Artificial Intelligence 294 (2021) 103457

Contents lists available at ScienceDirect

Artificial Intelligence

www.elsevier.com/locate/artint

GLocalX - From Local to Global Explanations of Black Box AI Models ☆

Mattia Setzu [a,*], Riccardo Guidotti [a], Anna Monreale [a], Franco Turini [a], Dino Pedreschi [a], Fosca Giannotti [b]

[a] University of Pisa, Largo B. Pontecorvo, Pisa, Italy
[b] ISTI-CNR, Via G. Moruzzi, Pisa, Italy

ARTICLE INFO

Article history:
Received 29 February 2020
Received in revised form 24 November 2020
Accepted 20 January 2021
Available online 22 January 2021

Keywords:
Explainable AI
Global explanation
Local explanations
Interpretable models
Open the black box

ABSTRACT

Artificial Intelligence (AI) has come to prominence as one of the major components of our society, with applications in most aspects of our lives. In this field, complex and highly nonlinear machine learning models such as ensemble models, deep neural networks, and Support Vector Machines have consistently shown remarkable accuracy in solving complex tasks. Although accurate, AI models often are "black boxes" which we are not able to understand. Relying on these models has a multifaceted impact and raises significant concerns about their transparency. Applications in sensitive and critical domains are a strong motivational factor in trying to understand the behavior of black boxes. We propose to address this issue by providing an interpretable layer on top of black box models by aggregating *"local"* explanations. We present GLocalX, a *"local-first"* model agnostic explanation method. Starting from local explanations expressed in form of local decision rules, GLocalX iteratively generalizes them into global explanations by hierarchically aggregating them. Our goal is to learn accurate yet simple interpretable models to emulate the given black box, and, if possible, replace it entirely. We validate GLocalX in a set of experiments in standard and constrained settings with limited or no access to either data or local explanations. Experiments show that GLocalX is able to accurately emulate several models with simple and small models, reaching state-of-the-art performance against natively global solutions. Our findings show how it is often possible to achieve a high level of both accuracy and comprehensibility of classification models, even in complex domains with high-dimensional data, without necessarily trading one property for the other. This is a key requirement for a trustworthy AI, necessary for adoption in high-stakes decision making applications.

© 2021 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

In the last decade, Artificial Intelligence (AI) decision systems have been widely used in a plethora of applications such as credit score, insurance risk, and health monitoring, in which accuracy is of the utmost importance [1]. Complex nonlinear

☆ This paper is part of the Special Issue on Explainable AI.

* Corresponding author.
E-mail addresses: mattia.setzu@phd.unipi.it (M. Setzu), riccardo.guidotti@unipi.it (R. Guidotti), anna.monreale@unipi.it (A. Monreale), turini@di.unipi.it (F. Turini), dino.pedreschi@unipi.it (D. Pedreschi), fosca.giannotti@isti.cnr.it (F. Giannotti).

https://doi.org/10.1016/j.artint.2021.103457
0004-3702/© 2021 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

machine learning models such as ensemble models, deep neural networks (DNN) and Support Vector Machines (SVM) have shown remarkable performance in these tasks, and have made their way into a large number of systems [2]. Unfortunately, their state-of-the-art performance comes at the cost of a clear interpretation of their inner workings [3]. These "black box" models have an opaque, hidden internal structure that humans do not understand [4]. Relying on black box systems is becoming increasingly risky both for their lack of transparency and the systematic bias they have shown in real-world scenarios [5]. The lack of proper explanations also has ethical implications, legally reported in the *General Data Protection Regulation* (GDPR), approved by the European Parliament in May 2018. The GDPR provides restrictions and guidelines for automated black box decision-making processes which, for the first time, introduces a "right to explanation" on the decisions of the system. More specifically, the GDPR introduces a right to meaningful explanations when one is subject to automated AI systems [6–8].

Given the great interest on the topic [9,2,4,10], several works in the literature try to explain opaque models with one of two goals: either providing instance explanations for a given decision by using a *local* approach [11–14], or providing *global* explanations able to describe the overall logic of the black box [15,16]. These approaches differ both on their task and their use case, with local approaches having the upper hand in several scenarios. As stated, local approaches [11,33] provide explanations of single instances and are beneficial to a plethora of users with different needs and resources at their disposal. Model developers, who have access both to data and black box, can directly inspect the model; analysts can audit the model on a small sample of instances; users can retrieve an explanation on a decision that involves them directly, gain trust in the model and, possibly, an actionable recourse to follow up the decision. On the other hand, global approaches have stricter requirements. Users may require access both to the model and the data used to train and validate the model [17,15,18]. On the other hand, most local approaches are *model agnostic* [11,33], i.e., they do not assume knowledge of the black box or its internals, while global approaches may depend on the black box model [4]. In spite of this additional layer of abstraction, local approaches have repeatedly shown competitive performances similar to global approaches, mainly due to the smaller set of decisions to explain. Finally, local approaches inherently enjoy a high degree of explanation independence, as they provide different views on data and are able to grasp qualitatively different patterns, providing different explanations for similar instances when confounding factors are at play [19,20]. This behavior is reminiscent of *bagging* estimators, as local approaches are a peculiar bagging in which each model is fit on one sample. Feature transformations are also typical of local approaches [11,13] as they are for bagging approaches.

Following these premises, we posit that leveraging *local* approaches for tackling *global* tasks can yield the benefits of the former and overcome the constraints of the latter. In this paper, we decline the local-global dichotomy in favor of a *Local to Global* interpretation and propose a formal definition of this problem and an algorithm to solve it. Three assumptions [21] underpin our approach:

1. **Logical Explanability.** We believe that the cognitive vehicle for offering explanations should be close to the language of reasoning, that is *logic*. For this reason we adopt rule-based local explanations [12,22].
2. **Local Explainability.** While a black box can be arbitrarily complex, we assume that in the neighborhood of each specific instance there is a high chance that the decision boundary of the black box is simple enough to be accurately approximated by an explanation [11].
3. **Explanation Composition.** We assume that similar instances admit similar explanations [23,24] and that similar explanations are likely to be composed together into slightly more general ones.

We support these hypotheses with GLocalX (GLObal to loCAL eXplainer), a model-agnostic *"local-first"* explanation algorithm. GLocalX is based on the idea of deriving global explanations by inference on a set of logical rules representing local explanations. GLocalX aggregates local explanations expressed in form of logical rules into a global explanation by iteratively "merging" the rules in a growing hierarchical structure while accounting for both fidelity, i.e., accuracy in emulating black box predictions, and complexity of the rules. The merge procedure estimates a distance between explanations and yields a set of sorted candidate pairs to merge. Then, the pair with minimum distance satisfying constraints on both fidelity and complexity is processed to guarantee generalization and the updated explanation replaces the selected pairs. Constraint satisfaction is ensured on each merge. This guarantees high fidelity and low complexity on the final explanation yielded by GLocalX.

We showcase the *Local to Global* formulation in two constrained scenarios typical of real-world use cases: in the former, we consider as available input a restricted number of local explanation rules; while in the latter, we consider no data available for the global explanation construction. These settings can occur when the model is proprietary or data is inaccessible due to privacy concerns. Empirical results over different black box models and datasets indicate that GLocalX achieves both a high *fidelity* and a low *complexity* of the rule set representing the model explanation. Compared to transparent models that either optimize model complexity or fidelity, but not both, GLocalX reaches simultaneously high fidelity and low complexity. The high accuracy in prediction tasks also suggests that GLocalX might be used directly as a transparent model to replace global classifiers adopted in AI systems.

The rest of the paper is organized as follows. Section 2 discusses related work. Section 3 presents the local to global problem formulation and the idea adopted for solving it. Section 4 describes the proposed hierarchical approach for the global explanation. Experimental results are presented in Section 5. Finally, Section 6 concludes the paper and discusses new research directions.

2. Related work

To the best of our knowledge, no global frameworks merging local explanations are present in the literature. However, related work comprises of a set of both *global* and *local* explanation methods [4]. In addition, we can distinguish between methods that explain the black box *post-hoc*, and methods that are, on the other hand, *explainable by design* [21].

Explainable by design algorithms directly solve the classification problem [4] and yield an interpretable and global model. The most well known interpretable by design families of models comprise of the *Decision Tree* [25] (DT) and the rule-based classifiers (RBC), such as CPAR [26], *Decision Sets* [27,28] and CORELS [29]. CPAR [26] combines the positive aspects of both associative classification and traditional rule-based classification, greedily generating a small set of rules directly from training data. *Decision Sets* [28] are "collections of if-then rules that can be considered in any order". Given a set of rules, a set of functions is used to identify a subset which enjoys low internal conflict and several other desirable properties such as coverage, shortness, and accuracy. CORELS [29] also generates compact explanations with a optimality guarantees by using a discrete optimization technique for building sorted *rule lists* over categorical feature spaces. We observe different approaches to tackling the complexity of the interpretable global model, with some families ignoring it [26], partially acknowledging it [30] or fully including it [27] in the construction phase.

