International Journal of Data Science and Analytics
https://doi.org/10.1007/s41060-023-00445-1

REGULAR PAPER

Through the looking glass: evaluating post hoc explanations using transparent models

Mythreyi Velmurugan[1,2] · **Chun Ouyang**[1,2] · **Renuka Sindhgatta**[3] · **Catarina Moreira**[4]

Received: 13 April 2023 / Accepted: 9 August 2023

Abstract
Modern machine learning methods allow for complex and in-depth analytics, but the predictive models generated by these methods are often highly complex and lack transparency. Explainable Artificial Intelligence (XAI) methods are used to improve the interpretability of these complex "black box" models, thereby increasing transparency and enabling informed decision-making. However, the inherent fitness of these explainable methods, particularly the faithfulness of explanations to the decision-making processes of the model, can be hard to evaluate. In this work, we examine and evaluate the explanations provided by four XAI methods, using fully transparent "glass box" models trained on tabular data. Our results suggest that the fidelity of explanations is determined by the types of variables used, as well as the linearity of the relationship between variables and model prediction. We find that each XAI method evaluated has its own strengths and weaknesses, determined by the assumptions inherent in the explanation mechanism. Thus, though such methods are model-agnostic, we find significant differences in explanation quality across different technical setups. Given the numerous factors that determine the quality of explanations, including the specific explanation-generation procedures implemented by XAI methods, we suggest that model-agnostic XAI methods may still require expert guidance for implementation.

Keywords Predictive models · Explainable AI (XAI) · Performance evaluation of algorithms and systems · Explanation fidelity

1 Introduction

Modern machine learning and deep learning techniques have allowed the analysis and modelling of complex data to enable decision-making. However, these advanced machine learning techniques are often not human interpretable, and therefore, lack transparency and hamper responsible and accountable human decision-making [23]. Explainable AI (XAI) methods are used to improve the interpretability of these complex "black box" models, thereby increasing transparency and enabling informed decision-making [18]. A key category of XAI techniques is *post hoc explainable methods*; that is, external mechanisms that provide explanations for a given predictive model, rather than an interpretation provided directly by the model.

Despite the increasing use of post hoc XAI methods in the literature, evaluations and comparisons of post hoc explainable methods are so far under-explored. In particular, past evaluations of *explanation fidelity*, a measure to assess the correctness and completeness of an explanation with respect to the underlying model [59], are specific to particular classes of datasets or specific explainable methods. A number of easily accessible, open source XAI techniques, such as LIME [41] and SHAP [27] are commonly used, as they are model-agnostic and can be easily applied to pre-existing predictive models. However, the ability of such techniques to mimic the

✉ Mythreyi Velmurugan
m.velmurugan@qut.edu.au

Chun Ouyang
c.ouyang@qut.edu.au

Renuka Sindhgatta
renuka.sr@ibm.com

Catarina Moreira
catarina.pintomoreira@uts.edu.au

1 School of Information Systems, Queensland University of Technology, Brisbane, Australia

2 Centre for Data Science, Queensland University of Technology, Brisbane, Australia

3 IBM Research, Bangalore, India

4 Human Technology Institute, University of Technology Sydney, Sydney, Australia

Published online: 12 September 2023

workings of predictive models of different classes remains unknown.

Evaluations of fidelity in the literature can generally be classified as one of the following: *external fidelity*, which assesses how well the prediction of the underlying model and the prediction implied by the explanation and/or explanation generation mechanism agree; and *internal fidelity*, which assesses how well the explanation matches the decision-making processes of the underlying model [32]. While methods to evaluate external fidelity are relatively common in the literature [19, 25, 33, 46], internal fidelity evaluations are generally limited to text and image data [13, 16, 26, 38, 44], rather than tabular data. Thus, the capability of an explanation to mimic the decision-making processes of the underlying model, particularly models built on tabular data, is unknown.

In this paper, we attempt to understand the strengths and limitations of explanations provided by four XAI methods. We generate explanations for fully transparent "glass box" predictive models and compare explanations directly against the underlying model. This facilitates the evaluation of an explanation's ability to reflect the behaviour of the model. In particular, we focus on local explanations, though the techniques used to generate these explanations often vary between, and even within, explainable method types. The goal of our evaluation is not to identify "the best" explainable method, but to understand the strengths and weaknesses of different explanation-generating mechanisms and the contexts that each explainable method can support.

