Fuzzy Rule-Based Explainer Systems for Deep Neural Networks: From Local Explainability to Global Understanding

Fatemeh Aghaeipoor, Mohammad Sabokrou, and Alberto Fernández

***Abstract*—Explainability of deep neural networks has been receiving increasing attention with regard to auditability and trustworthiness purposes. Of the various post-hoc explainability approaches, rule extraction methods assist to understand the logic that underpins their functioning. Whereas the rule-based solutions are directly managed and understood by practitioners, the use of intervals or crisp values in the antecedents that rely on numerical values might not be intuitive enough. In this case, the benefits of a linguistic representation based on fuzzy sets/rules are straightforward, as these semantically meaningful components ease the model understanding. This article proposes fuzzy rule-based explainer systems for deep neural networks. The algorithm learns a compact yet accurate set of fuzzy rules based on features' importance (i.e., attribution values) distilled from the trained networks. These systems can be used for both local and global explainability purposes. The evaluation results of different applications revealed that the fuzzy explainers maintained the fidelity and accuracy of the original deep neural networks while implying lower complexity and better comprehensibility.**

***Index Terms*—Attribution methods, deep neural network (DNN), EXplainable artificial intelligence (XAI), features importance, fuzzy rule-based systems, trustworthy.**

I. Introduction

EXPLAINABLE Artificial Intelligence (XAI) is the emerging resurgence of AI that attempts to fulfill the demand of understanding intelligent models [1]. This demand is even becoming more crucial in the applications such as healthcare or criminal justice, where the model's outcome may directly influence the mental or physical health of human beings [2], [3]. In this context, deep neural networks (DNNs), due to their remarkable feats in many of these sensitive areas, are greatly facing different challenges [4], [5]. DNNs are usually considered black-box models and they are often unable to comprehensibly explain how, why, and when they make a particular prediction and whether their high accuracy can be relied upon or not [6]. This information is critical to ensuring transparency, trustworthiness, and accountability of the intelligent methods and matters from both viewpoints of expert and especially nonexpert users [7].

Among different explainability strategies of DNN, we can mention post-hoc explanations, which try to explain the existing models without sacrificing their accuracy [8]. These methods take an already trained model and process it subsequently to uncover its underlying decision-making logic. Recent efforts in this area can generally be categorized into two groups [4]: 1) attribution methods, and 2) sequential decision processes. While the former lies in the methods that basically explain the model's prediction by discovering which features influence the predictions most, the latter focuses on explaining the logic of making the decisions. Attribution methods, such as Lime and SHAP [9], attribute an importance score to each feature of every single data example, which is why they are referred to as attribution algorithms. Local surrogate models, occlusion analysis, and gradient-based techniques are some of the most well-known approaches in this area [10].

Attribution methods are able to address some application requirements. These methods, especially the gradient-based ones, can effectively be used to explain image data. They are utilized to provide an understandable visualization of the most significant parts of an image in the prediction. In addition, they can be used to analyze decisions, contrast for the absence of bias, and determine the importance of the different features. Despite these benefits and the wide range of attribution methods proposed to date, these methods are unable to account for the model's decision-making process; thereby some recent studies have concluded that they do not necessarily result in a better grasp of the model's behavior [11], [12]. Moreover, they are not able to provide a straightforward comprehension of the tabular data, unlike their visual perception of the pixel-based data. This motivated us to consider ways to leverage the knowledge of these methods for tabular data, robustly and meaningfully.

By way of example, saliency maps, which denote how interesting a particular part of an image is, make no attempt to explain misclassifications and the reason(s) behind them. They may also be fragile and vulnerable to adversarial attacks. Moreover, from the explainability point of view, these saliency maps are totally local methods offering no global views over the model's functionality. Indeed, features that are important in

Manuscript received 10 October 2022; revised 18 December 2022; accepted 2 February 2023. Date of publication 10 February 2023; date of current version 1 September 2023. This work is partially supported by Spanish Ministry of Science and Technology under project PID2020-116518GB-I00 financed by MCIN/AEI/10.13039/501100011033 and by the Andalusian frontier regional project A-TIC-434-UGR20. *(Corresponding author: Fatemeh Aghaeipoor.)*

Fatemeh Aghaeipoor and Mohammad Sabokrou are with the School of Computer Science, Institute for Research in Fundamental Sciences, Tehran 19538-33511, Iran (e-mail: f.aghaei@ipm.ir; sabokro@ipm.ir).

Alberto Fernández is with the Department of Computer Science and Artificial Intelligence, DaSCI, University of Granada, 18071 Granada, Spain (e-mail: alberto@decsai.ugr.es).

Digital Object Identifier 10.1109/TFUZZ.2023.3243935

the local context may have little relevance to the global setting. Nevertheless, their full knowledge may guide us through a global understanding of the model.

An alternative to further explore the system behaviors is the sequential decision methods, which are dissecting predictions into semantically meaningful components, as carried out in the rule-based systems [13], [14]. These models assist to express the input–output relations in the form of convenient IF–THEN rules. To this end, there are different rule extraction techniques aiming at creating supportive or surrogate rule-based models over the trained artificial neural networks (ANNs) in general or DNN in particular [15]. These techniques are generally categorized into three groups:

1) decompositional methods, which process the networks neuron by neuron;
2) pedagogical techniques that treat the entire network as a black-box model; and
3) the eclectic ones which fuse both strategies [16].

The use of rule extractors can be a good approach to boost XAI in DNN models. However, in order to strengthen the comprehensibility of these rule-based surrogate models, we propose the use of fuzzy linguistic representation, which models the semantic knowledge of the input space in a way similar to human cognition [17]. Fuzzy rules, as the key elements of fuzzy rule-based systems (FRBSs), are constructed using fuzzy linguistic labels inspired by the human language. In addition, fuzzy variables provide smoother and wider coverage of the case studies. These characteristics make FRBSs as flexible and adaptable solutions to be applied surrogately for the DNNs' explanations. Indeed, the simple and transparent inner mechanism of the FRBSs can supplementarily support understanding of the decision-making process of the DNNs [18], [19].

Despite the aforementioned good properties of fuzzy rules and the potential of FRBSs for XAI, they are not very commonly used in the general scope of machine learning (ML). Specifically, there are a few works on the ANNs' explainability using fuzzy rules [20] and, to the best of our knowledge, no work for the DNNs. While the fusion of the highly informative features' attribution into FRBSs could assist in providing post-hoc explanation methods, which are neither as computational as decompositional methods nor as unconcerned as pedagogical ones with the DNNs' latent space. In this way, we are able to provide well-performing explanation methods for the DNNs in any scale/type without computation overhead.

In this study, we propose to directly exploit features' attribution values and distill them into fuzzy rule-based classifiers, resulting in what we called fuzzy rule-based explainer systems (FRBESs) for DNNs. These systems are able to imitate the performance of their corresponding DNNs and are generated in two main stages:

1) The operation relating to training the DNNs and obtaining the most important local features.
2) The procedure associated with creating fuzzy classifiers and optimizing them (yet preserving the fidelity of the original networks) to ensure low complexity systems acting in two different directions: 1) to limit the length of the antecedents of the rules, and 2) to get rid of redundant rules and promote those that truly cover the problem space.

To show the good behavior and robustness of the proposed FRBESs, they were constructed using three different attribution methods, as well as two aggregated versions of them. They were individually evaluated using six classification case studies in terms of interpretability, accuracy, and fidelity to the original networks. They were also validated by comparison to the basic FRBSs and the original DNNs. In addition, several statistical tests were conducted in order to bolster the findings derived from the analysis. Robustness of the proposed FRBESs against some adversarial attacks was evaluated as well. The obtained results, especially in terms of the models' fidelity, revealed that they can effectively be applied as post-hoc explainers of the DNNs and make their black-box functionality more transparent.

It is worth mentioning that, although the DNNs are widely used in the case of image and text data, they have also shown great capabilities in heterogeneous tabular and nonimage data. Taking this into account, and for the sake of consolidating this work scenario in ML and DNNs, the case studies of FRBESs will be focused on in this context.

