Learning and Nonlinear Models - Journal of the Brazilian Society on Computational Intelligence (SBIC), Vol. 21, Iss. 2, pp. 4-15, 2023

LOCAL RULE-BASED EXPLANATIONS METHOD BASED ON GENETIC ALGORITHMS WITH FITNESS SHARING

Daniel A. Santos, José A. Baranauskas, Renato Tinós
Department of Computing and Mathematics, University of São Paulo, Ribeirão Preto, SP, Brazil
{daniel.asantos@usp.br, augusto@usp.br, rtinos@ffclrp.usp.br}

Abstract – The Local Rule Based Explanations method (LORE) explains decisions of black-box classifiers by using an interpretable model (Decision Tree - DT). The DT is trained with an artificial dataset generated by Genetic Algorithms (GAs). The primary objective of this approach is to replicate the decision boundaries of the black-box model in proximity to the instance under explanation. We show that the artificial examples generated by the GAs in LORE are not necessarily diverse. Consequently, we propose the integration of GAs with fitness sharing in LORE to generate a more diversified subset of artificial examples. The underlying motivation is to ensure that the local decision boundaries of the DT more closely resemble those of the black-box classifier. Experimental results with two classifiers (Multilayer Perceptron and Random Forests), and four classification problems, indicate that LORE with fitness sharing yields more diverse GA populations, consequently leading to improved local explanations. These findings underscore the effectiveness of incorporating fitness sharing into the LORE methodology for enhancing the explainability of black-box classifiers.

Keywords – Explainable Artificial Intelligence, Genetic Algorithms, Fitness Sharing.

1 Introduction

There is a great need for interpretable or explanatory Artificial Intelligence (XAI) models in many areas [1]. Despite the remarkable performance of *Machine Learning* (ML) models in many problems, the decisions made by some of them are not easily interpreted by humans [2]. *Artificial Neural Networks* (ANNs), and *Random Forests* (RFs) are examples of ML models that behave like black-boxes, i.e., models that generate input-output map approximations not easily interpreted by humans. Black-box models are not easily explainable either because its internal mechanisms are obscure, e.g., proprietary software, or because its decisions are not easily interpreted by humans. On the other hand, XAI models allow explaining decisions when input data is transformed into outputs. These explanations are important to ensure that the algorithms are suitable for domain experts, and to ensure that the algorithms work as expected [3].

Some efficient ML techniques, such as *Convolutional Neural Networks* (CNNs), do not have an internal logical mechanism to explain how a given result is obtained. However, the interpretability is an essential question for ML applied to many areas, e.g., Medicine and Healthcare [4,5]. Despite giving remarkable results in many medical applications, successful intelligent algorithms are not easily accepted in Medicine because they lack interpretability. For example, in medical image analysis, CNNs outperform radiologists in many diagnostic cases [6]. However, patients, and medical professionals will rarely accept their results (without a human intervention) because CNNs will not explain how the result was achieved in the same way radiologists easily do.

The lack of interpretability still can impact the possibilities of integrating human and artificial intelligence. After training the ML models, the knowledge of humans in many areas is rarely used to improve decisions taken by intelligent algorithms. On other hand, the knowledge obtained by ML algorithms from large and complex datasets are little used to improve the knowledge in particular areas. In addition, with the popularization of the use of ML, new regulatory laws emerge. An example is the *General Data Protection Regulation* (GDPR) [7] of the European Union. In its Article 22, GDPR guarantees that anyone affected by ML algorithms decisions has the right to know why that decision was taken.

Recent scientific works try to solve the interpretability problem by proposing algorithms that explain the decisions taken by ML models [8–10]. Guidotti et al. [11] survey XAI methods with different properties, pointing out their advantages, and disadvantages. Local XAI models explain particular decisions of black-box in terms that are understandable for humans. Some of the local XAI methods are specific, i.e., they explain decisions only for a particular black-box model. For example, saliency masks can be used in CNNs to indicate parts of images or sentences in a text that are most important for making a decision. On other hand, there are agnostic models, meaning that they can be used to explain decisions taken by any ML model; in practice, adaptations in ML models can be required in some applications, e.g., in image analysis.

One of the most popular agnostic XAI models is the *Local Interpretable Model-agnostic Explanations* method (LIME), proposed by Ribeiro et al. [9]. LIME is a linear explicability model, i.e., it assumes that every ML model is locally linear. LIME fits a simple linear model in the neighborhood of a single instance whose black-box model decision must be explained. The adjustment of the linear model is done by generating an artificial dataset through the perturbation of the instance to be explained. Each instance of the artificial dataset is classified by the black-box model and its (Euclidean) distance from the instance to be explained is calculated. A selection of the most relevant points for the adjustment of the linear model is then made according to the distance. The smaller the distance is, the higher the probability that a point belongs to the neighborhood of the instance to

be explained; therefore, it is more relevant for a local explanation. The linear model in LIME locally mimics the behavior of the black-box model; therefore, it can be used to locally explain its decisions. In classification problems, the decision boundaries of the black-box model are reproduced by the linear model in the neighborhood of the instance to be explained.

In [10], Guidotti et al. propose the *Local Rule-based Explanations* method (LORE). LORE is agnostic and local. It aims to explain the decision of an ML model for a given instance by creating a surrogate model that is interpretable. When using LORE, the dataset must be tabular, and the classification must be binary. In experiments presented in [10], LORE outperformed LIME when the classification accuracy and the quality of the explanations are considered. Local explanations in LORE are produced by training a *Decision Tree* (DT) that locally reproduce the black-box model decision boundaries. To do this, a dataset is generated by optimizing populations of artificial instances in standard *Genetic Algorithms* (GAs).

In [12], we showed that the GAs used in LORE do not necessarily generate a diverse artificial dataset. The population diversity is important when training DTs to accurately reproduce the local decision boundaries of the black-box model. Then, we proposed that fitness sharing can be used to preserve the diversity of the population. Fitness sharing is a niching strategy that penalizes solutions on densely populated regions, increasing the diversification of the population [13]. Consequently, decision boundaries of the DT are expected to be more similar to the local decision boundaries of the black-box model in the modified XAI model.

In this paper, we extend the work presented in [12]. First, a method is presented for empirically finding parameter σ, that controls the size of the niches in fitness sharing. The parameter σ has a great impact in the diversity of the GA population and in the quality of the decision boundaries generated by the DT. Second, we use a measure to quantify the diversity of the final populations generated by the GAs in LORE; in the previous work, the diversity was analyzed only qualitatively by visual inspection of the population distribution for some runs of the GAs. Finally, we extend the number and type of experiments. Previously, experimental results with one classifier (Multilayer Perceptron) and three classification datasets were analyzed. Here, new results are added: i) for finding parameter σ; ii) with RFs as black-box models; iii) with an additional dataset from the Medicine area.

This paper is organized as follows. The proposed LORE with fitness sharing is presented in Section 2. Section 3 shows the experimental results; standard LORE is compared to LORE with fitness sharing in a series of experiments. The paper is concluded in Section 4.

2 Methodology

We propose using a GA with fitness sharing in LORE. LORE is presented in Section 2.1, while Section 2.2 presents LORE with fitness sharing.