Concerning post-hoc black box explanation, several foundational works rely on training explainable by design models leveraging queries to black box models. An example of this family is the *Conj Rules* approach [31], which interprets a neural network by framing rule extraction as a learning problem. Along the same lines is TREPAN [15], a refinement of the previous approach based on a decision tree specifically built for explaining the behavior of a black box. The original training data and a randomized extension of it are labeled by the black box and used as training data for an interpretable model. The model learns a Decision Tree that maximizes both the gain ratio [32] and the current model fidelity to the black box. An advantage of TREPAN with respect to common tree classifiers [25] is that by enriching the dataset all splits are performed on a considerable amount of data. Another approach using a single tree approximation as an interpretable explanation of the global model is INTREES [16]. It extracts, measures, prunes and selects the final explanation rules from tree ensembles, and calculates frequent variable interactions. We underline that the INTREES approach is model-dependent and cannot be generalized for explaining every black box, while in TREPAN we can plug in any query-able black box, making TREPAN a more powerful and flexible algorithm.

More recently, we are observing a shift of paradigm towards *local* approaches for explaining the decision of black box classifiers for a single instance. These approaches assume that complex black box models such as deep neural networks implement an overall global logic too complex to be explained and/or understood. Conversely, they assume that local predictions of a single instance can be explained. We call this assumption of explainability in a local neighborhood the *local explainability* assumption. LIME [11] is the first attempt to derive a local explanation, as it relies on instances randomly generated in the neighborhood of the instance to be explained. The authors propose a feature importance framework in which instances are mapped to a simpler interpretable space on which a *linear model* is used to compute an importance score for each interpretable feature. Interpretable features are then mapped back to the original feature space to provide an explanation. LIME also provides a global feature importance tool to assess the feature relevance to discriminate in classification tasks, LIME-SP. Feature importance is the target of another local explainability model, SHAP [33], which frames feature importance as a collaborative game in which features are rewarded according to their contribution to the black box prediction. The game is framed in a formal game theory setting in which features approximate the provably unique Shapley values [34]. Based on the neighborhood generation premise is also LORE [13], which populates neighborhoods via genetic programming, optimizing both for the neighborhood distance and its label distribution. As in LIME, the neighborhood is then used to train an explainable model from which an explanation is extracted. LORE employs Decision Trees as explainable models, hence, other than returning a rule as an explanation, it is able to generate a set of counterfactual explanations, i.e., a set of rules similar to the one returned, but with a different outcome. This feature is particularly valuable in actionable settings, in which the model user may understand what changes to apply to data to comply with the black box predictions. The authors of LIME also propose ANCHORS [12], an algorithm that generates explanations in the form of decision rules by iteratively guessing premises and optimizing their precision. An anchor explanation consists of a minimal set of premises that guarantees a baseline of accuracy even when new premises are added. To quote the authors, an anchor "is a rule [..] such that changes to the rest of the feature values of the instance do not matter." Local explanation methods have repeatedly shown high accuracy, seldom outperforming global models. This, jointly with the *local explainability* assumption, prompts us to ask whether we can leverage *local* models to learn *global* ones while preserving their high fidelity.

3. Local to global explanation problem

In the following we introduce basic notations of classification on tabular data and we define the notion of *explanation* and the *local to global explanation problem* for which we propose a solution.

A classifier is a function $f : \mathcal{X}^{(m)} \rightarrow \mathcal{Y}$ which maps data instances (tuples) x from a feature space $\mathcal{X}^{(m)}$ with m input features to a decision y in a label space $\mathcal{Y}$. We write $f(x) = y$ to denote the decision y given by f, and $f(X) = Y$ as a shorthand for $\langle f(x) \mid x \in X \rangle$. We assume that any classifier can be queried at will. Here we restrict to binary classification but the formulation and the solution can be easily extended to multi-class and multi-label problems. An instance x consists of a set of m attribute-value pairs (a_i, v_i), where a_i is a categorical or continuous feature and v_i is a value from the

domain of a_i. We denote with b a *black box* classifier whose internals are either unknown to the observer or known but uninterpretable by humans. Examples include deep neural networks, SVMs, and ensemble classifiers like Random Forest and AdaBoost classifiers [32,2,4].

As *explanation* e we consider a decision rule r, i.e., $e = \langle r = P \rightarrow y \rangle$. The decision rule r describes the reason for the decision $y = b(x)$, $P = \{p_1, \ldots, p_s\}$ is a set of *premises* in conjunctive form, and y is the rule *outcome*. As an example, let us consider the following explanation for a loan request for a user $x = \{(\text{age}{=}22), (\text{job} = \textit{unemployed}), (\text{amount}{=}\ 10\text{k}), (\text{car} = \text{no})\}$:

$$e = \langle r = \{\text{age} \geq 25,\ \text{job} = \text{unemployed},\ \text{amount} \leq 10\text{k}\} \rightarrow \text{deny} \rangle$$

We name *explanation theory* $E = \{e_1, \ldots, e_n\}$ a set of explanations, i.e., a set of decision rules. We indicate with $\mathbb{E} = \{E_1, \ldots, E_N\}$ a set of explanation theories. We account for logical explanation theories as the explanations are expressed in the form of logical rules.

According to [4], the *local explanation problem* consists in retrieving an explanation which describes the reason behind a decision taken by a black box model b for a single instance x (local). On the other hand, the *global explanation problem* consists in finding the reasons for the classification for any instance in X taken by a black box model b. In this setting, the *local to global problem* consists in exploiting a set of local explanations, describing the reasons behind single (local) decisions, to understand the overall (global) logic of an opaque classifier used in an AI system. Formally, we define the problem as follows:

Definition 1 *(Local to Global Explanation Problem).* Let $e_1, \ldots, e_n \in \mathcal{E}$ be a set of local explanations for a black box classifier b defined in a human-interpretable domain $\mathcal{E}$. The *local to global explanation problem* consists in finding a function g yielding an *explanation theory* $E = g(e_1, \ldots, e_n) \subset \mathcal{E}$, such that E describes the overall logic according to which b makes decisions.

In order to solve the local to global explanation problem we need to generalize the local explanations, as they are accurate and faithful *locally* but not *globally*. Given a black box b adopted in an AI system, and a set of instances $X_{le} = \{x_1, \ldots, x_n\}$ explained locally, and their local explanations $\{e_1, \ldots, e_n\}$, we aim to solve the problem by deriving an explanation theory $E = \{e'_1, \ldots, e'_k\}$ by refining with an aggregation function g the local explanations into an explanation theory emulating the global decision logic of the black box b. Thus, the human-interpretable domain $\mathcal{E}$ consists in a set of logical decision rules.

4. Local to global hierarchy of explanation theories

GLocalX (GLObal to loCAL eXplainer) is an explanation method that hierarchically merges local explanations into a global explanation theory. In particular, GLocalX takes as input a set of local explanations in form of explanation theories $\mathbb{E} = \{E_1, \ldots, E_n\}$ where each theory $E_i = \{e_i\}$ is formed by a single explanation, i.e., $|E_i| = 1\ \forall E_i \in \mathbb{E}$. GLocalX iteratively *merges* the explanation theories and finally returns an explanation theory $E = \{e'_1, \ldots, e'_k\}$ which emulates the global behavior of the black box b simultaneously maintaining the overall model simple and interpretable.

At each iteration, GLocalX merges the *closest* pair of explanation theories E_i, E_j by using a notion of *similarity* between logical theories. The pairs are filtered out according to merge quality criterion: if no pair satisfying the criterion is found, GLocalX halts prematurely without building the full hierarchy. The resulting hierarchy of explanation theories can be represented by using a tree-like diagram called *dendrogram* [35]. There are two key elements in the GLocalX approach: *(i) similarity search*, which allows to select which theories to merge and refine, and *(ii)* a *merge function*, which allows to refine the explanations. We explore our choices in Section 4.1 and Section 4.2, respectively. Finally Section 4.3 describes how to use for a classification task an explanation theory capturing the global behavior of the black box model.

GLocalX is detailed in Algorithm 1. Given a set of explanation theories E, a pairwise similarity function and a quality criterion, we sort logical theories by similarity in a queue $\mathbb{Q}$[1] (line 3). Then, we sample a batch of data to merge the candidate theories (line 5). Using batches instead of the whole training dataset favors diverse merges, as the merge procedure has different behaviors according to the data at hand. In the merge loop, we pop the queue to find the most similar pair of theories whose merge satisfies a quality criterion (line 7) and, we run the merge operation (line 8), and if the merge is advantageous (line 9), the merged theory is kept.[2] As a quality criterion we have selected the Bayesian Information Criterion (BIC) [36], as it rewards models for their simplicity and accuracy. BIC has been successfully adopted in various techniques, i.e., clustering, adopting bisecting hierarchical refinement of the model [37,38]. After a successful merge we replace the two mergees E_i, E_j with the merged theory E_{i+j} (line 13). If no advantageous merge is found, GLocalX halts. This process is iterated until no more merges are possible (line 14). Finally, explanations with low fidelity are filtered out to reduce the output size (line 15): we use a parameter α to indicate this per-class trimming threshold. Specifically, we select the top-α explanations by fidelity, $\lceil \alpha^{-1} \rceil$ for positive and negative class, respectively. In addition to this, we introduce α_q, a relative trimming criterion discarding rules with fidelities under the α_qth fidelity percentile. Trimming explanations with low fidelity allows us to retain only the best explanations to provide in output.

