Artificial Intelligence In Medicine 135 (2023) 102471

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com/locate/artmed

Research paper

Why did AI get this one wrong? — Tree-based explanations of machine learning model predictions

Enea Parimbelli [a,c,*], Tommaso Mario Buonocore [a,1], Giovanna Nicora [a,b,1], Wojtek Michalowski [c], Szymon Wilk [d], Riccardo Bellazzi [a]

[a] Department of Electric, Computer and Biomedical Engineering, University of Pavia, Pavia, Italy
[b] enGenome srl, Pavia, Italy
[c] Telfer school of Management, University of Ottawa, Ottawa, Ontario, Canada
[d] Division of Intelligent Decision Support Systems, Institute of Computing Science, Poznan University of Technology, Poznan, Poland

ARTICLE INFO

Keywords:
XAI
Black-box
Explanation
Local explanation
Interpretable
Explainable
Fidelity
Reliability
Post-hoc
Model agnostic
Surrogate model

ABSTRACT

Increasingly complex learning methods such as boosting, bagging and deep learning have made ML models more accurate, but harder to interpret and explain, culminating in black-box machine learning models. Model developers and users alike are often presented with a trade-off between performance and intelligibility, especially in high-stakes applications like medicine. In the present article we propose a novel methodological approach for generating explanations for the predictions of a generic machine learning model, given a specific instance for which the prediction has been made. The method, named AraucanaXAI, is based on surrogate, locally-fitted classification and regression trees that are used to provide post-hoc explanations of the prediction of a generic machine learning model. Advantages of the proposed XAI approach include superior fidelity to the original model, ability to deal with non-linear decision boundaries, and native support to both classification and regression problems. We provide a packaged, open-source implementation of the AraucanaXAI method and evaluate its behaviour in a number of different settings that are commonly encountered in medical applications of AI. These include potential disagreement between the model prediction and physician's expert opinion and low reliability of the prediction due to data scarcity.

1. Introduction

In recent years, an interesting trend in artificial intelligence (AI), now largely known as eXplainable AI (XAI), has been gaining traction. Increasingly complex learning methods such as boosting, bagging and deep learning have made ML models more accurate, but harder to understand and interpret, often imposing a trade-off between performance and intelligibility (e.g. using a model which is a white-box, despite suboptimal performance) [1]. In particular, the last ten years have seen a resurgence of the popularity of neural networks, following the "deep learning revolution" [2], as well as highlighted superior performance of ensemble algorithms like gradient boosting [3] in a number of applications. These advancements are often associated with increased complexity of the ML models, which results in increased difficulty to understand their internal functioning and outputs.

In order to respond to this comprehensibility challenge, a number of researchers are actively working on ways to interpret or explain the model's output [4]. XAI is indeed a re-emerging research trend, boosted by recently introduced regulations such as the European Union's GDPR2 and its "right to an explanation" [5], the very recent EU Artificial Intelligence Act [6], as well as the US government's Algorithmic Accountability Act of 2019, and the U.S. Department of Defense's Ethical Principles for Artificial Intelligence. Fairness-, accountability-, and transparency- related requirements are even more essential in AI applications to medicine (AIM) where AI often supports high-stakes decisions in diagnosis, prognosis and treatment [7,8]. A testimony to this fact is the embedding of explanation facilities as an integral part of very early examples of AIM expert systems such as MYCIN [9].

The concepts of "interpretability" and "explainability" are often used interchangeably, denoting the still incomplete process of laying common ground in XAI terminology. However, an interesting distinction is proposed by Holzinger et al. [10] who define interpretable AI as ante-hoc property of "glass box" (also known as "transparent") ML models, and explainable AI as post-hoc explanation or inspection methods for "black box" ML models.

* Correspondence to: Department of Electrical, Computer and Biomedical Engineering, Via Ferrata 5, 27100, Pavia, Italy.
E-mail address: enea.parimbelli@gmail.com (E. Parimbelli).
[1] Equal contribution.

https://doi.org/10.1016/j.artmed.2022.102471
Received 26 January 2022; Received in revised form 25 November 2022; Accepted 28 November 2022
Available online 1 December 2022

Following Holzinger's categorization, in this paper we propose a novel post-hoc, model-agnostic, local explanation method for potentially "black box" ML models — namely an explainable AI (XAI) method. We name our proposed approach AraucanaXAI since the generated explanations are based on Classification And Regression *Trees* (CART), with a reference to the *Araucaria Araucana* tree species, native to central and southern Chile, which is commonly known as the "monkey puzzle tree" in English-speaking countries.

1.1. Related work

We scope our overview of related work to directly comparable approaches for model-agnostic, post-hoc local explanation. Several approaches have indeed been proposed to tackle the post-hoc local explainability problem, and we refer the reader to [4,11–15] for more comprehensive literature reviews.

Among others, the most used and cited methods are LIME [16] and SHAP [17].

SHAP (SHapley Additive exPlanations) uses a principle from coalitional game theory, namely Shapley values, to fairly distribute the payout of a game among a set of players. In the ML context, the game is the prediction for a single instance, while the players are the feature values of the instance cooperating to receive the gain. This gain consists of the difference between the Shapley value of the prediction and the average of the Shapley values of the predictions among the feature values of the instance to be explained [17].

LIME's underlying hypothesis is that the behaviour of a complex, black-box ML model can be locally approximated by a simpler, more interpretable model. Specifically, given a single instance, a local explanation of the prediction is obtained by perturbing the instance and by training a linear model on the perturbed samples. The linear model estimated coefficients represent the local post-hoc explainability of the more complex model. LIME has been widely extended to interpret image classification [18,19].

Another local interpretability approach is the rule-based approach **Anchors** [20]. Compared to LIME, Anchors has clear coverage, guaranteeing that the predictions of instances in the same area are almost the same. Anchors interpretation result is a simple IF–THEN rule, which is therefore more intuitive in comparison with LIME's coefficients interpretation.

The intuition behind using decision trees as better interpretable surrogate models is not new. [21] proposes a novel model extraction algorithm to learn an axis-aligned decision tree as a better interpretable alternative to various black-box models including random forests, neural networks and control policies learned by the means of reinforcement learning. Similarly, [22] proposes recursive partitioning and binary trees as a tool to provide insights on the most important features for classification. Both the above-mentioned approaches however target global explainability, i.e. provide an explanation for an ML model as a whole, instead of justifying a specific prediction. Some vertical applications for the explanations of specific classes of problems, like image classification, also rely on regression trees [23]. Finally, a somehow different line of research aims at interpreting complex models (mostly ensemble models) using decision trees as base learners, such as Random Forest or Gradient Boosting, using a heuristic for quicker calculation of Shapley values [24] or by efficiently extracting rules from the weak learners (again tree-based models like decision stumps) in the ensemble [25,26].

