Experimental Study on Generating Multi-modal Explanations of Black-box Classifiers in terms of Gray-box Classifiers

Jose M. Alonso, J. Toja-Alamancos, A. Bugarín
Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS)
Universidade de Santiago de Compostela
15782, Santiago de Compostela, SPAIN
{josemaria.alonso.moral, javier.toja, alberto.bugarin.diz}@usc.es

Abstract—**Artificial Intelligence (AI) is a first class citizen in the cities of the 21st century. In addition, trust, fairness, accountability, transparency and ethical issues are considered as hot topics regarding AI-based systems under the umbrella of Explainable AI (XAI). In this paper we have conducted an experimental study with 15 datasets to validate the feasibility of using a pool of gray-box classifiers (i.e., decision trees and fuzzy rule-based classifiers) to automatically explain a black-box classifier (i.e., Random Forest). Reported results validate our approach. They confirm the complementarity and diversity among the gray-box classifiers under study, which are able to provide users with plausible multi-modal explanations of the considered black-box classifier for all given datasets.**

Index Terms—**Explainable Artificial Intelligence, Interpretable Machine Learning, Classification, Open Source Software**

I. Introduction

Applications of Artificial Intelligence (AI) based systems are becoming popular and somehow essential in our daily life. The performance of these applications has led to an unseen growth due to the fact that AI is applicable to many different problems and contexts, including medical diagnostic, autonomous driving, and other decision-based tasks. One of the reasons of this growth is the increasing computational power available to be used by people, not only by institutions or big companies. Associated to this growth, trust issues have arisen within the professional community and people, who now use AI-based technologies and applications as part of their daily work and life. Many of these applications are considered as black boxes, and indeed many of them are based on opaque systems which do not provide any explanation about their behavior. So, they can be deemed as hazardous, and people are mistrustful about their use. For example, in fields like medicine, where a patient's life can be at stake, physicians need to trust a diagnostic decision support system, before applying its recommendations. Having AI-based systems operating as black boxes is quite controversial, because those systems might generate an unexpected side effect in society.

The European Commission (EC) identified AI as the most strategic technology of the 21st century [1]. In addition, the EC took a decision to guarantee the rights of the people and issued the General Data Protection Regulation (GDPR) which takes into account AI-based systems [2]. In particular, systems automatically generated with Machine Learning (ML) techniques are trained using data which may contain patterns that reflect prejudices towards certain people (e.g., because of their gender or ethnicity). In the example provided in [3], an AI-based system infers certain properties, like social status from addresses, generating a bias for/against certain people who live in a specific neighborhood.

Accordingly, there is an increasing attention to AI Fairness, Accountability, Transparency and Ethics (FATE in short) [4]–[7]. Within the research community there are events dedicated to FATE such us *ACM FAT* conference* which main subject is tackling the FATE in algorithms and ML. Also, the USA Defense Advanced Research Projects Agency (DARPA) promotes research on FATE through the XAI DARPA challenge [8] which has among main goals the generation of ML techniques and human-computer interfaces expected to lead the future development of XAI systems.

However, there is no agreement yet about how to build XAI systems in practice. On the one hand, many researchers focus on how to open black-box models [3] and this has become a hot open problem in the XAI field. On the other hand, some authors are for designing new XAI systems from scratch but supported by interpretable models [9].

We opt for a hybrid XAI approach which combines both black-box and gray-box models. In particular, we propose to combine expert knowledge with knowledge automatically extracted from data through ML techniques in a common framework [10]. In our previous work, we showed how to use our approach to build explainable classifiers for a real-world problem related to classification of beer styles. Firstly, we build an opaque AI system (based on Random Forest [11] or Neural Networks [12]) which acts as an "oracle" ready to identify the most plausible beer style. Secondly, we build several interpretable AI systems which are endowed with good interpretability-accuracy trade-offs but also with explanation capabilities. Then, given an unknown test data instance, the "oracle" identifies the output class and the related explanation comes from the simplest interpretable system which points out such class. Then, a multi-modal (i.e., a mixture of graphical

978-1-7281-6932-3/20/$31.00 ©2020 IEEE

and textual representations of the embedded knowledge) explanation is generated with the software ExpliClas [13].

In this work, we explore the chance of applying our hybrid XAI approach to a wide range of classification problems. Therefore, we have conducted an experimental study with the aim of validating the feasibility of using interpretable gray-box (i.e., decision tree and fuzzy rule-based) classifiers to automatically generate explanations associated to classifications made by a black-box "oracle".

The rest of the manuscript is organized as follows. Section II briefly introduces related work. Section III presents the datasets and classifiers under study. Section IV goes in depth with the experimental study and concludes with an illustrative use case. Section V summarizes the main conclusions and points out future work.

