Explaining Multi-Label Black-Box Classifiers for Health Applications

Cecilia Panigutti[1], Riccardo Guidotti[2,3], Anna Monreale[3], and Dino Pedreschi[3]

[1] Scuola Normale Superiore, Pisa, Italy, cecilia.panigutti@sns.it
[2] ISTI-CNR, Pisa, Italy, guidotti@isti.cnr.it
[3] University of Pisa, Italy, {name.surname}@di.unipi.it

Abstract. Today the state-of-the-art performance in classification is achieved by the so-called black boxes", i.e., decision-making systems whose internal logic is obscure. Such models could revolutionize the health-care system, however their deployment in real-world diagnosis decision support systems is subject to several risks and limitations due to the lack of transparency. The typical classification problem in health-care requires a multi-label approach since the possible labels are not mutually exclusive, e.g. diagnoses. We propose MARLENA, a model-agnostic method which explains multi-label black box decisions. MARLENA explains an individual decision in three steps. First, it generates a synthetic neighborhood around the instance to be explained using a strategy suitable for multi-label decisions. It then learns a decision tree on such neighborhood and finally derives from it a decision rule that explains the black box decision. Our experiments show that MARLENA performs well in terms of mimicking the black box behavior while gaining at the same time a notable amount of interpretability through compact decision rules, i.e., rules with limited length.

1 Introduction

Machine learning algorithms are often the heart of many opaque decision systems that take critical decisions that heavily impact on our life and society. Thanks to the ability of machine learning algorithms to leverage large volumes of health-related data, decision systems have the potential to help doctors in their diagnosis, in predicting the spread of diseases and in identifying groups of high-risk patients with high performance [7]. To this end, machine learning algorithms learn patterns from this available data in order to construct predictive models mapping features into a decision [6, 18, 21]. Unfortunately, real historical data used for the learning process may contain human biases which could lead to wrong or unfair decisions. The lack of transparency in the behavior of machine learning algorithms and the inability of explaining the logic involved in their decision process may limit the social acceptance and trust on their adoption in many sensitive contexts. Moreover, the lack of explanations for the decisions of black box systems is also a legal issue addressed in the *General Data Protection Regulation* approved by the European Parliament in May 2018. Besides giving people control over their personal data, it also provides restrictions and guidelines for automated decision-making processes (prediction models in this case) which, for the first time, introduce a right of explanation. This means that an individual has the right to obtain meaningful explanations about the logic involved when automated decision making takes place [25, 15, 12].

Some machine learning techniques aiming at learning predictive model in health-care, rather than specialize in predicting a particular outcome (heart-failure, in-hospital mortality, etc), focus on developing *generic* predictive models able to forecast any kind of future diagnosis. This task is called multi-label classification problem since diagnoses are not mutually exclusive, so a multilabel classifier has to assign to each sample a set of target labels (decisions). For example, in [6] a RNN is trained to implement a temporal model to predict the patient's next visit time, diagnosis and medication order.

In this paper we address the problem of explaining the decision taken by a multi-label black box classifier by providing "meaningful explanations" of the logic involved in the decision process. This task is particularly relevant in health-care applications since machine learning-based diagnosis decision support systems able to tackle mixed scenarios solve a multi-label classification problem. To this end, we propose a model agnostic solution called **MARLENA** (for **M**ulti-l**a**bel **R**u**l**e-based **E**xpla**NA**tions). Given any kind of multi-label black box predictor b and a specific instance x labeled with outcome y by b, we build an interpretable multi-label predictor by first generating a set of synthetic neighbor instances of the given instance x through an ad-hoc strategy, and then extracting from such a set a multi-label decision tree classifier. A local explanation represented by a decision rule is then extracted from the obtained decision tree. For the generation of the neighborhood of x we propose two alternative strategies based on the idea of generating neighbors close to x with respect to the feature values and the decision assigned by the black box b. The idea of miming the local behaviour of a black box is common with other approaches such as LIME [19] and LORE [10]. However, none of these approaches is applicable to explain multi-label black box classifiers. We validate our explanation method with experiments on real datasets to assess quantitatively its accuracy in miming a black box and the complexity of the produced explanations.

The rest of this work is organized as follows. In the *Related Work* Section 2 we discuss relevant works on multi-label classification for health applications and black box decision explanation. Then, *Setting the Stage* Section 3 introduces important notions as multi-label classification, black box classifier and interpretable classifier. Section 4 *Multi-label black box Outcome Explanation* presents a problem formalization and *Multi-label Explainer* Section 5 describes the details of the proposed explanation method. In the *Experiments* Section 6 we report a deep experimentation using datasets concerning health applications. Finally, we conclude the paper by discussing strengths and weaknesses of the proposed solutions and future research directions.

