Toward a Framework for Capturing Interpretability of Hierarchical Fuzzy Systems—A Participatory Design Approach

Tajul Rosli Razak, *Student Member, IEEE*, Jonathan M. Garibaldi, *Senior Member, IEEE*,
Christian Wagner, *Senior Member, IEEE*, Amir Pourabdollah, *Member, IEEE*,
and Daniele Soria, *Member, IEEE*

***Abstract*—Hierarchical fuzzy systems (HFSs) have been shown to have the potential to improve the interpretability of fuzzy logic systems (FLSs). However, challenges remain, such as "How can we measure their interpretability?" "How can we make an informed assessment of how HFSs should be designed to enhance interpretability?" The challenges consist of measuring the interpretability of HFSs include issues such as their topological structure, the number of layers, the meaning of intermediate variables, and so on. In this article, an initial framework to measure the interpretability of HFSs is proposed, combined with a participatory user design process to create a specific instance of the framework for an application context. This approach enables the subjective views of a range of practitioners, experts in the design and creation of FLSs, to be taken into account in shaping the design of a generic framework for measuring interpretability in HFSs. This design process and framework are demonstrated through two classification application examples, showing the ability of the resulting index to appropriately capture interpretability as perceived by system design experts.**

***Index Terms*—Fuzzy logic systems, hierarchical fuzzy systems, interpretability assessments, participatory design.**

I. Introduction

INTERPRETABILITY is related to the capability of expressing something in an understandable way [1]. That is, people may say that something is interpretable if they can easily understand it. One of the strengths of fuzzy logic systems (FLSs) is claimed to be their interpretability [2], particularly in applications such as knowledge extraction and decision support [3], [4]. However, key challenges remain in the design of FLSs, such as the fact that the number of rules (NR) required commonly increases exponentially with the number of input variables [5]. This challenge also known as rule explosion, sometimes referred to as the *curse of dimensionality*, can reduce the transparency and interpretability of FLSs [6].

Hierarchical fuzzy systems (HFSs) could be a practical approach to overcome rule explosion arising in conventional FLSs [7], [8]. In HFSs, the original FLSs are decomposed into a series of low-dimensional subsystems (see Section II-B). As a result, the rules in HFSs commonly have antecedents with fewer variables than the rules in "flat" FLSs with equivalent function, since the number of input variables of each subsystem is lower [9], [10]. HFSs can thereby address rule explosion and, thus, provide a potentially valuable pathway to interpretability in FLSs [6], [11]–[15], [16]. However, whilst the number of rules can be reduced, it is an open question as to how interpretability is affected when systems are hierarchical, featuring various subsystems, layers and topologies. A wide range of basic interpretability indices have been proposed to measure the interpretability of standard FLSs [17]–[33].

However, to determine which of these possible interpretability measurements is best used in practice remains an open discussion. The problem is that interpretability is a very difficult concept, because of its subjective nature in the sense that it is challenging to know how people perceive interpretability. Whilst an index can be relatively easily calculated, it is extremely difficult to validate any such index even for FLSs. This makes the creation of a measure for HFSs even more difficult. Perhaps as a consequence of this, very little (if any) work has been carried out till date in exploring how interpretability can be measured in HFSs.

Participatory design is an approach that involves the participation of users in the design development process to help ensure

Manuscript received March 23, 2018; revised November 15, 2018, May 31, 2019, and November 21, 2019; accepted January 22, 2020. Date of publication January 28, 2020; date of current version May 5, 2021. The work of T. R. Razak was supported in part by the University of Nottingham, Universiti Teknologi Mara (UiTM) and in part by the Ministry of Higher Education, Malaysia. *(Corresponding author: Tajul Rosli Razak.)*

T. R. Razak is with the Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Perlis Branch, Arau 02600, Malaysia, and also with the Laboratory for Uncertainty in Data and Decision Making (LUCID) and the Intelligent Modelling and Analysis Research Group, School of Computer Science, University of Nottingham, Nottingham NG8 1BB, U.K. (e-mail: tajulrosli@uitm.edu.my).

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

A. Pourabdollah is with the School of Science and Technology, Nottingham Trent University, Nottingham NG11 8NS, U.K. (e-mail: amir.pourabdollah@ntu.ac.uk).

D. Soria is with the School of Computing, University of Kent, Medway Campus, Canterbury ME4 4AG, U.K. (e-mail: d.soria@kent.ac.uk).

This article has supplementary downloadable material available at https://ieeexplore.ieee.org, provided by the authors.

Color versions of one or more of the figures in this article are available online at https://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TFUZZ.2020.2969901

that the result meets their needs and is usable in practice [34]. Participatory design has been used to develop solutions to complex problems, especially when dealing with people, such as in control systems [35], educational [36], and medical [37] fields. It provides a methodology toward making the design process co-operative and efficient. Hence, it may provide a method of assessing the interpretability of HFSs.

This article introduces a framework for an index to measure the interpretability of HFSs. A participatory design approach is then used to guide the development of this framework for capturing the interpretability of HFSs, building on initial work to measure the interpretability of HFSs previously proposed by the authors [38]. Naturally, a variety of aspects should be considered in capturing interpretability of HFSs, such as semantic interpretability in the sense of the meaningfulness of the constituent fuzzy sets and intermediate variables. For example, as also discussed by Magdelena [39], if the hierarchical decomposition in the fuzzy system reflects a well-understood hierarchical decomposition in the real world, then this is conducive to interpretability. However, in this article, the framework focuses on addressing key challenges arising from the structure of HFSs. Specifically, it incorporates an elementary index for assessing the interpretability of each subsystem, an aggregation strategy for combining the indices of the various subsystems within a single layer, and a layer-weighting strategy that combines layers while capturing the topology of the HFS. Initial demonstration and evaluation using the participatory design approach is presented to compare and configure the framework so as to allow its implementation in practice.

The rest of this article is organised as follows. Section II discusses relevant background on the interpretability of FLSs, HFSs, and the use of user studies. The framework for interpretability of HFSs is discussed in Section III, followed by an outline of how the framework is demonstrated in principle in Section IV. Section V introduces the participatory design process, consisting of the following two key experiments: 1) comparing *H* measure with other aggregation strategies in order to capture overall interpretability of HFSs; and 2) refinement of the *H* framework around, particularly, the aggregation strategy for combining the subsystem indices within a single layer and the strategy for assigning weights to the layers. Discussions are presented in Sections VI. Finally, Section VII concludes this article.

II. Background

A. Interpretability of FLSs

In recent years, the interest of researchers in obtaining interpretable FLSs has increased. Substantial research on interpretability measures has proposed a range of alternative interpretability indices for FLSs [17]–[33]. The most common interpretability indices are the Nauck [17] and the fuzzy index [19].

1) Nauck Index: This is a numerical index introduced by Nauck [17] to measure the interpretability of fuzzy rule-based classification systems. It is computed as the product of three terms (for details of these, see [38])

$$\text{Nauck index} = \text{comp} \times \overline{\text{cov}} \times \overline{\text{part}} \tag{1}$$

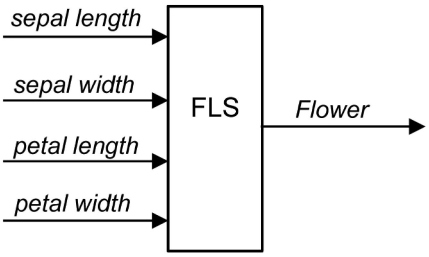

Fig. 1. Iris topology: A conventional four-input one-output FLS.

where

- comp complexity of FLSs measured as the number membership functions (MFs) of output variables divided by the number of input variables in FLSs rules;
- $\overline{\text{cov}}$ average normalized coverage degree of the fuzzy partition. It is equal to one for strong fuzzy partitions that satisfy all constraints (coverage, distinguishability, normality, etc.);
- $\overline{\text{part}}$ average normalized partition index. The partition index that is computed as the inverse of the number of MFs minus one for each input variable.

An FLS model is said to be less interpretable when its Nauck index is closer to 0 and more interpretable when its Nauck index is closer to 1.

2) Fuzzy Index: As discussed in [19] and [21], the fuzzy index, which is inspired by Nauck's index, has been proposed in interpretability assessment, particularly for fuzzy rule-based classification systems. Six variables are taken as the input of an HFS and combined into a single index. The six variables are as follows:

1) the total number of rules (NR);
2) the total number of premises in all the rules (NP)—In a complete rule-set, this equals the number of rules multiplied by the number of input variables;
3) the number of rules that use one input variable ($\text{NR}_{i=1}$);
4) the number of rules that use two input variables ($\text{NR}_{i=2}$);
5) the number of rules that use three or more input variables ($\text{NR}_{i\geq 3}$);
6) the average number of linguistic terms defined for each input variable ($\overline{\text{terms}}$).

