Confident Interpretations of Black Box Classifiers

Nedeljko Radulovic
LTCI, Télécom Paris
Institut Polytechnique de Paris
Paris, France
nedeljko.radulovic@telecom-paris.fr

Albert Bifet
LTCI, Télécom Paris
AI Institute,University of Waikato
France, New Zealand
albert.bifet@telecom-paris.fr

Fabian Suchanek
LTCI, Télécom Paris
Institut Polytechnique de Paris
Paris, France
suchanek@telecom-paris.fr

***Abstract*—Deep Learning models provide state of the art classification results, but are not human-interpretable. We propose a novel method to interpret the classification results of a black box model a posteriori. We emulate the complex classifier by surrogate decision trees. Each tree mimics the behavior of the complex classifier by overestimating one of the classes. This yields a global, interpretable approximation of the black box classifier. Our method provides interpretations that are at the same time general (applying to many data points), confident (generalizing well to other data points), faithful to the original model (making the same predictions), and simple (easy to understand). Our experiments show that our method beats competing methods in these desiderata, and our user study shows that users prefer this type of interpretations over others.**

I. Introduction

Recent years have seen the rise of powerful predictive models. These include, e.g., Neural Networks (NN) or Random Forests (RF). In some tasks, these models have proven to have better performance than humans. The problem with these models is that they are black box models, i.e., it is not possible to understand the logic behind their decision-making process. This means that we cannot assess whether the model is performing well just because of a lucky guess or because the model has learned the patterns and dependencies in the data. For this reason, these models cannot be applied to critical tasks in security, health, or justice, or more generally in any situation where the user or citizen needs some level of understanding of the algorithmic prediction. Indeed, the General Data Protection Regulation [1] (GDPR) of the European Union and other legislation are increasingly imposing some form of explanation in algorithmic recommendations.

We are thus trapped in a dilemma where we have powerful predictive models at our disposal, but cannot use them since they are not interpretable. For this reason, the scientific community now studies Explainable Artificial Intelligence (`xAI`) – the attempt to make predictive models interpretable. We focus here on classification models, which take as input a data point, and classify it into one out of several predefined classes. There exist several techniques to make such models interpretable. One popular family of approaches builds *surrogate models* – simple and interpretable models that mimic the behavior of the black box model. These can be, e.g., decision trees [2]–[5] or local linear surrogates [6]. These models are more amenable to human understanding.

There is considerable debate about what constitutes a "good" surrogate model [7]–[13]. A common criterion is *fidelity*, i.e., the percentage of the predictions on which the surrogate model agrees with the black box model. A second common metric is *complexity*: the interpretation provided by the surrogate model has to be less complex than the black box model. In the case of a decision tree, e.g., the complexity of the interpretation is the depth of the tree [8]. In this paper, we argue that a third important metric is *generality*: the number of data points that the interpretation applies to. If an interpretation applies to more data points, it appears less ad-hoc to a user. The interpretation should also have a high *confidence*, i.e., all data points concerned by the interpretation should be classified in the same way.

It is intuitively clear that these desiderata are pitted against each other: Higher fidelity means higher complexity. This problem is known as the *comprehensibility-complexity trade-off*. In the same spirit, higher generality means lower confidence (akin to the precision-recall trade-off). Finally, higher generality at low complexity also means lower fidelity.

In this paper, we propose a new methodology that addresses this impasse: We propose to mimic a given black box classifier not by a single surrogate model, but by several – one for each class. In this way, each of the models can be simple while their combination still has high fidelity. Our method has not just a high fidelity and a low complexity, it also provides very general interpretations at high confidence. Our main contributions are as follows:

- We develop an abstraction of surrogate models, and formalise quality metrics of surrogate models.
- We present our method STACI[1], which learns surrogate models that are at the same time simple, general, confident, and faithful to the original.
- We perform an extensive empirical evaluation on several popular datasets from the UCI Machine Learning Repository, showing that STACI outperforms other state-of-the-art methodologies in these desiderata.
- We perform a user study that shows that users prefer the interpretations of our method over others.

The rest of the paper is organised as follows. Section II discusses related work. Section III presents our new method,

[1]Surrogate Trees for A posteriori Confident Interpretations

STACI. Section IV evaluates our method on several datasets and compares it to a baseline and the state of the art, before Section V concludes.

II. Related work

Several approaches aim to produce human-interpretable ML models. Some approaches propose readily interpretable models, such as decision trees (CART [14]), rule-based models [15], Scalable Bayesian Rule Lists [16], or linear models [17].

Other approaches approximate a given black box model *post-hoc* by an interpretable model. Among these, local post-hoc models provide an interpretation for a given input data point. LIME [6], e.g., provides local explanations by training an interpretable linear model around individual data points. The explanation is given in the form of a list of the most relevant features with their weights. The weights are proportional to the feature's contribution to the outcome probability. LIME also provides a sampling algorithm that combines multiple explanations for individual data points to provide a global explanation. The same authors proposed another approach [18], called Anchors. Anchors are also local approximations of the complex model that provide explanations in form of sufficient *if ... then* rules. These rules include only the features that influence the outcome, i.e. changing the value of other features will not have an impact on the outcome.