The remainder of this article is structured as follows. Section II gives a general review of the fundamental concepts and related works necessary to support the proposed FRBESs. Section III introduces the proposed method and details the generation process of the components. Section IV includes the experimental study, results, and discussions. Finally, Section V concludes this study.

II. Preliminaries

In this section, the key concepts and principles that benefit understanding of the proposed algorithm are introduced. First, the fundamentals of fuzzy linguistic models are presented, and these systems are analyzed in terms of how they can potentially support explainable systems. Then, the attribution algorithms that are the basis to obtain explanation systems, and in particular the ones employed in this study are described. Finally, a review of some related works is presented in order to delve deeper into the available methodologies.

A. Fuzzy Linguistic Models

1) Components and Structure: Fuzzy rule-based classification systems, FRBCSs, as their name indicates, are built on fuzzy IF–THEN rules. In a typical FRBCS, the task of classification is accomplished by the interaction of two main components, namely a knowledge base (KB) and an inference module. The KB contains a rule base (RB) and a data base (DB) that are made of fuzzy rules and fuzzy sets, respectively. On the other hand, there is the inference module that includes a fuzzy reasoning method and the required fuzzification and/or defuzzification interfaces.

Suppose that we have dataset X with n data samples, m input variables, and l class labels. Each data sample is denoted as $X_i = (x_i^1, x_i^2, \ldots, x_i^m)$, in which $i = 1, \ldots, n$. This example belongs to the class label $y_i =\in C = \{c_1, \ldots, c_l\}$. This dataset is utilized by different rule learning strategies to generate fuzzy rules [21],

[22], [23]. These rules are typically created as the following format:

$$\text{Rule}_j : \textbf{If } x^1 \text{ is } A_j^1 \text{ and} \ldots, \text{ and } x^m \text{ is } A_j^m$$
$$\textbf{Then} \text{ class is } c_j : \text{RW}_j \tag{1}$$

where A_j^k is the corresponding linguistic label of the kth input variable $(k = 1, 2, \ldots, m)$, and c_j and RW_j are the class label and the rule weight, respectively. Alternatively, this rule can be interpreted as a fuzzy association rule such as the following structure:

$$R_j : A_j \rightarrow c_j : \text{RW}_j \quad ; \quad A_j = \{A_j^1, \ldots, A_j^m\} \tag{2}$$

where A_j is the set of antecedents and c_j is the consequence part of this rule. In this study, Mamdani fuzzy rules, which employ linguistic labels for both antecedent and consequence parts, are taken into account. This type of rule is more intuitive and interpretable, resulting in better fulfilling the XAI requirements [24]

$$\text{RW}_j = \frac{\text{matchClass}_j - \text{matchNotClass}_j}{\text{matchClass}_j + \text{matchNotClass}_j} \tag{3}$$

where match Class_j aggregates the matching degree of all the examples that are in the same class of rule j and it is obtained as

$$\text{matchClass}_j = \sum_{x_i \in c_j} \mu_{A_j}(x_i) \times \text{cost}(y_i). \tag{4}$$

Similarly, for inconsistent examples whose class labels are not matched with rule j, matchNotClass_j is calculated as follows:

$$\text{matchNotClass}_j = \sum_{x_i \notin c_j} \mu_{A_j}(x_i) \times \text{cost}(y_i) \tag{5}$$

in which $\text{cost}(y_i)$ is the misclassification cost associated with the class label y_i, and it is calculated by counting the frequency of the class labels (σ) as follows:

$$\text{cost}(y_i) = \frac{\max(\ \{\sigma(c_k) \mid c_k \in C\}\ )}{\sigma(y_i)}. \tag{6}$$

2) Fuzzy Linguistic Models to Improve XAI: In light of XAI and its human-centered nature, fuzzy modeling has open doors to greatly benefit from. They provide natural knowledge representation by employing linguistic and human-like terms. This assists to enhance the semantic knowledge of the models and facilitates human interactions. In particular, FRBCSs demonstrate their inference logic by way of simple and cognitively understandable If–Then rules and consequently provide direct insight into the prediction process [24].

In the context of XAI, further emphasis must be placed on the compactness of the RB and the semantic comprehensibility of the DB. Indeed, two main kinds of approaches are available in the literature to take into account the interpretability of linguistic FRBSs [24], [25]: Complexity-based interpretability and semantics-based interpretability. The former is devoted to decreasing the complexity of the obtained model, i.e., a few rules with good coverage and high level of confidence are desirable. Furthermore, rules with long antecedent parts are difficult to be handled by human cognition and they may cause interpretability issues. Conversely, the semantics-based approaches are devoted to preserving the semantics and the comprehensibility associated with the DB, in which a small number of linguistic fuzzy sets with homogeneous distribution are more meaningful and straightforward to be interacted with [17]. Given these statements, the explainability potentials of fuzzy systems can be leveraged to construct simple and transparent models, even to be accompanied by the other black-box models such as the DNNs and alleviate their transparency limitations.

B. Attribution Algorithms

In respect of covering scope, explanation methods are classified into two groups: global and local [1]. The former provides insight into how a model functions as a whole, that is, they explain the model's behavior for a range of input data. Conversely, local methods are trying to explain individual decisions made by a classifier for every single data example, e.g., in a medical scenario, imagine understanding of the model's prediction for a certain patient, where the doctor needs to make reliable judgments based on the outputs of the model.

Different local explainers are available in the literature [4]. They typically work by processing features-based data. Among them, attribution algorithms are designed to explore how much each input feature contributes to the model prediction, i.e., the input examples are assigned featurewise scores (attribution values), either positive or negative values. In this context, the positive values indicate that the corresponding features improved the class probability of that given output, whereas the negative values show that those features deteriorated the output class probability [26].

To compute the attribution values in ANNs, different strategies are applied. For instance, perturbation-based methods remove, mask, or alert the input features and then run a forward pass on the newly obtained data and measure the differences with the original output. In this way, the importance of those manipulated features is revealed. On the other hand, there are backpropagation-based methods, which compute the attribution values in a single forward and backward pass through the network. Sometimes, these methods are referred to as the gradient-based ones [27].

Among backpropagation-based methods, we can mention integrated gradients [28], DeepLIFT [29], and gradient SHAP [9] as some of the best-performing ones. In the former, gradients are computed with respect to the inputs and then integrated along the path from a given baseline to an input. DeepLIFT is one of the most efficient and noncostly attribution methods, which, instead of gradients, computes the differences with respect to a reference data point. This makes information can be propagated even when activation saturation happens and the gradients are zero. As a result of this as well as differentiating between positive and negative contributions, DeepLIFT is capable of uncovering dependencies that other methods do not recognize. Gradient SHAP is also a gradient-based method that computes SHAP values, which are based on Shapley values proposed in cooperative game theory.

C. *Related Works*

Among different rule extraction algorithms from ANNs, DeepRED [30] was one of the earliest that focused on DNNs with several hidden layers. This method decompositionally processes a network and generates intermediate rules for every single layer. In the second step, the rules are substituted backward to demonstrate the networks' behavior from inputs to outputs. This method was later modified in REM-D [31] by merging the rules incrementally and employing C5.0 as the intermediate rule extraction strategy, leading to a more accurate rule set and a more efficient algorithm in terms of time and memory. In both of these methods, the substitution step was accomplished in a termwise manner, resulting in an exponential postprocessing substitution step. In a quite recent contribution, ECLAIRE [32] proposes a clausewise substitution and consequently a polynomial rule extraction method from DNNs. Despite these efforts, the decompositional nature of these algorithms and processing the output of every single neuron cause some serious challenges such as excessive memory usage, a large set of rules, and high running times, especially in the case of big datasets and huge networks. These issues inspired the authors to wonder whether we can take advantage of the neurons' outputs indirectly, rather than going through them one by one e.g., utilizing the features' importance obtained in a single backward or forward pass using the neurons' outputs.