[1] $\mathbb{Q}$ is the set of sorted candidates theories.

[2] E_{i+j} indicates the merged theories and does not refer to the sum of the indexes.

Algorithm 1 GLocalX($\mathbb{E}, \alpha$).

Input: $\mathbb{E}$ explanation theories, α filter threshold
Output: E explanation theory

1: $\mathrm{E} \leftarrow \emptyset$
2: **repeat**
3: $\mathbb{Q} \leftarrow$ SORT($\mathbb{E}$) ▷ sort pairs of theories by similarity
4: *merged* $\leftarrow$ False
5: $X' \leftarrow$ batch(X)
6: **while** $\neg$ *merged* $\wedge$ $\mathbb{Q} \neq \emptyset$ **do**
7: $E_i, E_j \leftarrow$ POP($\mathbb{Q}$) ▷ select most similar theories
8: $E_{i+j} \leftarrow$ MERGE(E_i, E_j, X') ▷ merge theories
9: **if** BIC(E_{i+j}) $\leq$ BIC($E_i \cup E_j$) **then** ▷ verify improvement
10: *merged* $\leftarrow$ True
11: **break**
12: **if** *merged* **then** ▷ merge occurred
13: $\mathbb{E} \leftarrow$ UPDATE(E_i, E_j, E_{i+j}) ▷ update hierarchy
14: **until** $|\ \mathrm{E}\ | > 1 \wedge$ *merged* ▷ until the merge is successful
15: $E \leftarrow$ FILTER(E, α) ▷ Filter final theory
16: **return** E

4.1. Finding similar theories

Selecting pairs of theories to merge requires the definition of a pairwise similarity function on logical explanation theories (line 2, Algorithm 1). To this aim, we define the similarity of two theories E_1, E_2 as the *Jaccard similarity* [32] of their *coverage* on a given instance set X:

$$similarity_X(E_i, E_j) = \frac{|coverage(E_i, X) \cap coverage(E_j, X)|}{|coverage(E_i, X) \cup coverage(E_j, X)|}.$$

An explanation $e = \langle r = P \rightarrow y \rangle$ *covers* an instance x if the premise P of r is satisfied by x. We extend the notion of coverage to explanation theories by saying that an explanation theory E *covers* an instance x if there is at least an explanation $e \in E$ covering x. $coverage(E, X)$ returns the set of records in X covered by the set of explanations in E, i.e., $coverage(E, X) = \{x \in X \mid \exists\, e \in E.\ e \text{ covers } x\}$. Conversely, $covered(x, E)$ returns the set of explanations in E with coverage on x, i.e., $covered(x, E) = \{e \in E \mid x \in coverage(e, \{x\})\}$. As for $coverage$, we extend this notion to sets of records by saying that a record x is *covered* by an explanation theory E if there is at least an explanation $e \in E$ that *covers* x.

The larger is the shared coverage of E_i and E_j on X, the more similar the two logical explanation theories are. Coverage similarity is a two-faceted similarity measure that captures *(i)* the premise similarity and *(ii)* the coverage similarity. The former is straightforward: rules with similar premises will have similar coverage. The latter balances the premise similarity to avoid that rules with similar premises but low coverage sway the similarity score.

4.2. Merging explanation theories

The merge function allows GLocalX to *generalize* a set of explanation theories while balancing fidelity and complexity through approximate logical entailment. As detailed in the following, the merge involves two operators, *join* and *cut*, to simultaneously generalize and preserve a high level of fidelity. In particular, in the logical domain, generalization seldom involves premises relaxation or outright removal [39]. Thus, GLocalX advances the state-of-the-art in exploiting also this kind of generalization. We highlight that generalization comes at a fidelity cost, as the more general a set of premises is, the more likely it is to capture unwanted instances. Pushing for generalization may pull down fidelity. These contrasting behaviors are the focal point in a *Local to Global* setting and must be dealt with accordingly. We tackle this double-faced problem with a merge function that handles both rule generalization and fidelity.

We illustrate our proposal with a trivial example. Suppose we have two explanation theories, $E_1 = \{e_1, e_2\}$ and $E_2 = \{e'_1, e'_2\}$, with e_1, e'_1 explaining a record x_1 and e_2, e'_2 explaining a record x_2.

$$E_1 = \{e_1 = \{\text{age} \geq 25, \text{job} = \text{unemployed}, \text{amount} \leq 10k\} \rightarrow \text{deny}$$
$$e_2 = \{\text{age} \geq 50, \text{job} = \text{office clerk}\} \rightarrow \text{deny}\}$$
$$E_2 = \{e'_1 = \{\text{age} \geq 20, \text{job} = \text{manager}, \text{amount} > 8k\} \rightarrow \text{accept}$$
$$e'_2 = \{\text{age} \geq 40, \text{job} = \text{office clerk}, \text{amount} > 5k\} \rightarrow \text{deny}\}$$

The resulting merge yields the following rules E_{1+2}:

$$E_{1+2} = \{e''_1 = \{\text{age} \geq 25, \text{job} = \text{unemployed}, \text{amount} \leq 10k\} \rightarrow \text{deny}$$
$$e''_2 = \{\text{age} \in [20, 25], \text{job} = \text{manager}, \text{amount} \in [8k, 10k]\} \rightarrow \text{accept}$$

Algorithm 2 MERGE(E_i, E_j, X).

Input: E_i, E_j explanation theories, X batch
Output: $E_{(i+j)}$ explanation theory
1: E $\leftarrow E_i \cup E_j$
2: **for** $x \in X$ **do**
3: $C_i \leftarrow$ COVERED(x, E_j) ▷ retrieve rules in E_i covering x
4: $C_j \leftarrow$ COVERED(x, E_j) ▷ retrieve rules in E_j covering x
5: $C_= \leftarrow$ NON-CONFLICTING(x, C_i, C_j) ▷ non-conflicting rules in C_i, C_j and covering x
6: $C_\neq \leftarrow$ CONFLICTING(x, C_i, C_j) ▷ non-conflicting rules in C_i, C_j covering x
7: E $\leftarrow$ E $\setminus (E^i \cup E^j)$
8: $E_= \leftarrow$ JOIN($C_=$)
9: $E_\neq \leftarrow$ CUT($C_\neq$, X)
10: E $\leftarrow$ E $\cup E_= \cup E_\neq$
11: **return** E

$$e_3'' = \{\text{age} \geq 40\} \rightarrow \text{deny}\}$$

In E_{1+2}, rules with equal predictions from different explanation theories have been generalized by relaxing their premises (rule e_3'') while rules with different predictions have been specialized by further constraining them (rule e_2''). Specifically, e_1'', e_2'' results from a cut on e_1, e_1', that is $e_1 - e_1'$.

More formally, given two explanation theories E_i, E_j, the *merge* function applies two operators on the explanation theories to derive a new theory approximately entailed by the two: the JOIN and the CUT operators. The former allows merging *non-conflicting* rules while the latter allows merging *conflicting* rules [27]. A set of explanations $E = \{e_1, \ldots, e_n\}$ part of a logical explanation theory is considered *conflicting* on an instance x if two or more of them cover an instance x but lead to two different outcomes. The merge of two logical explanation theories E_i and E_j applies the JOIN operator on *non-conflicting* explanations and the CUT operator on *conflicting* explanations iteratively on each instance in the batch. We detail this process in Algorithm 2. The resulting set of explanations composes the new explanation theory E_{i+j} (line 8, Algorithm 1). The candidate merge is then tested for considering the equilibrium between fidelity and complexity with the Bayesian Information Criterion (BIC) [36] computation. In our case, the model log-likelihood is computed as the rules fidelity, and the model complexity as the average rule length.

In the following we provide details of the JOIN and CUT operators. The two operators move in opposite directions: JOIN generalizes explanations, possibly at a fidelity cost, while the CUT specializes explanations, possibly at a generalization cost. In other words, the former allows generalization while the latter regularizes it. Inspired by [40], we define the JOIN and the CUT operators through an alternative representation of decision rules. Let $\widehat{\mathcal{X}}_i$ be the set of subspaces on the feature i. Given a decision rule $r = P \rightarrow y$ we have that any $P_i \in \widehat{\mathcal{X}}_i$ is a subspace on the feature i. The premise P of the rule identifies a *quasi-polyhedron* defined as a subspace of $\widehat{\mathcal{X}}^{(m)}$:

$$P = \{P_i, \ldots, P_j\} \in \mathcal{P}(\widehat{\mathcal{X}}_1 \times \cdots \times \widehat{\mathcal{X}}_m)$$

We say that an instance x *satisfies* P if $\forall P_i \in P.\ x_i \in P_i$, thus, x satisfies P if it lies in the subspace defined by the quasi-polyhedron. We define the operators *join* ($\oplus$) and *cut* ($\ominus$) exploited by the *merge* function based on this quasi-polyhedron interpretation.