SHAP (SHapley Additive exPlanations) [19] proposes using Shapley values from game theory as a unified measure of feature importance, and shows that previous methods use an approximation of this measure. [20] is the first work that addresses the quality of the explanations, providing counterfactual explanations for each data point and measuring the fidelity of each explanation. It uses the idea of *b-counterfactuals*, which represent the minimal change in the feature in order to gauge the prediction of the complex model. Then, a regression model is fitted in the neighborhood of the data point to find the best explanation.

Global approaches, in contrast, approximate the black box model by a single global surrogate model. One of the first global methods was TREPAN [4], which queries the complex model in order to train a decision tree that mimics its behavior. Dectext [3] refined this idea by using different types of splits in the decision tree and a specific tree pruning strategy to improve its fidelity. Another method based on decision trees [5] uses a genetic programming algorithm to sample new data points, which are then used to learn the behavior of the black box model. The most recent global approach [2] proposes DTExtract, a method that first fits a mixture of axis-aligned Gaussians to estimate the input distribution over features of the training dataset. Then the method builds binary decision trees iteratively, using an active sampling strategy.

As mentioned before, the main challenge when building surrogate models is the trade-off between fidelity and complexity. There are different approaches to tackle this trade-off: limiting the number of nodes in the surrogate tree [2], [4], applying specific pruning algorithms [3], or stopping the growth of the tree when a node covers the instances of only one class with high probability [4].

In contrast to all of these works, our approach builds not a single surrogate model, but one per class. This allows us to achieve high fidelity and low complexity without corrupting the resulting models through pruning. In our experiments, we compare our approach to the state of the art global post-hoc method, DTExtract [2]. We also compare to a model that is interpretable by design (Scalable Bayesian Rule Lists [16]), to LIME [6], and to CART [14] as a baseline.

III. Our Approach

A. Interpretation Models

Post-hoc Interpretation. We are given a black box multi-class classification model. We wish to make it interpretable *post-hoc*, i.e., after the model has been trained. There is considerable debate about the meanings of the terms "explainable" and "interpretable" [7]–[11]. In this paper, we aim at interpretability in the following sense: *We want to provide a meaning for the results of a model in terms that are understandable to humans* [11].

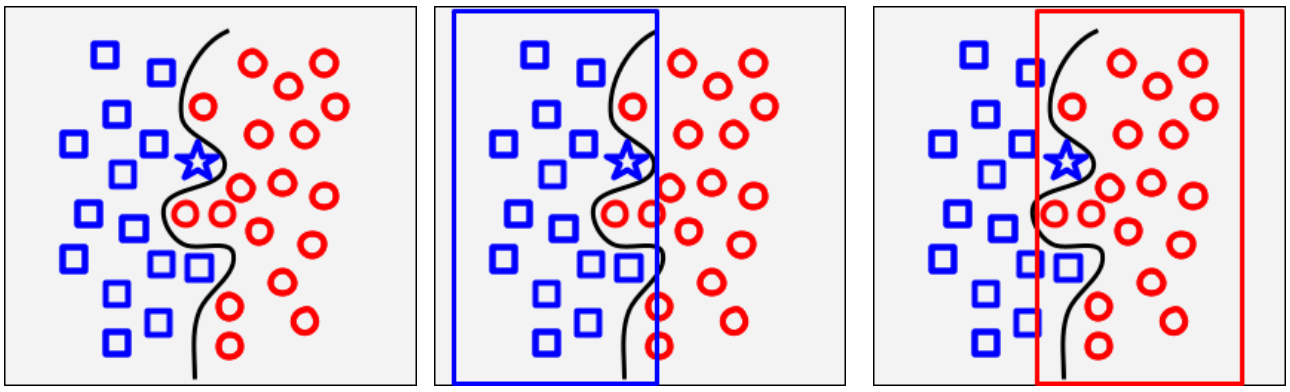
Local vs. global interpretations. Local interpretations help us understand the classification of one given input data point ("Why does the model predict that this particular patient should undergo chemotherapy?"). While these interpretations can be very tailored, they are less well adapted for scenarios where the model is used repeatedly: Local interpretations can be unstable, and can provide very different interpretations even in a very close neighborhood [21]. Individual interpretations may also be contradictory to each other [22]. Global interpretations, in turn, consider the model as a whole ("What are the criteria that make patients more likely to be recommended chemotherapy in general?"). Such interpretations can also help understand an individual classification, but they are more geared towards an understanding of the model as a whole. In this paper, we study global interpretations.

Interpretation Models. Formally, we aim at global post-hoc interpretations of the following form:

Definition III.1. *Given a set $\mathcal{S}$ of labeled data points and a labeled input point $i \in \mathcal{S}$, an* interpretation *of point i is a set of conditions that i satisfies, so that the majority of the data points in $\mathcal{S}$ that satisfy these conditions carry the same label as i.*

An *interpretation model* is then a model that can provide such interpretations. This definition applies to a wide variety of models, be it decision trees, Bayesian rule lists, linear models, or our own method. Let us now discuss some quality metrics of such models.