2.1 LORE

In LORE, a symbolic surrogate model (DT) is used to explain a black-box model decision for a given instance. The DT is trained by using an artificial dataset optimized by GAs and initially generated by randomly disturbing the instance to be explained. According to [14, 15], the construction of a DT (decision tree) is carried out as follows. Using the training set, an attribute is chosen to partition the instances into subsets, according to values of this attribute. For each subset, another attribute is chosen to repartition each one. This process continues if one of the subsets contains a mixture of instances belonging to different classes. Once a uniform subset is obtained in which all instances in that subset belong to the same class, a leaf node is created and labeled with the respective class name. When a new instance must be classified, starting at the root of the DT, the classifier tests and branches to each node with the respective attribute until it reaches a leaf. The class of this leaf node will be assigned to the new instance. A DT is an inherently understandable model: each path from the root to a leaf of the decision tree can be easily converted into an if-then rule, also an understandable model. Detailed surveys on the explainability of models can be found in [16, 17].

The artificial dataset generated in LORE is composed of the final populations optimized by two GAs. Alg. 1 presents the pseudo-code of LORE. Given a binary classification black-box model b, its decision $y_b(\mathbf{x})$ for instance $\mathbf{x}$ is explained by $e_c(\mathbf{x})$, generated by DT c. The surrogate model (DT c) is trained using an artificial dataset Z. The artificial dataset Z is composed of the populations (Z_1 and Z_2) optimized by the GAs, each one with a different fitness function. While GA_1 generates artificial instances with label $y_b(\mathbf{x})$, GA_2 generates instances with the opposite label. The label $y_b(\mathbf{x})$ is generated by running the black-box model b, presenting instance $\mathbf{x}$ in its inputs.

The explanation of the decision of black-box model b for instance $\mathbf{x}$ is formed from the subset of logical rules r extracted from DT c. Rules r are extracted by traversing the path for deciding instance $\mathbf{x}$ in DT c. Counterfactual subset Φ, with conditions that change the label $y_b(\mathbf{x})$, is extracted by traversing the paths in c that result in a different classification. Each GA optimizes an initial population composed of multiple copies of instance $\mathbf{x}$. The population then evolves through standard selection, mutation and crossover operators [13, 18]. The fitness of an artificial instance (individual) $\mathbf{z}$ is respectively computed in the first and second GAs by:

$$f_1(\mathbf{z}) = I_{(y_b(\mathbf{x})=y_b(\mathbf{z}))} + (1 - d(\mathbf{x}, \mathbf{z})) - I_{(\mathbf{x}=\mathbf{z})} \tag{1}$$

and

$$f_2(\mathbf{z}) = I_{(y_b(\mathbf{x})\neq y_b(\mathbf{z}))} + (1 - d(\mathbf{x}, \mathbf{z})) - I_{(\mathbf{x}=\mathbf{z})} \tag{2}$$

Algorithm 1 LORE [10]

Input: $\mathbf{x}$ - instance to explain, b - black-box model, N - population size,
G - number of generations, p_c - crossover rate, p_m - mutation rate
Output: $e_c(\mathbf{x})$ - explanation of decision $y_b(\mathbf{x})$

$Z_1 \leftarrow GA_1(\mathbf{x}, b, N/2, G, p_c, p_m)$;
$Z_2 \leftarrow GA_2(\mathbf{x}, b, N/2, G, p_c, p_m)$;
$Z \leftarrow Z_1 \cup Z_2$;
$c \leftarrow DecisionTree(Z)$;
$r_c(\mathbf{x}) \leftarrow ExtractRule(c, \mathbf{x})$;
$\Phi_c(\mathbf{x}) \leftarrow ExtractCounterfactuals(c, r_c(\mathbf{x}), \mathbf{x})$;
$e_c(\mathbf{x}) = \langle r_c(\mathbf{x}), \Phi_c(\mathbf{x}) \rangle$;
return $e_c(\mathbf{x})$

where $I_{(true)} = 1$ and $I_{(false)} = 0$, and $d(\mathbf{x}, \mathbf{z}) \in [0, 1]$ is a distance function. The Normalized Euclidean Distance is used for continuous features, where the variance of the dataset is used to normalize the values. For categorical features, the Simple Match function is used. The sum of the second and third terms of the fitness functions are maximized for artificial instances (solutions) $\mathbf{z}$ different from $\mathbf{x}$, but close to it. The first term, that is different in the two functions, maximizes instances $\mathbf{z}$ with the same (Eq. 1), or different (Eq. 2), label $y_b(\mathbf{x})$.

2.2 LORE with Fitness Sharing

In LORE, the final populations of the GAs compose the artificial dataset that will be used to train the DT. The artificial dataset should allow generating decision boundaries for the DT c that reproduce those of the black-box model b in the neighborhood of instance $\mathbf{x}$. Thus, the diversity of the GAs final populations is very important for generating good surrogate models. For populations with low diversity, the decision boundaries around instance $\mathbf{x}$ will not be properly reproduced. Consequently, the explanation given by the DT can be affected. This can also affect the counterfactual subset generation. The impact of low diversity populations is generally stronger for high-dimensional datasets because much more instances can be necessary to generate the local decision boundaries.

The authors in LORE suggest optimizing the GAs populations for a small number of generations (10). They also suggest setting a high mutation rate (0.2) [10]. Both strategies are used to preserve the diversity of the population. Section 3 shows results of experiments investigating the impact of the mutation rate and number of generations in the diversity of the artificial datasets optimized by GAs. The results indicate that the diversity of individuals around instance $\mathbf{x}$ is low in several runs. Besides, if high mutation rates are adopted in runs with small number of generations, some regions close to the decision boundaries of the black-box model around instance $\mathbf{x}$ are not properly covered by artificial instances. As a result, unsatisfactory local explanations for the classification of instance $\mathbf{x}$ can be generated.

We propose using fitness sharing in the GAs of LORE. Fitness sharing is used to preserve the diversity of the population along the generations. It is a niching strategy that penalizes solutions (individuals) on densely populated regions [13, 19]. Niching strategies encourage diversification of the population. By using fitness sharing, we expect that running the GAs for more generations will result in local decision boundaries of the DT more like to those of the black-box model.

In *LORE with fitness sharing* (LOREfs), the fitness of the i-th individual $\mathbf{z}_i$ of population P for GA_1 is given by:

$$f_1^{\text{new}}(\mathbf{z}_i) = \frac{f_1(\mathbf{z}_i)}{\sum_{j=1}^{N} \psi\big(d(\mathbf{z}_i, \mathbf{z}_j)\big)} \tag{3}$$

where N is the population size, $d(\mathbf{z}_i, \mathbf{z}_j)$ is the distance between individuals $\mathbf{z}_i$ and $\mathbf{z}_j$, and $\psi(\cdot)$ is a sharing function that measures the similarity between the solutions. The original fitness $f_1(\mathbf{z}_i)$ of solution $\mathbf{z}_i$ is given by Eq. 1. For GA_2, the fitness is given by:

$$f_2^{\text{new}}(\mathbf{z}_i) = \frac{f_2(\mathbf{z}_i)}{\sum_{j=1}^{N} \psi\big(d(\mathbf{z}_i, \mathbf{z}_j)\big)} \tag{4}$$

where the original fitness $f_2(\mathbf{z}_i)$ of solution $\mathbf{z}_i$ is given by Eq. 2.

The sharing function [13] is given by:

$$\psi(u) = \begin{cases} 1 - (\frac{u}{\sigma})^{\beta}, & \text{if } u < \sigma \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

where parameter σ defines the niche radius and β controls the shape of the sharing function. Here, $\beta = 1$, i.e., the sharing function is linear. In fitness sharing, solutions within a distance smaller than σ are located in the same niche and, therefore, are penalized. Eq. 5 is the far the most prevalent sharing function [20]. However, other functions can also be utilized. In LOREfs, functions GA_1 and GA_2 of Alg. 1 are modified by computing the fitness using the fitness sharing equations 3 and 4, instead of equations 1 and 2. Alg 2 shows the pseudo-code of GA_1, where the fitness of the individuals is given by Eq. 3. The pseudo-code for GA_2 is the same but using Eq. 4 for computing the fitness of the solutions.