Another group of rule extraction methods is related to the fuzzy-based ones [20], which are in a very fewer range. They mostly focus on shallow networks. For instance, Tan et al. [18] proposed extraction of rules from a two-layer feed-forward neural network and Jin et al. [33] proposed an algorithm to extract fuzzy rules from the trained radial basis function (RBF) networks. In another work, an approach was developed in order to the fuzzy discretization of continuous input parameters and then extract the most dominant fuzzy rules from the trained binary single-layer neural networks [34]. That study took advantage of an adaptive weight-sharing algorithm and a neural network regularization technique. In [35], the antecedent parts of the fuzzy rules were created using the similarity of the input data and the networks' weight vectors. In this way, the fuzzy models were able to better uncover hidden knowledge of the networks. Even with these works, to the best of our knowledge, it is surprising that this field of study comprises so little research yet, especially for the DNNs. This can be probably due to the lack of knowledge of fuzzy from the general ML community.

III. FRBESs: Fuzzy Rule-Based Explainer Systems for DNNs

FRBESs are simpler and more manageable models having similar predictive performance to surrogately explain the DNNs functionality. These systems are built taking advantage of an already trained DNN rather than from scratch. Indeed, they utilize the trained networks to find out the most informative features, getting rid of usually expensive and nonprecise feature selection operations in conventional fuzzy modeling. To do so, the working procedure of creating FRBESs is composed of two main stages: one to create, configure, train, and fine-tune the DNNs (Section III-A); and the other to extract and optimize the fuzzy classifiers based on those trained DNNs (Section III-B). In view of the fact that the whole of this procedure is developed to satisfy the priorities of XAI, it also strives to keep the surrogate fuzzy classifiers as compact as possible, especially in terms of complexity at the level of DB and RB, as discussed in Section II-A. Algorithm 1 shows the details of the FRBESs generation, which is described in the following.

A. *Stage 1: Training DNN and Obtaining Features Importance*

Considering the general assumption of Section II-A, a DNN f_θ is trained on dataset X and for each input sample X_i, f_θ predicts the probability distribution of each class label as $f_\theta(X_i) = (p_i^{c1}, p_i^{c2}, \ldots, p_i^{cl})$. Then, the final class is predicted as

$$\widehat{y}_i = \arg\max\nolimits_{c_k \in C} \left(p_i^{c_k} \right). \tag{7}$$

In order to create a compact and efficient FRBES having short antecedent rules, we must reduce dimensionality of the RB by considering the subset of the most important features. For this purpose, we can make a great use of f_θ. Indeed, the knowledge of f_θ is exploited using the attribution algorithms and the importance degrees (attribution values) are assigned to the input features, i.e., as line 1 in Algorithm 1 shows, the value of Features' Importance is calculated for every single dimension of each example. These values will be the basis for the establishment of the FRBESs in the next stage.

B. *Stage 2: Establishment of the FRBESs*

In order to construct a comprehensive linguistic FRBS, two fundamental components of the systems, DB and RB, must be well defined. DB contains parameters of membership functions (MFs) to transform crisp values into fuzzy degrees and RB includes fuzzy linguistic If–Then rules to make inferences. In the following, the creating procedure of these two components as well as the output prediction process are described.

1) Generating DB: In this framework, triangular MFs and uniform fuzzy partitioning[1] are employed to fuzzify the input values. Indeed, a set of homogeneous fuzzy labels is defined on the domain of each variable of dataset X. That way, variable X^f, which is in the range of $U = [a^f, b^f]$, has $\mathrm{NS}(X^f)$ uniform fuzzy sets, in which $\mathrm{NS}(.)$ counts the number of fuzzy sets. Fig. 1 illustrates an example of this distribution with $\mathrm{NS}(X^f) = 3$.

2) Generating RB: The process of creating RB follows a standard learning procedure from examples. Specifically, it starts with developing initial rules that are then optimized to promote the interpretability perspectives of the system as well as the predictive performance of the model. This process is conducted through the three steps of generating initial candidate rules, pruning redundant rules, and selecting best rules, which are detailed in what follows.

Generating Initial Candidate Rules: The initial RB includes several candidate rules, which may potentially perform well

[1] The fuzzy partitioning methodology is considered linguistic, that is, all rules share the same fuzzy partitions for all variables [24].

Algorithm 1: Fuzzy Rule-Based Explainer Systems (FRBESs) for DNNs.

```
Input:
 DNN $f_\theta$, $X_{n\times m}$, $C = \{c_1, c_2, \ldots, c_l\}$, $maxLen$, $\alpha$ : $X_i = (x_i^1, x_i^2, \ldots, x_i^m)$, $i = 1, 2, .., n$
Output : FRBES (DB, RB)
1: Apply attribution algorithm on $f_\theta(X)$ and obtain $FI_X$ : $FI_{X_i} = (FI_{x_i^1}, FI_{x_i^2}, \ldots, FI_{x_i^m})$, $i = 1, 2, .., n$
2:     Build DB on top of $X$
 /* Create initial rule set */
3: Initialize RB $= \emptyset$
4: for $i \leftarrow 1$ to $n$ do
5:     for $k \leftarrow 1$ to $maxLen$ do
6:         $X'_i = \{(x'^1_i, x'^2_i, \ldots, x'^k_i) \mid x'_i \in X_i$ and $FI_{x'^1_i} \geq FI_{x'^2_i} \geq \ldots \geq FI_{x'^k_i} \ldots \geq FI_{x'^m_i}\}$
7:         $R_{i,k} =$ CHI($X'_i$, DB) with PCF-CS as $\mathrm{R}W_{R_{i,k}}$
8:         RB = RB $\cup \{R_{i,k}\}$
9:     end for
10: end for
11: Resolve rules' conflict in RB
12: RB$_{\text{Initial}} = \{R_i \in$ RB $\mid \mathrm{R}W_{R_i} > 0\}$
      /* Prune redundant rules */
13:     RB$_{\text{Pruned}} = \{R_j \in$ RB$_{\text{Initial}} \mid \nexists R_i \in$ RB$_{\text{Initial}}$ : $A_i \subset A_j$ and $c_i = c_j$ and $\mathrm{Conf}(R_i) \geq \mathrm{Conf}(R_j)\}$
      /* Select best rules */
14:     RB$_{\mathrm{c_k}}$ = Sort_and_Select( $\{R_i \in$ RB$_{\mathrm{Pruned}} \mid c_i = c_k\}$ , $\alpha$ ) : $\forall c_k \in C$
15:     RB$_{\text{Final}}$ = RB$_{\mathrm{c_1}} \cup$ RB$_{\mathrm{c_2}} \cup \ldots \cup$ RB$_{\mathrm{c_l}}$
16:     return FRBES (DB, RB$_{\mathbf{Final}}$)
```

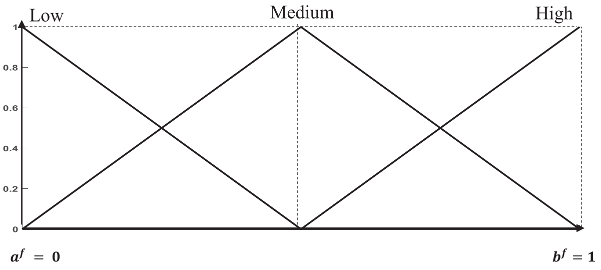

Fig. 1. Fuzzy partitioning of variable X^f, where $U = [0, 1]$.

considering all the data examples. To generate these rules, an adaptation of Chi et al.'s algorithm [22] is employed and fuzzy rules are built from the examples. However, unlike the conventional fuzzy algorithms, the rules are not built on top of all the input dimensions and they are created using the most fruitful features learned with the help of the DNN.