Fidelity. All post-hoc approaches have the problem that the interpretation model usually deviates from the black box model, because it has to be simpler than the black box model. If the model deviates for a given point, the approach cannot deliver an interpretation for this point. We formalize this notion by the concept of *fidelity*.



(a) Complex model (b) Interpretable model for left class (c) Interpretable model for right class

Fig. 1: In order to approximate the complex decision boundary of the black box model (Figure 1a), we train two specialised surrogate models: One, shown in Figure 1b, overgeneralizes the "square" class. The other, shown in Figure 1c, overgeneralizes the "circle" class.

Definition III.2. *Given an interpretation model $\mathcal{M}$, and given a labeled dataset $\mathcal{S}$, the* fidelity *is the ratio of points of $\mathcal{S}$ for which the model $\mathcal{M}$ can provide interpretations.*

Thus, the fidelity is just the ratio of points on which the surrogate model agrees with the complex model. In the case of a decision tree trained on $\mathcal{S}$, the fidelity is just the weighted average confidence of the leaf nodes.

Confidence. An interpretation will identify some characteristics of the input point, and say that the majority of points with these characteristics are classified in a certain way. Naturally, such an interpretation is more convincing when that majority is larger. To quantify this intuition, we define the notion of confidence:

Definition III.3. *Given an interpretation model $\mathcal{M}$, a labeled dataset $\mathcal{S}$, and a labeled input data point $i \in \mathcal{S}$ that the model can interpret, the* confidence *at point i is the ratio of data points of $\mathcal{S}$ that satisfy the interpretation of i and share the label of i over the data points of $\mathcal{S}$ that satisfy the interpretation of i.*

This definition can be generalized to the *average confidence of the model $\mathcal{M}$* on the set $\mathcal{S}$, which is simply the average confidence for all points of $\mathcal{S}$. In the case of a decision tree trained on $\mathcal{S}$, the average confidence is just the weighted average confidence of the leaf nodes (and thus identical to fidelity). In general, however, the fidelity gauges the percentage of data points where the model applies (no matter their class). The confidence, in contrast, measures whether the data points concerned by the interpretation are of the same class. In this way, confidence corresponds to the precision of the interpretation on the set of cases to which it applies. Optimizing only confidence for a given class may reduce the fidelity.

Generality. An interpretation will be more convincing if it applies to more input data points. For example, imagine an interpretation model in the medical domain that provides interpretations based on the social security number of the patient. This model will have a high fidelity and a high confidence, because it just "explains" the illness of the patient by her social security number. To avoid such interpretations, we need the notion of generality:

Definition III.4. *Given an interpretation model $\mathcal{M}$, a labeled dataset $\mathcal{S}$, and a labeled input data point $i \in \mathcal{S}$ that the model can interpret, the* generality *at point i is the ratio of data points in $\mathcal{S}$ that satisfy the interpretation of i and share the label of i over all the data points of $\mathcal{S}$ that share the label of i.*

The larger that percentage, the more satisfactory the interpretation will be. We define generality as a measure relative to the class size in order to be scale-invariant, and in order to guard against skewed class distributions. Generality thus corresponds to the recall of the interpretation on the set of all data points with the same label.

Complexity. The goal of an interpretation model is to provide interpretations that are as simple as possible. Intuitively, the complexity of an interpretation corresponds to the number of conditions it contains: Simpler interpretations have fewer conditions. Formally, the *complexity* of an interpretation depends on the type of the interpretation model. For decision trees, the complexity of an interpretation is usually the length of the path from the root to the leaf node [8]. The worst case complexity of a tree is the maximal depth.

B. Our Approach

Goal. Our approach receives as input a black box classification model. It produces as output a surrogate model, which makes (by and large) the same predictions as the black box model, but is simpler and thus easier to understand. When this surrogate model is presented with a new data point of a class C, it will produce an interpretation such as: *"This data point has the characteristics X, Y, Z and was classified as C. There are 500 other data points with these characteristics, and 80% of them are also classified as C."*

Approach. The key idea of our method is to generate not a single surrogate model, but one specialized surrogate model for each class. Each specialized model is trained to overgeneralize its class. Figure 1 exemplifies this for a binary

classification model: the model (a) is the black box model. The model (b) overgeneralizes the "square" class, while the model (c) overgeneralizes the "circle" class. We then interpret an incoming data point in two steps:

1) We first classify the data point by the black box model. This may appear to be cheating, but the goal is not to replicate the black box model in its absence. Rather, the goal is to interpret a prediction of the black box model. Thus, this prediction is necessarily available. In Figure 1, if we receive the point marked by a star, we classify it as "square".
2) We then use the specialized surrogate model for that class to provide an interpretation. In the example, we use the model of Figure 1 (b). The result is an interpretation such as: This data point has the characteristics of Model (b), and was classified as "square"; there are 30 other data points with these characteristics, and 80% of them are also classified as "square".