II. Related Work

XAI has become a hot topic within the AI research community, and over the last years, different algorithms have addressed the problem of generating explanations associated to black-box models [3]. Regarding post-hoc local explanations, i.e., explanations associated to classification of single data instances, it is assumed that around a data point the decision boundary can be captured by an interpretable model, no matter how complex is the global decision boundary for the problem under consideration. The most popular approaches are:

- **Local Interpretable Model-agnostic Explanations** (LIME) [14]. Firstly, a linear model is trained with data around the input to be classified by the black-box model. Then, local interpretations are provided in terms of such linear model. More precisely, LIME defines simplified inputs as *interpretable inputs*, and a mapping function which converts a binary vector of interpretable inputs into the original input space. In the text classification example provided in [14], LIME provides a visual explanation based on a bar chart with the relevance of each individual feature for the given test data instance. The color of the bars indicates which class the feature contributes to. Notice that LIME is model-agnostic, i.e., it does not depend on the black-box ML model, but it is quite depending on the underlying linear model used to generate the local approximations as well as on the way how local training samples are generated. In addition, the output of LIME is not self-explanatory but it is the user who has to do the final interpretation. Anyway, LIME has become one of the most popular algorithms in the XAI community and some authors have already developed variants and extensions to it. For example, Forrest et al. generated explanations in natural language as automatic interpretation of the output given by LIME [15].
- **Model-agnostic Explanations based on IF-THEN Rules** (ANCHOR) [16]. This approach can be seen as an extension of LIME to produce high-precision model-agnostic explanations. By construction, ANCHOR rules are computed incrementally adding conditions until the estimated precision is at least 95%. Thus, each rule represents sufficient local conditions for the explanation of the prediction inferred for a given data instance.
- **Local Rule-based Explanation** (LORE) [17] provides users with a factual explanation of reasons why a decision is made for a given test data instance, along with one or more counterfactual explanations which hypothesize what may have happened in case some features in the given data instance were modified. LORE first applies a genetic algorithm to learn synthetic data around the test instance. Then, this synthetic training dataset is used to grow a local decision tree classifier. Then, factual and counterfactual explanations of the given test instance are derived from the automatic interpretation of this decision tree. In comparison with ANCHOR, LORE manages naturally continuous features thanks to the capabilities inherent to decision trees while ANCHOR requires the a priori discretization of continuous features.
- **Shapley Additive exPlanations** (SHAP) [18] are similar to LIME in the sense that they are also local explanations. Nevertheless, each explanation is a linear function of binary variables. The so-called SHAP values measure the importance of each feature. They are computed by removing each feature individually and analyzing the variation produced on the output. Then, the aggregation of all SHAP values yields an additive explanation.
- **Growing Spheres** [19]. This is a model agnostic approach based on non-linear inverse classification. Given a test data instance, the challenge is to determine the minimal changes needed to alter the current prediction, i.e., the algorithm grows a sphere around the given data point with the minimum radius until finding out a close neighbor which is classified differently.

In addition to local explanations, it is also possible to generate global explanations. Let us briefly introduce two software tools which manage both local and global explanations:

- **iForest** [20] is a tool for visual analytic of models generated with the Random Forest (RF) [21] algorithm. As explained in [22], RF is able to get high accuracy in most classification problems. This is the reason why it has become very popular in the ML community. However, RF is considered as a black-box classifier because it combines hundreds of local decision trees through an ensemble mechanism, what makes hard to interpret the global output.
  This software addresses the challenge of automatically interpreting RF models and provides users with graphical representations (e.g., data overview, feature view and decision path view) about how the output class is generated for a given test data instance. It is worthy to note that iForest assists users to understand how RF works but it is not self-explanatory, i.e., given a data instance, the final interpretation is done by a human in the light of the given visualizations.
- **ExpliClas** [13] is a web service aimed to provide users with multi-modal (textual + graphical) explanations as-

sociated to gray-box classifiers. The current version of ExpliClas manages the following algorithms implemented in the framework Weka [23]: three decision trees (J48, REPTree, RandomTree) which are actually different implementations of the C4.5 algorithm first introduced by Quinlan [24]; and one fuzzy rule-based classifier (FURIA [25]). All these four algorithms are characterized by local semantics, i.e., the conditions in each node of a tree as well as the fuzzy sets in each rule are determined locally (without taking care of any global semantics associated to the entire model), what jeopardizes linguistic interpretability and makes hard automatic generation of textual explanations. Thus, the only way to provide users with textual explanations implies the need to do linguistic approximations associated to the models under consideration.

More precisely, the Explainer module in ExpliClas implements a linguistic layer on top of the gray-box classifiers. As a result, ExpliClas is endowed with global semantics what makes feasible not only to generate textual explanations for each single model but also to compare at linguistic level the classifications (and related explanations) carried out by different classifiers. In the core of the linguistic layer there are Strong Fuzzy Partitions (SFPs) associated to each feature in a given dataset. By default, a SFP is made up of three linguistic terms (*Low*, *Medium*, *High*) which correspond to triangular fuzzy sets uniformly distributed in the universe of discourse of the given feature, but it is editable by the user. SFPs were first introduced by Ruspini [26] and they satisfy all mathematical properties (e.g., coverage, distinguishability, etc.) required for designing fuzzy partitions endowed with linguistic interpretability [27].

III. Material and methods

In this section we describe the materials (datasets) and methods (classifiers) used to develop the experimental study.

A. Datasets

Table I presents the 15 data sets under consideration. For each dataset, we have reported the number of features, with the specific number of numerical features given in brackets. In addition, the table includes the number of instances and classes for all datasets. It is worthy to note that for the sake of generality in our experimental study we have selected a wide range of datasets with varied structural complexity and size. Namely, we have taken into account both variety and diversity in the number of attributes, instances and classes but also in the application domain. With the aim of highlighting the huge difference among datasets, we give average, standard deviation, minimum and maximum values for all datasets at the bottom of the table. Notice that all these datasets are freely available in the well-known UCI ML [28] and Weka [29] repositories.