Chi algorithm is a fast rule learning approach that generates one fuzzy rule per example. It assigns antecedent labels using variables having maximum membership values. Indeed, after fuzzifying input values, the top $k \ll m$ features of every single example are identified based on the attribution values and then, the Chi algorithm is applied to make the corresponding rules and add them to the initial RB, i.e., for data example $X'_i = (x'^1_i, x'^2_i, \ldots, x'^k_i)$ with target class $y_i = c_i$ (see line 6 in Algorithm 1), fuzzy rule $R_{i,k}$ is generated as follows:

$$R_{i,k} : A_{i,k} \rightarrow c_i : \mathrm{R}W_{i,k} \quad ; \quad A_{i,k} = \{A_i^1, \ldots, A_i^k\} \tag{8}$$

where $A_{i,k}$ is the antecedent set of this rule and c_i is its consequence. This rule is exactly according to the original rule structure (2), in which the absent features[2] are marked as "don' care" linguistic labels, meaning that the classifier disregards those dimensions they pertain to. This rule has weight $\mathrm{R}W_{i,k}$, which is calculated with the method of penalized cost-sensitive certainty factor (PCF-CS), as (3).

Regarding the value of k, as lines 4-10 Algorithm 1 indicate, different combinations of top features (for each certain data example) are taken into account, ensuring having variant-length rules as well as not losing informative relations. Number of these combinations is specified using $maxLen$, which is a hyperparameter and determines the maximum length of the (initial) rule's antecedents. Setting $maxLen$ to the proper values is a case-dependent task and results in creating rules that are sufficiently representative.

Up to this step, we may have an RB with some possible conflicting rules, which have the same antecedent parts but different consequences. To resolve these conflicts, the rules must be evaluated to find out how strong they are with respect to their class labels. This task is straightforwardly accomplished using the RWs, i.e., that rule with the highest $\mathrm{R}W$ is kept in the candidate RB, and the others are removed.

The last modification before RB optimization is removing rules with nonpositive RWs. These rules may be destructive for their corresponding class labels in the whole dataset, as the contribution of inconsistent examples (examples whose class labels are not as the same as this particular rule) is higher than the consistent ones.

From the interpretability perspective of fuzzy systems, more compact RBs having short antecedent and minimum number of rules are desirable [24]. In this framework, the former was attained considering the aforementioned $maxLen$ parameter, and the latter is achieved through the two following optimization tasks.

Pruning Redundant Rules: In the initial rule learning process, we may face the case of covering rules, which are defined as rules that exclusively contain some antecedent part of another rule. These cases may happen for two reasons: 1) we considered variant-length combinations of top features for every single example, and 2) two different examples may have common top features. These rules, either the covered or the covering ones, may be redundant in the rule set and should be removed. Toward this end, in a pair of covering and covered rules, the most confident one is recognized by the measure of (fuzzy) confidence and chosen to remain, but the other to remove. In this context, the confidence degree of the jth rule is calculated by taking advantage of the previously computed matching degrees as follows:

$$\mathrm{Conf}(R_j) = \frac{\mathrm{matchClass}_j}{\mathrm{matchClass}_j + \mathrm{matchNotClass}_j} \tag{9}$$

[2] Features that have lower importance degrees and are not present in X'.

where $\mathrm{matchClass}_j$ and $\mathrm{matchNotClass}_j$ are estimated using (4) and (5), respectively.

As a means of clarifying the pruning process, suppose these two candidate rules, either generated from one example or two different examples, as follows:

$$\begin{cases} \mathrm{R}_i : \textbf{If } x^1 \text{ is } \mathit{Low} \text{ and } x^5 \text{ is } \mathit{High}, \\ \quad \textbf{Then} \text{ class is } c_1 : \mathrm{Conf}(R_i) = 0.899 \\ \mathrm{R}_j : \textbf{If } x^1 \text{ is } \mathit{Low} \text{ and } x^5 \text{ is } \mathit{High} \text{ and } x^9 \text{ is } \mathit{Low}, \\ \quad \textbf{Then} \text{ class is } c_1 : \mathrm{Conf}(R_j) = 0.765. \end{cases}$$

These rules are in the same class and the antecedent part of R_i is entirely covered by R_j. Additionally, $\mathrm{Conf}(R_i) \geq \mathrm{Conf}(R_j)$. In this case, R_j can be truly pruned, because there is a shorter and more confident rule that covers all the covered examples of R_j. As line 13 of Algorithm 1 shows, all such redundant rules are discarded and only those rules that do not have a better covering pair are kept in the RB. In this way, it is assured that we have the most “informative” and “shortest-length” combinations in the rules.

Selecting Best Rules: Finally, a selection process is carried out to retain more confident rules and improve the global interpretability of the model. Since the RB of the previous stage $(\mathrm{RB}_{\mathrm{Pruned}})$ may still include less-effective rules, it is purified in a classwise manner, i.e., the rules are descendingly sorted based on their confidence values, and the top $\alpha\%$ of each class are only retained in the final RB and the rest are discarded (lines 14 and 15 in Algorithm 1).

The idea of doing selection in a classwise manner is to ensure that the final RB is constructed with the top confident rules of every single class present in the initial dataset. This avoids removing rules of underrepresented classes, or those difficult rules that are overlapped and therefore are given lower confidences. Parameter α is a user-defined threshold to directly control the model complexity and provide different tradeoffs between the system interpretability and the model performance. Along the experiments, it was set following similar studies [23] as well as empirical trials.

C. Making Predictions

Having the DB and RB generated, FRBESs are ready to be used. In this step, a fuzzy reasoning method must be applied to make new predictions using the learned components. One of the alternative reasoning methods that provides a great level of explainability is the winning rule [36]. This method predicts the final class label using rule having the highest matching degree among all, and since there is only one rule involved, it is straightforward to determine the specific attributes and values that contribute to the decision. In this way, for the new example X_i, the output $\widehat{y_i}$ is predicted c_o, where c_o is the class label of the winning rule R_o, which is obtained as follows:

$$R_o = \arg\max_{R_j \in RB} \left\{\mu_{A_j}\left(X_i\right) \cdot \mathrm{RW}_j\right\} \tag{10}$$

in which $\mu_{A_j}(X_i)$ shows the matching degree of example X_i with rule R_j (1), and it is computed as

$$\mu_{A_j}\left(X_i\right) = \prod_{k=1}^{m} \mu_{A_j^k}\left(x_i^k\right) \tag{11}$$

where $\mu_{A_j^k}(x_i^k)$ returns the degree of membership for input value x_i^k in fuzzy set A_j^k. For the sake of simplicity, as (11) shows, the *t*-norm of product was applied as the aggregation function in all the experiments of this study.

IV. Experiments and Results

In this section, a detailed discussion of the conducted experiments is presented. First, the experimental framework, including the datasets, the configuration of the DNNs and the fuzzy systems, the evaluation criteria, and the comparing methods, is described. Next, the performance of the FRBESs in terms of accuracy and interpretability as well as their fidelity to the original DNNs is discussed. Finally, the robustness evaluation and statistical tests are supplementary provided to conclude the findings of this section.

A. Experimental Setup

1) Datasets: In this study, six classification datasets from different areas of real life were employed to execute the experiments [37], [38]. These datasets have different numbers of features and samples summarized in Table I. Some of these datasets are related to medical applications in which providing explainable yet accurate models is certainly a critical issue. The fivefold cross-validation mechanism was applied to generate the training and test data of the experiments and the average of the five trials performed on the obtained folds have been reported as the final results.

2) Configuration of the DNNs: A fully connected network, with three hidden layers and `relu` activation function in between, was considered to learn each task. The networks were examined over different hyperparameters to find out the best architecture and configuration in terms of number of neurons per layer, batch size, and learning rate. Indeed, a grid search with the following search space and early stopping of the evaluation of bad trials was performed for each task; $\{256, 128\}$, $\{64, 32\}$, and $\{16, 8, 4\}$ for size of first, second, and third hidden layers, respectively, $\{32, 64\}$ for the batch size, and $\{0.01, 0.001\}$ for the learning rate. Furthermore, the networks were trained using an Adam optimizer for 150 epochs to minimize the weighted classification cross-entropy loss. All these settings are in line with encoding architectures of several state-of-the-art rule extraction methods of the DNNs [31], [32]. The training processes were run using the GPU resources offered by Google’s Colab services.