Reasoning in the polyhedral space. The polyhedral interpretation lends itself to a straightforward approximate inference algorithm, which we call "of inclusion". Given the equivalence of decision rules and quasi-polyhedra, a rule s with quasi-polyhedron P_s is inferred by another rule r with quasi-polyhedron P_r ($r \rightarrow s$) if and only if its premises (and thus the record satisfying them) are implied by the other. Simply put, all instances satisfied by s are also satisfied by r. Quasi-polyhedra-wise, it follows that $P_s \subseteq P_r$. Hence, inference by inclusion produces flat reasoning paths in which the implied local rules are elevated to global explanations, yielding a "global" model just as local as before. With this consideration in mind, we reject exact inference in favor of approximate inference in which *join* and *cut* perform approximate rule entailment.

Join. The JOIN operator aims to generalize a set of non-conflicting explanations relaxing their premises, hence generalizing the associated rules. Given two quasi-polyhedra P and Q, the JOIN ($\oplus$) is defined as follows:

$$P \oplus Q = \{P_1 + Q_1, \ldots, P_m + Q_m\}$$

where:

$$P_i + Q_i = \begin{cases} P_i \cup Q_i & \text{non-empty intersection} \\ [\min\{P_i \cup Q_i\}, \max\{P_i \cup Q_i\}) & \text{empty intersection} \\ \emptyset & P_i \text{ is empty} \vee Q_i \text{ empty} \end{cases}$$

Example. Consider the following two explanations:

$$e_1 = \{\text{age} \geq 50, \text{job} = \text{office clerk}\} \rightarrow \text{deny}$$

$$e_2 = \{\text{age} \geq 40\} \rightarrow \text{deny}$$

The merge function applies the JOIN operator $e_1 \oplus e_2$ and returns

$$e'_1 = \{\text{age} \geq 40\} \rightarrow \text{deny}$$

The shared feature *age* has a *non-empty intersection* ($\text{age} \geq 50 \cap \text{age} \geq 40 \neq \emptyset$), hence it is generalized to encompass both premises according to the first case. On the non shared feature job = office clerk we have one empty quasi-polyhedron, hence it is removed according to the third case.

Cut. The CUT operator acts in a complementary fashion by slicing quasi-polyhedra. Here, the goal is to preserve the better rules, and confine the lesser ones to subspaces in which they have high fidelity. In other words, the aim is to remove overlaps between rules by subtracting them. This directly translates to subtracting quasi-polyhedra. Formally, given two quasi-polyhedra P and Q, the CUT ($\ominus$) is defined as:

$$P \ominus Q = \{P_1 - Q_1, \ldots, P_m - Q_m\}$$

where:

$$P_i - Q_i = \begin{cases} \{P_i, \emptyset\} & Q_i \text{ empty} \\ \{P_i, Q_i \setminus Q_i\} & \text{otherwise} \end{cases}$$

Note that, unlike the JOIN operator, the CUT operator is not symmetric, hence $P \ominus Q \neq Q \ominus P$. With our goal of preserving high fidelity rules and restrict lower fidelity rules in mind, it is straightforward to select subtracted and subtracting polyhedron, with the two rules being the one with higher and lower fidelity, respectively.
Example. Consider the following two explanations:

$$e_1 = \{age \geq 25, job = unemployed, amount \geq 10k\} \rightarrow deny$$

$$e_2 = \{age \geq 20, job = manager, amount > 8k\} \rightarrow accept$$

where the first one is the dominant one, that is, the one with highest fidelity. The merge function applies the CUT operator $e_1 \ominus e_2$ and returns

$$e_1 = \{age \geq 25, job = unemployed, amount \geq 10k\} \rightarrow deny$$

$$e'_1 = \{age \in [20, 25], job = manager, amount \in [8k, 10k]\} \rightarrow accept$$

We note that e_1 is preserved as-is while the premises of e_2 are further constrained to reduce overlap. Namely, *age* is constrained to remove overlap on $age \geq 25$ and *amount* is constrained to remove overlap on $amount \geq 10$k.

We highlight that $\oplus$ and $\ominus$ strictly operate on the premises of rules, ignoring their outcome, which is preserved after the merge. Moreover, JOIN and CUT produce sets of different cardinality. While CUT preserves the number of rules, join indeed lessens it by subsuming it in a smaller more general explanation set.

4.3. Interpretable classification

The global explanation theory E mimes the black box decision logic for predicting a set of instances. By construction, E is not exhaustive nor mutually exclusive, i.e., not all records are necessarily covered by at least one rule while covered records may be covered by more than one rule. We address the former case with a default majority rule and the latter with a voting schema.

We exploit E as a rule-based classifier to replace the opaque black box b in the prediction task. Moreover, E may also be used for explaining the decision of the black box b on a single instance x. Clearly, as a global explanation, it can be less accurate than the local one, due to manipulations applied to capture the global behavior of the black box.

It is important to highlight that, given an instance x we could have more than one decision rule (explanation) in E that satisfies x, and some of them can also be conflicting. As a consequence, to apply E for classifying x or explaining $b(x)$ we need a mechanism for deriving a decision when multiple covering rules have different outcomes. In line with the literature [17], we employ the following: given a record x, we select the rule with the highest accuracy among the ones covering x, and use its outcome as prediction. Note that this mechanism replicates the Laplacian scoring schema proposed in [17] and the *Falling Rule List* schema introduced by [18] setting a number of voting rules to one.

5. Experiments

In this section, after presenting the experimental setup, we report an analysis of GLOCALX on a set of standard benchmark datasets and a real-world proprietary dataset, and compare GLOCALX to native baselines and state-of-the-art global explainers. Moreover, we analyze GLOCALX in two settings: in the former, we provide GLOCALX with a restricted number

Table 1
Datasets statistics.

	instances	features	$\lvert X_{bb}\rvert$	$\lvert X_{le}\rvert$	$\lvert X_{ts}\rvert$
adult	32,560	10	22,791	6,837	2,930
compas	7,214	18	5,048	1,514	649
german	1,000	19	699	209	89
diva	8,000	88	4800	1600	1600

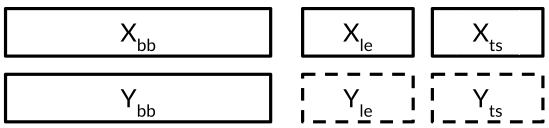

Fig. 1. Validation schema: X_{bb} is used for training the black box, X_{le} is the partition to explain, while X_{ts} is reserved for validating the fidelity and the accuracy. The dashed line indicates the labels predicted by the black box and on which the model fidelity is estimated.

of rules while in the latter, we do not provide GLocalX with any data. GLocalX has been developed in Python.[3] The experiments were performed on `16-core Intel(R) Xeon(R) CPU E5-2620 v4 @ 2.10GHz, 128 GB` of RAM. For the local rule extraction we selected LORE [22] due to its high fidelity and stability [24]. Alternatives are ANCHORS [12] or generating a random local neighborhood like in LIME [11] and the using locally a rule based classifier.

5.1. Experimental setup

We showcase the proposed approach on three benchmark datasets frequently used in the literature, namely `adult`,[4] `compas`,[5] and `german`,[6] and a real world proprietary dataset, `diva`.[7]

The `adult` dataset includes 48,842 instances with information like age, job, marital status, race, capital loss, capital gain, etc. The labels have values "<=50K" or ">50K", indicating whether the person will earn more or less than 50k\$ in this fiscal year. The `compas` recidivism dataset contains the features used by the COMPAS algorithm for scoring defendants and their risk of recidivism (Low, Medium and High), for over 4,000 individuals. We have considered the two classes "Low-Medium" and "Medium" as equivalent to map the task into a binary classification task. The dataset includes features like: age, sex, race, priors count, days before screening arrest, length of sentence, charge degree, etc. In `german` each one of the 1,000 persons is classified as a "good" or "bad" creditor according to attributes like age, sex, checking account, credit amount, loan purpose, etc. Finally, `diva` is a privately released dataset on fraud evasion, periodically issued by the Italian Ministry of Economics.[8] `diva` records financial activities for more than 12,000 citizens, including their past financial credit score, declared income and property value, debt and several taxation detailed infos. The labels mark fraudulent citizens. These datasets contain both categorical and continuous features. Missing values, if present, were replaced by the mean for continuous features and by the mode for categorical ones. Details are reported in Table 1. Similarly to the training/test classical split, we split the dataset into three partitions: X_{bb}, the set of records to train a black box model; X_{le}, the set of locally explained records which is also used as reference set by GLocalX for calculating coverage and fidelity in training phase, and by the global competitors for training; X_{ts}, the held-out set of records to validate the fidelity and accuracy of GLocalX. Fig. 1 depicts this three-way split.

We validate GLocalX on its ability to mime Deep Neural Networks (DNN), Random Forests (RF) and Support Vector Machines (SVM). The black boxes have been trained on X_{bb} with grid searches on a 3-fold cross-validation schema. Given the poor performances of DNNs and SVMs on `diva` and `german` datasets, they were experimented upon only on Random Forests. Performance in terms of accuracy of these black boxes are reported in Table 2 (3rd column). Given the novelty of the problem and the lack of local to global explainer in the literature,[9] we compare GLocalX against natively global frameworks:

[3] The Python implementation is available at https://github.com/msetzu/glocalx.
[4] https://archive.ics.uci.edu/ml/datasets/Adult.
[5] https://github.com/propublica/compas-analysis.
[6] https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data).
[7] https://kdd.isti.cnr.it/project/diva.
[8] Due to privacy and legal concerns, we are not allowed to publicly release the dataset.
[9] LIME [11] and some other recent works extending LIME claim to obtain a global explanation by joining local feature importance but in fact the resulting model is just a set of numbers and cannot be used to replace the black box, nor expresses in logical form the logic adopted by the black box for taking decisions.