Quite interestingly, despite the relevant number of XAI methods developed in the last years, a clear methodology and performance metrics for their validation and quantitative evaluation are still lacking. As an exception to this general observation, a recent comparison between LIME, Anchors and SHAP in the medical domain has been performed [27]. The article proposes an evaluation of XAI methods according to a set of metrics, such as:

- Identity: if there are 2 identical instances, they must have identical explanations
- Fidelity: concordance of the predictions between the XAI proxy model and the complex model.
- Separability: if there are 2 dissimilar instances, they must have dissimilar explanations
- Similarity: the more similar the instances to be explained, the closer the explanations should be
- Time: average time used by the XAI method to output an explanation across the entire test set
- Bias detection: ability to detect bias in training data

SHAP was found to be the fastest algorithm to output an explanation, also being able to detect bias. On the other hand, LIME has the lowest performance on identity, but the highest for separability.

1.2. Objective and original contribution

We propose AraucanaXAI as a novel XAI methodological approach, and provide a reusable, readily-accessible python implementation of such method.[2]

A recent review of XAI methods applied to data coming from electronic health records [28] highlighted how reproducibility of many of the works analysed is still a relevant issue, with many papers not only failing to share datasets used in the experiments, but also the experimental setup and pipeline as well as the open-source implementation of their XAI method and evaluation experiments. We provide all of the previous in this article, following guidelines [29] to assess the reproducibility of our research. Moreover, the same review [28] points to the fact that adversarial attacks, or even slight perturbations of the original dataset, can significantly harm XAI approaches based on feature importance and feature ranking like SHAP and LIME. This is confirmed by our results about identity (see Table 2) where only AraucanaXAI, which is not based on a feature importance strategy alone, scores perfect results across the whole span of evaluation setups.

Methodologically, what distinguishes our novel approach from the related work discussed above is its use of a simple, yet non-linear, interpretable model such as CART to generate locally accurate explanations for predictions. Also, differently from LIME and SHAP, AraucanaXAI is not a feature importance extraction method, but rather allows to extract explanations that are more similar to "clinical rules" than scores of importance and feature ranking. Feature-importance based XAI methods indeed, albeit being able to inspect feature importance is often presented as the main feature making an ML model "explainable" (e.g. Random Forests and Gradient boosting algorithms implementations do have embedded feature-importance calculation facilities), are however only one facet of the more complex XAI concept [8,30]. Also, the reliance on tree-based surrogate models translates into virtually perfect identity and fidelity to the original model (see Table 2). Furthermore, AraucanaXAI provides the possibility of presenting the generated explanations in different forms (e.g. IF–THEN rules vs. feature-importance scores vs. navigable tree structure) and accounting for the reliability of the model prediction when producing an explanation. These aspects will be further detailed and discussed in the following sections of the article.

2. Methods

Our AraucanaXAI approach is based on a relatively small set of general principles: given a new instance for which we want to generate an explanation of prediction we: (i) generate a local set of neighbouring instances, coming from the original training set augmented

[2] Available through the pip package manager https://github.com/detsutut/AraucanaXAI#installation.

with oversampling, re-labelled with the predictions of the *explained* model, (ii) grow an unpruned (or lightly-pruned) tree to learn the *explainer* as a white-box model, with the ability to deal with non-linear decision boundaries, (iii) navigate the *explainer* tree according to feature values of the instance to be explained and use it as the proposed explanation. Algorithm 1 presents the pseudocode of AraucanaXAI, which implements these principles.

2.1. Dependencies and parametrization

AraucanaXAI has some dependencies from well-established methods and implementations described in the ML literature. The local tree model surrogates that our method uses to generate the explanations are Brieman's classification and regression trees (CART) [31], and in particular we rely on scikit-learn [32] optimized implementation of CART. Similarly, the optional pruning step of AraucanaXAI's algorithm is minimal cost-complexity pruning, also defined in [31]. The distance metric, albeit configurable as a hyperparameter of the method, is by default set to use the Gower distance [33]. Finally, for the default oversampling step described in Algorithm 1, the currently available choices rely on SMOTE [34], or random oversampling either from a uniform distribution or a normal distribution.

Some of the behaviour of our XAI method have defaults, but AraucanaXAI allows for a higher level of customization through hyperparameters tuning (e.g. degree of oversampling, or pruning of the *explainer* tree). Such design choices, together with guidance on how to optimize them for AraucanaXAI users, are presented in Algorithm 1 and further analysed in the discussion section.

Algorithm 1 Local Tree-based explanation

Require:

predictive function f, instance x, distance function *dist* (hyperparameter), number of local neighbours N (hyperparameter), oversampling method *Omethod* (hyperparameter), pruning criteria P (hyperparameter)

1: Compute $D = dist(x, z)$ for each training element z
2: Select the subset T_n consisting of the N training samples with lowest distance from x
3: Create a set S of additional examples, generated from T_n using *Omethod* oversampling (optional).
4: Re-label the samples in T_n and S with the class predicted by f. Define the explainer set E as $T_n \cup x \cup S$
5: Train e as a decision tree on E. Optionally prune e according to pruning hyperparameter P
6: Navigate explainer tree e according to feature values of x and provide explanation of the prediction for instance x made by f

2.2. Experimental setup

We set up an experiment where we apply AraucanaXAI to synthetic data with specific properties in order to provide insights into the behaviour of our method in circumstances that are relevant to biomedical applications. As an additional experiment that might be closer to real-world conditions, we also test our approach on the MIMIC dataset [35,36], and compare its performance with other state-of-the-art XAI tools as LIME and SHAP. The source code for the experimental setup is available on github[3] and all experiments have been run on a Google Colab pro high-RAM instance with no hardware (GPU or TPU) acceleration.

To generate synthetic datasets that fit our purposes, we employed a Bayesian Network (BN) as a generative model. A Bayesian network is a graphical model of the joint probability distribution for a set of variables. By modelling the conditional dependencies of a set of attributes and outcomes, BNs can be used to generate realistic synthetic health data [37,38]. In particular, we simulated a cohort of patients using the BN described in [39], where authors developed a probabilistic causal model for the diagnosis of liver disorders. A complete implementation of this network, quantified with proper parameters, is available in the `bnlearn` R package.[4] The network has 70 nodes and 123 arcs. Examples of nodes reported to describe the liver disorder problem are age, gender, cholesterol level, hospitalization. We generated a dataset for the prediction of the outcome hospitalization, based on the remaining 69 independent variables. In particular, 10,000 patients were sampled (53% resulting in class 1 for hospitalization). Then, ML models were trained on 95% of the set of 10,000 generated patients, while the remaining 5% (500 instances) are reserved to be used as a test set to evaluate the performance of the models, but most importantly to evaluate our proposed XAI method. Also, to be able to evaluate the identity XAI metric, we duplicated 100 instances of the test set to have identical patients on which to evaluate identity of the explanations generated by the different XAI methods. Raw synthetic data we generated and used in the evaluation experiments are available on Zenodo, and the associated Data in Brief publication.[5]

Subsequently, we simulated a shift in the population by changing the BN prior probabilities of some of the nodes, such as gender, age and hospitalization. Dataset shifts are frequent in health data since patients' populations can change for a variety of reasons, from different patient selection strategies to evolving treatments and guidelines [40]. We sample 500 out-of-distribution (o.o.d) instances from the perturbed network and we evaluate the performance of our XAI method, LIME and SHAP both on identically distributed (i.i.d) (test set) and on o.o.d samples. The same procedure of adding 100 duplicated examples to the test set to evaluate identity was also performed on the ood training set.