Fulfilment of the Desiderata. For the specialized models, we use decision trees with limited depth. This entails that our interpretations have a *low complexity*. Since we have one specialized tree per class, their combination is still sufficiently complex to approximate the black box model. We thus achieve a *high fidelity*. Finally, our trees are constructed in such a way that they maximize both the percentage of correctly classified points (the *confidence*) and the percentage of points of the target class (the *generality*).

In general, training more trees leads to higher confidence, less complexity per tree, and higher fidelity. Thus, one could think that one should simply train many more trees. However, more trees lead to a decrease in generality. If we train only one tree per class, in contrast, this does not lead to a decrease in generality. This is because generality is computed per class. Furthermore, one tree per class satisfies our goal of providing global interpretations: A single tree provides a human-understandable interpretation of a single, entire class.

Let us now detail the construction of our trees.

C. Training algorithm

We are given a black box N-class classification model, and a set $\mathcal{S}$ of data points. We want to construct a surrogate model that can interpret these points. As is customary, we use the black box model to label the data points $\mathcal{S}$. On this labeled dataset, we train one decision tree per class in a *one-vs-all* fashion.

Our goal is to train our class-specific trees in such a way that they maximize confidence and generality. For this purpose, we do not use standard metrics such as the Gini Impurity Index or the Information Gain when deciding a split on a node in the decision tree. Rather, we employ the *F1 score*.

The F1 score is a measure of the model's accuracy on the dataset and is defined as a harmonic mean of precision and recall:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{1}$$

Precision and recall are defined for a binary classification scenario. Precision is the ratio of true positives (TP), i.e. of positive data points that model has classified as such, over all data points that the model has classified as positive. Recall is the ratio of true positives, i.e., the ratio of positive data points that model has classified as such, over all positive data points in the dataset:

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

where FP and FN represent false positives and false negatives, respectively.

If we recall the definitions of *confidence* (Def. III.3) and *generality* (Def. III.4), we can see that they directly correspond to the definitions of precision and recall, respectively, in our scenario of one tree per class. It is worth mentioning that optimizing only one of these two metrics would lead to pathological solutions. Optimizing only precision would result in splits that are highly confident but that would not cover a lot of data points (i.e., the generality would be low). In the extreme case, the split would single out only one data point, resulting in a confidence of 100%. On the other hand, optimizing only for recall, the split would tend to encapsulate as many points of one class as possible, including many points of the other classes, thus significantly degrading the confidence. The F1 score is the harmonic mean of precision and recall, and is thus a natural metric to optimize both desiderata together. It complies with our approach in multiple aspects:

- It is an asymmetric metric (false positives and false negatives count differently), and thus it fits our strategy of training one surrogate model per class in a *one-vs-all* manner.
- It is a widely used metric for the evaluation of classifiers, and optimizing it will thus, by definition, not harm *fidelity*.
- It optimizes both *confidence* and *generality*.
- It does not require any user-defined parameters.

In fact, our training algorithm has just one user defined input: the desired complexity of the interpretation, i.e. the maximal depth of the surrogate decision trees. Our method trains one decision tree per class on the labeled $\mathcal{S}$, using the F1 score to decide node splits, and limiting the depth of the trees as specified by the user. The tree growth stops when the gain in the split metric is less or equal to zero or when the maximal depth has been reached. The output of our algorithm is set of decision trees, one for each class.

Let us now discuss how we can provide an interpretation for a given input data point. Algorithm 1 receives as input the black box model M, a data point x, and the surrogate trees $\{T_c\}_{c=1}^{N}$ for each of the N classes. We first use M to classify x. Then we check if the corresponding surrogate tree T_c agrees with M. If so, we produce an interpretation: The characteristics of the data point can be read off the path

TABLE I: Datasets

Dataset	Features	Num.	Cat.	Classes	Instances
Heart	13	6	7	2	303
Breast	31	31	0	2	569
Diabetes	8	8	0	2	768
Voting	16	0	16	2	435
Sick	29	7	22	2	2800
Hypothyroid	25	7	18	2	3163
Adult[2]	11	5	6	2	30162
Wine	13	13	0	3	179
Dermatology	34	33	1	6	358
Vehicle	14	18	0	4	846

from the root of T_c to the leaf for x. The generality and the confidence can be found directly at the leaf node. If T_c does not agree with M on x, we are in an area that fell victim to the constraint of limited complexity, and we cannot provide an interpretation. (If many data points fall in this area, fidelity suffers.)

Algorithm 1 Interpret with STACI

Input: black box Model: M
Data point x
Surrogate trees $\{T_c\}_{c=1}^{N}$

1: $c = M(x)$
2: **if** $T_c(x) = c$ **then**
3: **return** "This data point has the characteristics T_c.*pathFor*(x), and is classified as c. There are T_c.*numSamples*(x)-1 other data points with these characteristics, and T_c.*leafConf*(x)% of them are also classified as c."
4: **end if**
5: **return** "STACI cannot provide an interpretation."