Table 2
Black box accuracy on the various dataset (3rd column), and performance of a decision tree (DT) trained on the black box models and DT are trained on X_{bb}, y_{bb} where y_{bb} is the ground truth. Both the black box models and the decision tree are evaluated on the X_{ts}.

Dataset	Black Box	Accuracy	DT Accuracy	DT Size	DT Length
adult	DNN	0.868			
	RF	0.860	0.813	5452	16.611 ± 6.150
	SVM	0.860			
compas	DNN	0.611			
	RF	0.548	0.587	1514	6.271 ± 1.985
	SVM	0.557			
german	RF	0.753	0.896	68	5.294 ± 2.065
diva	RF	0.900	0.848	934	11.233 ± 5.268

Decision Tree [25] (DT), Pruned Decision Tree[10] (PDT) in a C4.5 [25] implementation, and CPAR [17], a rule-based classifier (RB) implemented in the *LUCS-KDD* library.[11]

In addition, we also experiment with GLocalX in a particular setting in which only local explanations $e_1, \ldots, e_n$ are provided, and no data X_{le} is available. We call this setting *synthetic* and name it GLocalX* for short. In this case, we assume to have some information on the feature distribution, and employ a data generation and sampling schema similar to [15] to construct a training dataset for GLocalX of the same size of X_{le}. Specifically, the joint distribution has been estimated with a Gaussian density estimator, which has then been randomly sampled to build a training set for GLocalX*. This dataset is then provided as input, and GLocalX is executed as previously defined. With GLocalX*, we wish to grasp the learning abilities of GLocalX when a minimal input comprised exclusively of the local rules and data distributions is provided. A minimal assumption of knowledge on the data distribution is needed to validate the merging procedure. Relaxing this assumption would reduce the merging problem in a logical inference one, as no quantitative evaluation of any merge can be performed, leaving only intra-rule implication as a form of rule generalization. In other words, when no knowledge on the data is available, the *Local to Global* problem is equivalent to a logic inference problem.

5.2. Evaluation measures

Given a black box b and the explanation theory E returned either by GLocalX or by a global transparent model, we consider the following properties in evaluating its performance:

- *fidelity*$(Y, \widehat{Y}) \in [0, 1]$ where Y and $\widehat{Y}$ are the predictions returned by the rules in E or by the black box b, respectively. The fidelity is the accuracy of the transparent global model in approximating the behavior of b [3,18].
- *size*$(E) = |E|$ is the *number of explanations* in the explanation theory E.
- *length*$(E) \in \mathbb{R}^+$ is the *mean number of premises* of the rules in E.
- *accuracy*$(Y, Y^*) \in [0, 1]$ where Y is the classification returned by the rules in E (or by the black box b), and Y^* are the real labels. It answers the questions: how good is the transparent model represented by E in solving the classification problem? Can we replace b with E?

If not differently specified, the results in the rest of this section use X_{bb} to train the black box classifiers, X_{le} to extract local explanations,[12] run GLocalX and learn the global interpretable models DT, PDT and RB, and X_{ts} for evaluating the fidelity. Thus, the *fidelity* and the *accuracy* refer to X_{ts}, while the *size* and *length* refer to models learned on X_{le}. Experiments have been run with a batch size of 128.[13]

5.3. Empirical motivation of the local to global explanation problem

Besides the motivation presented in the introduction and in the problem definition section we show here an empirical reason for the local to global explanation problem. Table 2 shows the performance of an interpretable model, i.e., a Decision Tree (DT), trained on the same training set of the black box, i.e., X_{bb}, Y_{bb} where Y_{bb} are the real labels. The results highlight that although DT has a high accuracy (4th column), it is usually lower than the accuracy of the black box classifiers (3rd

[10] A pruned decision tree is a decision tree with maximum dept equals to four. We adopt four as maximum dept as it is the measure used in Optimal Decision trees [41]. We do not compare with Optimal Decision trees due to the complexity of running the models that requires a particular architecture and the non public availability of the code.

[11] https://cgi.csc.liv.ac.uk/~frans/KDD/Software.

[12] Specifically, one explanation per record: the number of local explanations is directly inferred by the size of X_{le} in Table 1.

[13] Smaller batches are not large enough for performing reasonably accurate inference, while larger sets tend to reduce the diversity and number of merges. In a preliminary stage of the research we have tested several batch sizes in $\{2^i\}, i \in \{1, \ldots, 8\}$, with 128 yielding satisfying results on all datasets.

Table 3
Sensitive features for the recidivous and non-recidivous group. Average number of `prior offenses` and percentage of past recidivists, violent recidivists, african-americans and caucasians.

Feature	Recidivous	Non-recidivous
prior offenses	**9.89 ± 5.18**	2.75 ± 4.23
past recidivous (%)	**74 ± 0.43**	0.42 ± 0.49
violent recidivous (%)	**19 ± 0.39**	0.09 ± 0.29
african-american (%)	**70 ± 0.45**	0.50 ± 0.50
caucasian (%)	21 ± 0.41	**33 ± 0.47**

column) in most of the cases (justifying the black box usage). At the same time, despite being interpretable, the DT does not guarantee a real explicability. Indeed, the good levels of accuracy come at the cost of a very high *complexity* (5th, 6th columns) of the learned model. This is testified by the high *size* (number of rules derived from the DT) and average rule length of the DT, making the overall transparent model non-interpretable in practice. We show in the following how GLocalX, that works on a much smaller portion of data, is able to reach comparable performance in terms of fidelity while maintaining an admissible complexity of the explanation theory returned.

5.4. Qualitative evaluation

In this section, we show an example of explanation theory E yielded by GLocalX. We use as example the explanation theory returned by GLocalX, for explaining the behavior of the RF black box on `compas`, $\alpha = 6$. GLocalX achieves a fidelity of 0.86 with a set of 6 rules inferred from 1,515 starting local rules:

$$
\begin{aligned}
E = \{ & e_1 = \{\textit{prior offenses} \geq 4, \textit{age} < 45, \\
& \quad \textit{sex = male}\} \rightarrow \textit{Non-recidivous}; \\
& e_2 = \{\textit{prior offenses} \geq 4, \textit{age} < 45, \textit{sex = male}, \\
& \quad \textit{charge degree} = \textit{light}\} \rightarrow \textit{Non-recidivous}; \\
& e_3 = \{\textit{age} \leq 43, \textit{prior offenses} < 4, \textit{past recidivous} = \textit{no}, \\
& \quad \textit{sex = male}, \textit{charge degree} = \textit{light}\} \rightarrow \textit{Non-recidivous}; \\
& e_4 = \{\textit{past time in prison} < 8\} \rightarrow \textit{Recidivous}; \\
& e_5 = \{\textit{priors count} < 4, \textit{past time in prison} < 1\} \rightarrow \textit{Recidivous}; \\
& e_6 = \{\textit{priors count} < 2, \textit{past time in prison} < 8\} \rightarrow \textit{Recidivous}; \\
& \}
\end{aligned}
$$

The explanations for the two classes show very different behaviors. For the `Non-recidivous` class, explanations are rather lengthy, and account for the defendant's prior offenses, age, current charge and prior recidivous behavior. Young men with a handful of light crimes and previous non-recidivous behavior appear to be the demographic of the non-recidivous behavior. The `Recidivous` class shows different explanations, with brief rules involving exclusively the previous time in prison and the previous offenses.

On first sight, these explanations appear to be more coarse and do not show any self-evident bias towards sensitive features, e.g., race, of the defendants. Table 3 reports average values for the records covered by the two sets of explanations. The recidivous explanations target, as expected, defendants with a high number of prior offenses (9.89 on average), high past recidivism (74%) and past violent recidivism (19%). Intuitively, these are highly predictive features which one might expect to lead to future recidivism. Moreover, the explanations indicate a possible bias against african-americans, which appear to be recidivous at a much higher rate (70%) than caucasians (50%).