The index also depends on the number of classes (NC), also referred to as number of output terms. It should be noted that although the fuzzy Index is generated using an HFS, it is only designed to measure the interpretability of standard FLSs, and has not previously been applied to HFSs. A fuzzy index closer to 0 implies that a given FLS is less interpretable, while values closer to 1 imply higher interpretability.

B. Hierarchical Fuzzy Systems

HFSs are characterized by structuring the input variables into a collection of low-dimensional fuzzy logic subsystems, in which the output of each layer is an input to the following layer [7], [8]. Consider a standard FLS consisting of a single layer, as shown in Fig. 1. This can be transformed into one of several alternative HFSs, two of which are shown in Figs. 2 and 3.

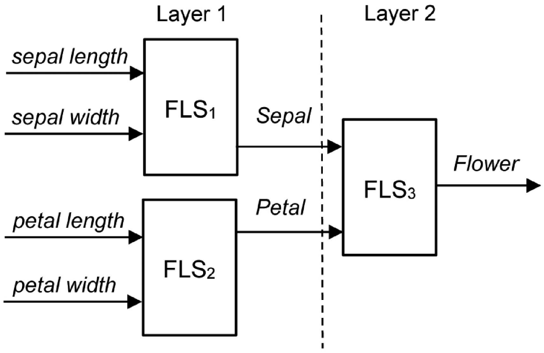

Fig. 2. Iris topology: parallel HFS.

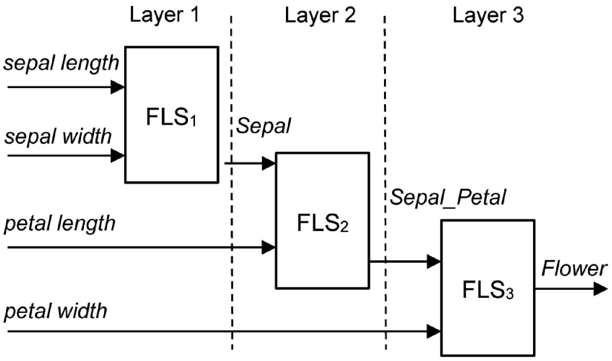

Fig. 3. Iris topology: serial HFS.

An FLS that is transformed from a one-layer FLS into a multilayer HFS has a smaller number of rules when considering a fully specified rule base. The most extreme reduction of rules occurs if the structure of the HFS has two input variables for each layer.

In conventional FLSs, the number of rules increases exponentially with the increase in the number of input variables [7], [40]. Suppose there are n input variables and m fuzzy sets for each input variable, then the number of rules (R_{FLS}) needed to construct a complete fuzzy system with a fully specified rule base (using the "AND" logical connective) is

$$R_{\text{FLS}} = m^n. \tag{2}$$

In contrast, in an HFS that is fully decomposed into subsystems consisting of two inputs and one output, if we define m fuzzy sets for each input variable and each of the intermediate output variables $y_1, \ldots, y_{n-2}$, the total number of rules (R_{HFS}) is a linear function of the number of input variables n [41], and can be expressed as

$$R_{\text{HFS}} = (n-1)m^2. \tag{3}$$

From (2) and (3), it is clear that the total number of rules in the FLSs (R_{FLS}) is always higher than or equal to the number in the HFSs (R_{HFS}). For example, Figs. 1 and 3 show an FLS and HFS with $n = 4$ input variables and, assuming that three fuzzy sets are defined for each input variable (i.e., $m = 3$), the total number of rules for this FLS is $R_{\text{FLS}} = m^n = 3^4 = 81$, whereas for the HFS the total number of rules is $R_{\text{HFS}} = (n-1)m^2 = (4-1)3^2 = 27$. Previous research has shown that HFSs can be used to reduce the number of rules in this manner, and claiming to, thus, improve interpretability [6], [11]–[14]. However, indices for actually measuring interpretability of HFSs were not discussed by any of these authors.

As mentioned in [38], there are several challenges in creating methodologies for measuring the interpretability of HFSs.

1) Multiple Individual Subsystems: As mentioned above, HFSs are produced by structuring the input variables in FLSs into multiple subsystems. Each subsystem commonly has a small number of inputs and outputs and a small rule base, and serves commonly a single purpose [42]. The first challenge may be expressed as "How can the interpretability of each subsystem in an HFS be measured using an index?" This challenge is akin to the principal challenge of capturing standard FLS interpretability using an index.

2) Aggregation: The second challenge is the choice of aggregation strategy to combine the indices of the various subsystems in an HFS. Several aggregation strategies may be suitable, such as *mean*, *min*, *max*, and order weighted average (*OWA*) [43], [44]. An *OWA* is calculated by reordered subsystems in descending order before multiplying them by the weights. Yager [45] introduced the linguistic quantifier to calculate weights (w) in which he defines certain values of alpha (α) to capture labels such as "at least one" ($\alpha = 0.0$), "at least a few" ($\alpha = 0.1$), "a few" ($\alpha = 0.5$), "half" ($\alpha = 1.0$), "most" ($\alpha = 2.0$), "almost all" ($\alpha = 10.0$), and "all" ($\alpha = \infty$). This is done by assigning weights according to

$$w_i = \left(\frac{i}{s}\right)^{\alpha} - \left(\frac{i-1}{s}\right)^{\alpha}, \quad i = 1, \ldots, s \tag{4}$$

where s is the total number of subsystems.

The specific attractiveness of the *OWA* is that it enables dynamic weighting of the individual interpretability of subsystems (based on the individual interpretability of subsystems such as established by the traditional FLS indices). For example, choosing $\alpha = 0.1$ results in a weighting strategy closely resembling the *max*, in which the most interpretable subsystem in the layer is given the highest weight, the second-most interpretable a substantially lower weight, and so on.

3) Topology and Layering: Based on the same input variables, HFSs with different topologies may be produced, such as the serial and parallel HFSs shown in [9]. A parallel HFS can have more than one subsystem per layer (e.g., Fig. 2), while serial HFSs use strictly one subsystem per layer (e.g., Fig. 3). Thus, these topologies commonly have a different number of layers. For example, Figs. 2 and 3 show two different topologies of HFSs using the same four input variables. Both topologies use the same number of subsystems, but with different numbers of layers in their structure. Thus, this challenge can be expressed as "How can the interpretability of HFSs with different topologies and number of layers be measured systematically?"

C. Assessing Interpretability: User Studies

A user study allows researchers to identify specific variables that are interesting and observe the impact of varying the values of those variables [46]. Examples of user studies include that of Balazs and Koczy [47] who conducted interviews to ask users to define fuzzy sets, i.e., to get to know what a user meant by "hot." Based on the user-defined linguistic terms, fuzzy rules and rule bases can be constructed easily. This was claimed to lead to complexity reduction and improved interpretability.

Mencar and Fanelli [48] conducted a survey with the following aim:

1) give a homogeneous description of all interpretability constraints;
2) provide a critical review of such constraints;
3) identify potentially different meanings of interpretability.

Alonso *et al.* [23] evaluated the most common interpretability indices with a user study (in the form of a web poll) to extract useful information regarding interpretability assessment. The results showed that the fuzzy index was more easily adapted to the context of each problem as well as the quality criteria of the users.

Here, we conduct a user study, inspired by Alonso *et al.* [23], asking users how the interpretability of given FLSs and HFSs is perceived. However, rather than using the results of the user study to directly evaluate the framework, this article describes how the data obtained from the user study has been used to guide the development of our framework through a participatory design approach.

III. Framework for Interpretability of HFSs

A key aspect toward a framework for interpretability of HFSs is the need to assess the interpretability of each of its constituent subsystems, present across its layers (as illustrated in Figs. 2 and 3), and then combine these together into a single overall measure of interpretability of the whole system. Clearly, there are many alternative operators that could be selected. For example, it is reasonable to use an aggregation operator that selects something between min and max values [49]. Alternatively, operators that generate results beyond the min and max, such as t-norms or t-conorms, may be applicable. In this article, our aim is not to identify the best (set of) operator(s); but to put forward one viable strategy toward a flexible framework modeling interpretability in HFSs.

A. Overall Framework

Following the discussion mentioned above, we propose the following high-level structure for the framework. Consider *H*, the interpretability of an HFS, as follows

$$H = \sum_{j=1}^{q} \left(l_j \bigoplus_{k=1}^{s_j} E_{jk} \right) \tag{5}$$

where

- E_{jk} underlying (standard) FLS index associated with the subsystem k at layer j;
- $\bigoplus$ represents a general aggregation operator;
- l_j weight associated with layer j of the HFS (see below);
- s_j number of subsystems located in layer j and s is the total number of subsystems;
- q number of layers of the HFS.