2 Related Work

The recent availability of large amounts of electronic health records (EHRs) provides an opportunity for training classification algorithms to develop health applications. EHRs are usually noisy, sparse, have high dimensionality and nonlinear relationships among variables [26]. Deep Learning ability to model non-linear relationships [14] led to successful applications of such technologies to clinical tasks based on EHR data [21]. Deep Learning techniques have been proven useful for patients and medical concepts representation [16], outcome prediction [6, 18, 21] and new phenotype discovery [4, 13].

Consequence of the wide use of black box techniques is a remarkable interest in developing interpretable predictive systems for health applications. To give insights to the behavior of their model, the authors of [6] studied the relationship between the length of the patient medical history and the prediction performance. However, their finding do not help in explaining how the system reasons. In [9] the authors propose a multichannel convolutional neural network based on embeddings of medical concepts to examine the effect of patient characteristics on total hospital costs and length of stay. Despite the good performance the proposed method is completely obscure. A partially interpretable solution to the same problem is described in [2]. The authors propose a model based on the fact that different patient conditions have different temporal progression patterns. The model learns time decay factors for every medical code and allows to analyze the attention weights and disease progression for interpreting the predictions and understand how risks of future visits change over time. However, this approach still depends on a neural network and is not reusable for other applications. In line with [19, 10], our proposal is not to develop interpretable solutions specifically designed for some applications, but to provide an agnostic-approach able to deal with multiple applications and to explain the predictions of high performance classifiers. In [5] the authors compress the knowledge learned by several deep networks into a more interpretable model (gradient boosting trees) which mimics the global behavior of the black box achieving similar performance. In contrast, our approach explains the black box local behavior.

Concerning multi-label prediction, in the literature, there are various approaches using transparent or obscure models. In [24, 3] are proposed variants of decision trees to deal with multi-labels organized into a hierarchy. On the other hand, yet to deal with the multi-label problem, in [1, 20] are presented respectively a fuzzy SVM and a fuzzy neural network. Despite the usage of interpretable models, these work do not offer any specific clue on how to employ them for explanability purposes.

To the best of our knowledge our work is the first attempt to solve local explanation [11] for agnostic health applications in with multi-label classification.

3 Setting the Stage

We recall basic notations on *multi-label* classification [23], the definition of the *outcome explanation problem* [11], and then, we define the notion of *explanation* for multi-label classifiers for which we propose a solution.

A multi-label classifier, is a function $b{:}\mathcal{X}^{(m)}{\rightarrow}\mathcal{Y}^{(l)}$ which maps data instances (tuples) x from a feature space $\mathcal{X}^{(m)}$ with m input features to a decision vector y in a target space $\mathcal{Y}^{(l)}{=}\{0,1\}^l$.

Note that, $y_i{=}1$ if the i^{th} label is associated with the instance x, $y_i{=}0$ otherwise.

We use $b(x){=}y$ to denote the decision y predicted by b, and $b(X){=}Y$ as a shorthand for $\{b(x) \mid x \in X\}{=}Y$.

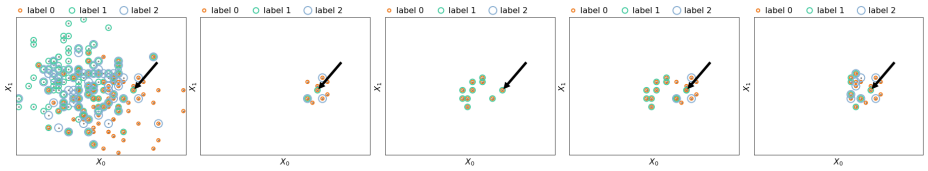

Fig. 1. *(1st)* dataset sample, the arrow points out the instance to explain, Mixed neighborhood generation: *(2nd)* real instances close to x w.r.t. the feature space; *(3rd)* real instances close to x w.r.t. the target space; *(4th)* merge of the previous sets of instances. Unified core real neighborhood: *(5th)* real instances close to x w.r.t. feature and target spaces, i.e., the real core neighborhood.

An instance x consists of a set of m attribute-value pairs (a_i, v_i), where a_i is a feature (or attribute) and v_i is a value from the domain of a_i. The domain of a feature can be continuous or categorical. A predictor can be a machine learning model, a domain-expert rule-based system, or any combination of algorithmic and human knowledge processing. We assume that a classifier can be queried at will. We denote by b a *black box* classifier, whose internals are either unknown to the observer or they are uninterpretable by humans. Examples include neural networks, SVMs, ensemble classifiers, etc. Instead, we denote with c an *interpretable classifier*, whose internal processing yielding a decision $c(x)=y$ has a symbolic interpretation understandable by a human. Examples include rule-based classifiers, decision trees, decision sets, etc.

4 Multi-label black box Outcome Explanation

Given a black box classifier b and an instance x, the *outcome explanation problem*, introduced in [11], consists in providing for the decision $b(x) = y$ an explanation e belonging to a human interpretable domain E.