Regarding ML models we selected Logistic Regression (LR), Random Forest (RF), Gradient Boosting (GB), and Multi-Layer Perceptron (NN). Not to influence the experiment with subjective design choices we used default settings for all the models, except for NN where we employed 3 hidden layers (instead of the default, which is 1 hidden layer) in order to emulate the behaviour of a "deep" NN. We also performed hyperparameter tuning for each of the four models employing a simple grid-search for optimal parameters employing a nested 5-fold cross-validation strategy, and optimizing parameters to maximize F1 score. The source code for the parameters optimization phase is available on AraucanaXAI's github repository. Optimal parameters have been in turn used in subsequent model training, predictive performance, and XAI evaluation experiments.

For our comparisons in Tables 2–5 we choose to employ the default hyperparameters for all XAI methods, in order to allow the fairest possible comparison between them assuming usage from a naive (i.e. not expert) user. AraucanaXAI has a number of hyperparameters to allow fine-grained control by the advanced user. However, our packaged implementation does provide defaults that can be used out-of-the-box and get valid local explanations. The same applies to SHAP and, to a larger extent, to LIME where also hyperparameters like size of the neighbourhood from which the surrogate models for explanations are derived are tunable by the user. The only exception to our commitment to use defaults is that we set num_samples = 200 (i.e. the number of samples from the training set from which the surrogate linear model is derived) for LIME since this setting resulted in overly penalizing

[3] https://github.com/detsutut/AraucanaXAI/blob/master/araucana_AI_special_issue.ipynb.

[4] https://www.bnlearn.com/bnrepository/discrete-large.html#hepar2.

[5] Giovanna Nicora. (2022). AraucanaXAI - HEPAR synthetic datasets [Data set]. In Data in Brief: Vol. under evaluation (1.0). Zenodo. https://doi.org/10.5281/zenodo.6726768.

results for identity and separability, probably due to the fact that LIME's default number of 5000 is both (a) an absolute number (while Araucana's default is 0.01 of the size of the original training set) and (b) very large when compared to the size of our dataset, resulting in the derivation of largely the same surrogate explainer model, which negatively impacted separability and fidelity.

The MIMIC-III dataset contains clinical data and vital signs of thousands of patients in the ICU. In particular, we used the preprocessed dataset made available by the PhysioNet 2012 challenge [36]. Details about the features set can be found in [36] and at https://physionet.org/content/challenge-2012/1.0.0/. The binary target outcome is in-hospital death. As a preprocessing step, we removed features with at least 90% of missing values, leading us to filter 76 out of 125 clinical features. Examples of retained features are age, gender and the first and last glucose measurement for each patient. We also removed patients with at least one missing value. The resulting dataset contains 4480 patients that survived ("In-Hospital death" = 0) and 768 that died in the hospital ("In-Hospital death" = 1). We are interested in developing a model to predict in-hospital mortality from clinical data and provide explanations with our proposed methodology.

3. Results

3.1. Open-source implementation

We implemented the described method for AraucanaXAI in Python and make it available as a PyPi package.[6] along with its source code[7] In turn, we then used our implementation to carry out a comparative evaluation of AraucanaXAI versus LIME and SHAP.

3.2. Comparative evaluation of XAI methods performance

Table 1 shows the predictive performance of the four ML models on the test set (i.i.d.), the o.o.d. samples and the MIMIC dataset. It can be observed how, as expected, almost all the relevant performance metrics degrade in the o.o.d. set. It is also worth mentioning that the modest results achieved in the MIMIC dataset case are in line with the highly challenging nature of the task (i.e. predicting in-hospital mortality from the available features), as testified by the organizers of the original PhysioNet challenge in 2012. [36]. Tables 2–5 show the full set of results from our comparative evaluation of XAI methods performance on both the two synthetic datasets (i.i.d and o.o.d. respectively) as well as on the MIMIC dataset.

The identity metric (Table 3) shows consistent performance for AraucanaXAI, which has an ideal score across the board. This means that two identical instances always have the same explanation generated by AraucanaXAI. This is not the case for either LIME, which highlights that this method generates potentially unstable explanations (probably due to the random perturbation of examples that the algorithm includes as part of its linear surrogate model generation), and interestingly also for SHAP. In particular, SHAP performs with an identity lower than 1, and close to 0 for the case of NN, highlighting a possible implementation flaw for models that are not tree-based like RF and GB, for which an optimized heuristic calculation of SHAP values is available [24].

Also, Table 2 shows how both SHAP and AraucanaXAI, when oversampling is turned off, present a fidelity of 1 (i.e. ability to predict the same class as the more complex model being explained) across all the experiments, regardless of the predictive ML model to explain (GB, RF, LR or NN). This is marginally better than the performance of LIME, while on par with SHAP. This observation is, intuitively (since separability and identity are often in a trade-off similar to precision and recall), balanced by the fact that separability (Table 4 is ideal for LIME and SHAP (i.e. two different examples must have different explanations) while significantly lower for certain uses of AraucanaXAI, highlighting the fact that our method may actually output the same explanation for two different instances. We stress the fact that the trade-off between identity and separability can be controlled in AraucanaXAI by acting on a number of the method hyperparameters, and in particular the size of the neighbourhood considered. We elaborate on this aspect further in the discussion section.

Finally, Table 5 shows how, in agreement with the observations in [27], our experimental results confirm that SHAP is the fastest algorithm in all settings involving the synthetic dataset, as well as on the MIMIC dataset when tree-based ensemble algorithms are used.

4. Discussion

AraucanaXAI is proposed as our original contribution to the range of available local, model-agnostic, post-hoc XAI methods. Some properties of our method, and implementation choices, make it particularly fit for the purpose of generating locally-valid explanations for predictions of biomedical ML-based predictive models [8]. We discuss such properties of AraucanaXAI, and their implications, in the following subsections.

4.1. Hyperparameters selection

As previously pointed out, no hyperparameter optimization for the XAI methods examined in the experimental evaluation was performed, using the default parameters provided by the available methods implementations. However, AraucanaXAI offers the largest potential for customization for the advanced user, while still providing good off-the-shelf usability for the novice.

First of all, the default distance function employed by the algorithm is the Gower distance, which is applicable to both numeric and categorical features and thus covers the most typical use cases. However, a more specific choice of a distance function may be useful in scenarios where certain features have greater weight in defining what cases are more "similar" (i.e. closer according to the distance function to be considered) [41]. This may prove useful where disease subtypes (e.g. based on some genetic variant or other biomarkers) or previously known outcome classes (e.g. from risk stratification models) are previously known and the user wants to account for that during AraucanaXAI's neighbourhood identification step. Similarly, the number of local neighbours to be included during the generation of the explanation influences how much the user wants the explanation to focus on globally relevant factors (e.g. male sex and high blood pressure are well-known risk factors for the risk of cardiac events in the general population) compared to locally important ones (e.g. when it happens that a great part of the neighbouring examples all happens to be taking a blood-thinner medication, which turns out to be a determining factor for the predictive model prediction in this narrow subset of instances).