IV. EXPERIMENTS

We evaluate our approach on the datasets of the UCI Machine Learning repository. We compare the performance of our approach with the state of the art method DTExtract [2], the interpretable-by-design method SBRL [16], LIME [6], and CART [14] as a baseline. We have also conducted an user study to validate the desiderata defined in Section III-A and to determine the users' preferences.

A. Settings

Datasets. All the datasets used in our experiments are publicly available at the UCI Machine Learning Repository [23]. Table I shows their key characteristics.

Metrics. We report the four metrics from Section III-A: The complexity is the average depth of the path for the decision trees, the number of rules for SBRL, and number of features in the explanation for LIME. The confidence is the average confidence on the decision path of the decision tree, or the confidence of the firing rule in case of SBRL, or the confidence of the set of conditions in case of LIME. The fidelity is the percentage of data points where the model agrees with the black box model. The generality is the percentage of the data points of one class that is covered by an interpretation.

Black box Models. For every dataset, we train two different black box models, a *Neural Network* and a *Random Forest*. We use the implementations of the scikit-learn Python package [24]. We train the *Multi Layer Perceptron* classifier with 500 nodes in the hidden layer and the *Random Forest* classifier with 1000 trees. We use 10% of the data as test set.

Systems. For DTExtract[3] and SBRL[4], we use the code published by the authors. For CART, we train the decision tree with limited depth, as we do for our method. We train SBRL using the default parameters and setting the length of the rule list to 10. Since this method supports only categorical features, we discretize the numerical features. Also, SBRL doesn't support the multi-class scenario, and we thus cannot provide results for the Wine, Dermatology and Vehicle datasets. For LIME[5], we use its global version, where `Submodular pick` algorithm is used to provide multiple explanations that represent the black box model as a whole. We fix the number of features to the maximal number of conditions. We allow an arbitrary number of explanations for each dataset, except for Adult, Sick and Hypothyroid datasets, where we limit the number to 1000. To validate our method, we also evaluate the fidelity of a modified STACI method, called STACI'. This method does not have access to the black box model at testing time. For each prediction, it computes the average confidence along the decision path of each surrogate tree. The confidence of a node in the tree is computed as the ratio of correctly classified data points by the split on that node over the total number of data points on that node. Finally, the method uses the tree with the highest average confidence to make the prediction. Thus, STACI' is a kind of disadvantaged variant of STACI, which has to make do without access to the black box model at testing time.

We train our models on 90% of the dataset, run all experiments on 20 random train/test splits, and report the averaged results. The code for our approach, as well as all experimental data, is available at https://github.com/nedRad88/STACI.

Fidelity. The results of the fidelity comparison are shown in Table II for the NN model, and in Table III for the RF model. We can see that our method, STACI, outperforms all competitors in most of the cases. This is not surprising, because STACI has one tree per class to ensure fidelity. Even our disadvantaged method STACI' outperforms competitors in several cases, while in others it has comparable performance. This means that our training algorithm successfully ensured a high fidelity of the specialized surrogate trees.

Complexity. Table IV shows the average complexity of the surrogate models. CART and LIME always have the same

[2]We used a subset of Adult dataset, removing less relevant features (race and native-country)

[3]https://github.com/obastani/dtextract

[4]https://github.com/Hongyuy/sbrl-python-wrapper

[5]https://github.com/marcotcr/lime

TABLE II: Fidelity (%) with NN as black box model

Dataset	DTE	SBRL	LIME	CART	STACI'	STACI
Heart	**87.34**	85.88	84.84	80.97	79.68	84.84
Breast	**94.93**	91.57	87.28	89.65	91.05	93.16
Diabetes	80.58	83.38	71.49	75.19	76.23	**84.55**
Voting	**95.91**	94.55	95.34	95.34	94.55	95.00
Sick	97.88	97.25	75.36	96.66	97.79	**98.46**
Hypo.	96.39	97.88	94.32	98.99	**98.45**	**99.31**
Adult	92.35	93.88	87.56	73.53	**98.23**	**99.58**
Wine	91.11	**N/A**	52.78	66.67	86.67	**97.78**
Derma.	94.86	**N/A**	82.70	80.28	**95.28**	**96.11**
Vehicle	74.47	**N/A**	54.71	69.06	68.24	**86.35**

TABLE III: Fidelity (%) with RF as black box model

Dataset	DTE	SBRL	LIME	CART	STACI'	STACI
Heart	87.10	88.06	91.13	86.94	83.87	**92.90**
Breast	96.32	92.21	89.82	96.49	92.63	**97.89**
Diabetes	87.86	87.01	71.56	85.00	81.82	**94.16**
Voting	97.96	96.59	98.07	97.05	**98.86**	**98.86**
Sick	**99.43**	94.61	73.93	99.20	97.86	98.46
Hypo.	98.97	95.93	93.85	99.45	99.31	**99.96**
Adult	83.89	87.56	80.36	89.69	85.73	**96.74**
Wine	91.67	**N/A**	62.78	93.89	91.11	**96.67**
Derma.	**96.75**	**N/A**	76.98	90.00	93.06	96.39
Vehicle	74.35	**N/A**	58.33	72.94	70.82	**86.11**

complexity – simply because we set the maximal depth of the trees (in CART) and the maximal number of conditions (in LIME) to the same value as maximal depth of the tree for our approach. We show that interpretations provided by our approach are usually shorter than the ones of DTExtract and SBRL.