3) Configuration of the FRBESs: In order to investigate the effect of attribution methods on the performance of the proposed systems, we constructed FRBESs on top of three attribution methods, namely DeepLIFT [29], integrated gradients [28], and gradient Shap [9] (in the following tables, these systems are abbreviated as FRBES_DL, FRBES_IG, and FRBES_GS,

TABLE I
PROPERTIES OF THE USED DATASETS

Dataset	Area	# Samples	# Classes	# Features	Majority Class(%)
XOR	Synthetic	1,000	2	10	52.6
MB_GE_ER	Healthcare	1,980	2	1,000	76
MB_Hist	Healthcare	1,695	2	1,004	91.3
MAGIC	Particle Physics	19,020	2	10	64.8
MiniBoo	Particle Physics	130,065	2	50	71.9
Letter	Recognition	20,000	26	16	11.7

respectively). These methods were chosen because they are some of the most well-established and best-performing attribution approaches that serve as good representative methods as well. In order to satisfy the conditions of equivalence, all these methods are among the backpropagation-based ones (see Section II-B for details). In addition, given the hypothesis that an ensemble of several multiple attribution methods would be more robust than any single method, we also employed two aggregated versions of them, namely AGG-Mean and AGG-Var (abbreviated as FRBES_Mean and FRBES_Var). The former takes the pointwise average over the attribution values of all the available methods and the latter considers the local variance as well (see [39] for details).

In all the FRBESs, $maxLen$ was set to 3 in accordance with the previous fuzzy models having appropriate complexity and accuracy tradeoffs [21], [23]. Regarding the selection parameter of α, it was set depending on the use case under study, trying to keep it as low as possible in the direction of XAI priorities. Nevertheless, in most tasks, it was observed that the values lower than 0.5 tend to give more compact and efficient rule sets.

4) Evaluation Criteria and Statistical Tests: In the following tables, the values of seven evaluation criteria, namely ACC, AUC, Fidelity, $\#\mathrm{F_c}$, $\#\mathrm{R}$, ARL, and time, have been reported to assess the accuracy, complexity,[3] and efficiency of the models. ACC and AUC were considered to measure the discrimination capability of the proposed systems. AUC was employed to demonstrate the classification success considering all the class labels and also cover the imbalance cases [40].

Another essential performance criterion is Fidelity, which signifies how much reliable the surrogate models are. In our context, Fidelity refers to how truthfully the fuzzy explainers portray the underlying networks. This measure is computed as the percentage of match between the predictions of the two models, namely the original DNN and its corresponding FRBES. Despite the fact that it is almost impossible for an explanation to be fully faithful unless it is a complete description of the model itself. In order for an explainer to be reliable, it must at least appear to be generally faithful and possess a relatively acceptable level of Fidelity, i.e., the higher the Fidelity, the closer the performance of the models. Last but not least, this measure is emphasized in alignment with the other accuracy criteria [41].

In terms of complexity measures, $\#\mathrm{F_c}$ shows the number of features contributing to the whole process of modeling, e.g., in FRBESs, features that appeared in the final rule sets are aggregated to compute $\#\mathrm{F_c}$. Indeed, with the functionality of DNNs in feature learning and utilizing the most important of these features in the construction of FRBESs, we expect a substantial reduction in the number of used features. This achievement is highly desirable in the context of XAI, especially for high-dimensional cases.

In addition to the above-mentioned criteria, the values of $\#\mathrm{R}$ and ARL were calculated to quantify the compactness of the rule sets. $\#\mathrm{R}$ shows the number of rules available in the final RB, and ARL indicates the average length of them. The amount of time (h:mm:ss) taken to provide the post-hoc explainers are also reported for each problem. These times are related to the whole process of learning and classifying.

To provide more comprehensive comparisons, several statistical tests, including Friedman's and Holm's, were conducted in this study [42]. Friedman's test ranks the algorithms according to a certain criterion considering all the datasets. This test begins by evaluating the equality hypothesis ($\mathrm{H_0}$) of all the algorithms, which can be either accepted or rejected. Additionally, the post-hoc Holm's test is conducted to compare the performance of two methods versus each other, namely a certain control method versus each of the remaining ones. These tests calculate the measure of `p-value` and determine whether the hypothesis is accepted/rejected. This evaluation is conducted using the significance level parameter α, which was set to 0.1 in this study.

5) Comparing Methods: To the best of our knowledge, there is no fuzzy rule-based algorithm to explain DNNs. Being a pioneering approach, we were unable to find a fuzzy-based baseline to compare with. Therefore, we chose ECLAIRE [32] as the comparison method, which is the state-of-the-art rule extraction algorithm for DNNs and creates efficient yet compact explainer systems composed of crisp rules. Although this method follows a different strategy and is not comparable to FRBESs in terms of time and memory, it helps us to obtain a good view of accuracy/fidelity and interpretability of the explainers.

Additionally, we have included the results of original DNNs and Chi_FRBCS. The latter is a fuzzy rule learning method based on the standard Chi et al.'s algorithm performing on all the input features (see Section III-B2). It assists to confirm the necessity of using DNNs rather than applying inherently explainable models (such as FRBSs) from scratch.

B. Results and Discussions

In the following, results of different applications are separately evaluated and discussed. In addition, several statistical and robustness tests are conducted to further analyze the results.

[3] The interpretability is considered based on the complexity (not semantics) of the FRBSs [24].

TABLE II
FINAL EXTRACTED RULES FROM THE TRAINED DNN OF THE XOR TASK (HIDDEN LAYER SIZES: {256,64,16}, LR: 0.001, AND BATCH_SIZE: 32)

Rule	Antecedents			Class	Conf
R1	x^1 is High	x^2 is High	-	0	0.927
R2	x^3 is High	x^9 is High	x^{10} is Low	0	0.618
R3	x^4 is High	x^5 is High	x^{10} is High	0	0.615
R4	x^1 is Low	x^2 is High	-	1	1.0
R5	x^1 is High	x^2 is Low	-	1	0.999

1) A Synthetic Case Study (XOR Dataset): In the first case, a synthetic dataset was employed to demonstrate the performance of FRBESs. We used a variant of XOR dataset having ten input features and 1000 data examples. This dataset has commonly been used for studying feature selection methods and is known as a challenging dataset for vanilla rule induction algorithms [32], [43]. In this problem, input values were independently generated using a uniform distribution in the range of $[0, 1]$, and binary output labels were assigned by performing the XOR operation on the rounded values of the first two features, namely x^1 and x^2. To efficiently perform this task, the classifier must be able to learn that only the first two dimensions are relevant for predicting the output label. In this regard, the DNN's latent space could effectively be utilized not only to guide the rule set construction but also to identify the most important and meaningful features for the classification task.

First, to illustrate the performance of the FRBESs, we indicate the rules extracted from the best-performing DNN (in terms of the validation's accuracy obtained by different grid search trails of a certain fold) in Table II. This FRBES was made using the DeepLIFT attribution values (FRBES_DL). As can be seen, the rule set is truly compact and has only five rules with maximum of three antecedents. Furthermore, in R1, R4, and R5, the main features x^1 and x^2, were finely recognized and the XOR patterns were properly described using fuzzy labels. These rules have also a greater degree of confidence rather than the less precise ones, namely R2 and R3. This indicates the ability of FRBES_DL in identifying the most reliable relations hidden in the knowledge of the DNNs.

Table III reports results of FRBESs as well as the other comparing methods. The best values of each dataset (among the explainers) have been indicated in bold. In the XOR rows of this table, there are considerable differences between the measures of the original DNN and the Chi_FRBCS. This implies first, the necessity of using the power of DNNs in detecting strong features of the problem space, and second, the potential advantages of the FRBESs to easily represent this knowledge. That is, fuzzy logic is needed for explainability purposes, but FRBESs are proposed to address the interpretability issues better than a simple FRBCS. The results also show that the FRBESs are able to efficiently follow the behavior of the DNNs, in which the accuracy measures are close to the original DNN and the Fidelity values are acceptably high. In this case, all the fuzzy explainers performed more or less similar but FRBES_DL was slightly better with ACC 96.50% and Fidelity 93%. In comparison to the other rule-based explainer, ECLAIRE, all the FRBESs performed better in terms of both accuracy and fidelity.