5.5. Impact of the filter parameter

First of all, we report an analysis of the impact of the filter parameter used to filter out rules from the final explanation theory returned by GLocalX. Here we use the relative trimming criterion α_q instead of the absolute one to minimize dataset-specific dependencies and study the effects of the filter across the fidelity distribution. Table 4 reports the *fidelity* and complexity (*size* and *length*) of GLocalX varying α_q for the various datasets and black box models analyzed. The fidelity is lower for values of α_q around 75 and peaks around $\alpha_q = 95$, suggesting that a large number of rules may mislead the predictions of the model. We can attribute this behavior to the use of batches in the construction phase: as stated in Section 4, the highest-fidelity rule is selected by verifying the fidelity of the rules on the batch at hand. Hence, poor rules with good performance only on single batches may leak into the final model.With respect to the complexity we observe a

Table 4
Impact of the filter parameter α_q on GLocalX performance in terms of fidelity, size, and length and various datasets and black box classifiers. We observe that when considering simultaneously the three evaluation measures the best performance are reached for $\alpha_q = 95$, suggesting that few rules are sufficient for mimicking the global black box behavior.

	b	α_q	fidelity	size	length		black box	α_q	fidelity	size	length
adult	DNN	75	.872	53	6.73 ± 1.84	compas	DNN	75	.769	16	2.93 ± 0.99
		90	.905	17	6.43 ± 1.76			90	.751	8	2.71 ± 0.69
		95	.911	10	5.33 ± 1.15			95	.757	5	2.25 ± 0.43
		99	.909	3	7.00 ± 0.00			99	.759	4	2.33 ± 0.47
	RF	75	.845	27	10.25 ± 1.03		RF	75	.829	27	5.57 ± 1.52
		90	.862	11	10.80 ± 0.60			90	.873	9	5.37 ± 1.49
		95	.865	7	11.00 ± 0.57			95	.874	6	5.20 ± 1.72
		99	.867	2	12.00 ± 0.00			99	.875	10	8.00 ± 0.00
	SVM	75	.859	56	4.67 ± 1.39		SVM	75	.827	19	6.22 ± 1.54
		90	.868	22	4.61 ± 1.25			90	.861	7	5.66 ± 0.47
		95	.902	10	4.22 ± 1.13			95	.860	4	5.66 ± 0.47
		99	.865	3	4.50 ± 0.50			99	.851	2	6.00 ± 0.00
diva	RF	75	.855	48	3.91 ± 0.96	german	RF	75	.786	2	4.00 ± 0.00
		90	.873	16	3.86 ± 0.71			90	.786	2	4.00 ± 0.00
		95	.806	9	3.75 ± 0.82			95	.786	2	4.00 ± 0.00
		99	.814	3	4.00 ± 1.00			99	.786	2	4.00 ± 0.00

consistent decrease in size and a slight increase in explanation length across datasets and black boxes. Thus, more complex rules seem to compensate for less rules in the explanation theory.

5.6. Effect of the number of local rules

As second experiment we analyze the effect on the performance of GLocalX of a reduced number or local explanation rules $e_1, \ldots, e_k$ $(k < n)$ is provided in input with respect to considering a large set or all the available rules form X_{le}. This scenario is of particular interest for applications where extracting rules is costly, there are additional constraints on the computation time, or simply there is a low number of available local explanations. In practice, we provide as input to GLocalX a subset of the rules available extracted with LORE from X_{le}. Specifically, we provide GLocalX with $\beta = \{1\%, 20\%, 40\%, 60\%, 80\%\}$ of the available rules, randomly sampling them in 10 independent trials.[14] Fig. 2 reports fidelity and the size averaged over trials. To lessen the impact of the α-trimming, we report results with $\alpha = q^{50}\%$, i.e., instead of selecting the top-α explanations, we trim those with fidelity under the 50th percentile. GLocalX shows small fluctuations in fidelity (left axis). In particular, varying β we observe an overall slight fidelity improvement. While there is a small drop in fidelity, these results suggest that GLocalX can be used with smaller sets of rules, at the cost of fidelity. On the other hand, the main effect of β is registered on the size (right axis) that significantly grows with the number of input rules. We remark that in this experiment we used $\alpha_q = 50$ that causes a smaller number of rules to be filtered out. Therefore, the α (α_q) parameter should be tuned not only according to the required fidelity and size, but also according to the number of input rules. In the rest of this section we report results using all the local rules returned by LORE from X_{le} if not differently specified.

5.7. Effectiveness of the global explainers

The final goal of GLocalX is to provide a global explanation of a black box classifier. In order to test the effectiveness in replicating the black box behavior we compare the fidelity and complexity of the explanation theories returned by GLocalX and GLocalX* with the rule sets returned by the interpretable global models (DT, CPAR, PDT) trained on X_{le} with the labels returned by the black box. In addition, as a baseline local to global method we compare against an approach (UNI) that simply performs the union of the local decision rule and adopts all of them as explanation theory. For GLocalX we select α to be lower than the smallest competitor, in this case PDT. Table 5 reports the results of this comparison. GLocalX and GLocalX* are shortened as GLX and GLX* for readability purposes. For each dataset and black box the highest fidelity and lowest size and length are underlined.

The results show the ability of GLocalX to find explanation theories with a high fidelity and low complexity. We observe that GLocalX has a competing fidelity which is comparable to the one of the best global explainer (generally the DT) as it is always only less than a 0.1 lower. The loss is more evident when explaining DNNs, while it is negligible for the other black box classifiers. At the same time, GLocalX or GLocalX yield the lowest size (number of rules) of the global explanation resulting in a simple and compact but effective model. It is important to notice that GLocalX learns sets of rules one order of magnitude smaller than CPAR and one/two orders of magnitude smaller than the DT. This suggests that accounting for

[14] german, which has a low number of rules, has been tested for $\beta = \{20\%, 40\%, 60\%, 80\%\}$.

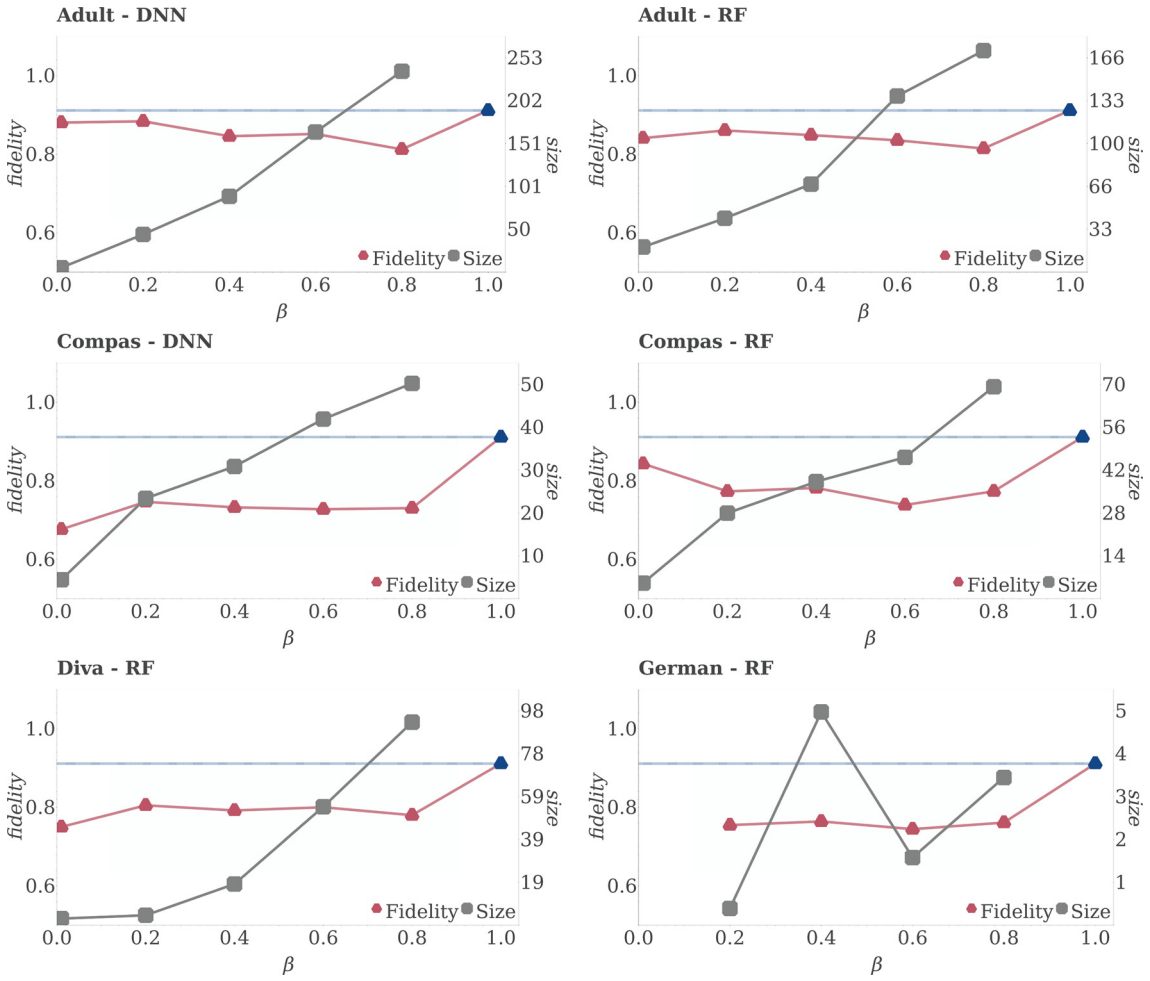

Fig. 2. Effect of the number of local rules in GLocalX with $\alpha_q = 50$. The red line denotes the fidelity, while the gray the size of GLocalXvarying the number of rules. The blue line denotes the fidelity of GLocalX using all the available rules.

complexity and fidelity at the same time in the merging process can yield good results in both metrics. The DT has a comparable number of rules with GLocalX and a fidelity within 8% of the one of GLocalX. In terms of rule length, GLocalX learns rules consistently shorter than the DT, but longer than PDT and CPAR who are the best performers according to this evaluation metric. However, as previously stated, the sets of rules learned by CPAR are one/two orders of magnitude larger than the size of the explanation theory returned by GLocalX. On the other hand, the DT has on average more than two times the number of rules of GLocalX and in a real scenario this can cause confusion, especially if every feature model a complicated concept.