Algorithm 2 GA_1

Input: $\mathbf{x}$ - instance to explain, b - black-box model, N - population size,
G - number of generations, p_c - crossover rate, p_m - mutation rate
σ - niche radius
Output: Z_1 - artificial dataset 1

$P_0 \leftarrow initializePopulation(\mathbf{x})$;
$evaluate(P_0, b, \mathbf{x})$; // Eq. 1
for $i \leftarrow 0$ **to** $G - 1$ **do**
  $P_{i+1} \leftarrow select(P_i)$;
  $P'_{i+1} \leftarrow crossover(P_{i+1}, p_c)$;
  $P''_{i+1} \leftarrow mutate(P'_{i+1}, p_m)$;
  $evaluate(P''_{i+1}, b, \mathbf{x})$; // Eq. 1
  $fitnessSharing(P''_{i+1}, \sigma)$; // Eq. 3
  $P_{i+1} = P''_{i+1}$;
end for
$Z_1 \leftarrow P_{i+1}$
return Z_1

3 Experiments

LOREfs is compared to LORE in a series of experiments. Section 3.1 shows results of experiments evaluating the impact of the number of generations (G) and mutation rate (p_m) in LORE and LOREfs. In this experiment, the black-box model is a Multilayer Perceptron (MLP). LOREfs has a parameter, σ, that controls the size of the niches. In Section 3.2, we present a strategy that empirically find good values of σ. LOREfs is quantitatively compared to LORE in Section 3.3. The comparison is made regarding the classification error for instances of the original dataset close to the instance to be explained. In Section 3.3, results of RFs as black-box models are also presented. Finally, in Section 3.4, the comparison between LORE and LOREfs is made regarding the local diversity of the artificial datasets generated by the GAs.

Four datasets with real attributes were used to compare LORE and LOREfs (see next sections). A desktop with Intel Core i7-2600K (8 MB Cache, 4.20GHz) and 16 GB of RAM was used for running the Python codes. The Python implementation of LORE made available by the authors [10] was employed. In the implementation of LORE, the population size of the GAs is set to 1,000. In order to reduce the size of the artificial dataset used to train the DT, the following procedure is applied to each population of the GA in the original LORE: i. sort the list of individuals of the final population according to their fitness; ii. find the largest fitness difference between two consecutive individuals of the list; iii. define the fitness of the individual with the largest fitness difference (step ii) as a threshold; iv. remove from the artificial dataset all individuals with fitness smaller than the threshold defined in step iii. In initial experiments with LORE, we observed that, for some datasets, the number of individuals removed by using this procedure is too high. Consequently, a small number of examples is used to train the DT, resulting in very simple explanations for the decisions of the black-box model. Here, to avoid this effect, we modify the procedure by defining that at least 100 individuals generated by each GA compose the training set. The same procedure is used in LOREfs.

3.1 Impact of the GA parameters

Two shape datasets [21], $Jain$ and $Flame$, were used to evaluate the impact of G and p_m. These datasets were used because: i) they are two-dimensional (2 attributes), which allow visualizing the decision boundaries generated by the black-box and surrogate models, and ii) each one has 2 clusters with different shapes. $Jain$ has 373 instances and $Flame$ has 240 instances. The MLP is used as black-box model.

Results of LORE and LOREfs for different values of p_m (0.05, 0.10, 0.15, and 0.20) and G (10, 50, 100, and 150) are presented. When the mutation rate is changed, the number of generations is fixed ($G = 10$); when the number of generations is changed, the mutation rate is fixed ($p_m = 0.2$). In all experiments, the other parameters are set to the default values used in [10]: $p_c = 0.5$ and the population size for each GA is $N = 1,000$. In LOREfs, $\sigma = 1$ (see Section 3.2). In each experiment with a combination of GA parameters, the algorithms LORE and LOREfs run for five manually chosen instances of the dataset (instances to be explained).

The results of LORE are presented in Figures 1 and 2. Only results for changing p_m for the $Jain$ dataset and for changing G for the $Flame$ dataset are shown. Some observations can be made from these figures. There is a high concentration of artificial instances (individuals) in some regions of the classification space. The histograms on the top and right of each graph allow us to visualize better this concentration. Section 3.4 shows results (Table 4) where the diversity of the artificial datasets is quantitatively analyzed. The concentration of artificial instances in some regions is caused by the low diversity of the populations optimized by the GAs. The lack of diversity can be a problem when the GA fails to explore regions that allow to reproduce the decision boundaries of the black-box model. We can observe that the surrogate models fail to reproduce the decision boundaries close to the instance to be explained (represented by a red star) for some experiments.

Both number of generations and mutation rate have a great impact on the decision boundaries produced by the surrogate model. From Figure 1, it is possible to observe that lower mutation rates imply in lower diversity. Since the initial population of