The main contributions of this work are:

- the development of a method to compare local explanations against the decision-making of three classes of fully transparent predictive models;
- evaluation of four XAI methods in various experimental setups; and
- insights to support the choice of XAI methods to be used within a particular context.

This paper is structured as follows. We first present a number of related works, along with the motivations of this work, in Sect. 2. This is followed by a proposal of our evaluation method (Sect. 3) and design of experiments (Sect. 4). We present our results in Sect. 5 and analyse the results and draw insights in Sect. 6. Finally, we conclude our paper and discuss future work (Sect. 7).

2 Background and related works

2.1 Explainable AI

The "black box problem" of AI arises from the inherent complexity and sophisticated internal data representations of many modern machine learning algorithms [26]. Although more complex predictive models may produce more accurate results, this accuracy comes at the cost of human interpretability of algorithmic decisions [24]. XAI research attempts to provide human-understandable explanations for predictive models. This is necessary to ensure system quality and to facilitate informed human decision-making [8].

In this work, we consider "interpretability" to be the ability to provide meaning in terms understandable to a human and "explanations" to be the interface between the human and the predictive model [18]. Interpretability in machine learning is generally broken down into two categories: inherently interpretable predictive models and post hoc interpretation. Interpretable predictive models are those that are immediately interpretable by a human [18], though this often means that the models are simpler, and so may have reduced predictive power. This category includes simple decision trees and linear and logistic regression models, which are easily interpretable, and thus *fully transparent*.

Post hoc methods provide interpretations that are not inherent to the predictive model and are derived after its creation through an external mechanism [18]. Post hoc methods are usually applied to *opaque* models, which have complex internal mechanisms, such as a neural network, or are otherwise "black box". A sufficiently complex tree-ensemble model or a model protected by IP agreements may also fall into this category [43]. Thus, post hoc interpretability allows for the use of more complex and accurate predictive models without compromising interpretability. However, since post hoc interpretation is external to the predictive model, there is no guarantee that the explanations provided are fully and correctly representative of the underlying model [43].

2.2 Post hoc explainable methods

The purpose and scope of post hoc explanations often vary [29]. In this work, we assess four local (prediction–level) explanations, rather than global (whole–model) explanations. That is, when given an individual input $x \in X$ of length n and a prediction function f, a local XAI method would return an explanation $\phi(x)$, rather than $\phi(f)$. In this work, we use four methods that generate local, post hoc explanations.

Moreover, we examine XAI methods that are model-agnostic. These methods typically use some explanation-generation mechanism external to the underlying predictive model, and/or use multiple mechanisms suited for examining multiple types of predictive models. We describe these in more detail below.

2.2.1 LIME

Local Interpretable Model-Agnostic Explanations (LIME) is a local feature attribution method that highlights how important each feature is to a prediction [41]. Given an instance for which one wants to generate explanations, x, LIME first applies a set of permutations $x' \in X'$ and then, generates predictions over those permutations, $f(X')$, to train a more interpretable surrogate model. The surrogate model $g \in G$ that mimics the behaviour of f in the vicinity of x is obtained by:

$$\underset{g \in G}{\operatorname{argmin}}\, \mathcal{L}(f, g, \pi_x) + \Omega(g) \tag{1}$$

where $\Omega(g)$ is defined as the complexity of g (the number of nonzero coefficients in the linear surrogate model), π_x is a distance function centred on x and the loss function $\mathcal{L}$ is defined as:

$$\mathcal{L}(f, g, \pi_x) = \sum_{x' \in X'} \pi_x(x')[f(x') - g(x')]^2 \tag{2}$$

The coefficients of g are returned as the explanation $\phi(x)$.