GLocalX show remarkable higher performances than UNI. Also GLocalX*, that does not have access to data, obtains similar fidelity to GLocalX, with fidelities 3% lower in the worst configuration. This indicates that just aggregating all the local explanations together is not beneficial for obtaining an effective global explanation. The local explanation rules must be carefully processed in order to remove useless and/or misleading local aspects that do not help in understanding the global reasons for the classification. Finally, GLocalX* shows a slightly lower performance than GLocalX with fidelity scores lower than 2−3% and non-monotonic pattern on length and size, which are shorter/longer and smaller/larger across datasets and black boxes.

We adopt the non-parametric Friedman test [42] for comparing the average ranks of explanation methods over multiple datasets and black boxes with respect to the fidelity, size and length. The null hypothesis that all methods are equivalent is rejected for $p-value < 0.05$ for fidelity, $p-value < 0.0001$ for size, and $p-value < 0.0005$ for length.

5.8. Replacing the black box with a global explainer

As final experiment, we test the performance of GLocalX in terms of *accuracy*, i.e., we consider predictions on the real dataset labels, rather than the ones predicted by the black box. With this experiment, we aim to understand whether GLocalX can be used to replace the black box classifier instead of being used only for understanding the classification reasons.

Table 5
Effectiveness of the global explainers in terms of fidelity, size and length on different datasets and black box classifiers *b*. GLocalX and GLocalX* (having not access to the data) are indicated with GLX and GLX*, respectively. For each dataset and black box the highest fidelity and lowest size and length are underlined. GLocalX has a high fidelity comparable with the best performer, the lowest complexity in terms of size.

	b	method	fidelity	size	length		*b*	method	fidelity	size	length
adult, $\alpha = 10$	DNN	GLX	.912	5	6.00 ± 1.0	compas, $\alpha = 5$	DNN	GLX	.759	3	2.33 ± 0.4
		GLX*	.880	10	6.30 ± 2.5			GLX*	.756	6	4.50 ± 0.9
		CPAR	.929	100	3.78 ± 2.4			CPAR	.821	69	3.11 ± 1.4
		DT	.917	1068	7.22 ± 1.9			DT	.789	1014	6.00 ± 1.8
		PDT	.908	28	2.71 ± 0.8			PDT	.780	30	2.33 ± 0.6
		UNI	.880	6838	3.86 ± 2.2			UNI	.627	1515	4.65 ± 1.7
	RF	GLX	.902	10	4.00 ± 1.0		RF	GLX	.870	6	3.66 ± 2.4
		GLX*	.876	10	3.20 ± 1.5			GLX*	.862	6	4.33 ± 1.1
		CPAR	.944	107	2.57 ± 1.3			CPAR	.908	53	2.86 ± 1.6
		DT	.959	926	7.43 ± 1.6			DT	.906	452	4.91 ± 1.4
		PDT	.935	26	2.53 ± 0.8			PDT	.886	30	2.60 ± 0.7
		UNI	.876	838	5.00 ± 1.4			UNI	.658	1515	3.52 ± 1.2
	SVM	GLX	.865	10	7.70 ± 3.4		SVM	GLX	.860	6	4.33 ± 1.3
		GLX*	.854	10	6.10 ± 2.9			GLX*	.840	6	4.16 ± 0.8
		CPAR	.848	95	4.77 ± 3.0			CPAR	.858	70	3.04 ± 1.3
		DT	.875	2956	7.57 ± 1.6			DT	.875	824	4.58 ± 1.0
		PDT	.854	24	2.50 ± 0.6			PDT	.850	30	2.53 ± 0.7
		UNI	.854	6838	3.54 ± 2.4			UNI	.696	1515	4.08 ± 0.8
diva, $\alpha = 25$	RF	GLX	.854	26	3.26 ± 0.9	german, $\alpha = 2$	RF	GLX	.786	2	3.00 ± 1.0
		GLX*	.848	26	3.88 ± 1.3			GLX*	.766	2	5.87 ± 2.4
		CPAR	.850	221	2.03 ± 1.0			CPAR	.773	18	2.33 ± 1.4
		DT	.853	976	10.63 ± 3.9			DT	.830	76	4.50 ± 1.7
		PDT	.836	28	3.21 ± 0.9			PDT	.796	28	2.78 ± 0.8
		UNI	.794	2013	2.90 ± 1.0			UNI	.766	210	5.62 ± 2.4

Table 6
Accuracy of the global explainers on X_{ts} adopted as replacement of the black box models. For each dataset, the first line report the average accuracy of the black box classifiers indicated with *b*. The average *accuracy* (and standard deviation) among the various black box classifiers is reported for each global explainer. Δ_{acc} is calculated as the difference between the model fidelity and accuracy. The closer to 0, the better it is. The average (and standard deviation) size and length complete the table.

	method	accuracy	Δ_{acc}	size	length
adult	*b*	$.801 \pm 0.056$			
	CPAR	$.816 \pm 0.148$	-0.091 ± 0.042	139.1 ± 217.2	7.526 ± 1.464
	DT	$.945 \pm 0.062$	0.027 ± 0.039	1654.0 ± 1134.5	7.430 ± 0.159
	GLX	$.792 \pm 0.019$	0.101 ± 0.013	8.3 ± 2.8	5.533 ± 1.601
	PDT	$.894 \pm 0.057$	-0.005 ± 0.043	16.6 ± 14.4	2.519 ± 0.139
compas	*b*	$.673 \pm 0.033$			
	CPAR	$.790 \pm 0.108$	-0.072 ± 0.377	43.7 ± 79.5	4.948 ± 1.306
	DT	$.593 \pm 0.514$	-0.264 ± 0.366	762.0 ± 288.7	4.988 ± 0.225
	GLX	$.726 \pm 0.002$	0.103 ± 0.049	5.00 ± 1.73	4.611 ± 0.976
	PDT	$.868 \pm 0.025$	0.029 ± 0.363	20.0 ± 17.3	2.566 ± 0.004
diva	*b*	.908			
	CPAR	.850	0.020	221.0	2.031
	DT	.854	−0.0170	976.0	10.680
	GLX	.824	0.029	26.0	3.192
	PDT	.836	−0.021	28.0	3.214
german	*b*	.700			
	CPAR	.880	−0.006	5.2	2.103
	DT	.915	−0.000	26.0	2.789
	GLX	.726	0.006	2.0	3.000
	PDT	.898	−0.000	10.0	1.892

Table 6 reports the average accuracy values across the various black box classifiers for GLocalX and for the interpretable classifiers adopted as competitors on the held-out test set X_{ts}. Table 6 also reports the standard deviation of the accuracy and the *accuracy delta* $\Delta_{acc} \in [0, 1]$ calculated as the difference between the model fidelity and accuracy. An explainable model with minimum Δ_{acc} score ($\Delta_{acc} = 0$) is as accurate on the dataset labels as it is faithful to the black box labels. In other words, we can expect similar performances when the explainable model is deployed to predict the actual dataset labels. As the fidelity-accuracy gap grows (Δ_{acc} approaching 1), the explainable model is significantly more faithful to the black box and less accurate on the dataset, that is, it overfits the black box labels at the cost of the dataset labels. In other

Table 7
Average Δ_{acc} on X_{ts}. The lower the value, the better are the performance of a global explainer as it indicates stability between the capacity of mimicking the black box behavior and the ability of being adopted as a classifier.

	Δ_{acc}
adult	0.128 ± 0.008
compas	0.228 ± 0.017
diva	0.026 ± 0.023
german	0.009 ± 0.007

(a) Aggregation by dataset.

	Δ_{acc}
CPAR	-0.037 ± 0.052
DT	-0.063 ± 0.135
GLX	0.073 ± 0.015
PDT	0.005 ± 0.020

(b) Aggregation by method.

words, we should expect a worse performance when the explainable model is deployed the actual dataset labels. The first line of each dataset reports the average accuracy for the black box classifiers b.

GLocalX falls behind some of the competitors in terms of accuracy, with accuracy values lower up to 7% and 15% for compas and adult, respectively. The Δ_{acc} is low for most models, with the DT showing a peculiar behavior. Indeed, its Δ_{acc} is highly unstable, with values ranging from -0.264 to 0.027. GLocalX, CPAR and PDT, which all comprise of simpler models, show lower variance. These results are strengthened by the numbers in Table 7 that report the average Δ_{acc} aggregated respectively per dataset and per global explanation model, i.e., a zoom out from the previous Table. Datasets show a wildly different behavior, with $|\Delta_{acc}|$ as low as 0.006 (0.6% absolute increase) and as high as 0.086 (9% absolute increase). Neither dataset size nor average fidelity or accuracy appear to correlate with these deltas. Surprisingly, PDT shows a positive average Δ_{acc}, indicating a better accuracy than fidelity. While this confirms the above considerations on model simplicity, it should be noted that the PDT are actually showing better performance on a label distribution different than the one they were trained on. Among the other models, GLocalX shows the lowest Δ_{acc}, $\approx \times 1.5$ times lower than CPAR and $\approx 2.8\times$ lower than DT. Therefore, since the explanation theory E returned by GLocalX guarantees not only high fidelity and small complexity, but also an high accuracy in classification, it can successfully be used for replacing the original black box classifier.