Secondly, regarding the choice of oversampling (both the method, and the number of synthetic examples generated are controllable through parametrization) the default is to use none. However, in cases where the original training set is rather sparse in specific areas of the data space, it can be advisable to use one of the available techniques (e.g. SMOTE for numerical data or SMOTE-NC for mixed data) in order to increase the density of examples from which the surrogate explainer tree would be derived. Generation of additional examples for the explainer set (see Algorithm 1, in combination with their labelling using the original predictive model f, is a way to probe the model for predictions (thus potentially exposing its prediction "rationale") in otherwise unexplored parts of the data space, where the new, potentially unseen, instance to be explained may fall. For a further discussion of such cases, and their impact on prediction reliability and accuracy, we direct the reader to the following Section 4.3 on *reliability and oversampling*.

[6] https://pypi.org/project/araucanaxai/.
[7] https://github.com/detsutut/AraucanaXAI.

Table 1
Predictive performance of Logistic Regression (LR) Gradient Boosting (GB), Random Forest (RF) and Multi-layer Perceptron (NN) on HEPAR i.i.d test set (iid), on HEPAR o.o.d. samples (ood), and MIMIC (mim) datasets.

	Accuracy			Recall			Precision			F1		
	iid	ood	mim	iid	ood	mim	iid	ood	mim	iid	ood	mim
LR	0.74	0.64	0.86	0.65	0.68	0.29	0.82	0.72	0.53	0.72	0.70	0.38
RF	0.74	0.62	0.86	0.66	0.77	0.30	0.81	0.66	0.51	0.72	0.71	0.38
GB	0.75	0.66	0.85	0.65	0.72	0.30	0.84	0.73	0.49	0.73	0.72	0.37
NN	0.74	0.65	0.74	0.65	0.63	0.56	0.82	0.75	0.30	0.72	0.69	0.39

Table 2
Fidelity performance of AraucanaXAI, LIME and SHAP evaluated on synthetic and MIMIC datasets. Logistic regression (LR), Random Forest (RF), Gradient Boosting (GB) and Multi-Layer Perceptron (NN).

		Fidelity		
		iid	ood	mim
AraucanaXAI[a]	LR	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**
AraucanaXAI		0.979 ± 0.002	0.672 ± 0.049	0.939 ± 0.006
LIME		0.989 ± 0.001	0.967 ± 0.007	0.970 ± 0.001
SHAP		**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**
AraucanaXAI[a]	RF	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**
AraucanaXAI		0.998 ± 0.001	0.696 ± 0.026	0.928 ± 0.004
LIME		0.997 ± 0.001	0.946 ± 0.009	0.967 ± 0.003
SHAP		**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**
AraucanaXAI[a]	GB	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**
AraucanaXAI		0.973 ± 0.002	0.666 ± 0.046	0.747 ± 0.021
LIME		0.985 ± 0.001	0.947 ± 0.015	0.964 ± 0.000
SHAP		**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**
AraucanaXAI[a]	NN	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**
AraucanaXAI		0.963 ± 0.003	0.645 ± 0.043	0.938 ± 0.008
LIME		0.980 ± 0.001	0.986 ± 0.005	0.964 ± 0.002
SHAP		**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**

[a]Oversampling turned off for AraucanaXAI.

Table 3
Identity performance of AraucanaXAI, LIME and SHAP evaluated on synthetic and MIMIC datasets. Logistic regression (LR), Random Forest (RF), Gradient Boosting (GB) and Multi-Layer Perceptron (NN).

		Identity		
		iid	ood	mim
AraucanaXAI[a]	LR	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**
AraucanaXAI		**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**
LIME		0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
SHAP		0.714 ± 0.032	0.490 ± 0.029	0.003 ± 0.005
AraucanaXAI[a]	RF	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**
AraucanaXAI		**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**
LIME		0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
SHAP		**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**
AraucanaXAI[a]	GB	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**
AraucanaXAI		**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**
LIME		0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
SHAP		**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**
AraucanaXAI[a]	NN	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**
AraucanaXAI		**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**
LIME		0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
SHAP		0.283 ± 0.032	0.072 ± 0.015	0.000 ± 0.000

[a]Oversampling turned off for AraucanaXAI.

Table 4
Separability performance of AraucanaXAI, LIME and SHAP evaluated on synthetic and MIMIC datasets. Lower scores are better. Logistic regression (LR), Random Forest (RF), Gradient Boosting (GB) and Multi-Layer Perceptron (NN).

		Separability		
		iid	ood	mim
AraucanaXAI[a]	LR	0.469 ± 0.000	**0.000 ± 0.000**	0.027 ± 0.000
AraucanaXAI		0.106 ± 0.018	0.036 ± 0.028	0.432 ± 0.049
LIME		**0.000 ± 0.000**	**0.000 ± 0.000**	**0.000 ± 0.000**
SHAP		**0.000 ± 0.000**	**0.000 ± 0.000**	**0.000 ± 0.000**
AraucanaXAI[a]	RF	0.199 ± 0.000	**0.000 ± 0.000**	**0.000 ± 0.000**
AraucanaXAI		0.205 ± 0.016	0.046 ± 0.025	0.734 ± 0.023
LIME		**0.000 ± 0.000**	**0.000 ± 0.000**	**0.000 ± 0.000**
SHAP		**0.000 ± 0.000**	**0.000 ± 0.000**	**0.000 ± 0.000**
AraucanaXAI[a]	GB	0.433 ± 0.000	**0.000 ± 0.000**	**0.000 ± 0.000**
AraucanaXAI		0.095 ± 0.018	0.034 ± 0.038	0.024 ± 0.008
LIME		**0.000 ± 0.000**	**0.000 ± 0.000**	**0.000 ± 0.000**
SHAP		**0.000 ± 0.000**	**0.000 ± 0.000**	**0.000 ± 0.000**
AraucanaXAI[a]	NN	0.167 ± 0.000	**0.000 ± 0.000**	0.019 ± 0.000
AraucanaXAI		0.034 ± 0.009	0.030 ± 0.015	0.421 ± 0.057
LIME		**0.000 ± 0.000**	**0.000 ± 0.000**	**0.000 ± 0.000**
SHAP		**0.000 ± 0.000**	**0.000 ± 0.000**	**0.000 ± 0.000**

[a]Oversampling turned off for AraucanaXAI.

Table 5
Running time performance of AraucanaXAI, LIME and SHAP evaluated on synthetic and MIMIC datasets. Lower scores are better. Logistic regression (LR), Random Forest (RF), Gradient Boosting (GB) and Multi-Layer Perceptron (NN).