Note that E_{jk} could be any index used for measuring the interpretability of a non-HFS. In this article, we neither evaluate or advocate any specific index. However, to illustrate the framework, we use the Nauck (N) and fuzzy (F) indices on the basis that they are commonly used. Note that (5) returns the original FLS index when applied to a standard FLS because it has only one subsystem and one layer. Furthermore, a linear weighted aggregation strategy is used in (5) to combine layers as the simplest strategy to model varying degrees of importance in respect to interpretability across layers. In future, of course, more complex and nonlinear operators could be explored.

Layer-weights l_j are associated with each subsystem according to their layer, such that the sum of all *layer-weights* l_j is equal to one regardless of the number of layers q, i.e.,

$$\sum_{j=1}^{q} l_j = 1, \quad l_j \in [0,1]. \tag{6}$$

Based on the abovementioned equation, an HFS model is less interpretable when *H* is close to 0 and more interpretable when *H* is close to 1.

B. Layer-Weighting Strategy

A variety of weighting strategies for the individual layers within HFSs is possible. Here, we briefly introduce a key set of alternatives.

1) Layer-Weights Decreasing With Depth: l_j are arranged in descending order. This is intended to reflect the fact that the structure of most HFSs is formed by having the most influential input variables in the first layer of the hierarchy, the next most important inputs in the second layer, as, for example, in [7] and [8]. Hence,

$$l_1 > l_2 > \cdots > l_q, \quad j = 1, \ldots, q.$$

In order to achieve this and satisfy (6), l_j can be given by

$$l_j = \frac{2\,(q-j+1)}{q\,(q+1)}, \quad j = 1, \ldots, q. \tag{7}$$

2) Increasing With Depth: The same principle as mentioned above, but with the layer-weights *increasing* with layer depth. This is indicated that the input variables in the last layer of the hierarchy are most important, as given by

$$l_j = \frac{2j}{q(q+1)}, \quad j = 1, \ldots, q. \tag{8}$$

3) Equal Weighting: Assigning an equal weight for all layers, as given by

$$l_j = \frac{1}{q}. \tag{9}$$

IV. Framework Demonstration in Principle

Following the principle of least commitment, it is intuitive to initially explore the *mean* as an aggregation operator, to both demonstrate the functionality of the *H* framework generally and to explore the behavior of the resulting "mean-based" *H*

TABLE I
DESCRIPTION OF THE PARAMETERS OF THE SIX IRIS CLASSIFICATION SYSTEMS

Iris systems	Model Type	NI	NMF	NR	NS	NL	NRS$_1$	NRS$_2$	NRS$_3$
F-2	FLS	4	2	16	1	1	16	-	-
P-2	Parallel HFS	4	2	12	3	2	4	4	4
S-2	Serial HFS	4	2	12	3	3	4	4	4
F-3	FLS	4	3	81	1	1	81	-	-
P-3	Parallel HFS	4	3	27	3	2	9	9	9
S-3	Serial HFS	4	3	27	3	3	9	9	9

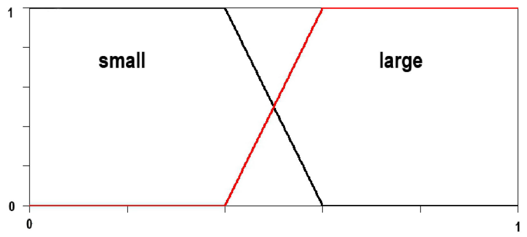

Fig. 4. MFs of input *sepal length* with two MFs.

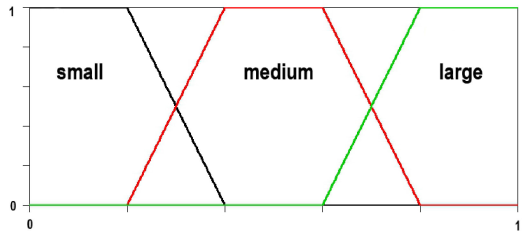

Fig. 5. MFs of input *sepal length* with three MFs.

in principle. We initially explored this approach in [38], and summarise the approach and results here.

Considering the *mean* as aggregation operator, (5) becomes

$$H_{\text{mean}} = \sum_{j=1}^{q} \left(l_j \sum_{k=1}^{s_j} E_{jk}/s_j \right). \tag{10}$$

To demonstrate the behavior of the resulting H_{mean}, we consider both the Nauck and the fuzzy indices (as the underlying indices applied to each subsystem) using the well-known Iris flower classification problem [50]. Note that the Iris classification example is used in this article because it is simple and well understood. It is used only to illustrate the proposed framework and not to show any benefits of a hierarchical approach over a nonhierarchical one. The Iris dataset has four attributes as input features, namely: *sepal length*, *sepal width*, *petal length*, and *petal width*; and three classes of iris flowers as output, namely: *Setosa*, *Versicolor*, and *Virginica*.

We design three individual systems to capture the variety of HFSs' architectures, namely a (standard) FLS (F), a parallel HFS (P), and serial HFS (S). The three systems were each designed in two configurations, where all variables have either two or three MFs—Termed F-2, P-2, S-2 (collectively referred to as MF-2) and F-3, P-3, S-3 (MF-3), respectively. The various systems are characterized by seven attributes as follows.

1) *Model type*: Type of fuzzy model, namely F, P, and S, as shown in Figs. 1, 2, and 3, respectively.
2) *NI*: Number of input variables.
3) *NMF*: Number of membership functions used for all the variables of each model, as shown in Figs. 4 and 5.
4) *NR*: Total number of rules.
5) *NS*: Number of subsystems in the model.
6) *NL*: Number of layers in the model.
7) *NRS*$_k$: Number of rules in subsystem k.

The attributes for the six systems are summarised in Table I. The complete rule set for each of the six variations of the systems are given Tables S-I– S-VI in Supplemental material.

A. Methods

In this section, the application of the H_{mean} framework to the six variations of the Iris system described earlier is shown in detail. Both the Nauck and fuzzy indices are used within the H_{mean} framework to enable their comparison. The six systems are then also used in the participatory design experiments described later.

The application of the H_{mean} framework to measure the interpretability of each of the six systems is carried out in the following steps.

1) Calculate Interpretability for Each Subsystem: First, the interpretability of each subsystem is calculated using both the Nauck and the fuzzy indices. For example, the values of the Nauck index for the three subsystems in P-2 (parallel HFS with two MFs) are $N_1 = 0.250$, $N_2 = 0.250$, and $N_3 = 0.375$ (the details of the calculations are shown in Table II).

2) Identify the Layer Weights: Next, the values of the layer weights are computed using (7). For instance, for P-2 and P-3 that consists of two layers, the values of the layer weights at each layer are $l_1 = 0.667$ and $l_2 = 0.333$; where for S-2 and S-3 that consist of three layers, the values of layer weights at each layer are $l_1 = 0.500$, $l_2 = 0.333$, and $l_3 = 0.167$.

3) Calculate the Overall Interpretability: Then, the overall interpretability can be calculated using the H_{mean} as given in (10). For example, the interpretability of model P-2 is computed as follows:

$$\begin{aligned} H_{\text{mean}} &= \sum_{j=1}^{q} \left(l_j \sum_{k=1}^{s_j} E_{jk}/s_j \right) \\ &= l_1(N_1 + N_2)/2 + l_2(N_3/1) \\ &= 0.667(0.250 + 0.250)/2 + 0.333(0.375) \\ &= 0.292. \end{aligned}$$

TABLE II
INTERPRETABILITY OF THE IRIS SYSTEMS; SET MF-2 (F-2, P-2, S-2) AND SET MF-3 (F-3, P-3, S-3) USING THE H_{MEAN}

Iris models	Nauck Index				Fuzzy Index							
	$comp$	$\overline{cov}$	$\overline{part}$	Index	NR	NP	NR$_{i=1}$	NR$_{i=2}$	NR$_{i\geq3}$	$\overline{terms}$	NC	Index
F-2	0.047	1	1	0.047	16	64	0	0	16	2	3	0.500
P-2:												
FLS$_1$	0.250	1	1	0.250	4	8	0	4	0	2	2	0.611
FLS$_2$	0.250	1	1	0.250	4	8	0	4	0	2	2	0.611
FLS$_3$	0.375	1	1	0.375	4	8	0	4	0	2	3	0.663
Hmean				0.292								0.628
S-2:												
FLS$_1$	0.250	1	1	0.250	4	8	0	4	0	2	2	0.611
FLS$_2$	0.250	1	1	0.250	4	8	0	4	0	2	2	0.611
FLS$_3$	0.375	1	1	0.375	4	8	0	4	0	2	3	0.663
Hmean				0.271								0.620
F-3	0.009	1	0.5	0.005	81	324	0	0	81	3	3	0.194
P-3:												
FLS$_1$	0.167	1	0.5	0.083	9	18	0	9	0	3	3	0.493
FLS$_2$	0.167	1	0.5	0.083	9	18	0	9	0	3	3	0.493
FLS$_3$	0.167	1	0.5	0.083	9	18	0	9	0	3	3	0.493
Hmean				0.083								0.493
S-3:												
FLS$_1$	0.167	1	0.5	0.083	9	18	0	9	0	3	3	0.493
FLS$_2$	0.167	1	0.5	0.083	9	18	0	9	0	3	3	0.493
FLS$_3$	0.167	1	0.5	0.083	9	18	0	9	0	3	3	0.493
Hmean				0.083								0.493

B. Results

The overall interpretability measurements of the six Iris classification systems calculated using the H_{mean} are shown in Table II. In general, it can be seen that the computed H_{mean} interpretability indices in the various hierarchical models are always larger (i.e., more interpretable) as compared to the interpretability of the flat FLSs, regardless of whether the hierarchical topology is parallel or serial, and regardless of the number of MFs.