6. Conclusions

In this paper we have proposed GLocalX, a model agnostic *Local to Global* explanation algorithm based on logic rules for AI systems using non interpretable machine learning models in the decision process. Starting from local explanations, GLocalX derives global explanations to describe the overall logic of a black box model. The proposed method applies a hierarchical approach to derive a global explanation from the local ones. GLocalX tackles both explanation complexity and *fidelity* in emulating the AI decision system behavior. The results suggest that GLocalX can be a valid Local to Global approach, as it tends to provide faithful and simple models. GLocalX outperforms the trivial union of rules and it is competitive with natively global explanators especially in terms of complexity. Finally, experiments also highlight that the explanation theories of GLocalX might be used directly as transparent predictors with performances similar to other global predictors.

The key advantage of GLocalX lies in its flexibility: merging and regularizing explanations as they are generated allows for a plethora of extensions. Among them we indicate direct human-guided regularization, with ad-hoc regularization penalizing reliance on some features rather than others, alignment to existing expert knowledge and balancing the fidelity-complexity equilibrium. In this paper, we defined an explanation with logical rules. A direct follow-up would be the extension to fuzzy and non-CNF rules, empowering reasoning with uncertainty. Adaptation to non-logical domains such as sequences, text and images is a primary objective, either by mapping them to logical rules, or re-defining the merge and similarity function in different domains. In addition, future investigations could be directed to the development of different merging functions and stopping criteria. An obvious extension is to study how to consider non-logic explanations and the application to other families of black boxes. Images and text may be good stride in this direction. Finally, an interesting future research direction is to study how to provide more informative *causal explanations*, able to capture the causal relationships among the (endogenous as well as exogenous) variables and the decision, based on data observed by appropriately querying the black box.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This work is partially supported by the European Community H2020 programme under the funding schemes: H2020-INFRAIA-2019-1: Research Infrastructure G.A. 871042 *SoBigData++* (sobigdata.eu), G.A. 952215 TAILOR, G.A. 952026 *Humane AI* NET (humane-ai.eu), G.A. 825619 *AI4EU* (ai4eu.eu), and the ERC-2018-ADG G.A. 834756 “XAI: Science and technology for the eXplanation of AI decision making”.

References

[1] F. Pasquale, The Black Box Society: The Secret Algorithms That Control Money and Information, Harvard University Press, 2015.
[2] T. Miller, Explanation in artificial intelligence: insights from the social sciences, Artif. Intell. 267 (2019) 1–38.
[3] A.A. Freitas, Comprehensible classification models: a position paper, ACM SIGKDD Explor. Newsl. 15 (1) (2014) 1–10.
[4] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, ACM Comput. Surv. 51 (5) (2018) 93.
[5] L.K. Jeff Larson, Surya Mattu, J. Angwin, How we analyzed the compas recidivism algorithm 2013.
[6] S. Wachter, et al., Why a right to explanation of automated decision-making does not exist in the general data protection regulation, Int. Data Priv. Law 7 (2) (2017) 76–99.
[7] G. Malgieri, G. Comandé, Why a right to legibility of automated decision-making exists in the general data protection regulation, Int. Data Priv. Law 7 (4) (2017) 243–265.
[8] R. Guidotti, et al., Helping your docker images to spread based on explainable models, in: ECML-PKDD, Springer, 2018.
[9] R. Goebel, A. Chander, K. Holzinger, F. Lecue, Z. Akata, S. Stumpf, P. Kieseberg, A. Holzinger, Explainable AI: the new 42?, in: International Cross-Domain Conference for Machine Learning and Knowledge Extraction, Springer, 2018, pp. 295–303.
[10] A. Adadi, M. Berrada, Peeking inside the black-box: a survey on explainable artificial intelligence (XAI), IEEE Access 6 (2018) 52138–52160.
[11] M.T. Ribeiro, S. Singh, C. Guestrin, Why should I trust you?: explaining the predictions of any classifier, in: KDD, ACM, 2016, pp. 1135–1144.
[12] M.T. Ribeiro, S. Singh, C. Guestrin Anchors, High-precision model-agnostic explanations, in: AAAI, 2018.
[13] R. Guidotti, A. Monreale, F. Giannotti, D. Pedreschi, S. Ruggieri, F. Turini, Factual and counterfactual explanations for black box decision making, IEEE Intell. Syst. 34 (6) (2019) 14–23, https://doi.org/10.1109/MIS.2019.2957223.
[14] C. Panigutti, et al., Explaining multi-label black box classifiers for health applications, in: W3PHIAI, Springer, 2019.
[15] M. Craven, J.W. Shavlik, Extracting tree-structured representations of trained networks, in: NIPS, 1996, pp. 24–30.
[16] H. Deng, Interpreting tree ensembles with intrees, arXiv preprint arXiv:1408.5456.
[17] X. Yin, J. Han, Cpar: classification based on predictive association rules, in: Proceedings of the 2003 SIAM International Conference on Data Mining, SIAM, 2003, pp. 331–335.
[18] H. Lakkaraju, et al., Interpretable decision sets: a joint framework for description and prediction, in: KDD, ACM, 2016, pp. 1675–1684.
[19] A.S. Ross, W. Pan, F. Doshi-Velez, Learning qualitatively diverse and interpretable rules for classification, arXiv preprint arXiv:1806.08716.
[20] A.S. Ross, W. Pan, L.A. Celi, F. Doshi-Velez, Ensembles of locally independent prediction models, arXiv preprint arXiv:1911.01291.
[21] D. Pedreschi, F. Giannotti, R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, Meaningful explanations of black box AI decision systems, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, 2019, pp. 9780–9784.
[22] R. Guidotti, A. Monreale, F. Giannotti, D. Pedreschi, S. Ruggieri, F. Turini, Factual and counterfactual explanations for black box decision making, IEEE Intelligent Systems.
[23] D. Alvarez-Melis, T.S. Jaakkola, On the robustness of interpretability methods, arXiv preprint arXiv:1806.08049.
[24] R. Guidotti, S. Ruggieri, On the stability of interpretable models, in: 2019 International Joint Conference on Neural Networks (IJCNN), IEEE, 2019, pp. 1–8.
[25] J.R. Quinlan, C4. 5: Programs for Machine Learning, Elsevier, 1993.
[26] J. Yoon, D.-W. Kim, Classification based on predictive association rules of incomplete data, IEICE Trans. Inf. Syst. 95 (5) (2012) 1531–1535.
[27] J. Schmidhuber, Deep learning in neural networks: an overview, Neural Netw. 61 (2015) 85–117.
[28] H. Lakkaraju, E. Kamar, R. Caruana, J. Leskovec, Interpretable & explorable approximations of black box models, arXiv preprint arXiv:1707.01154.
[29] E. Angelino, N. Larus-Stone, D. Alabi, M. Seltzer, C. Rudin, Learning certifiably optimal rule lists, in: KDD, ACM, 2017, pp. 35–44.
[30] S. Ruggieri, Yadt: yet another decision tree builder, in: Tools with Artificial Intelligence, ICTAI, IEEE, 2004, pp. 260–265.
[31] M.W. Craven, J.W. Shavlik, Using sampling and queries to extract rules from trained neural networks, in: JMLR, Elsevier, 1994, pp. 37–45.
[32] P.-N. Tan, et al., Introduction to Data Mining, Pearson Education India, 2007.
[33] S.M. Lundberg, S. Lee, A unified approach to interpreting model predictions, in: I. Guyon, U. von Luxburg, S. Bengio, H.M. Wallach, R. Fergus, S.V.N. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4–9 December 2017, Long Beach, CA, USA, 2017, pp. 4765–4774.
[34] L.S. Shapley, A value for n-person games, Contrib. Theor. Games 2 (28) (1953) 307–317.
[35] L. Rokach, O. Maimon, Clustering methods, in: Data Mining and Knowledge Discovery Handbook, Springer, 2005, pp. 321–352.
[36] E. Wit, E.v.d. Heuvel, J.-W. Romeijn, 'All models are wrong...': an introduction to model uncertainty, Stat. Neerl. 66 (3) (2012) 217–236.
[37] D. Pelleg, A.W. Moore, et al., X-means: extending k-means with efficient estimation of the number of clusters, in: ICML, vol. 1, 2000, pp. 727–734.
[38] R. Guidotti, R. Trasarti, M. Nanni, Tosca: two-steps clustering algorithm for personal locations detection, in: Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems, 2015, pp. 1–10.
[39] J. Fürnkranz, Pruning algorithms for rule learning, Mach. Learn. 27 (2) (1997) 139–172, https://doi.org/10.1023/A:1007329424533.
[40] S. Ruggieri, Learning from polyhedral sets, 2013.
[41] D. Bertsimas, J. Dunn, Optimal classification trees, Mach. Learn. 106 (7) (2017) 1039–1082.
[42] A. Richardson, Nonparametric statistics for non-statisticians: a step-by-step approach by Gregory W. Corder, dale I. Foreman, Int. Stat. Rev. 78 (3) (2010) 451–452.