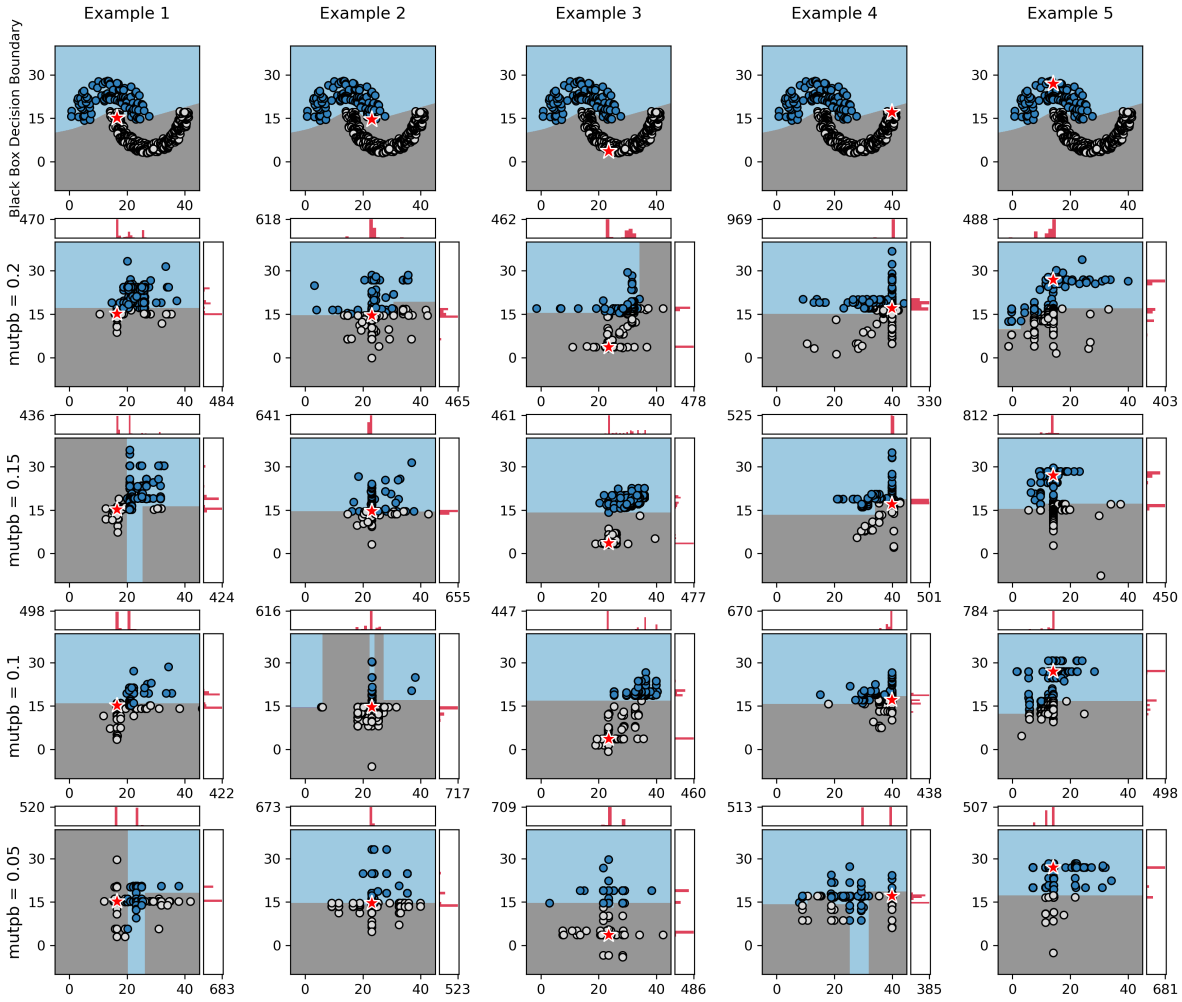

Figure 1: Decision boundaries generated by the black-box model (MLP) and by the surrogate model (DT) created by LORE for the $Jain$ Dataset. Each column represents the respective decision boundaries for one instance $\mathbf{x}$ (represented by the red star). The first row shows the decision boundaries generated by the black-box model. The $Jain$ dataset is also presented in the first row. Each row, after the first, shows the decision boundaries generated by a DT. LORE is run for each instance to be explained. Results for running LORE with different mutation rates for the GAs are presented. The artificial dataset generated by the GAs is also presented; this artificial dataset is used to infer the respective DT. The histograms on the top and right of each graph show the distribution of points along each axis.

the GAs have identical copies of the instance to be explained, the evolution along the generations depends mainly on the mutation operation; recombination has a small effect when the individuals are similar.

The results of LOREfs are presented in Figures 3 and 4. In these experiments, LOREfs produced decision boundaries for the surrogate model (DT) that are more similar to those of the black-box model. This can be explained by the higher population diversity when fitness sharing is adopted. The histograms indicate that more diverse artificial instances are generated by the GAs. Consequently, the decision boundaries close to the instance to be explained are more similar to those of the black-box model. We can observe that the parameters of the GAs also impact the decision boundaries. However, the impact is smaller when compared to LORE, especially regarding the number of generations. Thus, the GAs in LOREfs can be optimized for more generations, resulting in local decision boundaries of the DT more similar to those of the MLP. Figure 4 is particularly illustrative of the advantages of LOREfs. It shows the decision boundaries generated by the MLP (black-box model) and by the surrogate model (DT created by LOREfs) for Flame dataset and different number of generations. First, one can observe that for all experiments, the surrogate models are capable of reproducing the decision boundaries of the black-box model, particularly in the neighborhood of solution to be explained. Second, it shows that, unlike LORE (Figure 2), LOREfs produces decision boundaries that reproduces well the decision boundaries of the black-box model even when the number of generations increases. This occurs because of the diversity of the population is preserved, allowing to run the GAs for more generations, resulting in more optimized solutions.

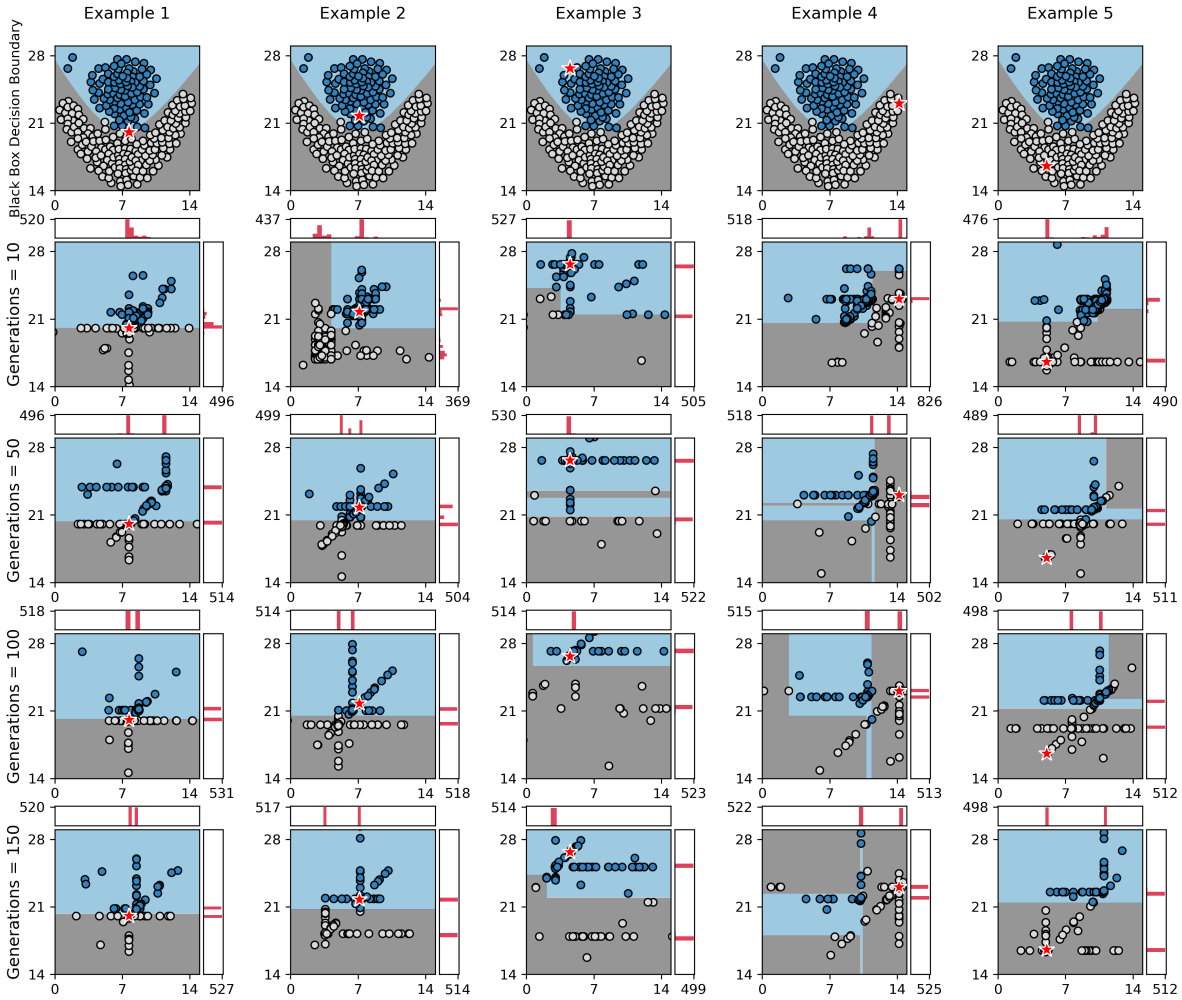

Figure 2: Decision boundaries generated by the black-box model (MLP) and by the DT created by LORE for the $Flame$ Dataset. Each row, after the first, shows the results for running LORE with different number of generations for the GAs.