As shown in Table II, considering the Nauck index for the two MF case, the resultant H_{mean} value (i.e., the calculated overall interpretability) is greatest for the parallel HFS model (P-2 $= 0.292$), followed by the serial HFS model (S-2 $= 0.271$), and finally the flat FLS (F-2 $= 0.047$). The same pattern is observed for the fuzzy index, although the absolute values of interpretability obtained are higher.

Furthermore, as seen in Table II, considering the Nauck index for the three MF case, the computed interpretabilities are higher for both the hierarchical models (P-3 = S-3 $= 0.083$) compared to the flat FLS (F-3 $= 0.005$). However, in this case, the interpretability of both the hierachical models are the same, i.e., the interpretability of the parallel and serial models featuring three MFs is the same. The same pattern is obtained with the fuzzy index, albeit with higher absolute values of interpretability.

The results generated for the H_{mean} follow intuition in the sense that the HFSs do have better interpretability than FLS for all systems. Furthermore, the parallel topology P-2 is seen to have a better interpretability than the serial topology S-2. This feature is actually due to a combination of the following three factors.

1) Both the Nauck and the fuzzy indices rate the interpretability of a (sub)system consisting of two inputs each with two MFs and an output with three MFs ($2 \times 2 \rightarrow 3$) *higher than* that of a (sub)system consisting of two inputs each with two MFs and an output with two MFs ($2 \times 2 \rightarrow 2$).
2) The proposed H_{mean} gives higher interpretability to subsystems in earlier layers.
3) P-2 features the ($2 \times 2 \rightarrow 3$) subsystem in layer 2, whereas S-2 features it in layer 3.

This is not repeated in the case of P-3 and S-3, as in these cases all subsystems are of form ($3 \times 3 \rightarrow 3$) and so all have equal interpretability; hence, the parallel and serial HFS topologies result in the same interpretability.

The fuzzy index is designed to provide a measurement of interpretability, which is closer to the user's point of view than the Nauck or other indices [23]. Given this and our finding that both produce similar results in our H_{mean} experiments, only the fuzzy index will be used for the remainder of this article in comparing and refining the H framework.

V. PARTICIPATORY DESIGN APPROACH

We propose a participatory design approach to compare and derive parameters of H within the framework. As mentioned earlier, participatory design is an approach that involves users in the design development process to ensure the result satisfies their needs [34]. In this section, a participatory design process consists of the following two main experiments:

1) to assess whether the approach of the H framework, taking into account the topology of connected layers, better

matches users perceptions of interpretability, rather than a nonlayered approach;

2) to guide the refinement of the *H* framework through the following: i) the aggregation strategy for combining the subsystem indices within a single layer; ii) the strategy for assigning weights to the layers.

These two experiments are now described in detail using the examples of the Iris classification application and rotary crane system (as used in [51]), respectively.

A. Experiment 1: The H *Framework Itself*

First, an experiment was conducted to examine the measurements of the interpretability of HFSs using the *H* framework and without the framework, from the point of view of users' interpretability within a participatory design approach.

1) Participatory User Study: Six of the Iris systems were classified into two groups. The first group was named Set MF-2, which consists of three Iris systems ("flat," "parallel," and "serial," with two MFs per variable), termed F-2, P-2, and S-2; the second group was named Set MF-3, which consists of three corresponding systems each with three MFs per variable, termed F-3, P-3, and S-3. Each of the Iris systems in Set MF-2 and MF-3 was printed on an A4 card. The topology, MFs and rule set of each system was summarised on these cards. For example, card F-2 (as shown in Fig. S-1 of the Supplemental material) contained the topology of FLS as shown in Fig. 1, the two MFs used in all input variables as shown in Fig. 4, and the complete 16 rules of the FLS.

We carried out this article-based survey at the Fuzz-IEEE 2017 Conference held in Naples, Italy, during which we asked a sample of participants at the conference to answer a set of questions concerning interpretability. The sample of 25 participants was selected from a range of academics (from doctoral students to full Professor), with a range of expertise in fuzzy system design and creation, recruited during the session "interpretable fuzzy systems" and also from other sessions at the conference. The participants were asked to separately rank order the three Iris systems in MF-2 and those in MF-3 based on the perceived interpretability. Users were asked to indicate a rank of 1, 2, or 3, for each of the three systems; with the refinement that they were free to indicate equal ranks for one or more system if they wished—that is, responses such as $1, 1, 1$ indicated that all three systems were ranked equally interpretable, or $1, 3, 3$ indicated that two of the systems were viewed as being equally less interpretable. Due to this, there may be more or fewer observations of each rank than the number of participants in the study.

The individual responses are shown in Table S-VII (in the Supplemental material), in which the first column indicates the 25 users (referred to as U-1 to U-25), while the second and third columns show the interpretability rankings for Set MF-2 and Set MF-3, respectively. These results are summarised in Tables III and IV, showing the frequency (count and percentage) of each ranking, together with the average rank, of each system. It can be seen that most of the users found the parallel HFS to be more interpretable than the flat FLS and serial HFS, in both Set MF-2 and Set MF-3, with 76% of the users selecting P-2 as the most interpretable of the systems with two MFs, and 72% selecting P-3 as the most interpretable of the systems with three MFs. In both cases of two and three MFs, the ranking of the flat and serial systems are less clear-cut; in the cae of MF-2, it appears that F-2 may be slightly more interpretable than S-2, whereas S-3 may be slightly more interpretable than F-3.

TABLE III
FREQUENCY OF THE INTERPRETABILITY RANKINGS FOR IRIS SYSTEMS IN SET MF-2, AS EXTRACTED FROM USER STUDY

Set MF-2	Rank						Average Rank
	1		2		3		
	Count	(%)	Count	(%)	Count	(%)	
F-2	5	20	8	32	12	48	2.3
P-2	19	76	4	16	2	8	1.3
S-2	2	8	9	36	14	56	2.5

TABLE IV
FREQUENCY OF THE INTERPRETABILITY RANKINGS FOR IRIS SYSTEMS IN SET MF-3, AS EXTRACTED FROM USER STUDY

Set MF-3	Rank						Average Rank
	1		2		3		
	Count	(%)	Count	(%)	Count	(%)	
F-3	5	20	5	20	15	60	2.4
P-3	18	72	4	16	3	12	1.4
S-3	4	16	11	44	10	40	2.2

TABLE V
SUMMARY OF THE INTERPRETABILITY FOR IRIS SYSTEMS USING THE H_{MEAN} AND *MEAN*

Set MF-2	F-2	P-2	S-2
Hmean	0.500 (3)	0.628 (1)	0.620 (2)
Mean	0.500 (3)	0.628 (1)	0.628 (1)
Set MF-3	F-3	P-3	S-3
Hmean	0.194 (3)	0.493 (1)	0.493 (1)
Mean	0.194 (3)	0.493 (1)	0.493 (1)

The Interpretability score (and rank) is shown for each set MF-2 and MF-3.

2) H*mean Versus "Mean":* This experiment explores measuring interpretability using the proposed *H* framework (H_{mean}) compared to not using a framework at all and instead just taking the mean of all the subsystems, regardless of topology (termed simply *Mean*). Note that our *H* framework performs averaging of individual interpretability of subsystem at each layer and then layer weighted at each layer, to obtain overall interpretability of HFSs. In contrast, without the framework, the *Mean* simply treats the interpretability of all subsystems with equal weight regardless of which layer each appears in, the number of layers, etc. That is, the *Mean* simply averages the interpretability of all subsystems to obtain the overall interpretability of an HFS.