2.2.2 SHAP

Much like LIME, Shapley Additive Explanations (SHAP) is a feature attribution method and produces SHAP values as a measure of a feature's contribution [27]. SHAP values for a single feature i can be calculated as:

$$\phi_i(f, x) = \sum_{z' \subseteq z} \frac{|z'|!(M - |z'| - 1)!}{|M|!}[f_x(z') - f_x(z' \backslash i)] \tag{3}$$

where:

- z is a simplified representation of x,
- $z' \subseteq z$ represents all z' vectors where the nonzero entries are a subset of the nonzero entries in z,
- $|z'|$ is the number of nonzero entries in z', and
- M is the number of features in z

Rather than attempting to estimate the importance of features to the model, SHAP defines the contribution of a feature ϕ_i such that:

$$f(x) = \phi_0 + \sum_{i=1}^{M} \phi_i z_i \tag{4}$$

where ϕ_0 is generally defined as $f(\varnothing)$ and calculated as the expected value of the prediction function f. Both model-agnostic [27] and model-specific [27, 28] approximations of SHAP values exist and are used in this work (see Sect. 4).

2.2.3 LINDA-BN

Unlike LIME and SHAP, Local Interpretation-Driven Abstract Bayesian Network (LINDA-BN) provides a profile-like explanation instead of a feature attribution explanation [34]. Similarly to LIME, LINDA-BN creates a surrogate model in a neighbourhood of x within a variance of ϵ. However, instead of using Linear Regression to fit the permuted data points, LINDA-BN learns a Bayesian network using the Greedy Hill Climbing approach. This Bayesian network is returned as the explanation for a prediction. Therefore, this method expresses explanations as correlations between features, rather than assuming that each feature individually contributes to the prediction. LINDA-BN is considered a profile-like explanation, as the extracted Bayesian network enables the identification of four rules to inform a decision-maker about the potential correctness of a single prediction (Fig. 1) [34].

Since LINDA-BN's explanations are grounded in probabilistic graphical models, they fulfil the axioms of probability theory and the rules that derive from them. Thus, the structure of the explanation, i.e. the Bayesian network, can be used to determine the model's confidence in and the likely correctness of a prediction [34].

Although LINDA-BN returns only the surrogate model, a quantified impact of each feature can be calculated by querying the returned Bayesian network (see Sect. 4).

2.2.4 ACV

Active Coalition of Variables (ACV) is a probabilistic method of explanations that aims to find some set of minimal suffi-

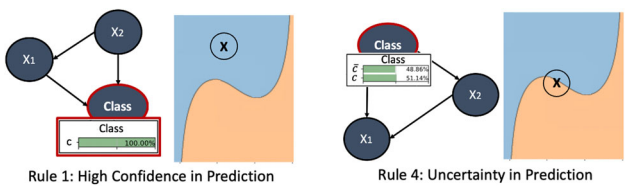


Fig. 1 Outputs from LINDA-BN [34]: when the data instance to be explained falls in a well-defined region of the decision space, then there is high confidence in the generated predictions (Rule 1). When the data instance falls near the decision boundary, the returned networks will show uncertainty towards the prediction (Rule 4)

cient features (minimal sufficient explanations, M-SE) such that retaining only the M-SE in x returns the same or similar $f(x)$ given a high probability π [5]. That is, given an instance x, prediction $f(x)$ and threshold t, $x_s \subseteq x$ is a Sufficient Explanation for probability π if the same decision probability $\text{SDP}_s(x, f(x), t) \geq \pi$ and no $x_z \subseteq x_s$ satisfies $\text{SDP}_z(x, f(x), t) \geq \pi$ where:

$$\text{SDP}_s(x, f(x), t) = \Pr((f(X) - f(x))^2 \leq t \mid X_s = x_s) \tag{5}$$

for regression, and for classification:

$$\text{SDP}_s(x, f(x)) = \Pr(f(X) = f(x) \mid X_s = x_s) \tag{6}$$

A M-SE is a Sufficient Explanation with minimal size, and the set of all Sufficient Explanations is denoted as A-SE.

ACV uses a Random Forest model as a global surrogate, i.e. a surrogate model that attempts to capture the entirety of model behaviour, instead of some subset of behaviour. The idea of Projected Forest, which is used to estimate $E[f(X) \mid X_s = x_s]$, is combined with a Quantile Regression Forest, which estimates $Pr(f(X) \leq f(x) \mid X = x)$ to compute the SDP [5]. As the number of possible subsets is exponential, to reduce the complexity of the computation, the search for Sufficient Explanations is contained to a maximum of ten features that are used most frequently in the surrogate model.