TABLE I
Datasets under study

Dataset	Features (Numerical)	Instances	Classes
AUDIOLOGY	69 (0)	226	24
AUTOS	25 (16)	205	7
BALANCE-SCALE	4 (4)	625	3
BREAST-CANCER	9 (0)	286	2
COLIC	22 (7)	368	2
CREDIT-A	15 (6)	690	2
DIABETES	8 (8)	768	2
GLASS	9 (9)	214	7
HEART-STATLOG	13 (13)	270	2
HYPOTHYROID	29 (7)	3772	4
IONOSPHERE	34 (34)	351	2
KR-VS-KP	36 (0)	3196	2
LETTER	16 (16)	20000	26
VEHICLE	18 (18)	846	4
WAVEFORM-5000	40 (40)	5000	3
Average	23.1 (11.9)	2401.5	6.1
Standard Deviation	16.8 (11.8)	5117.2	7.9
Minimum	4 (0)	205	2
Maximum	69 (40)	20000	26

The Weka Framework [23] is open source and coded in Java. It was initially developed as a collection of ML algorithms by researchers in the university of Waikato (New Zeland), but it has been later extended with more algorithms and functionality by researchers all around the world. Weka provides users with a user-friendly environment which allows to explore, analyze, and process datasets but also automatically learn (and evaluate) ML classifiers from such datasets.

B. Classifiers

In the experimental study, we have considered one RF black-box classifier [21]. This classifier is selected because of its ability to generate very accurate classifiers [22]. In addition, we already used RF as "oracle" in the proof of concept of the hybrid XAI approach [10] whose generality is under validation in this paper.

Regarding gray-box classifiers, we have considered the four classifiers (J48, REPTree, RandomTree and FURIA) which are already implemented in ExpliClas. This means that we can use later the ExpliClas Explainer module to generate multimodal explanations associated to them. Moreover, decision trees and fuzzy rule-based classifiers were also considered in [10]. In short, the gray-box classifiers under consideration are as follows:

- **J48** is a pruned C4.5 classifier [24]. It is deemed as an interpretable white box because it is possible to know the threshold values of each internal node splitting condition.
- **REPTree** uses regression tree logic and creates multiple C4.5 trees in different iterations. At the end, it selects the best one among all the generated trees. Because of the output of this model is just a simple decision tree classifier, it is also considered as a white box.
- **RandomTree** applies bagging to produce a random set of training data instances for the generation of several C4.5 decision trees. At the end, similarly to REPTree, it provides users with only one individual tree model, and because of that it is also deemed as a white box.

- **FURIA** is the acronym of Fuzzy Unordered Rule Induction Algorithm [25]. It generates fuzzy IF-THEN classification rules with fuzzy sets $A^i_h = \{a^1_h, a^2_h, a^3_h, a^4_h\}$ of trapezoidal shape in the antecedent of each rule. Notice that only the most relevant features are considered in each rule. For example, a rule with two antecedents may be:

$$\begin{aligned} R_i\text{: IF A1 in } [0.5, 3.7, \infty, \infty] \text{ AND A2 in } [-\infty, -\infty, 0.7, 10.8] \\ \text{THEN class is } C_1 \text{ with CF=0.85} \end{aligned} \tag{1}$$

  where CF is the *certainty factor* of rule R_i. The rule activation degree is computed as the multiplication of the CF by the rule firing degree which results of applying the usual inference mechanism of Mamdani fuzzy systems [30]. Given a test data instance, if there is no rule with activation degree greater than zero then FURIA offers to the user three options: 1) abstain (i.e., no output class is given); 2) voting for the a priori most frequent class in the dataset; and 3) rule stretching (i.e., a new set of rules is dynamically created from the initial rule base by removing antecedents in order one by one, rule by rule, until the instance is covered). FURIA is deemed as a gray-box classifier because it produces a set of rules which can be interpreted (at certain degree) by users.

For the sake of simplicity and generality of results, we limit our study to classifiers implemented in Weka [23]. The challenge here is to evaluate whether the considered gray-box classifiers are suitable for generating explanations associated to a black-box classifier.

IV. Experimental Study

In this section, we first introduce the experimental setting (see Section IV-A) and then we present and discuss the reported results (see Section IV-B).

A. Experimental setting

With the aim of evaluating the generality of the hybrid XAI approach first introduced in [10] just as a proof of concept, here the hypothesis to validate is as follows: Given a test data instance, if we had a gray-box classifier which were co-intensive with the black-box classifier that is considered as "oracle" in the hybrid XAI approach, i.e., both classifiers point out the same output class, then we may provide users with an explanation of the current classification in terms of the multi-modal explanation generated by ExpliClas for the selected gray-box classifier.

It is worthy to note that thanks to the global semantics imposed by the linguistic layer of the ExpliClas Explainer module, no matter the selected gray-box classifier, the given explanation is expected to be co-intensive with expert knowledge and thus comprehensible by the user.

The concept of co-intension was coined by Prof. Zadeh [31] and is in the core of the paradigm of Computing with Words (CWW) [32], [33] which is highly relevant in the context of providing humans with explanations in natural language. In short, two different concepts are deemed as co-intensive when they refer to the same entity at high degree. It is worth noting that the semantic co-intension approach [34] has been successfully applied to measure the interpretability of fuzzy systems and it can be generalized to consider gray-box classifiers such as decision trees and FURIA.

We assume that a given textual explanation is comprehensible only if its explicit semantics is co-intensive with the implicit semantics inferred by the user when reading such explanation. Therefore, we can measure how much similarly behave each gray-box classifier and the "oracle", i.e., how much co-intensive the two classifiers are, with the following co-intension metric:

$$COIN(A,B) = \frac{100}{N} \sum_{i=1}^{N} \left\{ \frac{W_i}{S_i} \right\} \in [0, 100] \tag{2}$$

where N is the number of folds in the experiment (N=10); W_i is the number of data instances classified with the same output value by both classifiers A and B, in the fold i; and S_i is the total number of data instances in the fold i. Accordingly, we count the number of data instances classified identically by both classifiers, no matter if the classification is right or wrong.