At the same time, the complexity of STACI remains always limited, even in the worst case. That is not the case for our competitors: Table V shows the maximal complexity of an interpretation. In the worst case, DTExtract and SBRL will deliver interpretations of more than 10 conditions. STACI, in contrast, always delivers simple interpretations, as the maximum tree depth is fixed. The method can still keep a high fidelity because it uses multiple surrogate trees.

Confidence. Tables VI and VII show the average confidence of the interpretations for the two different black box models. As we can see, STACI gives interpretations with higher confidence in most of the cases.

Generality. We compare the generality of our model and DTExtract in Table VIII. STACI has higher generality in most of the cases. Even though there is a clear trade-off between confidence and generality (or precision and recall), the results show that our specific training strategy and choice of $F1$ measure successfully solves this challenge.

In summary, our method outperforms the other methods in terms of all four aforementioned criteria: fidelity, complexity, confidence and generality. Thus, STACI overcomes the trade-off between confidence and generality, achieves higher fidelity, and still never delivers long interpretations. Figure 2 shows an example interpretation given by our system.

TABLE IV: Average Complexity

Dataset	Black	DTE	SBRL	LIME	CART	STACI
Heart	NN	3.15	3.90	3	3	**2.89**
	RF	3.11	**2.29**	4	4	3.28
Breast	NN	2.88	4.20	3	3	**1.9**
	RF	3.18	6.16	4	4	**2.88**
Diabetes	NN	2.89	5.78	3	3	**1.49**
	RF	2.75	7.21	4	4	**1.85**
Voting	NN	3.11	**1.57**	3	3	1.58
	RF	3.00	**1.63**	3	3	1.69
Sick	NN	2.40	3.64	3	3	**1.40**
	RF	2.25	3.77	3	3	**2.07**
Hypo.	NN	2.58	4.50	3	3	**1.20**
	RF	2.16	4.78	3	3	**1.09**
Adult	NN	3.25	8.49	4	4	**1.87**
	RF	2.75	7.22	4	4	**1.83**
Wine	NN	3.95	**N/A**	3	3	**2.42**
	RF	4.29	**N/A**	4	4	**2.93**
Derma.	NN	4.91	**N/A**	3	3	**2.24**
	RF	4.85	**N/A**	4	4	**2.36**
Vehicle	NN	3.99	**N/A**	3	3	**2.68**
	RF	4.50	**N/A**	4	4	**2.91**

TABLE V: Maximal Complexity

Dataset	Black	DTE	SBRL	LIME	CART	STACI
Heart	NN	9.30	7.00	3	3	**3**
	RF	10.80	5.4	4	4	**4**
Breast	NN	10.30	6.55	3	3	**3**
	RF	9.4	11.00	4	4	**3**
Diabetes	NN	9.50	10.40	3	3	**3**
	RF	10.2	11.90	4	4	**4**
Voting	NN	10.85	3.80	3	3	**3**
	RF	11.30	4.10	3	3	**3**
Sick	NN	10.50	5.60	3	3	**3**
	RF	10.60	7.60	3	3	**3**
Hypo.	NN	9.7	6.40	3	3	**3**
	RF	10.00	6.10	3	3	**3**
Adult	NN	10.6	12.20	4	4	**4**
	RF	10.8	18.85	4	4	**4**
Wine	NN	7.8	**N/A**	3	3	**3**
	RF	6.7	**N/A**	4	4	**4**
Derma.	NN	8	**N/A**	3	3	**3**
	RF	8.5	**N/A**	4	4	**4**
Vehicle	NN	6.4	**N/A**	3	3	**3**
	RF	6.4	**N/A**	3	3	**3**

TABLE VI: Confidence (%) of the interpretations (NN)

Dataset	DTE	SBRL	LIME	STACI
Heart	85.04	87.83	78.80	**88.57**
Breast	94.18	93.47	88.95	**95.57**
Diabetes	81.71	**84.50**	60.11	83.47
Voting	95.17	95.56	94.84	**95.91**
Sick	97.17	96.55	77.96	**97.83**
Hypo.	97.38	97.63	86.92	**98.98**
Adult	92.46	93.85	75.34	**98.20**
Wine	88.58	**N/A**	90.36	**92.23**
Derma.	78.63	**N/A**	53.12	**89.21**
Vehicle	66.02	**N/A**	40.41	**74.98**

TABLE VII: Confidence (%) of the interpretations (RF)

Dataset	DTE	SBRL	LIME	STACI
Heart	82.28	81.75	**85.74**	80.38
Breast	93.49	91.86	95.33	**95.52**
Diabetes	76.23	77.33	68.00	**80.22**
Voting	**96.50**	95.30	94.98	95.89
Sick	97.61	93.90	74.55	**97.79**
Hypo.	98.08	95.81	95.82	**99.07**
Adult	80.12	82.21	82.09	**87.53**
Wine	88.84	**N/A**	59.82	**93.48**
Derma.	78.83	**N/A**	64.48	**90.00**
Vehicle	59.45	**N/A**	52.99	**61.38**