Regarding the complexity of the systems, FRBESs generated significantly fewer rules than ECLAIRE and they made shorter rules as well. Such results indicate that FRBESs are able to properly capture the underlying classification process and provide straightforward and cognitively convenient explanations for what DNNs learn from the XOR dataset.

2) Medical Applications (MB_GE_ER and MB_Hist): FRBESs were evaluated using two real applications of medical areas where, as commented, both accuracy and explainability of predictive models play a critical role in making and/or supporting decisions. MB_GE_ER and MB_Hist datasets have been created based on the METABRIC data to predict the immunohistochemical and histological subtype of the breast cancer patients, respectively [31], [37]. These datasets are almost high-dimensional, comprising 1000 and 1004 mRNA expression patterns.

Results of these datasets have been reported in Table III. Like the previous case, the performance of Chi_FRBCSs in both datasets is obviously behind the other methods and we must take advantage of the DNNs. In the case of MB_GE_ER, FRBES_IG performed closer to the DNN and has the highest Fidelity among all. All the other FRBESs have also good results of accuracy and fidelity. However, the most notable achievement in this problem is in the interpretability measures, where the fuzzy systems are able to explain the underlying data with around 30 rules and two antecedents. Furthermore, the number of contributing features in the modeling process of FRBESs considerably reduced, i.e., from 1000 features to 10.2 in average (more than 98%). This matter helps to present more manageable and understandable explanations, and in this sense, FRBESs can be considered as a bridge that leads us from the most important local features to the most critical global ones.

The other case, MB_Hist, is a special and interesting case, where the efficiency of the aggregated methods is better revealed. This dataset is highly imbalanced. Among FRBESs, FRBES_IG and FRBES_GS obtained the highest values of Fidelity and generated 2 and 2.5 rules, respectively. However, it seems that these rules are not truly enough to cover all the class labels and the high values of Fidelity are not reliable, because they are not in companion with acceptable levels of AUC, and that is the circumstance that we mentioned Fidelity must be emphasized in alignment with the other accuracy criteria. On the other hand, we have FRBES_Mean and FRBES_Var with more robust results considering all the class labels, i.e., the higher performance measures and the higher number of rules are implying that the aggregation of attribution methods could potentially work better for the imbalance cases. It will, however, take a separate effort to study the performance of fuzzy explainers for imbalanced cases in the future. Finally, it is worth noting that ECLAIRE is outperformed in nearly all the measures for this dataset as well.

3) Particle Physics Tasks (MAGIC and MiniBoo): In order to examine the scalability of the models, two classification tasks from the particle physics area, namely MAGIC and MiniBoo, were chosen [38]. These datasets have a large number of training

TABLE III
RESULTS OF DIFFERENT METHODS

Dataset	Method	ACC(%)	AUC(%)	Fidelity(%)	#Fc	#R	ARL	Time
XOR ($\alpha = 0.1$)	Chi_FRBCS	60.90	64.25	-	10	752	10	0:00:15
	DNN	94.00	98.84	-	10	-	-	-
	ECLAIRE	91.80	91.40	91.40	N/A	87	3.03	0:00:04
	FRBES_DL	**96.50**	**99.47**	**93.00**	**8**	**8**	2.63	0:00:04
	FRBES_IG	95.00	98.41	92.31	8.8	9.2	2.81	0:00:05
	FRBES_GS	95.40	98.48	92.20	9	11.2	2.73	**0:00:03**
	FRBES_Mean	95.39	97.78	92.40	9.2	10.2	2.80	0:00:04
	FRBES_Var	96.41	97.32	92.86	9	12	**2.33**	0:00:04
MB_GE_ER ($\alpha = 0.1$)	Chi_FRBCS	53.03	67.47	-	1000	1584	1000	0:16:51
	DNN	95.30	98.74	-	1000	-	-	-
	ECLAIRE	**94.10**	91.80	94.70	N/A	48.4	2.84	0:01:00
	FRBES_DL	90.24	94.11	91.21	11	**30.**1	**1.79**	**0:00:10**
	FRBES_IG	91.16	**95.06**	**95.69**	**9**	33	2.03	0:00:12
	FRBES_GS	88.33	94.46	90.10	10.2	31.2	1.93	0:00:15
	FRBES_Mean	89.80	94.93	91.67	11.6	30.2	2.39	0:00:11
	FRBES_Var	90.80	93.97	90.67	9.2	32.2	1.89	**0:00:10**
MB_Hist ($\alpha = 0.3$)	Chi_FRBCS	88.66	61.30	-	1004	1355.2	1004	0:10:53
	DNN	91.32	83.54	-	1004	-	-	-
	ECLAIRE	**88.9**0	77.40	89.40	N/A	30	2.49	0:00:54
	FRBES_DL	84.32	80.02	91.21	13	28	**1.96**	**0:00:08**
	FRBES_IG	86.50	50.68	**96.31**	**5**	**2**	2.50	0:00:10
	FRBES_GS	86.73	52.73	95.36	6	2.5	2.58	0:00:11
	FRBES_Mean	88.50	**83.68**	91.53	12.8	29.5	2.50	**0:00:08**
	FRBES_Var	88.50	**83.68**	91.53	12.8	29.5	2.50	0:00:09
MAGIC ($\alpha = 0.2$)	Chi_FRBCS	76.87	81.99	-	10	309	10	0:03:12
	DNN	84.06	73.48	-	10	-	-	-
	ECLAIRE	**84.60**	80.20	87.40	N/A	396.2	3.82	**0:00:58**
	FRBES_DL	82.79	**85.96**	85.22	9.8	**35.8**	2.12	0:01:16
	FRBES_IG	83.17	85.10	**87.53**	10	42	2.06	0:01:23
	FRBES_GS	82.14	83.49	82.78	**9.5**	37	2.08	0:01:10
	FRBES_Mean	83.98	84.26	82.34	10	40	**2.04**	0:01:15
	FRBES_Var	83.12	84.66	82.34	9.6	42.4	**2.04**	0:01:11
MiniBoo_NE ($\alpha = 0.4$)	Chi_FRBCS	72.96	59.26	-	50	54.4	50	0:16:33
	DNN	93.54	48.61	-	50	-	-	-
	ECLAIRE	91.40	90.50	**94.60**	N/A	1484.8	5.81	0:48:51
	FRBES_DL	**91.57**	89.35	93.78	19	53	2.33	**0:17:05**
	FRBES_IG	91.23	**91.69**	92.71	**17**	**50**	2.28	0:18:03
	FRBES_GS	90.21	89.19	91.48	19.5	54	**2.26**	0:17:55
	FRBES_Mean	89.29	90.51	93.26	18	51	2.37	0:17:13
	FRBES_Var	90.27	89.99	93.15	18	52	2.32	0:18:01
Letter ($\alpha = 0.2$)	Chi_FRBCS	41.09	-	-	16	283.60	16	0:02:11
	DNN	57.25	-	-	16	-	-	-
	ECLAIRE	**55.70**	-	55.70	N/A	1219.4	5.41	0:07:53
	FRBES_DL	54.32	-	**56.11**	12.3	**31**	2.50	0:04:24
	FRBES_IG	53.76	-	54.72	12	37	2.42	0:04:03
	FRBES_GS	54.85	-	53.15	**11.5**	35	**2.41**	**0:03:39**
	FRBES_Mean	54.51	-	54.58	12	35	2.50	0:04:44
	FRBES_Var	53.43	-	54.43	12	41	2.50	0:04:51

examples, thereby prone to an increasing number of rules and high computation times. But, according to the results of the FRBESs in Table III, not only the number of rules is not high, but also they have extremely reduced, compared to ECLAIRE, and the obtained rules are shorter in both cases. In the case of MiniBoo, the number of contributing features decreased from 50 to 18.3 in average, and the process of fuzzy modeling was also faster. All these results were obtained while the accuracy and fidelity of DNNs, FRBESs, and ECLAIRE were mostly in the same range. It means that FRBESs provide less complex

TABLE IV
RESULTS OF FRIEDMAN'S TESTS ON THE ACC AND FIDELITY VALUES

Algorithm	ACC	Fidelity
	Rank	Rank
DNN	**2**	-
ECLAIRE	2.8	3.1
FRBES_DL	4.3	2.6
FRBES_IG	4.6	**2.5**
FRBES_GS	5.1	4.8
FRBES_Mean	4.6	3.6
FRBES_VAR	4.4	4.1
p-value	0.11	0.23
Result	H_0 is not rejected	H_0 is not rejected

yet accurate surrogate models in tolerable times, even for large datasets.