The focus is on determining if we can (or cannot) validate the hypothesis enunciated at the beginning of this section, no matter the dataset under consideration. To do so, we first train and test each individual classifier using a 10-fold cross-validation with the 15 datasets listed in Table I. Then, we have conducted the following three analyses:

- The Analysis 1 is aimed to evaluate how much co-intensive each individual gray-box classifier is with the black-box classifier. We computed the co-intension metric (see Eq. 2) between each gray-box classifier and the black-box classifier for each fold and reported the averaged results.
- The Analysis 2 is aimed to evaluate the co-intensive degree between the black-box classifier and the pool of gray-box classifiers seen as one unique classifier, i.e., we test if using a composition (but without applying any kind of voting mechanism) of gray-box classifiers we can get better results. Thus, this time we compared at once the output of the black-box classifier versus the output given by all individual gray-box classifiers. Then, we increased W_i in the Eq. 2 every time that at least one gray-box classifier was able to match the output given by the black-box classifier for the given test data instance, no matter the classification output provided by the rest of classifiers.
- The Analysis 3 is aimed to evaluate the complementarity and diversity among the pool of gray-box classifiers. Accordingly, we computed the co-intension metric for all pairs of gray-box classifiers.

The goodness of each classifier is evaluated under the 10-fold cross-validation provided by Weka [23]. The following quality metrics are reported (see Tables II and III):

- for **Accuracy**: the ratio of correctly classified instances (RCCI), and the root mean square error (RMSE).

TABLE II

ACCURACY METRICS FOR THE DATASETS AND CLASSIFIERS UNDER STUDY. RCCI STANDS FOR RATIO OF CORRECTLY CLASSIFIED INSTANCES AND RMSE IS THE ROOT MEAN SQUARE ERROR.

	RF		J48		REPTree		RandomTree		FURIA	
Dataset	RCCI	RMSE	RCCI	RMSE	RCCI	RMSE	RCCI	RMSE	RCCI	RMSE
AUDIOLOGY	72.36	**0.13**	**74.87**	**0.13**	67.84	0.14	56.78	0.18	72.86	**0.13**
AUTOS	**83.90**	**0.20**	83.41	**0.20**	62.44	0.28	76.58	0.25	77.07	0.23
BALANCE-SCALE	80.96	**0.29**	76.96	0.36	78.72	0.33	80.00	0.36	**82.88**	0.30
BREAST-CANCER	68.75	0.47	**74.63**	**0.44**	68.38	0.48	62.87	0.60	73.53	0.50
COLIC	84.87	**0.35**	**85.43**	0.36	84.03	0.37	77.03	0.43	82.63	0.39
CREDIT-A	86.23	**0.33**	85.80	0.34	84.78	0.35	75.65	0.48	**86.67**	0.35
DIABETES	**76.43**	**0.40**	75.91	0.43	74.22	0.43	71.09	0.54	75.39	0.48
GLASS	**79.81**	**0.21**	70.89	0.27	72.77	0.25	67.14	0.31	70.89	0.27
HEART-STATLOG	**83.33**	**0.35**	80.37	0.42	74.81	0.43	74.81	0.50	82.96	0.40
HYPOTHYROID	99.41	0.06	**99.49**	**0.05**	99.46	**0.05**	97.17	0.12	99.46	**0.05**
IONOSPHERE	**93.71**	**0.22**	91.14	0.29	88.57	0.31	87.71	0.35	89.43	0.29
KR-VS-KP	99.31	0.10	99.22	0.09	98.84	0.10	95.53	0.21	**99.47**	**0.06**
LETTER	**96.22**	**0.06**	87.00	0.09	83.24	0.10	85.03	0.11	90.48	0.07
VEHICLE	**75.41**	**0.27**	72.34	0.34	70.57	0.32	69.03	0.39	70.69	0.33
WAVEFORM-5000	**85.80**	**0.28**	75.72	0.39	76.66	0.34	72.96	0.42	82.36	0.32
Average	**84.43**	**0.25**	82.21	0.28	79.02	0.29	76.63	0.35	82.45	0.28
Standard Deviation	9.44	**0.12**	**9.12**	0.13	10.85	0.13	11.22	0.15	9.38	0.14
Minimum	68.75	0.06	**70.89**	**0.05**	62.44	**0.05**	56.78	0.11	70.69	**0.05**
Maximum	99.41	0.47	**99.49**	**0.44**	99.46	0.48	97.17	0.60	99.47	0.50

TABLE III

INTERPRETABILITY METRICS FOR THE DATASETS AND CLASSIFIERS UNDER STUDY. NR IS THE NUMBER OF LEAVES (IN DECISION TREES) AND THE NUMBER OF RULES (IN FURIA). TRL STANDS FOR TOTAL RULE LENGTH.