ACV is also distinct from the other XAI techniques used in this work, as it can also function as a self-explaining predictive model. When explaining a separate predictive model f, rather than taking the model as input, as LIME, SHAP and LINDA-BN do, ACV's surrogate model takes as input the training set X and the original model's predictions $f(X)$. However, if given X and the associated labels Y, ACV can also be trained as a predictive model, though we do not use it as such in this work.

2.3 Evaluating explanations

Evaluation methods for XAI can be categorised into three levels of evaluation [12]:

- *Application-grounded evaluation* wherein the evaluation is conducted with real end users in full context replicating a real-world application;
- *Human-grounded evaluation* wherein the evaluation is conducted with laymen in simpler simulated contexts or completing proxy tasks that reflect a target application context; and
- *Functionally-grounded evaluation* which requires no input from users and relies on evaluating the inherent abilities of the system via a formal definition of interpretability.

The former two categories are also often referred to as *cognitive metrics* of evaluation, and functionally-grounded evaluations fall into the complementary category of *computational metrics* [26].

Evaluations of this last category are often the first step in determining the quality of an XAI method [12]. This includes evaluation of *explanation fidelity* with respect to the original predictive model. The term "fidelity" is generally defined as a measure of how faithful and accurate an explanation is to the underlying black box model [56]. Two characteristics of explanations are relevant to explanation fidelity [30, 59]: the *completeness* of an explanation in capturing the dynamics of the underlying model; and the *correctness* and truthfulness of the explanation with respect to the underlying model.

Explanation fidelity can severely impact the safety and usefulness of a predictive system. The authors of [21] find that, when the user has deep knowledge of the context, local explanations enabled the identification of prediction errors and the predictive model's error boundaries. The study also finds that explanations affect the user's perception of model correctness. Similarly, the authors of [47] find that model transparency affects causability—the ability of the user to understand what features affect the prediction. In turn, causability is found to determine the user's trust in and perception of model performance. We must note that explanation fidelity does not play a direct role in causability or user perception; these are generally determined by explanation presentation and content [21, 58]. However, if a user's knowledge or perception is based on incorrect or incomplete explanations of the model, it may engender trust in a false prediction, or mistrust in a correct prediction, leading to unsafe or incorrect usage of a predictive system. Thus, explanation fidelity is necessary for predictive model safety and usefulness.

Although a number of methods have been proposed to assess fidelity, they are often highly specific to particular XAI methods or datasets. Fidelity evaluation approaches fall into two general measures: measurements of *external fidelity* and *internal fidelity* [32]. External fidelity approaches are those that compare how often the decision implied by the explanation and the decision made by the underlying predictive model agree. For example, when evaluating surrogate models that approximate the underlying predictive model, this method determines the accuracy of the surrogate model's predictions using the predictions of the underlying model as ground truth [46]. Explainable methods that produce decision rules are also often evaluated through this approach, and the predictions of the decision rule or decision rule sets are compared against the black box model's predictions [19, 25, 33].

Evaluating external fidelity provides an assessment of consistency between the explanation's predictions and the predictions of the underlying model, i.e. that the predictive model and the explanation mechanism can reach the same

decision. However, this does not guarantee that the explanations are faithful to the computations, or "decision-making", of the black box model, for example, weights given to each feature, or the decision path followed in a tree [32]. Such information may be necessary in order to change the predicted outcome to a desired outcome. For example, in the case of a risk prediction model, taking the most appropriate action to reduce the risk. Lack of full transparency can also hamper model inspection and diagnosis, which may be necessary to ensure that the model relies on appropriate features when computing a prediction [41]. Moreover, some explainable methods do not produce an output suitable for external fidelity evaluation approaches, including LINDA-BN and some optimisations of SHAP.