TABLE VIII: Generality comparison and counterfactuality

Dataset	Black	DTE	STACI	Counterfactuality
Heart	NN	59.21	**76.63**	66.81
	RF	58.83	**68.35**	64.63
Breast	NN	80.31	**92.59**	75.20
	RF	84.82	**88.67**	7.88
Diabetes	NN	66.92	**74.47**	72.33
	RF	64.23	**71.51**	90.99
Voting	NN	73.37	**95.01**	64.39
	RF	82.14	**95.15**	69.89
Sick	NN	**94.70**	94.18	17.14
	RF	93.39	**94.65**	30.44
Hypo.	NN	89.62	**97.08**	10.99
	RF	96.79	**96.94**	15.32
Adult	NN	92.06	**95.53**	52.35
	RF	**92.25**	73.84	42.12
Wine	NN	77.03	**86.67**	69.38
	RF	79.51	**85.12**	22.27
Derma.	NN	**91.74**	91.33	12.99
	RF	**91.54**	**91.54**	9.11
Vehicle	NN	53.98	**68.70**	48.21
	RF	46.16	**55.54**	27.39

B. Counterfactuality

In this section we discuss another property of interpretations: counterfactuality. An interpretation for a data point is *counterfactual* if the following is true: If we modify the data point in such a way that the conditions of the interpretation no longer hold, then the black box model classifies the data point differently. Counterfactuality is a very attractive property, and counterfactual interpretations have been considered tantamount to explanations [13], [25]. That said, counterfactuality alone is not sufficient: A counterfactual interpretation could just pose a condition that is so extreme that it is guaranteed to catapult the data point out of its current class – as in "You suffer from senility because you are not a baby. If you were a baby, you would not be senile." Such explanations are obviously absurd. Therefore, counterfactuality is always accompanied by the requirement to find the smallest modification that changes the class of the data point [25] – as in "You suffer from senility because you are older than 100 years." Counterfactuality in this sense is not possible in the global setting, because it inherently depends on the individual data point. Therefore, there are no guarantees that our approach will provide such explanations. However, we can assess counterfactuality a posteriori. We report the counterfactuality as the percentage of data points for which the prediction of the black box model changes when we modify the data point so that it no longer satisfies the conditions of the interpretation. The results are shown in Table VIII. As we can see, our approach is able to achieve a respectable ratio of counterfactuality, despite not being designed for it.

The datapoint

Pregnancies	5
Glucose	166
Blood pressure	72
Skin thickness	19
Insulin	175
BMI	25.8
Diabetes pedigree	0.59
Age	51

is classified as diabetic. It has these characteristics:

Glucose>154, Insulin>145, Age>30

There are 37 other data points with these characteristics, and 94.59% of them are also classified as diabetic.

Fig. 2: Example of a STACI interpretation

C. User study

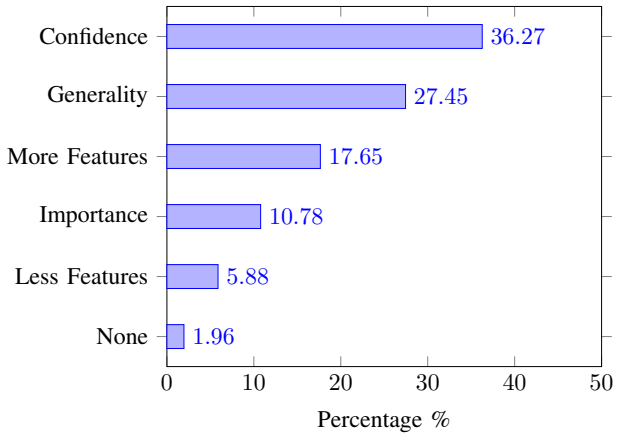
To evaluate which characteristics of the interpretations are subjectively most valuable, we conducted a user study. We used the Diabetes dataset, and trained a Random Forest with 1000 trees as black box model. Then, we trained three interpretable models: DTExtract, LIME and STACI. The length of the interpretation was set to 3 for both LIME and STACI. For LIME we show only the features that had positive influence (weight) on the outcome. Each participant of the user study was invited to look at a single patient. We showed the data of this patient and the model prediction, and we proposed the interpretations provided by the three approaches. In this study, we are not interested in the visual representation of the interpretations, and so we showed all three interpretations in textual form. Each interpretation consists of the set of conditions identified by the model. Since STACI also provides the confidence and generality, we computed the same measures for the other two systems as well, and provided them to the participant. For LIME, we also show the importance of each feature. We asked the participants to evaluate how satisfactory each interpretations was, using a scale from 1 (least satisfactory) to 5 (most satisfactory). We also asked them to identify which characteristics of the interpretations were important for their choice: the length of the interpretations, the feature importance, the confidence, or the generality. 55 people participated, and most of them have a background in computer science.