We address this problem in the specific case in which the black box is a *multi-label classifier*. Our approach is based on the idea, proposed in [10], of learning an interpretable classifier c that reproduces and accurately mimes the *local* behavior of the black box. An explanation for the decision is then derived from c. By *local*, we mean focusing on the behavior of the black box in the *neighborhood* of the specific instance x, without aiming at providing an overall description of the logic of the black box for all possible instances. The neighborhood of x has to be generated as part of the explanation process. We assume that some knowledge is available about the the feature space $\mathcal{X}^{(m)}$, like the ranges of admissible values for the domains of the features and, like in this work, the (empirical) distribution of the features. Nothing is instead assumed about the process of constructing the black box b. Let us formalize the problem of outcome explanation through interpretable models.

Definition 1 (Explanation Through Interpretable Models). *Let $c = \zeta(b, x)$ be an interpretable classifier derived from the black box b and the instance x using some process $\zeta(\cdot, \cdot)$. An explanation $e \in E$ is obtained through c, if $e = \varepsilon(c, x)$ for some explanation logic $\varepsilon(\cdot, \cdot)$ which reasons over c and x.*

In the next section we will describe the process $\zeta(\cdot, \cdot)$ we propose for obtaining an interpretable classifier c. As a consequence, like in [10], we adopt as explanation a

decision rule (simply, a rule) r of the form $p \rightarrow y$ describing the reason for the decision value $y = c(x)$. The decision y is the *consequence* of the rule, while the *premise* p is a boolean condition on feature values.

Definition 2 (Local Explanation). *Let x be an instance, and $c(x)$=y be the decision of an interpretable multi-label classifier c. A local explanation e is a a decision rule r=$(p \rightarrow y)$ consistent with c and satisfied by x.*

Let us consider as an example the following explanation for the diagnoses prediction of a patient:

$$\begin{aligned} e = \{ & 60 < age \leq 70, \\ & BMI{>}36.2, \\ & hyperglycemia{=}Yes, \\ & insulin{=}Up, \\ & systolicpressure{=}150/100mmHg\} \\ & \rightarrow [Diabetes, Hypertension, Hypothyroidism] \end{aligned} \tag{1}$$

$$e = \{60 < age \leq 70, BMI{>}36.2, hyperglycemia{=}Yes, insulin{=}Up,$$

$$systolicpressure{=}150/100mmHg\} \rightarrow [Diabetes, Hypertension, Hypothyroidism]$$

.

The meaning of this explanation is that the diagnoses of *diabetes*, *hypertension* and *hypothyroidism* are predicted by the black box because the patient is obese (BMI¿36.2), his systolic pressure is high, his age is in the $[60, 70)$ range and his blood test results show high levels of sugar (hyperglycemia) and insulin. For the sake of clarity, we only show the diseases that have been predicted by the black box, which correspond to non-zero elements of the binary label vector $y \in \mathcal{Y}^{(l)} = \{0, 1\}^l$.

We assume that p is the conjunction of split conditions sc of the form $a \in [v_1, v_2]$, where a is a feature and v_1, v_2 are values in the domain of a extended with $\pm\infty$. An instance x *satisfies* r, or r *covers* x, if the boolean condition p evaluates to true for x, i.e., if $sc(x)$ is true for every $sc \in p$. For example, the rule $r = \{60{<}age{\leq}70, BMI{>}36.2, hyperglycemia{=}Yes\} \rightarrow [Diabetes, Hypertension, Hypothyroidism]$ is satisfied by x_0=$\{age{=}63, BMI{=}36.5, hyperglycemia{=}Yes\}$ and not satisfied by x_1=$\{age{=}65, BMI{=}35, hyperglycemia{=}No\}$.

We say that r *is consistent* with c, if $c(x)$=y for every instance x that satisfies r. Consistency means that the rule specifies some conditions for which the classifier makes a specific decision. When the instance x for which we have to explain the decision satisfies p, the rule $p \rightarrow y$ represents *a motivation for taking* a decision value, i.e., p locally explains why b returned y. Therefore, a solution to the problem will consists of: *(i)* computing an interpretable predictor c for a black box b and an instance x, i.e., designing function $\zeta(\cdot, \cdot)$ according to Definition 1; *(ii)* deriving a local explanation e from c and x, i.e., defining the explanation logic $\varepsilon(\cdot, \cdot)$ according to Definition 2.

5 Multi-label Explainer

We propose **MARLENA** (**M**ulti-l**a**bel **R**u**l**e-based **E**xpla**NA**tions, as a solution to the multi-label black box outcome explanation problem. An interpretable decision tree classifier c is built for a given multi-label black box b and instance x by first generating a set of neighbor instances of x through the approach presented in the following, and then extracting from such a set a *decision tree* c. A local explanation, consisting of a single rule r, is then derived from the structure of c.

5.1 Neighborhood Generation

The goal of this phase is to identify a set of synthetic instances Z, with feature and/or label values close to the ones of x, in order to reproduce the local decision behavior of the multi-label black box b.