	J48		REPTree		RandomTree		FURIA	
Dataset	NR	TRL	NR	TRL	NR	TRL	NR	TRL
AUDIOLOGY	28.0	128.0	**19.4**	67.9	185.0	1439.6	21.6	**24.0**
AUTOS	45.0	160.6	33.9	56.3	168.0	737.0	**19.5**	**24.3**
BALANCE-SCALE	42.4	166.2	**21.2**	80.8	152.8	609.7	23.8	**26.4**
BREAST-CANCER	6.6	12.6	18.5	33.3	376.7	1723.6	**3.6**	**3.8**
COLIC	**6.2**	15.8	8.9	23.5	288.4	1702.6	7.9	**9.1**
CREDIT-A	22.4	101.6	17.0	56.0	276.1	1572.8	**7.5**	**10.0**
DIABETES	21.0	83.3	14.9	61.2	136.0	762.6	**7.7**	**8.9**
GLASS	24.4	116.3	**9.8**	31.8	47.2	226.6	14.9	**15.7**
HEART-STATLOG	17.2	77.3	**6.7**	21.4	54.6	322.0	7.2	**8.3**
HYPOTHYROID	14.7	70.3	**9.8**	41.3	182.5	1503.3	14.2	**20.9**
IONOSPHERE	14.6	75.8	**5.5**	**16.4**	29.8	180.2	10.8	18.0
KR-VS-KP	**28.2**	219.1	28.3	198.7	319.9	4213.7	**28.2**	**66.7**
LETTER	1189.5	10991.6	**613.2**	4652.2	2667.2	25239.7	660.7	**815.2**
VEHICLE	70.2	478.6	27.1	149.8	155.9	1237.0	**24.7**	**29.0**
WAVEFORM-5000	287.3	2539.2	**82.5**	560.3	612.1	5942.6	94.0	**115.0**
Average	121.2	1015.8	**61.1**	403.4	376.8	3160.9	63.1	**79.7**
Standard Deviation	303.5	2830.7	**153.9**	1183.3	650.8	6302.7	166.7	**205.5**
Minimum	6.2	12.6	5.5	16.4	29.8	180.2	**3.6**	**3.8**
Maximum	1189.5	10991.6	**613.2**	4652.2	2667.2	25239.7	660.7	**815.2**

- for **Interpretability**: the number of leaves in a tree or rules in a FURIA classifier (NR) and the total rule length (TRL). In the case of decision trees, TRL counts the number of non-leaf nodes in all the branches of the tree. In the case of FURIA classifiers, TRL counts the total number of conditions A_i in all the rules (e.g., TRL=2 in Eq. 1).

In Tables II and III, we have highlighted in bold the best reported values for each quality metric and dataset. It is easy to appreciate how RF usually turns up as the most accurate classifier (see Table II), as expected, that is the reason why it is considered as the "oracle" in our hybrid XAI approach. Anyway, in some cases, J48 or FURIA are not far from RF from the accuracy point of view. Moreover, sometimes J48 or FURIA are even more accurate than RF. This is in agreement with the no free lunch theorem [35], i.e., there is no classification algorithm to produce always the best classifier. Someone may argue that keeping RF as "oracle" in those cases is not the best choice. However, it is out of the scope of this paper to look for the most accurate "oracle". On the other hand, regarding interpretability (see Table III), REPTree is usually the best for NR while FURIA is the winner with respect to TRL. Again, it is out of the scope of this paper to look for the optimal interpretability-accuracy trade-off. Of

TABLE IV

Results for Analyses 1 and 2. The first columns show the co-intension metric (COIN) between each gray-box classifier (J48, REPTree, RandomTree, FURIA) and the black-box classifier (RF). The sixth column shows the COIN between the pool of gray-box classifiers (ALL=J48+REPTree+RandomTree+FURIA) and RF. GAIN = COIN(ALL,RF) - MAXCOIN, with MAXCOIN being the maximum COIN for single classifiers (i.e., the number in bold in each row).

Dataset	COIN (J48,RF)	COIN (REPTree,RF)	COIN (RandomTree,RF)	COIN (FURIA,RF)	COIN (ALL,RF)	GAIN
AUDIOLOGY	**76.38**	74.87	61.81	75.38	88.44	+12.06
AUTOS	82.44	70.24	**82.93**	77.07	96.59	+13.66
BALANCE-SCALE	86.88	81.28	**93.28**	83.04	99.20	+5.92
BREAST-CANCER	**85.29**	82.72	83.82	83.46	97.43	+12.14
COLIC	**96.64**	95.80	84.31	92.72	99.44	+2.80
CREDIT-A	**92.61**	92.17	81.01	91.16	99.42	+6.81
DIABETES	85.16	84.77	78.52	**86.72**	98.05	+11.33
GLASS	**84.51**	77.00	73.71	80.75	99.06	+14.55
HEART-STATLOG	88.89	81.11	79.63	**90.00**	97.78	+7.78
HYPOTYROID	99.43	**99.49**	97.39	99.41	99.81	+0.32
IONOSPHERE	**94.57**	91.43	91.14	94.00	99.71	+5.14
KR-VS-KP	**99.41**	99.09	95.71	**99.41**	**99.91**	+0.50
LETTER	87.70	84.33	85.68	**91.01**	98.41	+7.40
VEHICLE	77.66	76.48	76.36	**79.79**	98.11	**+18.32**
WAVEFORM-5000	80.84	82.50	76.20	**89.16**	98.98	+9.82
Average	**87.89**	84.88	82.76	87.54	98.02	8.57
Standard Deviation	**7.34**	8.93	9.30	7.49	2.82	5.22
Minimum	**76.38**	70.24	61.81	75.38	88.44	0.32
Maximum	99.43	**99.49**	97.39	99.41	99.91	18.32

course, the diversity among the datasets under study (see Table I) produces a difference in the inherent complexity of the different classifiers.