4) A Multiclass Case (Letter Recognition): The last case is a Letter Recognition problem [38], which identifies the class of images among 26 English capital letters (A to Z). This dataset has 20 000 representations of black-and-white images, which were employed to extract 16 statistical features to perform the classification task. Results of this task have been shown in the last rows of Table III. As indicated, all the explainers are reliably following the performance of the DNNs and the differences lie in the complexity criteria, where the results FRBESs are much more efficient than the best results of ECLAIRE in this perspective. Although these findings adequately justify the functionality of the DNNs and confirm the proposed algorithm is a general solution that can easily be applied even in multiclass problems, taking advantage of multiclass/imbalanced strategies [44], [45], [46] may result in better coverage of the rule set and consequently improve the efficiency of the explainers, which we leave as the future research to explore.

5) Statistical Tests: Throughout this section, results of the conducted statistical tests are reported. First, Table IV shows the ranking values of Friedman's test for the performance measures, namely ACC and Fidelity. In both of these tests, H_0 is not rejected (p-values are higher than the significance level 0.1), meaning that all the methods are statistically as accurate as the DNNs and preserve fidelity up to the same levels.

As the above results, the major differences probably lie in the complexity of the models. To investigate this, we performed Friedman's test for #R and ARL, as indicated in Table V. This time, H_0 is rejected for #R and is not rejected for ARL, implying that there is a significant difference in the compactness of the generated RBs in terms of number of rules, while the average length of the rules is more or less the same. The ranking results also emphasize that the FRBES_DL has the most compact RB (vertically), while ECLAIRE has the largest one (both vertically and horizontally). Therefore, it seems that DeepLIFT could be an appropriate option to be used in the explainability frameworks.

To take the last step, the Holm's post-hoc test was conducted for #R with ECLAIRE as the control method. Results of Table VI confirm that, in this perspective, ECLAIRE is considerably outperformed by nearly all the FRBESs (except the case of FRBES_Var). As these statements, FRBESs make effective use of linguistic fuzzy systems and provide more straightforward and meaningful yet accurate models to explain the predictive logic of the DNNs.

TABLE V
RESULTS OF FRIEDMAN'S TESTS ON THE COMPLEXITY MEASURES

Algorithm	#R	ARL
	Rank	Rank
ECLAIRE	6	5.3
FRBES_DL	**1.8**	2.8
FRBES_IG	2.8	3.3
FRBES_GS	3.1	3
FRBES_Mean	2.8	3.9
FRBES_Var	4.4	**2.6**
p-value	0.0001	0.105
Result	H_0 is rejected	H_0 is not rejected

TABLE VI
RESULTS OF HOLM'S TEST FOR #R WITH CONTROL METHOD ECLAIRE

Comparison	Statistic	Adjusted p-value	Result
ECLAIRE vs FRBES_DL	3.86	0.001	H_0 is rejected
ECLAIRE vs FRBES_IG	2.93	0.01	H_0 is rejected
ECLAIRE vs FRBES_Mean	2.93	0.01	H_0 is rejected
ECLAIRE vs FRBES_GS	2.70	0.01	H_0 is rejected
ECLAIRE vs FRBES_Var	1.46	0.14	H_0 is not rejected

TABLE VII
EVALUATING THE MODELS' ROBUSTNESS AGAINST ADVERSARIAL ATTACKS FOR THE XOR PROBLEM

Attack	Method	Train	Test	
		ACC(%)	ACC(%)	Fidelity(%)
PGD	DNN	51.98	58.30	53.42
	FRBES	63.60	68.30	62.70
SPSA	DNN	42.60	42.40	42.01
	FRBES	59.18	58.90	59.10

In addition to the above findings, Tables IV and V show that trials with different attribution algorithms statistically have close performance, implying that to form reliable FRBESs, distilling attribution values efficiently is more critical than the type of attribution methods to generate these values, i.e., the most influential factor on the systems' performances is the process of creating and optimizing FRBESs.

6) Robustness Tests: Finally, to find knowledge about robustness of the FRBESs, we examined the models against two of the most commonly used attacks that are stable and efficient at present, one gradient-based method namely projected gradient descent (PGD) [47] and one gradient-free method namely simultaneous perturbation stochastic approximation (SPSA) [48]. In this way, the generated adversarial examples of the methods were fed into the trained DNNs and their corresponding FRBES_DL, and then the performances were evaluated. Table VII shows the results of these attacks. As can be seen, in both cases, the accuracy values of the FRBESs deteriorated less than the DNNs. Additionally, given the Fidelity values, the FRBESs preserved the original models' performances to a higher extent than the DNNs, indicating the better robustness of the surrogate models against adversarial noisy data. Despite these results, studying the robustness of explainer systems from different points of view is a pivotal subject that needs to be approached in separate works.

In the end, it is worth mentioning that FRBESs are fully deployable models, which "simulates" the behavior of the original complex DNNs so that by means of "explaining" the decision/inference process of the DNNs, a complete decision support system is obtained, and this is not just explaining single decisions, but having new simplified models, as well as a methodology to generate them.

V. Conclusion

In this work, we proposed FRBESs to bridge between DNNs and FRBSs. We took advantage of attribution methods to supplementarily distill the DNNs knowledge into a set of fuzzy rules, which clarify the DNNs' decision process in the classification problems of tabular data.

In practice, we are interested in explaining how DNNs predict a single data point (interpretability) as well as their statistics for a whole dataset (explainability). In the proposed FRBESs, the former is achieved by observing the fired fuzzy rules and the latter by delving into the obtained fuzzy rules associated with every single class label. In addition, the synergy between the DNNs and the FRBESs obtained from them provides fast, robust, and competitive models that also have a straightforward understanding for the practitioner.

Our future work will focus on investigating how these fuzzy classifiers work when applied to homogeneous datasets like images. Convolutional neural networks could be among the best candidate for improving the discovery of the potential relationships between different parts of the learned feature maps, thereby facilitating the application of intelligent methods in analyzing medical images or other critical data.

References

[1] A. B. Arrieta et al., "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, 2020.

[2] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (XAI): Toward medical XAI," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 11, pp. 4793–4813, Nov. 2021.

[3] U. Ahmed, J. C.-W. Lin, and G. Srivastava, "Fuzzy explainable attention-based deep active learning on mental-health data," in *Proc. IEEE Int. Conf. Fuzzy Syst.*, 2021, pp. 1–6.

[4] W. Samek, G. Montavon, S. Lapuschkin, C. J. Anders, and K.-R. Müller, "Explaining deep neural networks and beyond: A review of methods and applications," *Proc. IEEE*, vol. 109, no. 3, pp. 247–278, 2021.

[5] J. P. Amorim, P. H. Abreu, A. Fernández, M. Reyes, J. Santos, and M. H. Abreu, "Interpreting deep machine learning models: An easy guide for oncologists," *IEEE Rev. Biomed. Eng.*, vol. 16, pp. 192–207, Nov. 30, 2021.

[6] W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, and K.-R. Müller, *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, vol. 11700. New York, NY, USA: Springer, 2019.