		Time (s)		
		iid	ood	mim
AraucanaXAI[a]	LR	144.912 ± 0.000	3.886 ± 0.000	85.310 ± 0.000
AraucanaXAI		191.979 ± 2.750	5.177 ± 0.689	66.748 ± 2.516
LIME		1704.014 ± 8.043	61.576 ± 0.675	**7.781 ± 0.373**
SHAP		**36.86 ± 2.908**	**2.267 ± 0.024**	36.860 ± 2.908
AraucanaXAI[a]	RF	2192.174 ± 0.000	6.151 ± 0.000	1020.710 ± 0.000
AraucanaXAI		2407.212 ± 12.725	62.462 ± 0.731	957.385 ± 19.711
LIME		1704.014 ± 8.043	92.523 ± 1.169	**228.627 ± 1.804**
SHAP		**250.471 ± 0.610**	**3.456 ± 0.012**	250.471 ± 0.610
AraucanaXAI[a]	GB	144.912 ± 0.000	6.151 ± 0.000	102.314 ± 0.000
AraucanaXAI		164.761 ± 2.393	3.633 ± 0.038	75.557 ± 2.347
LIME		1262.478 ± 47.103	62.543 ± 0.675	0.106 ± 0.006
SHAP		**0.045 ± 0.012**	**0.005 ± 0.000**	**0.045 ± 0.012**
AraucanaXAI[a]	NN	355.844 ± 0.000	5.981 ± 0.000	107.694 ± 0.000
AraucanaXAI		356.179 ± 4.430	5.894 ± 0.080	86.364 ± 2.995
LIME		1270.331 ± 53.890	62.346 ± 1.171	**18.698 ± 1.889**
SHAP		**81.853 ± 9.929**	**4.727 ± 1.259**	81.853 ± 9.929

[a]Oversampling turned off for AraucanaXAI.

Finally, the pruning hyperparameters, allow for control of the complexity of the explanations generated by AraucanaXAI. Whereas an unpruned (or very deep) tree would guarantee maximal fidelity and separability of generated explanation, it may also result in overly complex explainer trees that might damage human readability. We control the pruning of the generated explainer trees using the same set of parameters of the scikit-learn's CART implementation, both for leveraging its popularity in the ML and data-science communities, and accounting for different strategies for pruning (setting the max number of splits, vs. the maximum number of leaf nodes vs. the minimum decrease in impurity to add a further split, etc.) A deeper analysis of these aspects is presented in the following Section 4.4.

4.2. Fidelity to the original model

Our choice of relying on CART trees as surrogate models for AraucanaXAI explanations has been guided by the well-known high-variance of decision tree models. This high variance is the main factor, when combined with a light pruning (ideally no pruning) strategy, that guarantees a very high fidelity (i.e. the surrogate *explainer* model

outputs a prediction that is coherent to the prediction of the original *explained* model) to the original model predictions. This high fidelity is particularly relevant for medical applications where there is a good chance that explanations are more often sought in those cases where the physician and the ML model predictions are in disagreement [42]. In these cases, either the physician is looking to question his/her expert opinion and look for evidence that the ML model may be right (e.g. the explanation can highlight some evidence that the physician may have overlooked, but the ML model picked up), or he/she is looking at the explanation of the model prediction (e.g. what are the feature values that drove the model to predict the "wrong" class for this particular patient) to find "why did the AI get this one wrong". In both cases it is important that if the original model prediction was wrong we want the prediction of the *explainer* model to be the same, even if ultimately incorrect, not to correct the issue (i.e. explaining is not the same as finding causality [7,43]). In other words, we want the explanation to reflect what the original model "thought" as faithfully as possible, so that the insights coming from the explanation can effectively be used for model revision/improvement, or as feedback to the physician to challenge his initial opinion.

4.3. Reliability and oversampling

Recent work [44,45] and reports from the EU, such as the Ethics Guidelines for trustworthy AI (European Commission, Directorate-General for Communications Networks, Content and Technology, Ethics guidelines for trustworthy AI, Publications Office, 2019, https://data.europa.eu/doi/10.2759/177365) has analysed the importance of technical robustness and safety, which include as a key element *reliability* of a ML model on new unseen example, especially when significant dataset shift [46] exists. Analyses of ML prediction reliability are often aimed to update ML models in order to improve their ability to generalize or prevent degradation of calibration and discrimination [47]. However, reliability is also worth thorough consideration in the context of XAI. If a certain prediction is not reliable there is a good chance that the ML model may perform poorly, which is exactly when a good local explanation, such as the one obtained by AraucanaXAI or similar methods, becomes more valuable for model inspection, debugging and, ultimately, model update.

AraucanaXAI tackles such a problem with the introduction of a specific step in the algorithm dedicated to local neighbourhood enrichment through oversampling (see steps 3 and 4 in Algorithm 1). Generating additional examples, not originally included in the training set of the *explained* model, is essentially a way of probing the explained model on unseen instances. The usefulness of such an action when generating the *explainer* model is twofold. Firstly, it gives the *explainer* model learning procedure the chance to use information that is at a higher resolution (i.e. with more data points) than that used for the original model training. Secondly, relabelling of all the instances of the explainer set E with the predictions of the ML model is a way of making the decision boundary of the *explained* model more explicit, and thus more evident when inspecting the explanation.

Figs. 1 and 2 show examples of AraucanaXAI explanation on a Random Forest prediction for two samples in the Test set. The first example in Fig. 1 was incorrectly classified, while the second example (Fig. 2) was correctly classified in class 0. As we can see, the tree computed without oversampling (Fig. 1(a)) is very different from the tree computed with SMOTE oversampling (Fig. 1(b)). Instead, if we look at the second instance, the structure of the two trees, without oversampling and with SMOTE oversampling, is the same.

Although additional experiments are needed to better assess the relative merits and pitfalls of AraucanaXAI, these results may indicate that within more reliable regions (i.e. where the original model is confident in its prediction. See Section 4.2 for a more detailed discussion on reliability), also the explanations are more robust to changes in the hyperparameters. Instead, where the classifier is unreliable and errors occur, the explanation is more unstable. Fig. 3 gives further insights on how controlling the size of the neighbourhood where oversampling happens in the AraucanaXAI algorithm (see step 3 of Algorithm 1) impacts the XAI evaluation metrics. In particular, note how this parameter has a limited impact on AraucanaXAI's fidelity, while having a more notable effect on identity and separability, being able to also control the intrinsic trade-off between these two competing objectives.

Our implementation of AraucanaXAI provides the possibility to customize the oversampling strategy through a hyperparameter. An interesting future development may consist in providing guidance for hyperparameters tuning of AraucanaXAI on the basis of calculated prediction reliability, as defined in [44], of the model f on the specific instance x being explained.