Table V shows the interpretability values obtained using the H_{mean} and *Mean* (i.e., just averaging the subsystems) of the various Iris systems, Set MF-2 and Set MF-3. The resulting rank order of each of the systems is also shown. In general, as can be seen from Table V, the *Mean* measure produced the same interpretability result for P-2 and S-2; in contrast, the H_{mean} produced different values for P-2 and S-2, indicating that P-2 is more interpretable than S-2, in agreement with the results obtained from users. In the case of Set MF-3, both measures produced the same results in P-3 and S-3. This is because all the subsystems have similar structural characteristics, and hence, the same fuzzy index score (of 0.493, as can be seen in Table II). Thus, any aggregation operators and layer-weighting schemes will also result in the same overall result of 0.493. Whilst these results are insufficient to draw strong conclusions from, this is perhaps a reflection of the fact that the Iris system is too simple, with insufficient degrees of freedom to allow for much variation in alternative hierarchical systems. For this reason, we undertook a further set of experiments on a more complex system.

B. Experiment 2: Beyond the Mean

While aggregating the subsystems using the *mean* and *decreasing weight* layer-weight can be used as a default strategy, in order to capture the interpretability of HFSs as perceived by actual users, we propose another participatory design approach to derive *H* parameters within the framework. In this experiment, we use a more complex set of alternative HFSs, based on the rotary crane system as in [51].

1) Participatory User Study: A total of 12 rotary crane systems were constructed, termed A ,..., L. Each system has a different configuration such as the number of rules, number of subsystems, number of layers. Illustrations of the topology of each can be seen in Figs. S-2– S-13.

Similar to experiment 1 (see Section V-B1), each system was represented on an *A4* card. However, this time, we only presented the topology and rule structures. Users were asked to choose which system they favoured in terms of interpretability in a set of pairwise comparisons drawn from the total set of possible pairs. The combination of the pairwise comparisons (labeled PW-1 to PW-20) are as follows: (PW-1: A, B), (PW-2: A,C), (PW-3: B,C), (PW-4: B,D), (PW-5: B,E), (PW-6: C,D), (PW-7: C,E), (PW-8: D,E), (PW-9: D,F), (PW-10: E,F), (PW-11: F,G), (PW-12: F,H), (PW-13: G,H), (PW-14: H,I), (PW-15: I,J), (PW-16: I,K), (PW-17: I,L), (PW-18: J,K), (PW-19: J,L), and (PW-20: K,L). It should be noted that only 20 pairwise comparisons (out of a total of 132 possible pairs) were selected, as it was deemed impractical to ask users to provide a preference for all 132 pairs, due to the time and effort this would require. The selection of pairs to be used was based on consideration of whether they were felt to be "not obviously different from each other" and, hence, interesting and informative to gather preference opinion on. For instance, system A may be paired with all other systems B, . . ., L. However, only the pairs (A,B) and (A,C) were chosen, because they are not obviously different to each other in terms of their structure and number of rules. For instance, for PW-1, the participants were asked to choose between systems A and B, based on their perceived interpretability preference (see in Fig. S-14 for a mock-up of PW-1). If both systems seem equally interpretable, they could indicate "Equal" (EQ) as their answer. This experiment was carried out through an online-survey with 40 participants from a wide range of expertise.

TABLE VI
FREQUENCY OF THE USERS INTERPRETABILITY RATING FOR PAIRWISE COMPARISONS OF ROTARY CRANE, AS EXTRACTED FROM USER STUDY

Pairwise comparisons	Users Interpretability Rating												
	A	B	C	D	E	F	G	H	I	J	K	L	EQ
PW-1	21	11	-	-	-	-	-	-	-	-	-	-	8
PW-2	20	-	12	-	-	-	-	-	-	-	-	-	8
PW-3	-	17	8	-	-	-	-	-	-	-	-	-	15
PW-4	-	16	-	14	-	-	-	-	-	-	-	-	10
PW-5	-	16	-	-	13	-	-	-	-	-	-	-	11
PW-6	-	-	8	21	-	-	-	-	-	-	-	-	11
PW-7	-	-	10	-	19	-	-	-	-	-	-	-	11
PW-8	-	-	-	9	13	-	-	-	-	-	-	-	18
PW-9	-	-	-	13	-	25	-	-	-	-	-	-	2
PW-10	-	-	-	-	13	22	-	-	-	-	-	-	5
PW-11	-	-	-	-	-	26	11	-	-	-	-	-	3
PW-12	-	-	-	-	-	3	-	17	-	-	-	-	20
PW-13	-	-	-	-	-	-	6	26	-	-	-	-	8
PW-14	-	-	-	-	-	-	-	25	11	-	-	-	4
PW-15	-	-	-	-	-	-	-	-	18	5	-	-	17
PW-16	-	-	-	-	-	-	-	-	17	-	16	-	7
PW-17	-	-	-	-	-	-	-	-	20	-	-	14	6
PW-18	-	-	-	-	-	-	-	-	-	16	14	-	10
PW-19	-	-	-	-	-	-	-	-	-	15	-	19	6
PW-20	-	-	-	-	-	-	-	-	-	-	7	17	16

TABLE VII
BASE INTERPRETABILITY OF EACH OF THE ROTARY CRANE SYSTEMS MEASURED USING THE FUZZY INTERPRETABILITY INDEX

Rotary Crane Systems	Fuzzy Index for each HFS subsystem				
	FLS_1	FLS_2	FLS_3	FLS_4	FLS_5
System A	0.0216	-	-	-	-
System B	0.0647	0.4932	-	-	-
System C	0.4932	0.0647	-	-	-
System D	0.2408	0.1941	-	-	-
System E	0.1941	0.2408	-	-	-
System F	0.4932	0.1941	0.4932	-	-
System G	0.2408	0.4932	0.2408	-	-
System H	0.2408	0.2408	0.4932	-	-
System I	0.2408	0.4932	0.4932	0.4932	-
System J	0.4932	0.4932	0.2408	0.4932	-
System K	0.4932	0.4932	0.4932	0.2408	-
System L	0.4932	0.4932	0.4932	0.4932	0.4932

Table VI presents frequency of the users interpretability ratings for 20 pairwise comparisons. The detail of answers given by each participant to each of the pairwise comparisons are shown in Table S-VIII. From an initial observation, there is appreciable diversity of opinion in the 40 participants as to the interpretability of the various systems. This also shows that interpretability is very subjective because of each participant may perceive the interpretability differently.

2) Exploration of Alternatives Configurations of H*:* This section was conducted to explore various alternative aggregations and layer-weighting strategies as described in Section III. First, the fuzzy index for each of the subsystems present in the 12 different rotary crane system configurations was calculated, as shown in Table VII.

Five different aggregation strategies, *mean*, *min*, *max*, and two linguistic *OWAs* [using alpha (α) of 0.1 and 2] were explored. Each was used as a general aggregation operator $\bigoplus$ in the *H* framework presented in (5)—Note that only the *decreasing weight* layer-weighting strategy was used in conjunction with the various aggregation strategies. For example, for the case of Linguistic $OWA_{\alpha=0.1}$, the values of $\alpha = 0.1$ will be used in (4) to obtain its weights (w) before multiplying them by the reordered subsystems in descending order. Given that three fuzzy

TABLE VIII
INTERPRETABILITY OF THE ROTARY CRANE SYSTEMS USING *H* WITH DIFFERENT AGGREGATION STRATEGIES AND LAYER-WEIGHTING STRATEGIES

Rotary Crane Systems	*H* (using *decreasing weight* layer strategy)					*H* (using H_{mean} aggregation)		
	H_{mean}	H_{min}	H_{max}	$H_{\alpha=0.1}$	$H_{\alpha=2}$	*decreasing weight*	*increasing weight*	*equal weight*
System A	0.0216	0.0216	0.0216	0.0216	0.0216	0.0216	0.0216	0.0216
System B	0.2075	0.2075	0.2075	0.2075	0.2075	0.2075	0.3504	0.2789
System C	0.3504	0.3504	0.3504	0.3504	0.3504	0.3504	0.2075	0.2789
System D	0.2253	0.2253	0.2253	0.2253	0.2253	0.2252	0.2097	0.2175
System E	0.2097	0.2097	0.2097	0.2097	0.2097	0.2097	0.2252	0.2175
System F	0.3935	0.2938	0.4932	0.4798	0.3437	0.3935	0.4434	0.4184
System G	0.3249	0.2408	0.4091	0.2521	0.3670	0.3249	0.2829	0.3039
System H	0.3249	0.3249	0.3249	0.3249	0.3249	0.3249	0.4091	0.3670
System I	0.4301	0.3670	0.4932	0.3755	0.4617	0.4301	0.4722	0.4511
System J	0.4301	0.3670	0.4932	0.4847	0.3986	0.4301	0.4722	0.4511
System K	0.4679	0.4679	0.4679	0.4679	0.4679	0.4680	0.3922	0.4301
System L	0.4932	0.4932	0.4932	0.4932	0.4932	0.4932	0.4932	0.4932

index values for each subsystem in system F are $F_1 = 0.4932$, $F_2 = 0.1941$, and $F_3 = 0.4932$, the overall interpretability of rotary crane system F was computed using *H* with *OWA*$_{\alpha=0.1}$ and a *decreasing weight* layer-weighting, as follows:

$$\begin{aligned} H_{\alpha=0.1} &= \sum_{j=1}^{q} \left(l_j \sum_{k=1}^{s_j} E_{jk} w_k \right) \\ &= l_1(F_1 w_1 + F_2 w_2) + l_2(F_3) \\ &= 0.667(0.4932)(0.933) + (0.1941)(0.067)) \\ &\quad + 0.333(0.4931) \\ &= 0.4798. \end{aligned}$$