Since the objective is to learn a classifier, the neighborhood should be flexible enough to include instances with both decisions equal to $b(x)$, i.e. $b(z)$=$b(x)$ and decisions different from $b(x)$, i.e, $b(z)\neq b(x)$.

For the generation of Z we propose two approaches which

1. construct a *core real neighborhood* of x, useful for deriving the empirical distributions of features of x
2. randomly generate the set of *synthetic neighbors* Z according to these distributions

In order to derive the *core real neighbors* X^* these approaches assume as input a set of *known instances* $\hat{X}\in\mathcal{X}^{(m)}$ that may be a set of instances of the training set, a set of instances to be explained or in general, a set of instances belonging to the same domain of x. Given $\hat{X}$ the neighborhood X^* is built by identifying the instances of $\hat{X}$ which satisfy specific criteria. In our experiments, we setup $\hat{X}$ as the instances to explain in the test set.

Mixed Neighborhood This method selects from the given instances $\hat{X}$ a core of k real neighbors $X^* = X_f \cup X_l$, where

$$k = k_f + k_l$$

$$k_f{=}\alpha k \text{ and } k_l{=}(1{-}\alpha)$$

Figure 1 (2-4) shows a graphical representation of mixed neighborhood generation starting from a sample dataset with three different labels (left most plot).The arrow points out the instance to explain.

- The set X_f is composed of the k_f instances $\hat{x} \in \hat{X}$ closest to x with respect to the feature space $\mathcal{X}^{(m)}$, according to a distance function $d_f(x, \hat{x})$
- The set X_l comprises the k_l instances $\hat{x} \in \hat{X}$ closest to x with respect to the target space $\mathcal{Y}^{(l)}$, i.e., the black box decision, according to a distance function $d_l(b(x), b(\hat{x}))$.

In Figure 1, the set X_f is showed in the *(2nd)* plot, the set X_l is represented in the *(3rd)* plot and, the *(4th)* plot reports the *core real neighborhood*.
The parameter α is fundamental for the selection of the instances. Indeed, we underline that instances in X_l which are close to x with respect to the decision are not necessarily close to x in the feature space. Therefore, low values of α could bring to the generation of a sparse real core neighborhood in the feature space. This aspect is evident looking at Figure 1 where instances in (3) are sparser than the instances in (4).

Unified Neighborhood This method selects from the given instances $\hat{X}$ a core of k real neighbors X^* as the k instances $\hat{x} \in \hat{X}$ closest to x with respect to both the feature space $\mathcal{X}^{(m)}$and the target space $\mathcal{Y}^{(l)}$, according to a distance function $d_u(x, \hat{x}, b)$ which combines functions d_f and d_l:

$$d_u(x, \hat{x}, b) = \frac{m}{m+l} \cdot d_f(x, \hat{x}) + \frac{l}{m+l} \cdot d_l(b(x), b(\hat{x}))$$

. Figure 1 *(5th)* plot.

Both approaches are parametric with respect to the distance functions $d_f(\cdot, \cdot)$ and $d_l(\cdot, \cdot)$. Since we have binary vectors with length l, in the target space we use the Hamming distance as $d_l(\cdot, \cdot)$. On the other hand, in the feature space we account for the presence of mixed types of features by a weighted sum of the Hamming distance [22] for categorical features, and of the normalized Euclidean distance[4] for continuous features. Thus, assuming s categorical features and $m - s$ continuous ones, we use:
$d_f(x, \hat{x}) = \frac{s}{m} \cdot Hamming(x, \hat{x}) + \frac{m-s}{m} \cdot nEuclidean(x, \hat{x})$

In the following, we name **MARLENA-m** the MARLENA algorithm using the *mixed neighborhood* distance function, **MARLENA-u** the MARLENA algorithm using the *unified neighborhood* distance function.

5.2 Rule-based Explanation

Given the synthetic neighborhood Z of x, the second step is to build an interpretable classifier c trained on the instances $z \in Z$ labeled with the black box decision $b(z)$. Such a classifier is intended to mimic the behavior of b locally in the Z neighborhood. **MARLENA** adopts multi-label decision tree as interpretable classifier c as it makes easy the explanation extraction. Indeed, given the multi-label decision tree c, we derive the decision rule representing the explanation as a root-leaf path in the tree, i.e., the decision rule $r = (p \rightarrow y)$ is formed by including in p the split conditions on the path from the root to the leaf node that is satisfied by the instance x, and setting $y = c(x)$. By construction, the rule r is consistent with c and satisfied by x.

6 Experiments

In this section, we describe the experiments we carried out to evaluate the performance of **MARLENA**. We first present the experimental setup and then we show the results of our analyses which prove that the proposed multi-label local approach is more effective than a global one. We study the effect of the neighborhood generation parameter α on **MARLENA-m** performance, and we provide a qualitative and quantitative evaluation of the multi-label explanations[5]. **MARLENA** was developed in Python[6], we used the `sklearn` implementation of the multi-label decision tree as interpretable classifier.