[7] A. Weller, "Transparency: Motivations and challenges," in *Explainable AI: Interpreting, Explaining Visualizing Deep Learning*, New York, NY, USA: Springer, 2019, pp. 23–40.

[8] F.-L. Fan, J. Xiong, M. Li, and G. Wang, "On interpretability of artificial neural networks: A survey," *IEEE Trans. Radiat. Plasma Med. Sci.*, vol. 5, no. 6, pp. 741–760, Nov. 2021.

[9] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Adv. Neural Inf. Process. syst.*, vol. 30, pp. 4768–4777, 2017.

[10] C. Molnar, *Interpretable Machine Learning*, 2018. [Online]. Available: https://christophm.github.io/interpretable-ml-book/

[11] B. Kim et al., "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV)," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 2668–2677.

[12] Y. Zhou, S. Booth, M. T. Ribeiro, and J. Shah, "Do feature attribution methods correctly attribute features," in *Proc. 36th AAAI Conf. Artif. Intell.*, 2022, pp. 9623–9633.

[13] R. Guidotti, A. Monreale, S. Ruggieri, D. Pedreschi, F. Turini, and F. Giannotti, "Local rule-based explanations of black box decision systems," 2021, *arXiv:1805.10820*.

[14] T. Hailesilassie, "Rule extraction algorithm for deep neural networks: A review," 2016, *arXiv:1610.05267*.

[15] E. Soares, P. P. Angelov, B. Costa, M. P. G. Castro, S. Nageshrao, and D. Filev, "Explaining deep learning models through rule-based approximation and visualization," *IEEE Trans. Fuzzy Syst.*, vol. 29, no. 8, pp. 2399–2407, Aug. 2021.

[16] C. He, M. Ma, and P. Wang, "Extract interpretability-accuracy balanced rules from artificial neural networks: A review," *Neurocomputing*, vol. 387, pp. 346–358, 2020.

[17] A. Fernandez, F. Herrera, O. Cordon, M. J. del Jesus, and F. Marcelloni, "Evolutionary fuzzy systems for explainable artificial intelligence: Why, when, what for, and where to," *IEEE Comput. Intell. Mag.*, vol. 14, no. 1, pp. 69–81, Feb. 2019.

[18] X. Tan, Y. Zhou, Z. Ding, and Y. Liu, "Selecting correct methods to extract fuzzy rules from artificial neural network," *Mathematics*, vol. 9, no. 11, pp. 1164–1186, 2021.

[19] J. M. Mendel and P. P. Bonissone, "Critical thinking about explainable AI (XAI) for rule-based fuzzy systems," *IEEE Trans. Fuzzy Syst.*, vol. 29, no. 12, pp. 3579–3593, Dec. 2021.

[20] J. Li et al., "Explainable CNN with fuzzy tree regularization for respiratory sound analysis," *IEEE Trans. Fuzzy Syst.*, vol. 30, no. 6, pp. 1516–1528, Jun. 2022.

[21] M. Elkano, J. A. Sanz, E. Barrenechea, H. Bustince, and M. Galar, "CFM-BD: A distributed rule induction algorithm for building compact fuzzy models in Big Data classification problems," *IEEE Trans. Fuzzy Syst.*, vol. 28, no. 1, pp. 163–177, Jan. 2020.

[22] Z. Chi, H. Yan, and T. Pham, *Fuzzy Algorithms: With Applications to Image Processing and Pattern Recognition*, vol. 1. Singapore: World Scientific, 1996.

[23] F. Aghaeipoor, M. M. Javidi, and A. Fernandez, "IFC-BD: An interpretable fuzzy classifier for boosting explainable artificial intelligence in Big Data," *IEEE Trans. Fuzzy Syst.*, vol. 30, no. 3, pp. 830–840, Mar. 2022.

[24] J. Moral, C. Castiello, L. Magdalena, and C. Mencar, *Explainable Fuzzy Systems: Paving the Way From Interpretable Fuzzy Systems to Explainable AI Systems*. New York, NY, USA: Springer, 2021.

[25] 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.

[26] A. Das and P. Rad, "Opportunities and challenges in explainable artificial intelligence (XAI): A survey," 2020, *arXiv:2006.11371*.

[27] M. Ancona, E. Ceolini, C. Öztireli, and M. Gross, "Towards better understanding of gradient-based attribution methods for deep neural networks," 2017, *arXiv:1711.06104*.

[28] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 3319–3328.

[29] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 3145–3153.

[30] J. R. Zilke, E. L. Mencía, and F. Janssen, "DeepRED–rule extraction from deep neural networks," in *Proc. Int. Conf. Discov. Sci.*, 2016, pp. 457–473.

[31] Z. Shams et al., "REM: An integrative rule extraction methodology for explainable data analysis in healthcare," *bioRxiv*, 2021.

[32] M. E. Zarlenga, Z. Shams, and M. Jamnik, "Efficient decompositional rule extraction for deep neural networks," 2021, *arXiv:2111.12628*.

[33] Y. Jin and B. Sendhoff, "Extracting interpretable fuzzy rules from RBF networks," *Neural Process. Lett.*, vol. 17, no. 2, pp. 149–164, 2003.

[34] S. H. Huang and H. Xing, "Extract intelligible and concise fuzzy rules from neural networks," *Fuzzy Sets Syst.*, vol. 132, no. 2, pp. 233–243, 2002.

[35] C. J. Mantas, J. M. Puche, and J. M. Mantas, "Extraction of similarity based fuzzy rules from artificial neural networks," *Int. J. Approx. Reasoning*, vol. 43, no. 2, pp. 202–221, 2006.

[36] O. Cordón, M. J. Del Jesus, and F. Herrera, “A proposal on reasoning methods in fuzzy rule-based classification systems,” *Int. J. Approx. Reason.*, vol. 20, no. 1, pp. 21–45, 1999.

[37] B. Pereira et al., “The somatic mutation profiles of 2,433 breast cancers refines their genomic and transcriptomic landscapes,” *Nature Commun.*, vol. 7, 2016.

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

[39] L. Rieger and L. K. Hansen, “Aggregating explanation methods for stable and robust explainability,” 2019, *arXiv:1903.00519*.

[40] H. He and E. A. Garcia, “Learning from imbalanced data,” *IEEE Trans. Knowledge Data Eng.*, vol. 21, no. 9, pp. 1263–1284, 2009, doi: 10.1109/TKDE.2008.239.

[41] A. Papenmeier, G. Englebienne, and C. Seifert, “How model accuracy and explanation fidelity influence user trust,” 2019, *arXiv:1907.12652*.

[42] S. García, A. Fernández, J. Luengo, and F. Herrera, “Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power,” *Inf. Sci.*, vol. 180, no. 10, pp. 2044–2064, 2010.

[43] J. Chen, L. Song, M. Wainwright, and M. Jordan, “Learning to explain: An information-theoretic perspective on model interpretation,” in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 883–892.

[44] A. Fernández, C. J. Carmona, M. Jose del Jesus, and F. Herrera, “A pareto-based ensemble with feature and instance selection for learning from multi-class imbalanced datasets,” *Int. J. Neural Syst.*, vol. 27, no. 06, 2017, Art. no. 1750028.

[45] M. Galar, A. Fernández, E. Barrenechea, and F. Herrera, “Empowering difficult classes with a similarity-based aggregation in multi-class classification problems,” *Inf. Sci.*, vol. 264, pp. 135–157, 2014.

[46] A. Fernández et al., “Enhancing evolutionary fuzzy systems for multi-class problems: Distance-based relative competence weighting with truncated confidences (DRCW-TC),” *Int. J. Approx. Reasoning*, vol. 73, pp. 108–122, 2016.

[47] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, “Towards deep learning models resistant to adversarial attacks,” in *Proc. 6th Int. Conf. Learn. Representations*, 2018.

[48] J. Uesato, B. O’donoghue, P. Kohli, and A. Oord, “Adversarial risk and the dangers of evaluating against weak attacks,” in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 5025–5034.