3.2 Impact of Niche Size

We propose to empirically find the value of σ in LOREfs. Here, in addition to the experiments with two shape datasets ($Jain$ and $Flame$), experiments with two public datasets related to Medicine are performed. The two datasets are *Breast Cancer* and *Heart Disease*, both from the UCI Machine Learning Repository [22]. The Breast Cancer dataset has 699 instances and 10 attributes, while the Heart Disease dataset has 303 instances and 14 attributes. In the procedure to find σ, each dataset was randomly split in training (80%) and test sets (20%). The training set was used to train an MLP (black-box model), while the test set was used to explain the decision of the black-box model. Consequently, the datasets have a different number of instances (test set) to be explained: $Flame$ has 48 instances, $Jain$ has 75 instances, $HeartDisease$ has 61 instances, and $BreastCancer$ has 114 instances. For each instance $\mathbf{x}$ of the test set, a surrogate model is built and the averaged results are reported. The parameters of the GA are the same used in the original implementation of LORE: $p_m = 0.2$, $p_c = 0.5$, $G = 10$, and $N = 1,000$.

The comparison of LOREfs with different values of σ is made regarding the F1 score. The F1 score is computed by comparing the classification of the black-box (MLP) model b and the surrogate model (DT) c for a percentage of instances of the training set that are closest to instance $\mathbf{x}$. For example, when the percentage is 20%, the subset of instances to compute the F1 score is formed by the union of the 10% instances of the dataset that are closest to $\mathbf{x}$ and have the same label of $\mathbf{x}$ and 10% instances of the dataset that are closest to $\mathbf{x}$ but have the opposite label. This is done because we are interested in reproducing the decision boundaries that are close to the instance to be explained. The number of runs for each dataset is 20, one for each value of σ ranging from 1 to 20. The average F1 score is computed for all instances of the test set. The results are presented in Figure 5.

In most of the experiments, $\sigma = 1$ resulted in the best mean F1 score and in the minimum standard deviation. Even when the best F1 score is not obtained when $\sigma = 1$, the results are slightly worse. Anyway, the procedure presented here can be used to find good values of σ for different datasets. It is important to observe that the best values of σ can change due to different properties, e.g., distribution of instances in the classification space and maximum and minimum value of each attribute. In the experiments presented in the following sections, $\sigma = 1$.

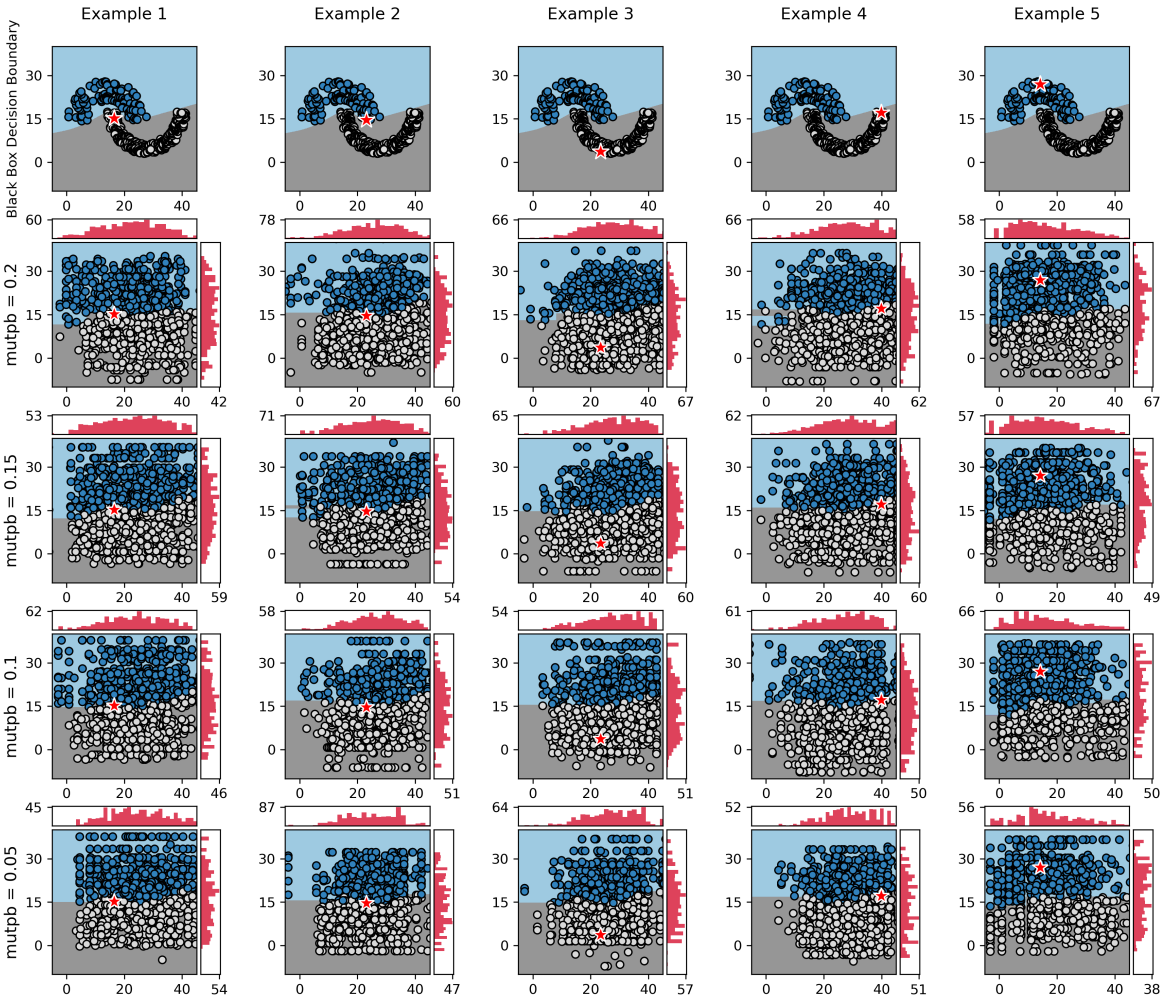

Figure 3: Decision boundaries generated by the black-box model (MLP) and by the decision tree created by LOREfs for the $Jain$ dataset and different mutation rates.

3.3 Comparing the classification produced by the black-box and surrogate models

Here, the experiment design is the same of the previous experiment (Section 3.2). Results of an RF as the black-box model are also presented. When training the MLP as the black-box model, the accuracy obtained was: 1 (Flame), 0.987 (Jain), 0.965 (Breast Cancer), 0.770 (Heart Disease). When training the RF as the black-box model, the accuracy obtained was: 0.937 (Flame), 0.920 (Jain), 0.965 (Breast Cancer), 0.787 (Heart Disease). In the experiments with LORE and LOREfs, different percentages of the dataset (training dataset) are considered for computing the F1 score. We also statistically compare the results of LORE and LOREfs. The results for the experiment with the MLP are presented in Table 1, while the results for the RF are presented in Table 2.

The proposed algorithm, LOREfs, presented significantly better results for F1 score in all cases for both MLP and RF experiments (tables 1 and 2). Therefore, LOREfs obtained a better average rank than LORE for all datasets. The better performance is explained by the higher diversity of the final populations of the GAs. Higher diversity resulted in decision boundaries around instance $\mathbf{x}$ more similar to those of the black-box model. The minimum and maximum time for explaining a decision of the black-box model in the experiments are presented in Table 3.

Examples for explaining instances black-box models for instances of Breast Cancer and Heart Disease datasets are respectively presented in figures 6 and 7. In these examples, the explanations r generated by LORE are very simple: only 1 feature is used to explain three of the decisions, while 3 features are used to explain one decision. The counterfactual rules Φ are very simple too: only 1 or 2 rule. On other hand, the explanation r and the counterfactual rules Φ produced by LOREfs are more complex. In the examples, 3, 5, or 6 features are used for explaining the decisions of the black-box models, indicating more complex decision regions of the surrogate model. Also, there are more counterfactual rules (2, 4, or 5). When compared to LORE, LOREfs generates local decision boundaries that are more similar to those produced by the black-box model. This is explained again by the higher diversity of the artificial datasets generated by the GAs with fitness sharing.