[4] `http://reference.wolfram.com/language/ref/NormalizedSquaredEuclideanDistance.html`

[5] For both neighborhood generation approaches *mixed* and *union*, the size of the synthetic neighborhood is 1000, and the size of the *core real neighborhood* X^* is $k = 0.5|\hat{X}|^{1/2}$

[6] Source code, datasets, and the scripts for reproducing experiments are publicly available at `https://github.com/riccotti/ExplainMultilabelClassifiers`

Dataset	instances	features	labels	avg. labels	RF	SVM	MLP
yeast	2,417	117	14	4.24	.62	.62	.64
women	14,644	44	14	3.53	.71	.72	.71
medical	978	1449	45	1.25	.37	.79	.77

Table 1. Real health-related dataset information and black box performance (F1-measure).

6.1 Experimental Setup

Datasets. We ran experiments on three real-world mulit-label tabular datasets: *yeast* [8], *woman*[7] and *medical* [17]. The *yeast* dataset is a collection of yeast microarray expressions and phylogenetic profiles which can be used to learn the yeast gene functional categories. One row of this dataset represents a gene, and the labels are its associated functional classes. Each gene might belong to more than one functional class. The *woman* dataset contains survey data about women health-care requirements gathered by a US non-profit organization. One row of this dataset contains the questionnaire replies of one woman concerning her demographics, pregnancies, family planning, use of health care services, and medical insurance. The labels of this dataset are the health-care requirements. The *medical* dataset contains a corpus of fully anonymized clinical text. Each document in the corpus is associated with a set of ICD-9 codes which represents the diagnosis associated with the clinical report. To each report might be assigned several ICD-9 codes. The *woman* dataset includes both categorical and continuous features, the *yeast* only continuous features and the *medical* dataset contains only binary features that represent the presence or absence of each word in each document.

Details of the datasets after missing values correction [8] and black box performance are reported in Table 1. To train the black boxes, we randomly split the *yeast* and *woman* dataset into a training and a test set containing respectively 70% and 30% of the instances. For the *medical* dataset we use the partitioning described in [17]. After the training phase we used the black boxes to classify the instances in the test set, denoted by X, and we used the **MARLENA** approach to explain such decisions. We denote by $\hat{Y}$ the decisions provided by the black box b on X, and with Y the decisions provided by the explainer c. We underline that the black box performance is not the focus of our work: we forget about the real label and we use the black box labels as target labels.

Black Box Classifiers. We experiment the following predictors as black boxes: Random Forests (**RF**), Support Vector Machines (**SVM**), and Multi-Layer Perceptron (**MLP**)[9]. For each black box, we perform hyper-parameters tuning using a five-fold cross-validation and a randomized search over a grid of parameters on the training set[10].

Evaluation Measures. We adopt the following metrics to evaluate **MARLENA**'s performance. Aggregated values[11] are reported in the experiments by averaging them.

[7] `https://tinyurl.com/y9maxnxr`, `https://tinyurl.com/yaz2lyrc`

[8] We replace the missing values with the mean for continuous variables and with the mode for categorical ones. We remove the features with more than 40% of missing values.

[9] Implementations are those of `scikit-learn` library.

[10] Details available at `https://github.com/riccotti/ExplainMultilabelClassifiers`.

[11] The performance reported consider only instances for which an explanation is returned. Indeed, for some instances of the *medical* dataset using the *RF* black box an explanation is not returned. We leave the investigation of this specific case fur future studies.

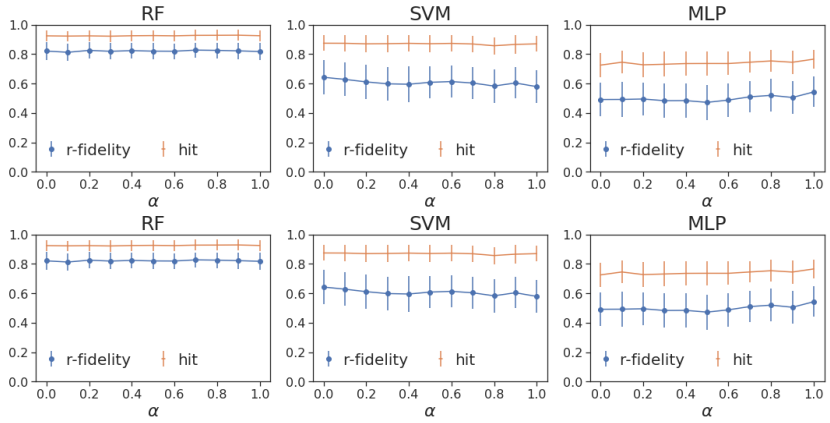

Fig. 2. Hit and *r-fidelity* varying α for *yeast* and *woman*, upper and lower figure respectively.

– ***fidelity***$(Y, \hat{Y}) \in [0, 1]$. It compares the decisions of the interpretable classifier c to those of the black box b on the set X. The *s-fidelity* measures the performance on the synthetic neighborhood, $X{=}Z$. The *r-fidelity* measures the performance on the core real neighborhood, $X{=}\hat{X}$. It answers the question:"how good is c at mimicking b in a neighborhood of x?". We measure it using the F1-measure [22].