Meanwhile, in the layer-weighting experiment, the aforementioned three layer-weighting strategies, *decreasing weight*, *increasing weight*, and *equal weight* as described in Section III-B were investigated. All these strategies were used as the layer-weight l_j in the *H* framework presented in (5)—Note that only the *mean* aggregation strategy was used in conjunction with these layer-weighting strategies. For instance, for the case of *increasing weight*, the values of layer weight can be computed using (8). For the parallel models that consists of two layers, the values of layer weights at each layer are $l_1 = 0.333$ and $l_2 = 0.667$. Meanwhile, for the serial models that consists of three layers, the values of layer weights at each layer are $l_1 = 0.167$, $l_2 = 0.333$, and $l_3 = 0.500$. Given that three fuzzy index values for each subsystem in system F are $F_1 = 0.4932$, $F_2 = 0.1941$, and $F_3 = 0.4932$, the overall interpretability of system F is computed using *H* with layer-weighting, *increasing weight*, and aggregation strategy, *mean* can be expressed as follows:

$$\begin{aligned} H_{\text{mean}} &= \sum_{j=1}^{q} \left(l_j \sum_{k=1}^{s_j} E_{jk}/s_j \right) \\ &= l_1(F_1 + F_2)/2 + l_2(F_3/1) \\ &= 0.333((0.4932 + 0.0.1941)/2) + 0.667(0.4932/1) \\ &= 0.4434. \end{aligned}$$

The results obtained are shown in Table VIII. From these results, we can see that *H* framework produced a diversity of answers for various systems, aggregations, and layer-weighting strategies. These results were then transformed to obtain the *H* scores for the 20 pairwise comparisons. For the case of the H_{mean} example in aggregation strategies, the first pairwise comparison is between systems A and B. In this example, system B was chosen as it scores higher than system A based on the overall interpretability, indicating that the *H* framework suggests that system B is more interpretable than system A. The complete results of pairwise comparison for the interpretability of the rotary crane systems obtained from the *H* with different aggregation and layer-weighted strategies can be seen in Table IX. Whilst the interpretability index is a real number, nevertheless sometimes it produces identical indices for two different systems—In this case, it is labeled in the Table as "EQ" (equal). In general, systems might be considered equal if the difference were below a certain threshold.

3) Matching H *to the Participatory Study:* This step was conducted to determine the level of agreement between the interpretability ratings provided by the participatory user study (as in Section V-B1) and various alternative configurations of the *H* framework (as shown in Section V-B2).

Specifically, we computed the agreement scores between the results in Table IX with those in Table VI. For example, for pairwise comparison PW-1, the user preferences are A$= 20$, B$= 11$, and EQ$= 8$, as shown in Table VI. Accordingly, from the fact that H_{mean} produces a higher interpretability score for B than A, we deduce that H_{mean} prefers B, and consequently the level agreement score obtained is 11 agreements (as B was preferred by 11 users). Full details of the agreement score are provided in Table X. The last two rows summarise the agreements, providing the mean and standard deviation (SD) for each column.

From Table X, it can be seen that the H_{min} aggregation strategy and *increasing weight* layer-weight strategy achieve the highest average agreement scores. That is, most of the answers given by users are closer to the ratings obtained using *H* with H_{min} aggregation strategy and *increasing weight* layer-weight strategy.

VI. DISCUSSION

We studied the newly proposed generic *H* framework through a participatory design process consisting of experiments of

TABLE IX
PAIRWISE COMPARISON FOR THE INTERPRETABILITY OF THE ROTARY CRANE SYSTEMS OBTAINED FROM THE H WITH DIFFERENT AGGREGATION AND LAYER-WEIGHTED STRATEGIES

Pairwise Comparision	H aggregation strategy					H layer-weighted strategy		
	H_{mean}	H_{min}	H_{max}	$H_{\alpha=0.1}$	$H_{\alpha=2}$	*decreasing weight*	*increasing weight*	*equal weight*
PW-1	B	B	B	B	B	B	B	B
PW-2	C	C	C	C	C	C	C	C
PW-3	C	C	C	C	C	C	B	EQ
PW-4	D	D	D	D	D	D	B	B
PW-5	E	E	E	E	E	E	B	B
PW-6	C	C	C	C	C	C	D	C
PW-7	C	C	C	C	C	C	E	C
PW-8	D	D	D	D	D	D	E	EQ
PW-9	F	F	F	F	F	F	F	F
PW-10	F	F	F	F	F	F	F	F
PW-11	F	F	F	F	G	F	F	F
PW-12	F	H	F	F	F	F	F	F
PW-13	EQ	H	G	H	G	EQ	H	H
PW-14	I	I	I	I	I	I	I	I
PW-15	EQ	EQ	EQ	J	I	EQ	EQ	EQ
PW-16	K	K	I	K	K	K	I	I
PW-17	L	L	L	L	L	L	L	L
PW-18	K	K	J	J	K	K	J	J
PW-19	L	L	L	L	L	L	L	L
PW-20	L	L	L	L	L	L	L	L

TABLE X
AGREEMENT SCORE BETWEEN THE PREFERENCES GIVEN BY EACH OF THE USERS (IN TABLE VI) AND THE PREFERENCE INDICATED BY H FRAMEWORK (USING DIFFERENT AGGREGATION AND LAYER-WEIGHTED STRATEGIES AS SHOWN IN TABLE IX)

	H aggregation strategies					H layer-weighted strategies		
	H_{mean}	H_{min}	H_{max}	$H_{\alpha=0.1}$	$H_{\alpha=2}$	*decreasing weight*	*increasing weight*	*equal weight*
PW-1	11	11	11	11	11	11	11	11
PW-2	12	12	12	12	12	12	12	12
PW-3	8	8	8	8	8	8	17	15
PW-4	14	14	14	14	14	14	16	16
PW-5	13	13	13	13	13	13	16	16
PW-6	8	8	8	8	8	8	21	8
PW-7	10	10	10	10	10	10	19	10
PW-8	9	9	9	9	9	9	13	18
PW-9	25	25	25	25	25	25	25	25
PW-10	22	22	22	22	22	22	22	22
PW-11	26	26	26	26	11	26	26	26
PW-12	3	17	3	3	3	3	3	3
PW-13	8	26	6	26	6	8	26	26
PW-14	11	11	11	11	11	11	11	11
PW-15	17	17	17	5	18	17	17	17
PW-16	16	16	17	16	16	16	17	17
PW-17	14	14	14	14	14	14	14	14
PW-18	14	14	16	16	14	14	16	16
PW-19	19	19	19	19	19	19	19	19
PW-20	17	17	17	17	17	17	17	17
Mean	14	**15**	14	14	13	14	**17**	16
SD	6	**6**	6	7	5	6	**6**	6

which the main aims are 1) to explore and compare the proposed H measure with other approaches to determining the overall interpretability of hierarchical systems; and 2) to refine the parameters of the proposed H measure.

In the first experiment, for the first step, a participatory user study was conducted to assess how users perceived the interpretability of the Iris systems. From the interpretability rankings provided by users, we found that the majority indicated that the parallel HFS was more interpretable than the flat FLS and serial HFS in Set MF-2 and Set MF-3 with a percentage of 76% and 72%, respectively (as shown in Tables III and IV). However, it was less clear cut as to whether the flat FLS was more interpretable than the serial HFS in Set MF-2 and Set MF-3.

Whilst, for the illustrative example, there is an absence of a clear relationship between the numerical results obtained for parallel and serial HFS systems, the comments of users (which can be seen in Supplementary material available with the digital copy of this article) indicate that the parallel form is more suited to the example of the Iris system. According to several users, it is more intuitive when *sepal* and *petal* are classified separately with the resulting outputs driving the classification of the Iris flowers. Similarly the participants expressed that fewer rules in each subsystem improved their readability. We do not believe that this preference is intrinsic to the parallel or serial form of decomposition, but is related to the natural structure (petals and sepals of flowers) inherent in this particular example.[1]

In the second step, we explored the interpretability of HFSs using the proposed *H* framework (H_{mean}) in comparison to that obtained without the framework, i.e., just using a plain average of subsystems (*Mean*). The results showed that while the *Mean* produced the same result for parallel and serial HFS (as it takes no account of the number of layers and topology), our framework produces results that are different depending on the topology of systems. The result obtained for the H_{mean} on the Iris system, particularly in configuration MF-2, produces a ranking that is closer to that given by users. Therefore, these observations and current evidence indicate that our *H* framework (H_{mean}) is better than a measure without the framework in capturing a natural concept of interpretability of HFSs.