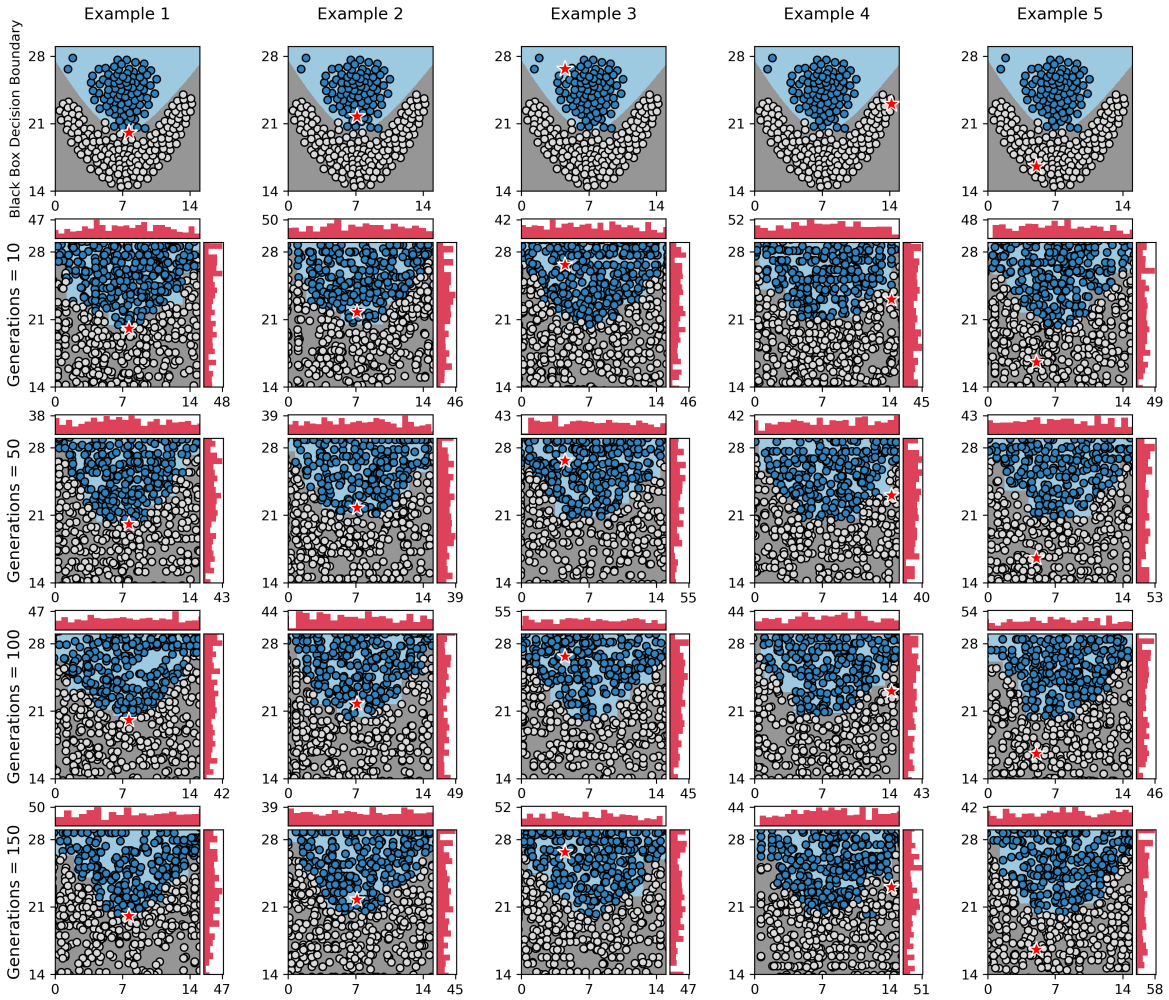

Figure 4: Decision boundaries generated by the black-box model (MLP) and by the DT created by LOREfs for the *Flame* dataset and different number of generations.

3.4 Diversity of the Artificial Dataset

In order to quantify the GA population diversity, we measured the mean distances between each pair of individuals generated by the GAs on each run of the experiments presented in Section 3.3; that is, after training the surrogate model (for test sets with 10%, 20% and 30% of the instances), distances between each pair of individuals generated by the GA are computed. All four datasets here have real features. Thus, the Euclidean distance is used as the distance metric. The mean and standard deviation of the Euclidean distance (for different instances to be explained) are presented in Table 4. LOREfs presented significantly better results for population diversity in all experiments. These results corroborate with previous results showed in Section 3.1.

4 Conclusions

In our experiments, we studied the impact of the number of generations and mutation rate on LORE's genetic algorithm (GA). We found that better results are achieved with a small number of generations and a high mutation rate. In a previous study [10], the authors suggested using GAs with 10 generations and a mutation rate of 0.2 in LORE to preserve population diversity. Instead, we propose using fitness sharing in LORE to generate more diverse artificial datasets, composed of individuals from the final GA populations. Our experimental results demonstrate that LORE with fitness sharing (LOREfs) produces artificial datasets with significantly higher diversity. As a result, the decision boundaries created by the surrogate model (DT) closely resemble the local decision boundaries generated by the black-box model. When compared to LORE, LOREfs achieves significantly better F1 scores in experiments with four classification datasets and two black-box models (MLP and RF).

In the experiments, we only evaluated datasets with real features. However, LOREfs can be adapted for problems involving categorical features in the future. LOREfs shows promise for datasets with a higher number of features compared to LORE. Investigating LOREfs' performance in problems with hundreds or thousands of features, as well as exploring non-binary classification problems, should be pursued in future research. Additionally, the impact of using different distance metrics to compute the sharing function should be investigated. Furthermore, the application of LOREfs in the classification of electroencephalogram signals for predicting stroke outcomes, and other medical and healthcare problems warrants exploration. Detecting and