For the sake of transparency, fairness and reproducibility of the experiments, both Java source code and complementary material are available online [36].

B. Experimental analysis

Table IV summarizes the experimental results to be discussed as part of the analyses 1 and 2:

- **Analysis 1**. Regarding columns 2 - 5 in Table IV, we observe that even though J48 turns up as the most co-intensive classifier with RF in many cases, none of the individual classifiers is able to perfectly match with RF. Notice that the best COIN value in columns 2 - 5, for each dataset, is highlighted in bold. In addition, only in 8 out of 15 datasets, the maximum COIN value (regarding all individual gray-box classifiers) is greater than 90. Moreover, RandomTree only gets COIN greater or equal than 90 for 4 datasets. We count 5 datasets in the case of J48 and REPTree, and 7 for FURIA. In the light of the reported results, we conclude that it seems not to be a good idea just to select one single gray-box classifier to generate all explanations associated to the "oracle".
- **Analysis 2**. If we pay attention to columns 6 and 7 in Table IV, we observe that COIN(ALL,RF) is always greater than the maximum COIN value regarding only one single classifier (which is highlighted in bold for each dataset). In addition, the gain is up to +18.32 in the best case (VEHICLE) and it is about +8.5 in average for all datasets. Moreover, the average COIN(ALL,RF) is about 98 what means that only about 2% of unknown test data instances remain without any associated explanation. This seems an acceptable rate because even human experts are not able to explain their decisions sometimes.

After Analysis 1 and Analysis 2, we can conclude the experimental study validates the feasibility of using a pool of gray-box classifiers to automatically generate explanations associated to classifications made by a black-box "oracle". In practice, given a test data instance, we will select the most suitable gray-box classifier (among all available ones) to generate the required explanation. However, the selection of the right classifier for each single instance remains a challenge out of the scope of this work.

Table V shows the reported results for the Analysis 3. They confirm the complementarity and diversity among gray-box classifiers under study, which are able to cover jointly most of the input space for a given dataset, as deduced from Table IV. Nevertheless, the average COIN values in Table V are always below 90 no matter the pair of classifiers under consideration. This means overlapping of classifiers in the input space is acceptable.

C. Illustrative Use Case

This section is aimed to show how to use the ExpliClas software [13] for generating multi-modal explanations associated to the classifiers built in the previous section.

We have selected the VEHICLE dataset because it exhibited the biggest gain value in the Analysis 2 (see the last column in Table IV). The classification task consists of identifying one out of four types of vehicles (i.e., Opel, Saab, Bus, Van) in terms of 18 numerical features (e.g., compactness, circularity, etc.) extracted from the silhouette of the vehicle, with data instances corresponding to vehicles that may be viewed from one of many different angles. The dataset includes 846 instances (see Table I). Each instance is labeled as one of the

TABLE V
RESULTS FOR ANALYSIS 3, SHOWING THE CO-INTENSION METRIC (COIN) BETWEEN EACH PAIR OF GRAY-BOX CLASSIFIERS.

Dataset	COIN (J48,REPTree)	COIN (J48,RandomTree)	COIN (J48,FURIA)	COIN (REPTree,RandomTree)	COIN (REPTree,FURIA)	COIN (RandomTree,FURIA)
AUDIOLOGY	**82.41**	60.30	81.41	59.30	75.38	53.77
AUTOS	63.90	77.07	**77.56**	68.29	60.98	73.17
BALANCE-SCALE	84.32	**85.92**	81.44	78.88	82.08	81.12
BREAST-CANCER	87.87	77.94	**95.22**	77.57	85.29	76.84
COLIC	**96.92**	84.31	93.28	83.47	93.00	83.75
CREDIT-A	91.16	79.42	92.46	77.54	**92.61**	78.26
DIABETES	**85.29**	78.26	83.59	75.00	83.98	76.17
GLASS	**73.71**	69.01	73.24	65.26	68.08	68.54
HEART-STATLOG	82.59	76.67	**90.00**	75.56	86.67	78.52
HYPOTYROID	**99.76**	97.22	**99.76**	97.22	99.68	97.25
IONOSPHERE	90.00	87.43	91.43	87.71	**91.71**	86.29
KR-VS-KP	99.37	95.68	**99.44**	95.37	99.12	95.49
LETTER	79.51	79.97	**84.20**	78.07	81.43	82.30
VEHICLE	**74.47**	72.46	71.04	71.39	72.34	69.27
WAVEFORM-5000	76.80	70.86	78.64	70.80	**80.94**	73.92
Average	84.54	79.50	**86.18**	77.43	83.55	78.31
Standard Deviation	10.10	9.72	**9.09**	10.38	11.04	10.69
Minimum	63.90	60.30	**71.04**	59.30	60.98	53.77
Maximum	**99.76**	97.22	**99.76**	97.22	99.68	97.25

four classes which are quite well balanced: 212 instances are Opel, 217 Saab, 218 Bus, and 199 Van.

For the sake of simplicity, we have just taken one of the folds of the 10-fold cross validation, with the dataset divided into 90% of instances for training and 10% for test. Then, we have used Weka to build the RF classifier with the training dataset. Then, we have uploaded to ExpliClas both training and test datasets. Then, we have built the four gray-box classifiers with ExpliClas. Then, we have picked up a couple of unknown test instances and gone deeper about: i) how each test instance is classified by the "oracle" RF; and ii) how this classification is explained by ExpliClas regarding each single gray-box classifier. With that aim, we have defined uniform SFPs with three linguistic terms (*Low*, *Medium*, and *High*) for each feature in the ExpliClas linguistic layer.

