interval approach for encoding words into interval type-2 fuzzy sets and its convergence analysis,” *IEEE Trans. Fuzzy Syst.*, vol. 20, no. 3, pp. 499–513, Jun. 2012.

[51] V. V. Cross and T. A. Sudkamp, *Similarity and Compatibility in Fuzzy Set Theory: Assessment and Applications*, Heidelberg, Germany: Springer, 2002.

[52] R. Yager, “A framework for multi-source data fusion,” *Inf. Sci.*, vol. 163, pp. 175–200, Jun. 2004.

[53] P. Jaccard, “Nouvelles recherches sur la distribution florale,” *Bull. de la Societe de Vaud des Sci. Naturelles*, vol. 44, pp. 223–270, Jan. 1908.

[54] J. M. Mendel, “Explaining the performance potential of rule-based fuzzy systems as a greater sculpting of the state space,” *IEEE Trans. Fuzzy Syst.*, vol. 26, no. 4, pp. 2362–2373, Aug. 2018.

[55] J. M. Mendel, I. Eyoh, and R. John, “Comparing perofrmance potentials of classical and intuitionistic fuzzy systems in terms of *sculpting the state space*,” *IEEE Trans. Fuzzy Syst.*, vol. 28, no. 9, pp. 2244–2254, Sep. 2020.

[56] P. P. Bonissone and K. H. Chang, “Fuzzy logic controllers: From development to deployment,” in *Proc. IEEE Conf. Neural Netw.*, 1993, pp. 610–619.

[57] J. M. Mendel, “Comparing the performance potentials of interval and general type-2 rule-based fuzzy systems in terms of sculpting the state space,” *IEEE Trans. Fuzzy Syst.*, vol. 27, no. 1. pp. 58–71, Jan. 2019.

[58] J. M. Mendel, R. Chimatapu, and H. Hagras, “Comparing the performance potentials of singleton and non-singleton type-1 and interval type-2 fuzzy systems in terms of *sculpting the state space*,” *IEEE Trans. Fuzzy Syst.*, vol. 28, no. 4, pp. 783–794, Apr. 2020.

[59] L. T. Kóczy and K. Hirota, “Size reduction by interpolation in fuzzy rule bases,” *IEEE Trans. Syst., Man, Cybern.,Part B: Cybern.*, vol. 27, no. 1, pp. 14–25, Feb. 1997.

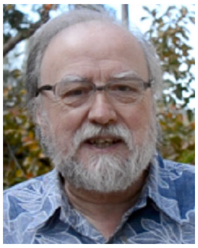
[60] B. Qin, F.-L. Chung, and S. Wang, "Biologically plausible fuzzy-knowledge-out and its induced wide learning of interpretable TSK fuzzy classifiers," *IEEE Trans. Fuzzy Syst.*, vol. 28, no. 7, pp. 1276–1290, Jul. 2020.

[61] S. Gu, F.-L. Chung, and S. Wang, "A novel deep fuzzy classifier by stacking adversarial interpretable TSK fuzzy sub-classifiers with smooth gradient information," *IEEE Trans. Fuzzy Syst.*, vol. 28, no. 7, pp. 1369–1382, Jul. 2020.

Jerry M. Mendel (Life Fellow, IEEE) received the Ph.D. degree in electrical engineering from the Polytechnic Institute of Brooklyn, Brooklyn, NY, USA.

He is currently an Emeritus Professor of electrical engineering with the University of Southern California in Los Angeles, Los Angeles, CA, USA. He has authored or coauthored more than 580 technical papers and is author and/or co-author of 12 books. He has more than 5700 citations to his publications on Google Scholar. His present research interests include type-2 fuzzy logic systems and XAI.

Dr. Mendel is a distinguished member of the IEEE Control Systems Society, and a Fellow of the International Fuzzy Systems Association. He was a member of the Administrative Committee of the IEEE Computational Intelligence Society for nine years, and Chairman of its Fuzzy Systems Technical Committee and the Computing with Words Task Force of that TC. Among his awards are four IEEE Transactions best/outstanding paper awards, a 1984 IEEE Centennial Medal, an IEEE Third Millenium Medal, and a Fuzzy Systems Pioneer Award (2008) from the IEEE Computational Intelligence Society.



Piero P. Bonissone (Life Fellow, IEEE) received the Ph.D. degree in electrical engineering and computer science from the University of California Berkeley, Berkeley, CA, USA.

He is currently an independent consultant specialized in using ML for Industrial AI. He was a Former Chief Scientist with GE Global Research, where he retired in 2014 after 34 years of service, and is a pioneer in the field of analytics, machine learning, fuzzy logic, AI, and soft computing applications. He was the recipient of 74 patents issued by the US Patent Office. He co-edited six books and has 180+ publications in refereed journals, book chapters, and conference proceedings, with 11600+ citations.

Dr. Bonissone is a Fellow of the Association for the Advancement of Artificial Intelligence, and the International Fuzzy Systems Association. He was the recipient of the 2012 Fuzzy Systems Pioneer Award and the 2005 Meritorious Service Award from the IEEE CIS, and the II Cajastur International Prize for Soft Computing from the European Centre for Soft Computing in 2008. He was Editor-in-Chief of the *International Journal of Approximate Reasoning* for 13 years and is an Editor at Large of the IEEE COMPUTATIONAL INTELLIGENCE MAGAZINE. In 2002, he was the President of the IEEE Neural Networks Society (now CIS). Since 1993, he was an Executive Committee member of NNC/NNS/CIS for 23 years. He is currently the Vice-Chair of the IEEE Fellow Committee.