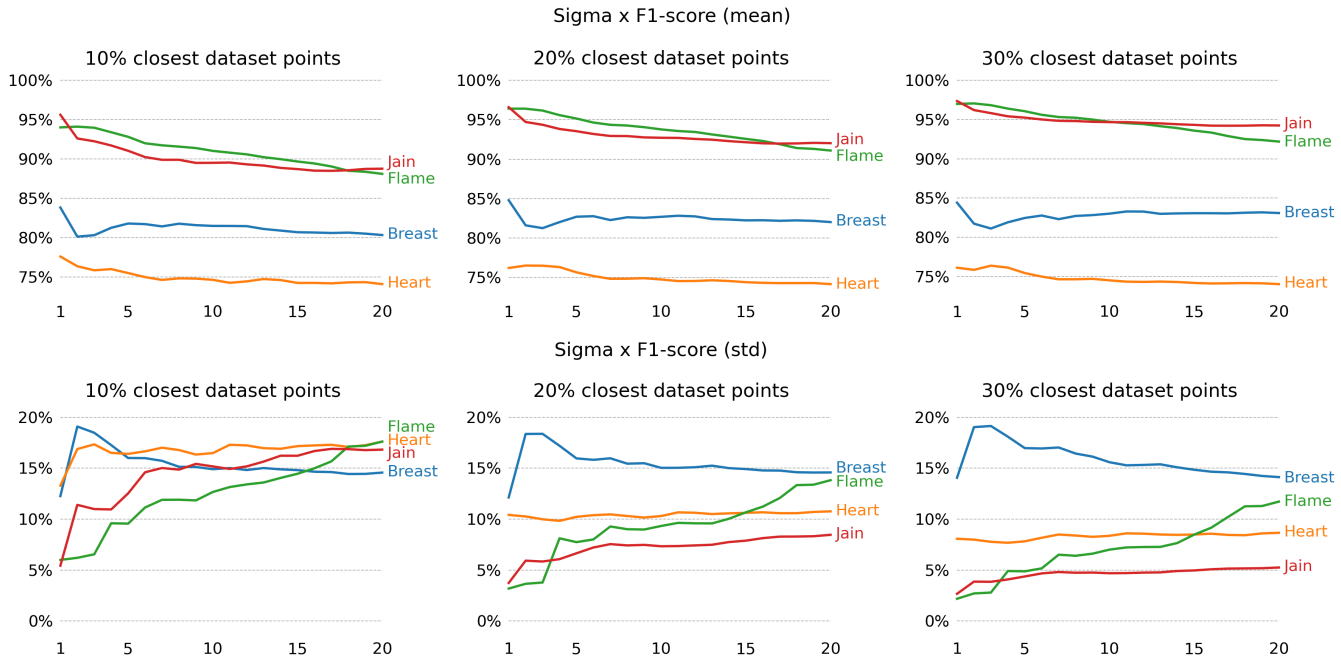

Figure 5: Mean and standard deviation (std) of F1 score for σ ranging from 1 to 20 for each dataset ($Jain$, $Flame$, $HeartDisease$ and $BreastCancer$) divided by percentages of training set instances closest to the instance to be explained.

Table 1: Comparison of LORE and LOREfs in the experiments with the MLP (black-box model). The mean and standard deviation of the F1 score are presented. The F1 score is computed comparing the classification of the black-box model b and surrogate model c. For explaining each instance $\mathbf{x}$ of the test set, a surrogate model is built. The F1 score is computed for models b and c applied to the following percentages of the dataset: 10%, 20%, and 30%. Instances of the dataset for each class that are closest to $\mathbf{x}$ are used. The symbols '$=$', '$+$', and '$-$' respectively indicate that the results of LOREfs are equal, better or worse than the results of LORE. The Wilcoxon signed rank test (non-parametric test), with $\alpha = 0.05$, is used to statistically compare the results. The letter s indicates that the $p - value$ is smaller than α.

	10% of the dataset		20% of the dataset		30% of the dataset	
Dataset	LORE	LOREfs	LORE	LOREfs	LORE	LOREfs
$Flame$	0.783±0.278	0.925±0.062(s+)	0.829±0.21	0.954±0.038(s+)	0.855±0.164	0.964±0.03(s+)
$Jain$	0.766±0.322	0.894±0.092(s+)	0.832±0.235	0.928±0.043(s+)	0.882±0.147	0.946±0.027(s+)
$BreastCancer$	0.736±0.243	0.859±0.142(s+)	0.717±0.246	0.865±0.136(s+)	0.707±0.249	0.855±0.138(s+)
$HeartDisease$	0.571±0.303	0.704±0.192(s+)	0.571±0.293	0.707±0.158(s+)	0.565±0.289	0.714±0.128(s+)

preserving borderline instances, which are crucial for defining decision boundaries in DT, can be accomplished by employing Tomek Links [23–25] as a potential avenue for future work. Considering two instances $\mathbf{z}_i$ and $\mathbf{z}_j$ from different classes, and a distance function $d(\mathbf{z}_i, \mathbf{z}_j)$ between $\mathbf{z}_i$ and $\mathbf{z}_j$, we say that the pair $(\mathbf{z}_i, \mathbf{z}_j)$ forms a Tomek link iff there is no such instance $\mathbf{z}_k$ satisfying $d(\mathbf{z}_k, \mathbf{z}_i) < d(\mathbf{z}_i, \mathbf{z}_j)$ or $d(\mathbf{z}_k, \mathbf{z}_j) < d(\mathbf{z}_i, \mathbf{z}_j)$. Therefore, if two instances form a Tomek link, then both are borderline instances. The use of efficient recombination [26] and mutation operators is also an additional future work, as well as changing other strategies used in LORE, e.g., the population initialization procedure.

Acknowledgments: This work was partially supported in Brazil by São Paulo Research Foundation - FAPESP (under grant #2021/09720-2), National Council for Scientific and Technological Development - CNPq (under grant #306689/2021-9) and Center for Artificial Intelligence - C4AI (supported by FAPESP, under grant #2019/07665-4, and IBM Corporation).

REFERENCES

[1] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter and L. Kagal. "Explaining Explanations: An Overview of Interpretability of Machine Learning". In *2018 IEEE 5th Int. Conf. on Data Science and Advanced Analytics (DSAA)*, pp. 80–89, 2018.

[2] F. Doshi-Velez and B. Kim. "Towards a rigorous science of interpretable machine learning". *arXiv preprint arXiv:1702.08608*, 2017.

[3] D. V. Carvalho, E. M. Pereira and J. S. Cardoso. "Machine Learning Interpretability: A Survey on Methods and Metrics". *Electronics*, vol. 8, no. 8, 2019.

Table 2: Comparison of LORE and LOREfs in the experiments with the RF (black-box model).

	10% of the dataset		20% of the dataset		30% of the dataset	
Dataset	LORE	LOREfs	LORE	LOREfs	LORE	LOREfs
$Flame$	0.810±0.274	0.932±0.08(s+)	0.854±0.226	0.955±0.044(s+)	0.873±0.177	0.965±0.031(s+)
$Jain$	0.834±0.253	0.969±0.062(s+)	0.878±0.196	0.978±0.038(s+)	0.901±0.176	0.983±0.025(s+)
$BreastCancer$	0.744±0.200	0.805±0.162(s+)	0.762±0.179	0.838±0.138(s+)	0.775±0.173	0.863±0.106(s+)
$HeartDisease$	0.542±0.305	0.697±0.169(s+)	0.539±0.278	0.747±0.096(s+)	0.546±0.272	0.750±0.088(s+)

Table 3: Minimum - maximum time (in seconds) for explaining a decision of the black-box model in the experiments. For each decision, the GAs are run, the DT is generated, and the explanation and counterfactual rules are produced.

	MLP		RF	
Dataset	LORE	LOREfs	LORE	LOREfs
$Flame$	3-4	12-14	13-17	27-31
$Jain$	3-4	12-17	14-17	27-33
$BreastCancer$	5-7	14-16	15-16	34-48
$HeartDisease$	3-4	11-14	12-14	27-30

[4] P. Lakhani, A. B. Prater, R. K. Hutson, K. P. Andriole, K. J. Dreyer, J. Morey, L. M. Prevedello, T. J. Clark, J. R. Geis, J. N. Itri *et al.*. "Machine learning in radiology: applications beyond image interpretation". *Journal of the American College of Radiology*, vol. 15, no. 2, pp. 350–359, 2018.

[5] R. ElShawi, Y. Sherif, M. Al-Mallah and S. Sakr. "Interpretability in healthcare: A comparative study of local machine learning interpretability techniques". *Computational Intelligence*, vol. 37, no. 4, pp. 1633–1650, 2021.

[6] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya *et al.*. "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning". *arXiv preprint arXiv:1711.05225*, 2017.

[7] B. Custers, A. M. Sears, F. Dechesne, I. Georgieva, T. Tani and S. van der Hof. *EU personal data protection in policy and practice*. Springer, Switzerland, 2019.

[8] S. M. Lundberg and S.-I. Lee. "A Unified Approach to Interpreting Model Predictions". In *Proc. of the 31st Int. Conf. on Neural Information Processing Systems*, NIPS'17, p. 4768–4777, 2017.