– ***hit***$(y, \hat{y}) \in [0, 1]$. It compares the prediction of c and b on the instance x under analysis. We use the simple match similarity to evaluate it, i.e., $1 - hamming(y, \hat{y})$. $hit(y, \hat{y}) = 1$ means that c correctly identifies all the labels returned by b, a value between 0 and 1 means that some labels are misclassifed.

	s-fidelity		*r-fidelity*	
Black Box	*mixed*	*unified*	*mixed*	*unified*
RF	.94 ± .02	.90 ± .05	.89 ± .09	.87 ± .11
SVM	.91 ± .05	.87 ± .07	.65 ± .20	.68 ± .21
MLP	.93 ± .07	.91 ± .11	.68 ± .22	.68 ± .21

Table 2. Fidelity (mean ± stddev) of *MARLENA-m* and *MARLENA-u* on all datasets.

Dataset	yeast		woman		medical	
Black Box	*mixed*	*union*	*mixed*	*union*	*mixed*	*union*
RF	.93 ± .03	.92 ± .04	.94 ± .02	.90 ± .05	.93 ± .06	.90 ± .12
SVM	.84 ± .07	.84 ± .08	.92 ± .03	.88 ± .05	.95 ± .05	.86 ± .14
MLP	.90 ± .05	.90 ± .06	.95 ± .02	.94 ± .04	.80 ± .12	.72 ± .20

Table 3. *s-fidelity* (mean ± stddev) of *MARLENA mixed* and *union* for each dataset.

Dataset	yeast		woman		medical	
Black Box	*mixed*	*union*	*mixed*	*union*	*mixed*	*union*
RF	.89 ± .06	.90 ± .06	.89 ± .09	.87 ± .12	.94 ± .09	.97 ± .06
SVM	.86 ± .08	.86 ± .08	.57 ± .16	.60 ± .18	.92 ± .12	.97 ± .06
MLP	.89 ± .06	.89 ± .07	.62 ± .21	.61 ± .19	.81 ± .20	.89 ± .14

Table 4. *r-fidelity* (mean ± stddev) of *MARLENA mixed* and *union* for each dataset.

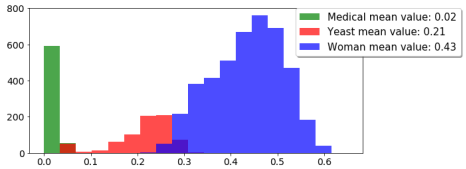

Fig. 3. Distributions of mean mixed distance among core real neighborhood points.

6.2 Results

We perform several experiments to assess how **MARLENA-m** performance are impacted by the neighborhood generation parameter α. We measure *r-fidelity* and *hit* for different values of α, the results are show in figure 2. We observe that the value of α does not have a noticeable impact on the **MARLENA-m** performance. Therefore, we can safely set α=0.7 for the following analyses, this guarantees the locality in the feature space of the core of real instances selected to generate the synthetic neighborhood. We recall that high values of α favorite neighbors close to x in the feature space.

To understand if one of the two approaches of neighborhood generation performs significantly better than the other, we compare them in terms of their *s-fidelity* and *r-fidelity* on the *woman* and *yeast* datasets. The results are reported in Tables 2. We observe that the two approaches have comparable performance, but the *mixed* approach performs slightly better on the synthetic neighborhood. We can also see how the aggregated performance on all datasets show lower values of *r-fidelity* when our methods are used to explain *SVM* and *MLP* decisions. Looking at *r-fidelity* values in Table 2, we observe that this behaviour is due to weak performance on the *woman* dataset. This gap of performance among the different datasets is due to the different levels of cohesion of the data points selected in the *core real neighborhood* in the feature space.

In order to quantitatively measure the level of cohesion of each neighborhood, we compute the SSE (Sum of Squared Errors [22]) employing distance function d_f defined in section 5.1. In Figure 3 we report the distribution of SSE values, i.e., the mean values of distances among the data points in the core real neighborhoods for each dataset. We observe how the data points in the *woman* dataset are more distant from the center of their neighborhood, compared to those of the other two datasets. This impacts the performance of the methods because selecting data points scattered in the feature space for the core real neighborhood generates a synthetic neighborhood which does not preserve the locality around the instance to be explained. The relationship between **MARLENA** performance and data points scatteredness in the *core real neighborhood* requires a detailed study and is left for future work.

For measuring the ability of **MARLENA** to mimic the black box behavior, we compare its *hit*-performance against those of a Global Decision Tree (GDT) learned on the

Dataset	yeast		woman		medical	
Black Box	MARLENA-m	GDT	MARLENA-m	GDT	MARLENA-m	GDT
RF	.97 ± .05	.98 ± .04	.95 ± .06	.99 ± .04	1.00 ± .01	1.00 ± .01
SVM	.95 ± .06	.93 ± .07	.87 ± .09	.99 ± .03	1.00 ± .01	.99 ± .01
MLP	.97 ± .05	.94 ± .07	.82 ± .13	.99 ± .03	.99 ± .01	.99 ± .01

Table 5. Hit performance comparison (mean and standard deviation).