Unfortunately, the first experiments undertaken on the Iris system did not have sufficient discriminatory power to help identify the most appropriate parameters (aggregation and layer-weighting strategies) of our framework. A second experiment was, therefore, carried out to derive the configuration of *H* framework using a more complex system, the rotary crane example. Note that this example has lower semantic meaning of its variables compared to the Iris classification used in the first experiment. Nevertheless, due to its inherently higher complexity (featuring six inputs), which means there are more possible hierarchical topologies, the second example has a higher discriminatory power to help identify the most appropriate parameters of *H* framework. In the first step, we carried out another user study to assess how people perceive the interpretability of 12 different configurations of the system through 20 pairwise system comparisons. Based on the opinions of 40 participants with a range of expertise, a diversity of perception regarding interpretability was found. The results imply that interpretability is very subjective and challenging to understand as views may vary greatly as to the interpretability of different system topologies. In the second step, alternative configurations of the *H* framework with various aggregation and layer-weighting strategies were used to measure interpretability. It can be seen from Table VIII that this more complex system produces different interpretability scores for almost all the different configurations of the system.

The final step is to examine the level of agreement in terms of interpretability between the pairwise comparison produced from aggregation and layer-weighting strategies as in Step 2, with the pairwise comparison obtained from participatory user study as in Step 1. The number of agreements between the users' views in Step 1 and *H* results in Step 2 show that the H_{min} aggregation strategy and *increasing weight* layer-weighted strategy produced the highest agreement score with a score of 15 and 17, respectively, when compared with the others. While the differences are relatively small, these results suggest that the *H* framework with configuration H_{min} aggregation strategy and *increasing weight* layer-weighted strategy as it produced the highest agreement with the users.

The proposed framework and user study raise some interesting issues that are worthy of further and more detailed study. One issue is "How is the experience of the participants measured?" and the associated question "Does it affect the results?" In our studies, we recruited a range of people from early stage Ph.D. students to Full Professors with many years experience of fuzzy systems. However, we did not formally assess their expertise. For obvious reasons, this might be a difficult matter to assess, as individuals may be reluctant to have their "expertise" measured! Nevertheless, it would surely be interesting to both attempt to measure actual expertise of fuzzy systems (rather than just self-reported expertise) and to explore whether this affects opinion of interpretability in any way. A second issue is "Is there a correlation between interpretability and the classification results?" It has been previously reported that there is a tradeoff between interpretability and accuracy [52], [53]. That is, the higher the interpretability of a given system, the lower its accuracy. Since accuracy concerns the ability of a model to make correct predictions, the same correlation may exist between interpretability and the classification results. For instance, if the classification results produce a higher accuracy, the classification result may have lessened their interpretability model. However, in this article, we are not showing any correlation between interpretability and classification results. We are focusing on introducing a general framework to capture interpretability of HFS.

The study of interpretability, particularly in the context of HFSs is an important area, which is likely to gain interest as it has clear relevance to explainable AI. The studies presented here show that there are sizeable differences in opinion between users as to the interpretability of various configurations of hierarchical systems, including with differing topologies and a range of sizes of rulebase.

VII. Conclusion

In conclusion, we have contributed a new generic framework for the measurement of the interpretability of HFSs, namely the *H* framework. This framework allows the use of any index for measuring the interpretability of a flat fuzzy system to be combined in any configuration of hierarchical systems with different numbers of subsystems, organised in differing topologies. We have then presented a participatory design process, consisting of two main experiments, which were aiming 1) to measure and compare the proposed *H* framework measure with others; and 2) to determine the selection of the best strategies for combining subsystems into an overall index of interpretability. Based on the current evidence, we tentatively suggest the use of the *min* operator to aggregate subsystems within a layer, together with the weighted mean operator using a *increasing weight* strategy

[1] An alternative approach to the one taken in this article would be to follow a data-driven design process for an interpretability index. This would require the acquisition of a large annotated data set from a well-balanced sample of people (where the annotation may be fine-grained and complex) together with the use of statistical optimisation techniques.

to combine layers, within the generic *H* framework for capturing the interpretability of HFSs.

Clearly, further work is also needed to explore the more general question of the wider meaning of interpretability of HFSs. Thus, in future, we expect further development of the *H* framework exploring other aspects of interpretability of HFSs, including the semantic interpretability of fuzzy sets, that of intermediate outputs and the logical complexity of the rules. For other future work, we will focus on conducting more experiments with different setting involving several case studies with more complex and varied hierarchical systems, including recruiting broader sets of participants from both within and outside the fuzzy community. Moreover, in future, we will also improve the agreement score, e.g., using the Spearman rank-order correlation with real numbers that may explore the difference between the HFS structure and considering the preferences indicated by the framework. In doing so, we would hope to gain further insight into different configurations of the framework, in order to ultimately gain a deeper understanding of the interpretability of HFSs, captured in a general index.

Acknowledgment

The authors would like to thank all participants for giving up their time for this article. Their answers and detailed comments were enormously valuable, and they have contributed enormously to this research.

References

[1] J. Casillas, O. Cordón, F. Herrera, and L. Magdalena, “Interpretability improvements to find the balance interpretability-accuracy in fuzzy modeling: An overview,” in *Interpretability Issues in Fuzzy Modeling*. Berlin, Germany: Springer, 2003, pp. 3–22.

[2] T. Furuhashi and T. Suzuki, “On interpretability of fuzzy models based on conciseness measure,” in *Proc. IEEE Int. Conf. Fuzzy Syst.*, 2001, vol. 1, 2001, pp. 284–287.

[3] D. Nauck and R. Kruse, “Obtaining interpretable fuzzy classification rules from medical data,” *Artif. Intell. Med.*, vol. 16, no. 2, pp. 149–169, 1999.

[4] R. Mikut, J. Jäkel, and L. Gröll, “Interpretability issues in data-based learning of fuzzy systems,” *Fuzzy Sets Syst.*, vol. 150, pp. 179–197, 2005.

[5] S.-M. Zhou, J. Garibaldi, R. John, and F. Chiclana, “On constructing parsimonious type-2 fuzzy logic systems via influential rule selection,” *IEEE Trans. Fuzzy Syst.*, vol. 17, no. 3, pp. 654–667, Jun. 2009.

[6] S. Jin and J. Peng, “Towards hierarchical fuzzy rule interpolation,” in *Proc. IEEE 14th Int. Conf. Cognitive Informat. Cognitive Comput.*, Jul. 2015, pp. 267–274.

[7] G. V. S. Raju, J. Zhou, and R. Kisner, “ Hierarchical fuzzy control,” *Int. J. Control*, vol. 54, no. 5, pp. 1201–1216, Nov. 1991.

[8] G. Raju and J. Zhou, “Adaptive hierarchical fuzzy controller,” *IEEE Trans. Syst., Man, Cybern.*, vol. 23, no. 4, pp. 973–980, Jul./Aug. 1993.

[9] A. D. Benítez and J. Casillas, “Multi-objective genetic learning of serial hierarchical fuzzy systems for large-scale problems,” *Soft Comput.*, vol. 17, no. 1, pp. 165–194, 2013.

[10] P. Salgado, “Rule generation for hierarchical collaborative fuzzy system,” *Appl. Math. Model.*, vol. 32, no. 7, pp. 1159–1178, Jul. 2008.

[11] R. Holve, “ Rule generation for hierarchical fuzzy systems,” in *Proc. Annu. Conf. North Amer. Fuzzy Inf. Process.*, 1997, pp. 444–449.

[12] M. Joo and J. Lee, “Hierarchical fuzzy control scheme using structured Takagi-Sugeno type fuzzy inference,” in *Proc. IEEE Int. Conf. Fuzzy Syst.*, 1999, vol. 1, pp. 78–83.

[13] O. Cordon, F. Herrera, and I. Zwir, “Linguistic modeling by hierarchical systems of linguistic rules,” *IEEE Trans. Fuzzy Syst.*, vol. 10, no. 1, pp. 2–20, Feb. 2002.

[14] A. Waldock, B. Carse, and C. Melhuish, “Hierarchical fuzzy rule based systems using an information theoretic approach,” *Soft Comput. Fusion Found., Methodologies Appl.*, vol. 10, no. 10, pp. 867–879, 2006.

[15] R. 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.