4.4. Optimizing the output of the explainer

An important aspect of the user-friendliness of XAI methods has to do with the way the generated explanations are presented to the user. The specific form of visualization of explanations is somewhat independent of the methodology used to generate them. Indeed, visual representations of feature importance scores are widely used (and embedded) in some popular ML algorithms such as random forests, XGB, and several attention-based deep neural networks working on text or images [48]. Also LIME and SHAP, as well as their implementation in popular ML libraries such as Orange (https://orangedatamining.com/blog/explain/) and Anaconda (https://www.anaconda.com) offer similar facilities. Other viable options consist in human-readable if–then rules, and statistical-inspired solutions like presenting odds-ratio for linear models or posterior probabilities in Bayesian settings. Currently, AraucanaXAI implementation is a barebone Python library, that does not provide a full-fledged user interface for inspecting and interacting with generated explanations. However, a few approaches can be applied ex-post to the *explainer* tree model to improve its presentation to the user and ultimately improve uptake and fitness for specific use cases in AIM. Options include: (i) presenting the entire *explainer* tree (currently supported solution), (ii) navigating the tree to extract the path that is relevant for instance x, and converting it to a sequence of "if–then" rules, (iii) use the *explainer* tree to calculate feature importance scores likewise LIME or SHAP (e.g. an optimized version of SHAP that runs in polynomial time on any tree-based ML model is available [49], and could be employed for the purpose) (4) interesting recent work [50] proposed a logic programming methodology to provide a representation of a decision tree in the form of a compact set of rules (e.g. compacting rules on the same variable, but at different levels of the tree structure, in one single rule).

Finally, the possibility of controlling the degree of tree pruning performed by AraucanaXAI, is a way to fine-tune the trade-off between a virtually perfect fidelity (i.e. when pruning of the *explained* tree is not performed at all) of the generated explanation in exchange for a more compact tree structure, which directly translates in a less complex explanation. This feature mitigates the current lack of a rich GUI for AraucanaXAI. Future work, involving clinical experts in the definition and empirical evaluation of what constitutes a "good explanation" in the medical context as well as HCI considerations, is needed to properly compare the alternative visualization options listed above and guide their implementation.

4.5. Limitations

The work described in this article, as well as the proposed AraucanaXAI method itself, have some limitations. Firstly, evaluation of what constitutes a "good" explanation for a user cannot be thoroughly assessed without clearly defined metrics (which is currently an acknowledged gap in the XAI literature [51]) and direct involvement of the physician users themselves in a properly designed evaluation study.

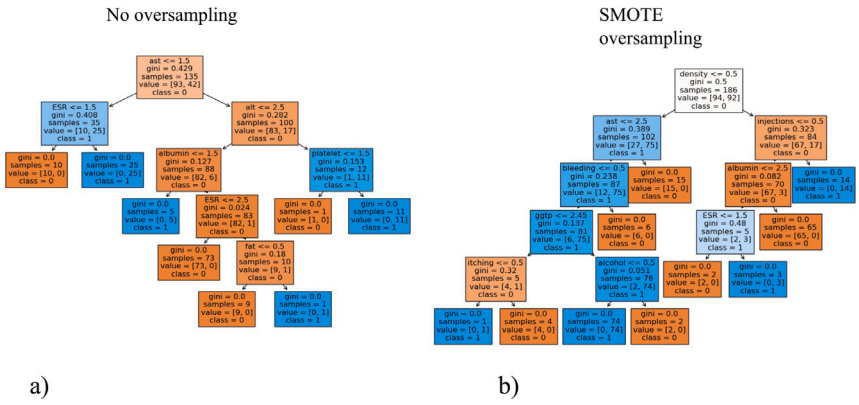

Fig. 1. Araucana explanation for a test sample in class 1, incorrectly predicted by the Random Forest as class 0. (a) explanation without oversampling. (b) explanation with SMOTE oversampling.

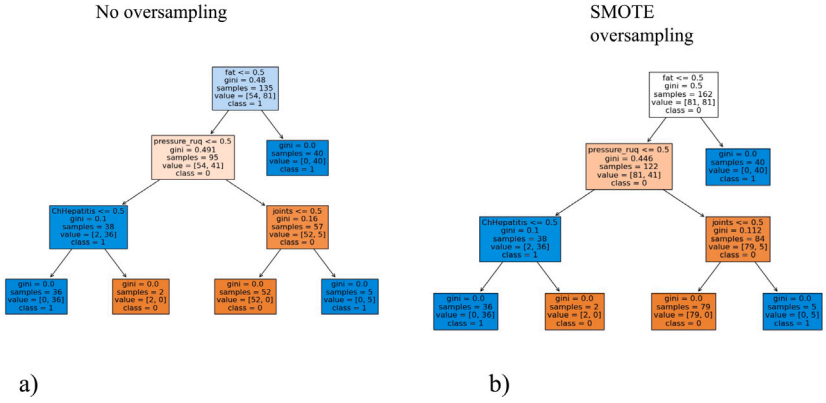

Fig. 2. Araucana explanation for a test sample in class 0, correctly predicted by the Random Forest as class 0. (a) explanation without oversampling. (b) explanation with SMOTE oversampling.

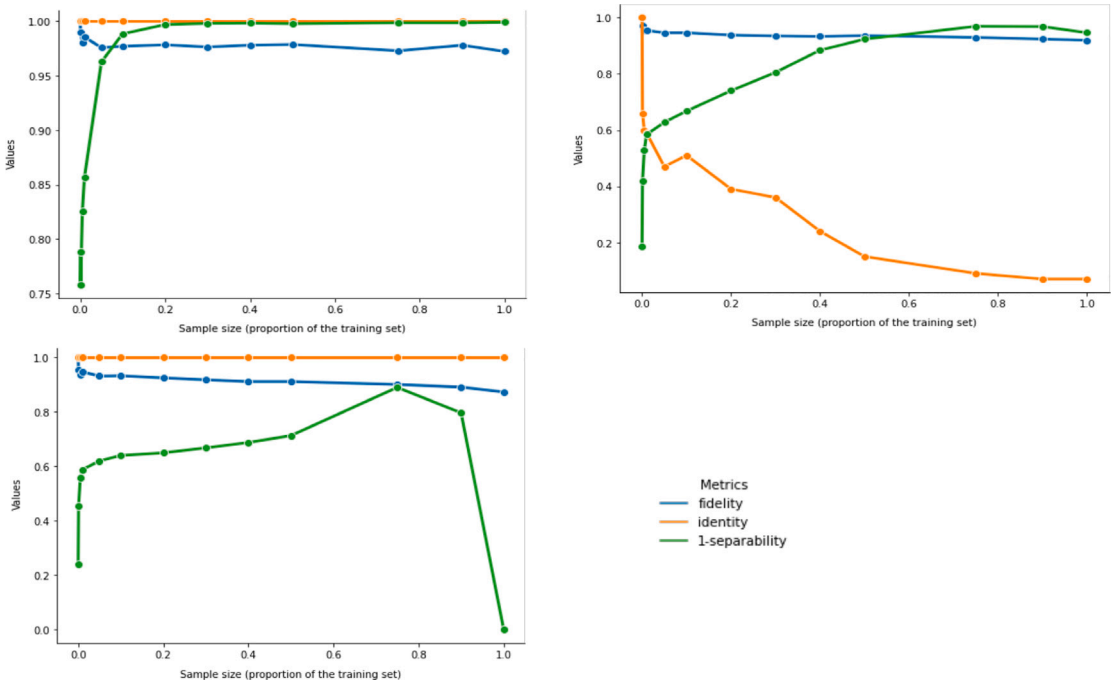

Fig. 3. Evaluation of the size of the local neighbourhood considered by AraucanaXAI on the metrics of fidelity, identity and separability (plotted as 1-separability to facilitate visual comparison with the other metrics as "higher is better"). Top: synthetic dataset i.i.d. (left) and o.o.d. (right), bottom-left MIMIC dataset.