Dataset	yeast		woman		medical	
Black Box	MARLENA-u	GDT	MARLENA-u	GDT	MARLENA-u	GDT
RF	.97 ± .05	.98 ± .04	.94 ± .07	.99 ± .04	1.00 ± .00	1.00 ± .01
SVM	.95 ± .06	.93 ± .07	.87 ± .09	.99 ± .03	1.00 ± .01	.99 ± .01
MLP	.96 ± .05	.94 ± .07	.81 ± .12	.99 ± .03	1.00 ± .01	.99 ± .01

Table 6. Hit performance comparison (mean and standard deviation).

set of instances to be explained with target labels given by the black box. The results for both the *mixed* and *unified* approaches are shown in Table 5 and Table 6, respectively. We underline how the comparison with such a global approach is not trivial, since the hit performance of the global decision tree (GDT) are high, all above 0.93. Our approaches outperform the global one in mimicking the SVM and the MLP black box on the *yeast* dataset. However, although **MARLENA** in some cases performs worse in terms of hit, it always greatly outperforms the GDT in terms of rule interpretability. Indeed, as shown in Tables 7 and 8, **MARLENA** always produces explanations (decision rules) with considerable lower number of conditions in the rule premise. The reduction of rule length is really important especially on *woman* dataset.

We now make a qualitative comparison of the explanations provided by **MARLENA-m** and the GDT. We consider explanations for black box behavior on the *medical* dataset since its features are easily comprehensible also by non-experts. What follows is an example of an explanation for the *SVM* black box where both **MARLENA-m** (e_M) and the GDT (e_G) predict the same labels as the black box. In the *medical* dataset the classification task is to map words coming from clinical notes to one or more diagnosis. The following explanations highlights which are the words that influenced more the black box decision with their presence or absence. We highlight words common to both explanations as they probably are the most important for the decision.

$$e_M = \{duplication{=}0, \textbf{\textit{reflux}}{=}0, \textbf{\textit{hydronephrosis}}{=}1, normal{=}1, \textbf{\textit{pyelectasis}}{=}1, mild{=}1\} \rightarrow [Urinaryincontinence, Hydronephrosis]$$

$$e_G = \{cough{=}0, \textbf{\textit{reflux}}{=}0, tract{=}0, neurogenic{=}0, \textbf{\textit{hydronephrosis}}{=}1, hydroureter{=}0, evaluate{=}0, \textbf{\textit{pyelectasis}}{=}1, follow{=}1\} \rightarrow [Urinaryincontinence, Hydronephrosis]$$

GDT's explanation is longer and more confusing as it contains words falling outside the context of kidney problems, like *cough*, and generic words like *evaluate* and *follow*.

7 Conclusion

We have proposed **MARLENA** a model agnostic approach to address the multi-label black box outcome explanation problem. Our approach learns a *local* classifier on a synthetic neighborhood generated by a strategy suitable for multi-label decisions. Then, it

Dataset	yeast		woman		medical	
Black Box	MARLENA-u	GDT	MARLENA	GDT	MARLENA-u	GDT
RF	2.92 ± 2.27	9.09 ± 3.35	$4.30 \pm .98$	13.20 ± 4.56	1.41 ± 1.90	7.70 ± 3.12
SVM	3.29 ± 2.24	5.68 ± 1.47	4.31 ± 1.51	16.30 ± 6.61	5.35 ± 1.67	11.76 ± 4.82
MLP	2.44 ± 1.99	6.70 ± 2.36	2.93 ± 1.17	14.85 ± 6.17	4.58 ± 1.40	10.77 ± 5.40

Table 7. Mean rule length and standard deviation comparison between MARLENA-m and GDT.

Dataset	yeast		woman		medical	
Black Box	MARLENA-u	GDT	MARLENA	GDT	MARLENA-u	GDT
RF	2.91 ± 2.44	9.09 ± 3.35	4.36 ± 1.19	13.20 ± 4.56	1.80 ± 2.01	7.70 ± 3.12
SVM	3.18 ± 1.99	5.68 ± 1.47	4.36 ± 1.62	16.30 ± 6.61	4.31 ± 2.32	11.76 ± 4.82
MLP	2.70 ± 2.30	6.70 ± 2.36	2.77 ± 1.42	14.85 ± 6.17	4.50 ± 1.75	10.77 ± 5.40

Table 8. Mean rule length and standard deviation comparison between MARLENA-u and GDT.

derives from the interpretable local prediction a meaningful explanation represented by a decision rule, explaining the reasons of the decision. We have proposed two strategies for the synthetic neighborhood generation that take into consideration the particular structure of the multi-label decision. Our experimentation shows that **MARLENA** presents acceptable performance in terms of accuracy in mimicking the black box and is able to produce explanations represented by compact rules.