- **Case1**:
  - Test instance 6: {Compactness, 96}; {Circularity, 47}; {Distance-circularity, 103}; {Radius-ratio, 215}; {Pr.axis-Aspect-ratio, 69}; {Max-length-aspect-ratio, 10}; {Scatter-ratio, 200}; {Elongatedness, 33}; {Pr.axis-rectangularity, 23}; {Max-length-rectangularity, 147}; {Scaled-variance-major, 220}; {Scaled-variance-minor, 598}; {Scaled-radius-of-gyration, 200}; {Skewness-about-major, 73}; {Skewness-about-minor, 6}; {Kurtosis-about-major, 6}; {Kurtosis-about-minor, 187}; {Hollows-ratio, 194}; {Class, Opel};
  - "Oracle" RF Classification: Opel
  - Gray-box Classification: Opel
    In this case, all the classifiers agree with the "oracle" output class. Therefore, we obtain four plausible explanations:
    "The vehicle is Opel because..."

J48 : "... compactness, **elongatedness**, scaled-variance-minor and hollows-ratio are low and max-length-aspect-ratio is medium. However, Saab is also possible due to the proximity of hollows-ratio with the related split value (196.0)".

REPTree : "... **elongatedness** and hollows-ratio are low and max-length-aspect-ratio is medium. However, Saab is also possible due to the proximity of hollows-ratio with the related split value (195.5)".

RandomTree : "... circularity and hollows-ratio are low and compactness, max-length-aspect-ratio, **elongatedness** and scaled-variance-minor are medium. However, Saab is also possible due to the proximity of hollows-ratio with the related split value (195.5)".

FURIA : "... **elongatedness** is low and skewness-about-minor is medium".

In this example, **elongatedness** seems to be the most informative feature as it is present in all plausible explanations.

- **Case2**:
  - Test instance 27: {Compactness, 108}; {Circularity, 54}; {Distance-circularity, 105}; {Radius-ratio, 203}; {Pr.axis-Aspect-ratio, 62}; {Max-length-aspect-ratio, 11}; {Scatter-ratio, 202}; {Elongatedness, 33}; {Pr.axis-rectangularity, 23}; {Max-length-rectangularity, 164}; {Scaled-variance-major, 216}; {Scaled-variance-minor, 608}; {Scaled-radius-of-gyration, 235}; {Skewness-about-major, 68}; {Skewness-about-minor, 12}; {Kurtosis-about-major, 3}; {Kurtosis-about-minor, 190}; {Hollows-ratio, 200}; {Class, Saab};
  - "Oracle" RF Classification: Saab
  - Gray-box Classification: Saab (J48), Saab (REPTree), Opel (RandomTree), and Saab (FURIA)
    In this case, three out of the four classifiers agree with the "oracle". Therefore, we obtain three plausible explanations (the explanation provided by RandomTree is discarded):
    "The vehicle is Saab because..."

J48 : "... compactness and hollows-ratio are high, elongatedness, scaled-variance-minor and scaled-radius-of-gyration are low and max-length-aspect-ratio and skewness-about-major are medium. However, Opel is also possible due to the proximity of skewness-about-major with the split value (67.0)".

REPTree : "... hollows-ratio is high, elongatedness and scaled-variance-minor are low and max-length-aspect-ratio is medium. However, Opel is also possilbe possible due to the proximity of hollows-ratio with the split value (195.5)".

FURIA : "... kurtosis-about-minor is low and scaled-variance-major is medium".

ExpliClas provides users with multi-modal explanations. In addition to the textual explanations given above, users are also provided with graphical visualizations of trees and rules, which are not included here for the sake of space. In addition, let us remind that assessing the naturalness and effectiveness of such explanations is a task in progress out of the scope of this paper.

V. Conclusions and Future Work

We have carried out an exhaustive experimental study with 15 datasets. As main conclusion, we have validated the generality of a previous proof of concept which only took into account one dataset. Accordingly, in the light of the experimental analysis in this paper, we can state that we are able to automatically generate multi-modal explanations associated to black-box classifiers in terms of gray-box classifiers, at least for the wide range of datasets under study.

Notice that explanations of gray-box classifiers are generated by the ExpliClas software and assessing their goodness and effectiveness is an ongoing work. In addition, we plan to enhance ExpliClas with our hybrid XAI approach for explaining black-box classifiers. This implies the definition and validation of a new procedure for generating one unique explanation instead of the pool of plausible explanations currently provided. As future work, we will also extend our analysis to more datasets and we will consider other black-box classifiers such as deep learning neural networks.

Acknowledgment

Jose M. Alonso is *Ramón y Cajal* Researcher (RYC-2016-19802). This research is supported by the Spanish Ministry of Science, Innovation and Universities (grants RTI2018-099646-B-I00, TIN2017-84796-C2-1-R, TIN2017-90773-REDT, and RED2018-102641-T) and the Galician Ministry of Education, University and Professional Training (grants ED431F 2018/02, ED431C 2018/29, ED431G/08, and ED431G2019/04). These grants were co-funded by the European Regional Development Fund (ERDF/FEDER program).

References

[1] European Commission, "Artificial Intelligence for Europe," European Commission, Brussels, Belgium, Tech. Rep., 2018, Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions (SWD(2018) 137 final). [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2018%3A237%3AFIN

[2] Parliament and Council of the European Union, "General data protection regulation (GDPR)," 2016, http://data.europa.eu/eli/reg/2016/679/oj.