Such studies constitute future work worth pursuing, with the potential to benefit the explainable AIM community at large [8].

Secondly, it would be interesting to provide the user with the ability to control the number of additional samples generated in S (see Algorithm 1) by the oversampling step. This would be particularly interesting to provide finer management of those predictions with suboptimal reliability, as discussed in Section 4.3. Currently, our implementation only allows the control of the oversampling strategy via a hyperparameter, while the number of generated samples is not tunable.

Finally, despite we provided an overview of different strategies for presenting explanations to a user in Section 4.4, we currently do not directly support these in our AraucanaXAI implementation. At the moment AraucanaXAI's way of presenting the generated explanations relies on scikit-learn facilities for decision tree visualization.

5. Conclusion

In the present paper we presented AraucanaXAI, a model-agnostic, post-hoc method for generating local explanations of ML model predictions. We also make an open-source implementation of the method available for use, and run a comparative evaluation experiment to highlight its strengths and limitations with respect to other comparable XAI methods. A comparative evaluation of our proposed method and its implementation is performed on both synthetic and real-world clinical data, allowing direct comparison with state-of-the-art XAI methods such as LIME and SHAP. AraucanaXAI's high fidelity and identity, combined with reasonably fast computation times make it a viable choice for contexts like human-in-the-loop [52,53] inspection and update of ML models, which is a promising direction for XAI research. AraucanaXAI's ability to account for, and manage low-reliability predictions, and its customizability through hyperparameters make it particularly fit for medical applications of AI that have a strong requirement for explainability in combination with cutting-edge predictive performance.

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.

Acknowledgments

E.P and R.B. are partially funded by the project Periscope (Pan-European Response to the ImpactS of COVID-19 and future Pandemics and Epidemics), H2020 GA 101016233, funded by the European Union.

References

[1] Caruana Rich, Lundberg Scott, Ribeiro Marco Tulio, Nori Harsha, Jenkins Samuel. Intelligible and explainable machine learning: Best practices and practical challenges. In: Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining. New York, NY, USA: Association for Computing Machinery; 2020, p. 3511–2.

[2] Sejnowski Terrence J. The deep learning revolution. MIT Press; 2018, Google-Books-ID: 9xZxDwAAQBAJ.

[3] Chen Tianqi, Guestrin Carlos. XGBoost: A scalable tree boosting system. In: Proceedings of the 22Nd ACM SIGKDD international conference on knowledge discovery and data mining. KDD '16, New York, NY, USA: ACM; 2016, p. 785–94.

[4] Guidotti Riccardo, Monreale Anna, Ruggieri Salvatore, Turini Franco, Giannotti Fosca, Pedreschi Dino. A survey of methods for explaining Black Box models. ACM Comput Surv 2018;51(5):93:1–42.

[5] Goodman Bryce, Flaxman Seth. European union regulations on algorithmic decision-making and a "right to explanation". AI Mag 2017;38(3):50–7.

[6] Kop Mauritz. EU artificial intelligence act: the european approach to AI. SSRN scholarly paper ID 3930959, Rochester, NY: Social Science Research Network; 2021.

[7] Holzinger Andreas, Langs Georg, Denk Helmut, Zatloukal Kurt, Müller Heimo. Causability and explainability of artificial intelligence in medicine. WIREs Data Min Knowl Discov 2019;9(4):e1312, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/widm.1312.

[8] Combi Carlo, Amico Beatrice, Bellazzi Riccardo, Holzinger Andreas, Moore Jason H, Zitnik Marinka, Holmes John H. A manifesto on explainability for artificial intelligence in medicine. Artif Intell Med 2022;102423.

[9] Shortliffe EH, Davis R, Axline SG, Buchanan BG, Green CC, Cohen SN. Computer-based consultations in clinical therapeutics: explanation and rule acquisition capabilities of the MYCIN system. Comput Biomed Res Int J 1975;8(4):303–20.

[10] Holzinger Andreas, Biemann Chris, Pattichis Constantinos S, Kell Douglas B. What do we need to build explainable AI systems for the medical domain?. 2017.

[11] Adadi Amina, Berrada Mohammed. Peeking inside the Black-Box: A survey on explainable artificial intelligence (XAI). IEEE Access 2018;6:52138–60, Conference Name: IEEE Access.

[12] Barredo Arrieta Alejandro, Díaz-Rodríguez Natalia, Del Ser Javier, Bennetot Adrien, Tabik Siham, Barbado Alberto, Garcia Salvador, Gil-Lopez Sergio, Molina Daniel, Benjamins Richard, Chatila Raja, Herrera Francisco. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fusion 2020;58:82–115.

[13] Chou Yu-Liang, Moreira Catarina, Bruza Peter, Ouyang Chun, Jorge Joaquim. Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms, and applications. 2021, arXiv:2103.04244 [Cs].

[14] Du Mengnan, Liu Ninghao, Hu Xia. Techniques for interpretable machine learning. Commun ACM 2019;63(1):68–77.

[15] Vilone Giulia, Longo Luca. Explainable artificial intelligence: a systematic review. 2020, arXiv:2006.00093 [Cs].

[16] Ribeiro Marco Tulio, Singh Sameer, Guestrin Carlos. "Why should I trust you?": Explaining the predictions of any classifier. 2016, arXiv:1602.04938 [Cs, Stat].

[17] Lundberg Scott, Lee Su-In. A unified approach to interpreting model predictions. 2017, arXiv:1705.07874 [Cs, Stat].

[18] Malolan Badhrinarayan, Parekh Ankit, Kazi Faruk. Explainable deep-fake detection using visual interpretability methods. In: 2020 3rd international conference on information and computer technologies (ICICT). 2020, p. 289–93.

[19] Zeiler Matthew D, Fergus Rob. Visualizing and understanding convolutional networks. 2013, arXiv:1311.2901 [Cs].

[20] Ribeiro Marco Tulio, Singh Sameer, Guestrin Carlos. Anchors: High-precision model-agnostic explanations. Proc AAAI Conf Artif Intell 2018;32(1). Number: 1.

[21] Bastani Osbert, Kim Carolyn, Bastani Hamsa. Interpretability via model extraction. 2018, arXiv:1706.09773 [Cs, Stat].

[22] Yang Chengliang, Rangarajan Anand, Ranka Sanjay. Global model interpretation via recursive partitioning. 2018, arXiv:1802.04253 [Cs, Stat].