A number of extensions and additional experiments can be considered for future works. An interesting future research direction is to design new approaches for the neighborhood generation for example methods based on the genetic programming. Second, another study might be focused on the possibility to generate a global explainer by composing the local explanations produced by **MARLENA**. Moreover, results in this paper show that it is necessary to extend the experiments by considering more datasets (even synthetic) characterized by different levels of density and to understand how this impact to the quality of neighborhood generation. Finally, it would be interesting to let domain experts evaluate and compare **MARLENA** explanations to the global ones.

Acknowledgements. This work is partially supported by the European H2020 Program under the funding scheme "INFRAIA-1-2014-2015: Research Infrastructures" g.a. 654024 *"SoBigData"*, `http://www.sobigdata.eu`.

References

1. S. Abe. Fuzzy support vector machines for multilabel classification. *PR*, (6):2110, 2015.
2. T. Bai et al. Interpretable representation learning for healthcare via capturing disease progression through time. In *KDD*, pages 43–51. ACM, 2018.
3. H. Blockeel, L. Schietgat, J. Struyf, A. Clare, and S. Dzeroski. Hierarchical multilabel classification trees for gene function prediction. In *MLSB*, pages 9–14, 2006.
4. Z. Che, D. Kale, W. Li, M. T. Bahadori, and Y. Liu. Deep computational phenotyping. In *KDD*, pages 507–516. ACM, 2015.
5. Z. Che, S. Purushotham, R. Khemani, and Y. Liu. Interpretable deep models for icu outcome prediction. In *AMIA Annual Symposium Proceedings*, volume 2016, page 371. American Medical Informatics Association, 2017.

6. E. Choi et al. Doctor ai: Predicting clinical events via recurrent neural networks. In *Machine Learning for Healthcare Conference*, pages 301–318, 2016.
7. M. Chui. Artificial intelligence the next digital frontier? *McKinsey and CGI*, page 47, 2017.
8. A. Elisseeff and J. Weston. A kernel method for multi-labelled classification. In *Advances in neural information processing systems*, pages 681–687, 2002.
9. Y. Feng et al. Patient outcome prediction via convolutional neural networks based on multi-granularity medical concept embedding. In *BIBM*, pages 770–777. IEEE, 2017.
10. R. Guidotti, A. Monreale, S. Ruggieri, D. Pedreschi, F. Turini, and F. Giannotti. Local rule-based explanations of black box decision systems. *CoRR*, abs/1805.10820, 2018.
11. R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, D. Pedreschi, and F. Giannotti. A survey of methods for explaining black box models. *ACM CSUR*, 51(5):93:1–93:42, Aug. 2018.
12. R. Guidotti, J. Soldani, D. Neri, A. Brogi, and D. Pedreschi. Helping your docker images to spread based on explainable models. In *ECML-PKDD*. Springer, 2018.
13. T. A. Lasko et al. Computational phenotype discovery using unsupervised feature learning over noisy, sparse, and irregular clinical data. *PloS one*, 8(6):e66341, 2013.
14. Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *nature*, 521(7553):436, 2015.
15. G. Malgieri and G. Comandé. Why a right to legibility of automated decision-making exists in the General Data Protection Regulation. *Int. Data Privacy Law*, 7(4):243–265, 2017.
16. R. Miotto et al. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Scientific reports*, 6:26094, 2016.
17. J. P. Pestian et al. A shared task involving multi-label classification of clinical free text. In *BioNLP*, pages 97–104. Association for Computational Linguistics, 2007.
18. A. Rajkomar et al. Scalable and accurate deep learning with ehr. *DM*, 1(1):18, 2018.
19. M. T. Ribeiro, S. Singh, and C. Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *KDD*, pages 1135–1144. ACM, 2016.
20. E. P. Sapozhnikova. Art-based neural networks for multi-label classification. In *International Symposium on Intelligent Data Analysis*, pages 167–177. Springer, 2009.
21. B. Shickel et al. Deep ehr: A survey of recent advances in deep learning techniques for ehr analysis. *Journal of biomedical and health informatics*, 22(5):1589–1604, 2018.
22. P.-N. Tan et al. *Introduction to data mining*. Pearson Education India, 2007.
23. G. Tsoumakas and I. Katakis. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM)*, 3(3):1–13, 2007.
24. C. Vens, J. Struyf, L. Schietgat, S. Džeroski, and H. Blockeel. Decision trees for hierarchical multi-label classification. *Machine learning*, 73(2):185, 2008.
25. S. Wachter et al. Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *Int. Data Privacy Law*, 7(2):76–99, 2017.
26. P. Yadav, M. Steinbach, V. Kumar, and G. Simon. Mining electronic health records (ehrs): a survey. *ACM Computing Surveys (CSUR)*, 50(6):85, 2018.