[9] M. T. Ribeiro, S. Singh and C. Guestrin. "" Why should i trust you?" Explaining the predictions of any classifier". In *Proc. of the 22nd ACM SIGKDD Int. Conf. on knowledge discovery and data mining*, pp. 1135–1144, 2016.

[10] 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*, vol. 34, no. 6, pp. 14–23, 2019.

[11] 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 (CSUR)*, vol. 51, no. 5, pp. 1–42, 2018.

[12] D. A. Santos, J. A. Baranauskas and R. Tinós. "Use of Fitness Sharing in the Local Rule-Based Explanations Method". In *Proc. of the 2021 IEEE Latin American Conference on Computational Intelligence (LA-CCI)*, pp. 1–6, 2021.

[13] A. E. Eiben and J. E. Smith. *Introduction to evolutionary computing*, volume 53. Springer, Germany, 2003.

[14] L. Breiman, J. Friedman, R. Olshen and C. Stone. *Classification and Regression Trees*. Wadsworth & Books, Pacific Grove, CA, 1984.

[15] J. R. Quinlan. *C4.5: Programs for Machine Learning*. MK, San Francisco, CA, 1993.

[16] P. Linardatos, V. Papastefanopoulos and S. Kotsiantis. "Explainable AI: A Review of Machine Learning Interpretability Methods". *Entropy*, vol. 23, no. 1, 2021.

[17] N. Burkart and M. F. Huber. "A Survey on the Explainability of Supervised Machine Learning". *Journal of Artificial Intelligence Research*, vol. 70, pp. 1–74, 2021.

[18] M. Mitchell. *An introduction to genetic algorithms*. MIT press, USA, 1998.

[19] B. Sareni and L. Krahenbuhl. "Fitness sharing and niching methods revisited". *IEEE Transactions on Evolutionary Computation*, vol. 2, no. 3, pp. 97–106, 1998.

[20] G. Singh and K. Deb. "Comparison of multi-modal optimization algorithms based on evolutionary algorithms". In *Proc. of the Genetic and Evolutionary Computation Conf.*, pp. 1305–1312, 2006.

Breast Cancer Dataset

Example 1

LORE

$r = (\{radius\_worst : > 12.014649,$
$perimeter\_mean : > 82.104949,$
$symmetry\_worst > 0.387108\} \rightarrow diagnosis = Benign)$
$\Phi = (\{radius\_worst : <= 12.014649\},$
$\{perimeter\_mean : <= 82.104949\})$

LOREfs

$r = (\{perimeter\_mean : > 85.201299,$
$perimeter\_worst : <= 106.167881,$
$texture\_se : <= 1.943543,$
$radius\_mean : > 11.824107,$
$area\_mean : <= 615.259162\} \rightarrow diagnosis = Benign)$
$\Phi = (\{perimeter\_mean : <= 85.201299, perimeter\_worst : > 99.62338\},$
$\{perimeter\_mean : <= 67.222484, perimeter\_worst : <= 99.62338\},$
$\{radius\_mean : <= 3.075754, compactness\_se : > 0.028771\},$
$\{area\_mean : > 615.259162, area\_se : > 110.491737\})$

Example 2

LORE

$r = (\{perimeter\_mean : <= 149.696084\} \rightarrow diagnosis = Malign)$
$\Phi = (\{perimeter\_mean : > 149.696084\})$

LOREfs

$r = (\{area\_se : > 138.741714,$
$area\_mean : > 599.103741,$
$perimeter\_worst : > 92.434470\} \rightarrow diagnosis = Malign)$
$\Phi = (\{area\_mean : 486.895625 < area\_mean <= 599.103741,$
$perimeter\_mean : > 89.620855\},$
$\{perimeter\_worst : <= 92.43447, perimeter\_mean : > 138.909976,$
$fractal\_dimension\_mean : <= 0.060839\})$

Figure 6: Explaining decisions of the black-box model for two instances of the Breast Cancer dataset. The black-box model here is an MLP. The local explanation for the decision of the black-box model on instance $\mathbf{x}$ is given by a subset of logical rules r extracted from the decision tree c. The subset of counterfactual rules Φ indicates conditions that change the label of $\mathbf{x}$.

Heart Disease Dataset

Example 1

LORE

$r = (\{cp : <= 2.00\} \rightarrow diagnosis = \text{" } > 50\% \ diameter \ narrowing\text{"})$
$\Phi = (\{cp : > 2.00\})$

LOREfs

$r = (\{thal : > 5.069187,$
$cp : <= 2.625233,$
$thalach : <= 159.433245,$
$fbs : -0.16829 < fbs <= 0.21,$
$ca : <= 1.597054,$
$chol : > 303.517165\} \rightarrow diagnosis = \text{" } > 50\% \ diameter \ narrowing\text{"})$
$\Phi = (\{thal : <= 5.069187, thalach : > 144.251838\},$
$\{fbs : 0.21 < fbs <= 0.234588\},$
$\{fbs : > 0.234588, oldpeak : > 1.870153\},$
$\{chol : 303.145116 < chol <= 303.517165\},$
$\{fbs : <= -0.16829, oldpeak : > 2.202191\})$

Example 2

LORE

$r = (\{oldpeak : <= 0.984268\} \rightarrow diagnosis = \text{" } < 50\% \ diameter \ narrowing\text{"})$
$\Phi = (\{oldpeak : > 0.984268\})$

LOREfs

$r = (\{thal : <= 4.504506,$
$oldpeak : <= 1.691220,$
$thalach : > 152.417628,$
$ca : <= 1.415176,$
$exang : <= 0.716649\} \rightarrow diagnosis = \text{" } < 50\% \ diameter \ narrowing\text{"})$
$\Phi = (\{exang : 0.716649 < exang <= 0.798229\},$
$\{thalach : <= 152.417628, exang : > 0.705035\})$

Figure 7: Explaining decisions of the black-box model for two instances of the Heart Disease dataset. The black-box model here is an RF.

[21] P. Fränti and S. Sieranoja. "K-means properties on six clustering benchmark datasets", 2018.

[22] D. Dua and C. Graff. "UCI Machine Learning Repository", 2017.

[23] I. Tomek. "An Experiment with the Edited Nearest-Neighbor Rule". *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-6, no. 6, pp. 448–452, 1976.

[24] M. Kubat and S. Matwin. "Addressing the Curse of Imbalanced Training Sets: One-Sided Selection". In *Proc. of the Fourteenth Int. Conf. on Machine Learning (ICML 1997)*, pp. 179–186, 1997.

[25] H. Sain and S. W. Purnami. "Combine Sampling Support Vector Machine for Imbalanced Data Classification". *Procedia Computer Science*, vol. 72, pp. 59–66, 2015. The Third Information Systems Int. Conf. 2015.

[26] R. Tinós, D. Whitley, F. Chicano and G. Ochoa. "Partition crossover for continuous optimization: ePX". In *Proc. of the Genetic and Evolutionary Computation Conf.*, pp. 627–635, 2021.

Table 4: Comparison of the population diversity in experiments with MLP and RF as black-box models. The mean and standard deviation of the Euclidean distance between all individuals generated by the GA are presented.

	MLP		RF	
Dataset	LORE	LOREfs	LORE	LOREfs
$Flame$	3.057±1.222	8.452±0.300(s+)	2.726±1.122	8.443±0.307(s+)
$Jain$	7.814±3.476	17.919±1.18(s+)	7.655±3.483	17.793±0.988(s+)
$BreastCancer$	131.609±159.83	364.751±203.236(s+)	164.384±155.722	444.99±208.967(s+)
$HeartDisease$	11.237±12.885	28.413±13.071(s+)	7.11±11.093	23.497±5.09(s+)