[23] Shi Sheng, Zhang Xinfeng, Fan Wei. Explaining the predictions of any image classifier via decision trees. 2020, arXiv:1911.01058 [Cs, Stat].

[24] Lundberg Scott M, Erion Gabriel, Chen Hugh, DeGrave Alex, Prutkin Jordan M, Nair Bala, Katz Ronit, Himmelfarb Jonathan, Bansal Nisha, Lee Su-In. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell 2020;2(1):56–67, Bandiera_abtest: a Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Computer science;Medical research;Software Subject_term_id: computer-science;medical-research;software.

[25] Deng Houtao. Interpreting tree ensembles with inTrees. Int J Data Sci Anal 2019;7(4):277–87.

[26] Hatwell Julian, Gaber Mohamed Medhat, Azad RMuhammad Atif. CHIRPS: Explaining random forest classification. Artif Intell Rev 2020;53(8):5747–88.

[27] El Shawi Radwa, Sherif Youssef, Al-Mallah Mouaz, Sakr Sherif. Interpretability in HealthCare A comparative study of local machine learning interpretability techniques. In: 2019 IEEE 32nd international symposium on computer-based medical systems (CBMS). 2019, p. 275–80, ISSN:2372-9198.

[28] Payrovnaziri Seyedeh Neelufar, Chen Zhaoyi, Rengifo-Moreno Pablo, Miller Tim, Bian Jiang, Chen Jonathan H, Liu Xiuwen, He Zhe. Explainable artificial intelligence models using real-world electronic health record data: a systematic scoping review. J Am Med Inform Assoc 2020;27(7):1173–85.

[29] Gundersen Odd Erik. Standing on the feet of Giants — Reproducibility in AI. AI Mag 2019;40(4):9–23, Number: 4.

[30] Adhikari Ajaya, Tax David MJ, Satta Riccardo, Faeth Matthias. LEAFAGE: Example-based and feature importance-based explanations for Black-box ML models. In: 2019 IEEE international conference on fuzzy systems (FUZZ-IEEE). IEEE; 2019, p. 1–7.

[31] Breiman Leo, Friedman Jerome H, Olshen Richard A, Stone Charles J. Classification and regression trees. Routledge; 2017.

[32] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E. Scikit-learn: Machine learning in Python. J Mach Learn Res 2011;12:2825–30.

[33] Gower John C. A general coefficient of similarity and some of its properties. Biometrics 1971;857–71.

[34] Chawla Nitesh V, Bowyer Kevin W, Hall Lawrence O, Kegelmeyer W Philip. SMOTE: synthetic minority over-sampling technique. J Artificial Intelligence Res 2002;16:321–57.

[35] Johnson Alistair EW, Pollard Tom J, Shen Lu, Lehman Li-wei H, Feng Mengling, Ghassemi Mohammad, Moody Benjamin, Szolovits Peter, Anthony Celi Leo, Mark Roger G. MIMIC-III, a freely accessible critical care database. Sci Data 2016;3(1):160035, Number: 1 Publisher: Nature Publishing Group.

[36] Silva Ikaro, Moody George, Scott Daniel J, Celi Leo A, Mark Roger G. Predicting in-hospital mortality of ICU patients: The PhysioNet/Computing in cardiology challenge 2012. Comput Cardiol 2012;39:245–8.

[37] Young Jim, Graham Patrick, Penny Richard. Using Bayesian networks to create synthetic data. J Off Statist 2009;25:549–67.

[38] Kaur Dhamanpreet, Sobiesk Matthew, Patil Shubham, Liu Jin, Bhagat Puran, Gupta Amar, Markuzon Natasha. Application of Bayesian networks to generate synthetic health data. J Am Med Inform Assoc: JAMIA 2021;28(4):801–11.

[39] Onisko Agnieszka, Druzdzel Marek J, Wasyluk Hanna, Onisko Agnieszka. A probabilistic causal model for diagnosis of liver disorders. 2005.

[40] Kelly Christopher J, Karthikesalingam Alan, Suleyman Mustafa, Corrado Greg, King Dominic. Key challenges for delivering clinical impact with artificial intelligence. BMC Med 2019;17(1):195.

[41] Parimbelli E, Marini S, Sacchi L, Bellazzi R. Patient similarity for precision medicine: A systematic review. J Biomed Inform 2018;83:87–96.

[42] McCoy Liam G, Brenna Connor TA, Chen Stacy S, Vold Karina, Das Sunit. Believing in black boxes: machine learning for healthcare does not need explainability to be evidence-based. J Clin Epidemiol 2021;S0895–4356(21)00354–1.

[43] Shmueli Galit. To explain or to predict? Statist Sci 2010;25(3):289–310, Publisher: Institute of Mathematical Statistics.

[44] Nicora Giovanna, Bellazzi Riccardo. A reliable machine learning approach applied to single-cell classification in acute myeloid leukemia. In: AMIA annual symposium proceedings, Vol. 2020. 2021, p. 925–32.

[45] Nicora Giovanna, Rios Miguel, Abu-Hanna Ameen, Bellazzi Riccardo. Evaluating pointwise reliability of machine learning prediction. J Biomed Inform 2022;103996.

[46] Finlayson Samuel G, Subbaswamy Adarsh, Singh Karandeep, Bowers John, Kupke Annabel, Zittrain Jonathan, Kohane Isaac S, Saria Suchi. The clinician and dataset shift in artificial intelligence. N Engl J Med 2021;385(3):283–6.

[47] Guo Lin Lawrence, Pfohl Stephen R, Fries Jason, Posada Jose, Fleming Scott Lanyon, Aftandilian Catherine, Shah Nigam, Sung Lillian. Systematic review of approaches to preserve machine learning performance in the presence of temporal dataset shift in clinical medicine. Appl Clin Inform 2021;12(4):808–15, Publisher: Georg Thieme Verlag KG.

[48] Selvaraju Ramprasaath R, Cogswell Michael, Das Abhishek, Vedantam Ramakrishna, Parikh Devi, Batra Dhruv. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision (ICCV). 2017.

[49] Lundberg Scott M, Erion Gabriel G, Lee Su-In. Consistent individualized feature attribution for tree ensembles. 2019, arXiv:1802.03888 [Cs, Stat].

[50] Cabalar Pedro, Muñiz Brais, Pérez Gilberto, Suárez Francisco. Explainable Machine Larning for liver transplantation. 2021, arXiv:2109.13893 [Cs].

[51] Guidotti Riccardo. Evaluating local explanation methods on ground truth. Artificial Intelligence 2021;291:103428.

[52] Holzinger Andreas. Interactive machine learning for health informatics: when do we need the human-in-the-loop? Brain Inform 2016;3(2):119–31.

[53] Ehsan Upol, Wintersberger Philipp, Liao Q Vera, Mara Martina, Streit Marc, Wachter Sandra, Riener Andreas, Riedl Mark O. Operationalizing human-centered perspectives in explainable AI. In: Extended abstracts of the 2021 CHI conference on human factors in computing systems. 2021, p. 1–6.