[16] E. Hüllermeier, “From knowledge-based to data-driven fuzzy modeling,” *Informatik-Spektrum*, vol. 38, no. 6, pp. 500–509, Dec. 2015.

[17] D. Nauck, “Measuring interpretability in rule-based classification systems,” in *Proc. IEEE 12th IEEE Int. Conf. Fuzzy Syst.*, 2003, pp. 196–201.

[18] S. Guillaume and B. Charnomordic, “A new method for inducing a set of interpretable fuzzy partitions and fuzzy inference systems from data,” in *Interpretability Issues in Fuzzy Modeling*. Berlin, Germany: Springer, 2003, pp. 148–175.

[19] J. M. Alonso, S. Guillaume, and L. Magdalena, “A hierarchical fuzzy system for assessing interpretability of linguistic knowledge bases in classification problems,” in *Proc. Inf. Process. Manage. Uncertainty Knowl. Based Syst.*, 2006, pp. 348–355.

[20] H. Ishibuchi and Y. Nojima, “Analysis of interpretability-accuracy tradeoff of fuzzy systems by multiobjective fuzzy genetics-based machine learning,” *Int. J. Approx. Reasoning*, vol. 44, no. 1, pp. 4–31, Jan. 2007.

[21] J. Alonso, L. Magdalena, and S. Guillaume, “HILK: A new methodology for designing highly interpretable linguistic knowledge bases using the fuzzy logic formalism,” *Int. J. Intell. Syst.*, vol. 23, no. 7, pp. 761–794, Jul. 2008.

[22] S.-M. Zhou and J. Gan, “Low-level interpretability and high-level interpretability: A unified view of data-driven interpretable fuzzy system modelling,” *Fuzzy Sets Syst.*, vol. 159, no. 23, pp. 3091–3131, Dec. 2008.

[23] J. M. Alonso, L. Magdalena, and G. González-Rodríguez, “Looking for a good fuzzy system interpretability index: An experimental approach,” *Int. J. Approx. Reasoning*, vol. 51, no. 1, pp. 115–134, Dec. 2009.

[24] A. Botta, B. Lazzerini, F. Marcelloni, and D. Stefanescu, “Context adaptation of fuzzy systems through a multi-objective evolutionary approach based on a novel interpretability index,” *Soft Comput.*, vol. 13, no. 5, pp. 437–449, Mar. 2009.

[25] J. M. Alonso and L. Magdalena, “Combining user's preferences and quality criteria into a new index for guiding the design of fuzzy systems with a good interpretability-accuracy trade-off,” in *Proc. IEEE Int. Conf. Fuzzy Syst.*, Jul. 2010, pp. 1–8.

[26] A. Riid and E. Rustern, “Interpretability improvement of fuzzy systems: Reducing the number of unique singletons in zeroth order Takagi–Sugeno systems,” in *Proc. IEEE Int. Conf. Fuzzy Syst.*, Jul. 2010, pp. 1–6.

[27] A. A. Marquez, F. A. Marquez, and A. Peregrin, “A multi-objective evolutionary algorithm with an interpretability improvement mechanism for linguistic fuzzy systems with adaptive defuzzification,” in *Proc. IEEE Int. Conf. Fuzzy Syst.*, Jul. 2010, pp. 1–7.

[28] C. Mencar, C. Castiello, R. Cannone, and A. M. Fanelli, “Interpretability assessment of fuzzy knowledge bases: A cointension based approach,” *Int. J. Approx. Reasoning*, vol. 52, pp. 501–518, 2011.

[29] M. J. Gacto, R. Alcalá, and F. Herrera, “Interpretability of linguistic fuzzy rule-based systems: An overview of interpretability measures,” *Inf. Sci.*, vol. 181, no. 20, pp. 4340–4360, 2011.

[30] A. R. De Soto, “A hierarchical model of a linguistic variable,” *Inf. Sci.*, vol. 181, pp. 4394–4408, 2011.

[31] D. Pancho, J. Alonso, O. Cordon, A. Quirin, and L. Magdalena, “FINGRAMS: Visual representations of fuzzy rule-based inference for expert analysis of comprehensibility,” *IEEE Trans. Fuzzy Syst.*, vol. 21, no. 6, pp. 1133–1149, Dec. 2013.

[32] D. P. Pancho, J. M. Alonso, and L. Magdalena, “Enhancing Fingrams to deal with precise fuzzy systems,” *Fuzzy Sets Syst.*, vol. 297, pp. 1–25, 2016.

[33] M. Pota, M. Esposito, and G. Pietro, “Interpretability indexes for fuzzy classification in cognitive systems,” in *Proc. IEEE Int. Conf. Fuzzy Syst.*, Jul. 2016, pp. 24–31.

[34] F. Kensing and J. Blomberg, “Participatory design: Issues and concerns,” *Comput. Supported Cooperative Work*, vol. 7, no. 3/4, pp. 167–185, Sep. 1998.

[35] M. Sugeno, M. Griffin, and A. Bastian, “Fuzzy hierarchical control of an unmanned helicopter,” in *Proc. 17th IFSA World Congr.*, 1993, pp. 179–182.

[36] F. Olga, H. Ioannis, K. Dimitris, and L. Spires, “Implementing participatory design for developing a constructivist e-learning activity,” in *Proc. 24th EAEEIE Annu. Conf.*, May 2013, pp. 157–162.

[37] C. Sjöberg and T. Timpka, “Participatory design of information systems in health care,” *J. Amer. Med. Informat. Assoc.*, vol. 5, no. 2, pp. 177–83, 1998.

[38] T. R. Razak, J. M. Garibaldi, C. Wagner, A. Pourabdollah, and D. Soria, "Interpretability indices for hierarchical fuzzy systems," in *Proc. IEEE Int. Conf. Fuzzy Syst.*, Jul. 2017, pp. 1–6.

[39] L. Magdalena, "Semantic interpretability in hierarchical fuzzy systems: Creating semantically decouplable hierarchies," *Inf. Sci.*, vol. 496, pp. 109–123, Sep. 2019.

[40] M.-L. Lee, H.-Y. Chung, and F.-M. Yu, "Modeling of hierarchical fuzzy systems," *Fuzzy Sets Syst.*, vol. 138, no. 2, pp. 343–361, Sep. 2003.

[41] L.-X. Wang, "Universal approximation by hierarchical fuzzy systems," *Fuzzy Sets Syst.*, vol. 93, no. 2, pp. 223–230, Jan. 1998.

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

[43] R. Yager, "On ordered weighted averaging aggregation operators in multicriteria decisionmaking," *IEEE Trans. Syst., Man, Cybern.*, vol. 18, no. 1, pp. 183–190, Jan./Feb. 1988.

[44] R. Yager, "Quantifier guided aggregation using OWA operators," *Int. J. Intell. Syst.*, vol. 11, no. 1, pp. 49–73, Dec. 1996.

[45] R. Yager, "Families of OWA operators," *Fuzzy Sets Syst.*, vol. 59, no. 2, pp. 125–148, Oct. 1993.

[46] K. J. Ostergaard, W. Wetmore, and J. D. Summers, "A methodology for the study of the effects of communication method on design review effectiveness," in *Proc. 29th ASME Des. Autom. Conf., Parts A B*, Jan. 2003, vol. 2, pp. 383–390.

[47] K. Balazs and L. T. Koczy, "New parameterizable search space narrowing technique for adjusting between accuracy and interpretability in fuzzy systems," in *Proc. IEEE 13th Int. Symp. Comput. Intell. Informat.*, Nov. 2012, pp. 323–328.

[48] C. Mencar and A. Fanelli, "Interpretability constraints for fuzzy information granulation," *Inf. Sci.*, vol. 178, no. 24, pp. 4585–4618, 2008.

[49] G. Klir and B. Yuan, *Fuzzy Sets and Fuzzy Logic: Theory and Applications*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1995.

[50] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Ann. Eugenics*, vol. 7, no. 2, pp. 179–188, Sep. 1936.

[51] N. Masmoudi, C. Rekik, M. Djemel, and N. Derbel, "Optimal control for discrete large scale nonlinear systems using hierarchical fuzzy systems," in *Proc. 2nd Int. Conf. Mach. Learn. Comput.*, 2010, pp. 91–95.

[52] D. Tikk and P. Baranyi, "Exact trade-off between approximation accuracy and interpretability: Solving the saturation problem for certain FRBSs," in *Interpretability Issues in Fuzzy Modeling*. Berlin, Germany: Springer, 2003, pp. 587–601.

[53] R. Alcalá, J. Alcalá-Fdez, J. Casillas, O. Cordón, and F. Herrera, "Hybrid learning models to get the interpretabilityaccuracy trade-off in fuzzy modeling," *Soft Comput.*, vol. 10, no. 9, pp. 717–734, Jul. 2006.