[3] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM Computing Surveys*, vol. 51, no. 5, pp. 93:1–93:42, 2018.

[4] T. Revell, "Computer says no comment," *New Scientist*, vol. 238, no. 3173, pp. 40–43, 2018.

[5] F. Pasquale, *The black box society*. Harvard University Press, 2015.

[6] C. C. Perez, *Invisible Women: Exposing Data Bias in a World Designed for Men*. Random House, 2019.

[7] S. Barocas and D. Boyd, "Engaging the ethics of data science in practice," *Commun. ACM*, vol. 60, no. 11, pp. 23–25, Oct. 2017.

[8] D. Gunning, "Explainable Artificial Intelligence (XAI)," DARPA, Defense Advanced Research Projects Agency, Arlington, USA, Tech. Rep., 2016, DARPA-BAA-16-53, https://www.darpa.mil/program/explainable-artificial-intelligence.

[9] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, pp. 206–215, 2019.

[10] J. M. Alonso, A. Ramos-Soto, C. Castiello, and C. Mencar, "Hybrid data-expert explainable beer style classifier," in *IJCAI/ECAI Workshop on Explainable Artificial Intelligence*, 2018, pp. 1–5.

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

[12] F. Rosenblatt, *Principles of neurodynamics: Perceptions and the theory of brain mechanism*. Spartan Books, Washington, DC, 1961.

[13] J. M. Alonso and A. Bugarín, "ExpliClas: Automatic generation of explanations in natural language for weka classifiers," in *IEEE International Conference on Fuzzy Systems*, 2019, pp. 1–6.

[14] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you?: Explaining the predictions of any classifier," in *22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16. New York, NY, USA: ACM, 2016, pp. 1135–1144.

[15] J. Forrest, S. Sripada, W. Pang, and M. Coghill, "Towards making NLG a voice for interpretable machine learning," in *11th International Natural Language Generation Conference (INLG)*. ACL, 2018, pp. 177–182.

[16] M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-precision model-agnostic explanations," in *32nd AAAI Conference on Artificial Intelligence*. AAAI Press, 2018, pp. 1527–1535.

[17] R. Guidotti, A. Monreale, F. Giannotti, D. Pedreschi, S. Ruggieri, and F. Turini, "Factual and counterfactual explanations for black box decision making," *IEEE Intelligent Systems*, 2019.

[18] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *31st International Conference on Neural Information Processing Systems*. Curran Associates Inc., 2017, pp. 4768–4777.

[19] T. Laugel, M.-J. Lesot, C. Marsala, X. Renard, and M. Detyniecki, "Comparison-based inverse classification for interpretability in machine learning," in *International Conference on Information Processing and Management of Uncertainty in Knowledge-based Systems (IPMU)*. Springer, 2018, pp. 100–111.

[20] X. Zhao, Y. Wu, D. L. Lee, and W. Cui, "iForest: Interpreting random forests via visual analytics," *IEEE transactions on visualization and computer graphics*, vol. 25, no. 1, pp. 407–416, 2018.

[21] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 2001.

[22] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim, "Do we need hundreds of classifiers to solve real world classification problems?" *Journal of Machine Learning Research*, vol. 15, pp. 3133–3181, 2014.

[23] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

[24] R. Quinlan, *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann Publishers, 1993.

[25] J. Hühn and E. Hüllermeier, "FURIA: an algorithm for unordered fuzzy rule induction," *Data Mining and Knowledge Discovery*, vol. 19, no. 3, pp. 293–319, 2009.

[26] E. Ruspini, "A new approach to clustering," *Information and Control*, vol. 15, no. 1, pp. 22–32, 1969.

[27] J. M. Alonso, C. Castiello, and C. Mencar, "Interpretability of Fuzzy Systems: Current Research Trends and Prospects," in *Springer Handbook of Computational Intelligence*, J. Kacprzyk and W. Pedrycz, Eds. Springer Berlin / Heidelberg, 2015, pp. 219–237.

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

[29] Waikato University, "Repository of datasets in the arff Weka format," 2020, https://waikato.github.io/weka-wiki/datasets.

[30] E. H. Mamdani, "Application of fuzzy logic to approximate reasoning using linguistic synthesis," *IEEE Transactions on Computers*, vol. 26, no. 12, pp. 1182–1191, 1977.

[31] L. A. Zadeh, "Is there a need for fuzzy logic?" *Information Sciences*, vol. 178, pp. 2751–2779, 2008.

[32] ——, "From computing with numbers to computing with words - From manipulation of measurements to manipulation of perceptions," *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, vol. 46, no. 1, pp. 105–119, 1999.

[33] J. Kacprzyk, "Computing with words is an implementable paradigm: Fuzzy queries, linguistic data summaries, and natural-language generation," *IEEE Transactions on Fuzzy Systems*, vol. 8, pp. 451–472, 2010.

[34] C. Mencar, C. Castiello, R. Cannone, and A. M. Fanelli, "Interpretability assessment of fuzzy knowledge bases: A cointension based approach," *International Journal of Approximate Reasoning*, vol. 52, no. 4, pp. 501–518, 2011.

[35] D. H. Wolpert and W. G. Macready, "No free lunch theorems for optimization," *IEEE Transactions on Evolutionary Computation*, vol. 1, pp. 67–82, 1997.

[36] J. Toja-Alamancos, "Java library for experimental analysis of explanations associated to black-box classifiers," 2020, University of Santiago de Compostela, https://gitlab.citius.usc.es/javier.toja/expliclas-model-comparator.