Table IX shows the average characteristics of the interpretations provided by each system, and the average rating by the users. Our method achieves the highest confidence, and ranks second in generality after DTExtract. This is because

TABLE IX: User study

System	Confidence(%)	Generality(%)	Length	Average Rating
DTExtract	76.61	74.68	1.16	3.12
LIME	59.18	0.41	1.86	1.93
STACI	85.23	42.42	2.52	**3.91**

Fig. 3: User preferences



DTExtract classified many instances on the root node. This entails shorter interpretations, and more generality, but comes at the cost of confidence. LIME achieves very low generality, which is because it provides local interpretations, which are not designed to regroup many data points.

Overall, STACI achieves the highest user rating. The reason for this good performance of STACI is shown in Figure 3: The users rated confidence and generality as the two most valuable properties of interpretations. These are exactly the metrics that we introduced in this work, and that STACI optimizes. This preference for more general interpretations also explains why LIME, with its local explanations, performs poorly: the participants did not like interpretations that appear tailored to a few data points. The reason why STACI outperforms DTExtract in the user rating is that users value confidence above everything else. Hence, DTExtract's strategy of providing short, general, but low-confidence interpretations falls behind STACI's more balanced approach of optimizing generality and confidence at the same time.

Overall, our study emphasizes the importance of generality and confidence, and shows that the interpretations by our method were considered the most satisfactory.

V. CONCLUSIONS

In this paper, we have presented STACI, a method for providing interpretations of black box classification models. Our method uses one surrogate decision tree per class, each trained using the *F1 score* as a metric to decide a split. The resulting models provide simple, confident, but general interpretations.

We have shown that our new method outperforms state of the art methods in terms of fidelity, maximal complexity, and confidence. Our user study confirms that the metrics we proposed, confidence and generality, are important features of an interpretation, and that users prefer our interpretations over others.

REFERENCES

[1] B. Goodman and S. Flaxman, "European union regulations on algorithmic decision-making and a "right to explanation"," *AI magazine*, vol. 38, no. 3, 2017.

[2] O. Bastani, C. Kim, and H. Bastani, "Interpreting blackbox models via model extraction," *arXiv preprint arXiv:1705.08504*, 2017.

[3] O. Boz, "Extracting decision trees from trained neural networks," in *SIGKDD*, 2002.

[4] M. Craven and J. W. Shavlik, "Extracting tree-structured representations of trained networks," in *Advances in neural information processing systems*, 1996.

[5] U. Johansson and L. Niklasson, "Evolving decision trees using oracle guides," in *Symp. on Computational Intelligence and Data Mining*, 2009.

[6] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should i trust you? – explaining the predictions of any classifier," in *SIGKDD*, 2016.

[7] V. Beaudouin, I. Bloch, D. Bounie, S. Clémençon, F. d'Alché Buc, J. Eagan, W. Maxwell, P. Mozharovskyi, and J. Parekh, "Flexible and context-specific ai explainability: a multidisciplinary approach," *SSRN 3559477*, 2020.

[8] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM computing surveys*, vol. 51, no. 5, 2018.

[9] F. K. Došilović, M. Brčić, and N. Hlupić, "Explainable artificial intelligence: A survey," in *MIPRO*, 2018.

[10] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (xai)," *Access*, vol. 6, 2018.

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

[12] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu, "Interpretable machine learning: definitions, methods, and applications," *arXiv preprint arXiv:1901.04592*, 2019.

[13] T. Miller, "Contrastive explanation: A structural-model approach," *arXiv preprint arXiv:1811.03163*, 2018.

[14] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and regression trees*. CRC press, 1984.

[15] B. Letham, C. Rudin, T. H. McCormick, D. Madigan *et al.*, "Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model," *The Annals of Applied Statistics*, vol. 9, no. 3, 2015.

[16] H. Yang, C. Rudin, and M. Seltzer, "Scalable bayesian rule lists," in *ICML*, 2017.

[17] B. Ustun and C. Rudin, "Supersparse linear integer models for optimized medical scoring systems," *Machine Learning*, vol. 102, no. 3, 2016.

[18] M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-precision model-agnostic explanations," in *AAAI*, 2018.

[19] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *NEURIPS*, 2017.

[20] A. White and A. d. Garcez, "Measurable counterfactual local explanations for any classifier," *arXiv preprint arXiv:1908.03020*, 2019.

[21] D. Alvarez-Melis and T. S. Jaakkola, "On the robustness of interpretability methods," *arXiv preprint arXiv:1806.08049*, 2018.

[22] S. Tan, R. Caruana, G. Hooker, P. Koch, and A. Gordo, "Learning global additive explanations for neural nets using model distillation," *arXiv preprint arXiv:1801.08640*, 2018.

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

[24] F. Pedregosa, G. Varoquaux, Gramfort *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, 2011.

[25] S. Wachter, B. D. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the GDPR," *CoRR*, vol. abs/1711.00399, 2017. [Online]. Available: http://arxiv.org/abs/1711.00399