As such, we aim to develop a method to assess the internal fidelity of feature attribution explainable methods, where the fidelity of the explanation regarding the underlying model's decision-making process is evaluated [32]. As noted in [6], given that the internal workings of a black box model are opaque and there is no "ground truth" of the model, it is near impossible to assess explanation correctness. Significant effort is required to assess the internal fidelity of an explanation or explainable method. Potential internal fidelity evaluation approaches include:

1. Generating explanations for a fully transparent, inherently interpretable "glass box" model, and comparing the explanation to the glass box model's decision-making [4, 41];
2. Using a post hoc explanation approach to explain both a surrogate model and the underlying predictive model, and comparing how often the explanations concur [32]; and
3. "Removing" or permuting the features identified as relevant by the explanation and measuring the change in model output [4, 13, 16, 26, 37, 38, 44].

It is important to note that the method in [32] is an evaluation of a surrogate model using a post hoc technique, rather than an evaluation of the post hoc technique. Thus, it is not suitable for this work. As another note, while the method of permuting features is relatively common in the literature, it is mostly applied to text [13, 38] or image data [16, 26, 44], in which case, the relevant features are often removed or blurred from the original input prior to encoding. This method has been applied to tabular data [4, 31, 37], typically by permuting rather than removing features at the input level [4, 37] or removing the feature as a whole from the dataset [31]. Moreover, we note that these works typically used few, standard benchmark datasets and/or limited predictive models. In our previous work, we adopted a local, perturbation method to evaluate explanations for models built on time series tabular data, which, however, provided poor results [49]. We theorise that the results of our previous work were likely influenced by the difficulty in choosing and permuting an appropriate subset of features given the complexity of time series data, and lack of understanding of model and XAI technique behaviour with respect to tabular data.

Tabular datasets are highly heterogeneous with respect to the quantity and type of features and may present challenges such as lack of locality and sparsity in data, as well as the need to manage mixed feature types [48]. Furthermore, there may be fewer correlations between features in tabular data than in other types of data such as images, as well as lack of semantic meaning and connection between features such as may be present in natural language data [7]. Although it is noted in [17] that simpler, glass-box models can generally provide accurate predictions for tabular data, the authors of the study also note that in some cases, particularly when the dataset is "noisy", more complicated black box models may be necessary. These factors affect the predictive techniques that may be used [7], but, as we noted in [49], the characteristics of tabular data also does not allow for use of explanation evaluation methods developed for image and text data.

2.4 Motivation

In this work, we use fully transparent, glass-box models to better understand interactions between data, model and XAI technique. This provides a "ground truth" for model decision-making, with which to assess explanations. Furthermore, we extend the approach used in [41], which relied on measuring only the completeness of the explanation in capturing a hard-coded, arbitrary number of features. Moreover, to the best of our knowledge, no work compares the performance of XAI techniques in different technical contexts to understand the factors that affect explanation quality. In particular, we aim to examine whether easily accessible, model-agnostic XAI methods, such as those described, work reliably and consistently in all technical contexts. Thus, in this work, we attempt to take a more holistic approach to evaluation, as described in Sect. 3.2.

We investigate properties of post hoc XAI methods by addressing the following questions:

- Local surrogate models are commonly used to generate local, post hoc explanations. *To what degree do permutation and the choice of local surrogate model affect the quality of explanation generated by the surrogate model?*
- Model-agnostic XAI methods often use some theoretical construct to define and calculate explanations. For example, principles from game theory are used to derive SHAP explanations [27, 28] and statistical models are applied by LINDA-BN [34]. However, some assumptions and design decisions may be needed to make these theoretical foundations suitable for application. *How do the design*

choices made in the implementation of the explanation generation mechanism affect explanation quality?

- While many local explanation methods use local surrogate models (at the data point level), a few also use global surrogate models (i.e. at the dataset level). *How does the use of a global surrogate model affect explanation quality?*

3 Method

3.1 Evaluation method

In this work, we aim to investigate the internal fidelity of model-agnostic explanations, where the faithfulness of the explanation to the underlying model's computational process is evaluated [32]. Given that the internal workings of a black box model are opaque, significant effort is required to assess the internal fidelity of an explanation or explainable method. In our previous work [49], we attempted to assess explanation fidelity for black box models trained on tabular data, by replicating the methods used to test the internal fidelity of explanations for models trained on image and text data. Our results suggested that explanation fidelity is poor when explaining tabular data, though we found that this may be a consequence of the evaluation method and the complexity of the datasets used. This lack of evaluation method for tabular data is notable given the volume and diversity of tabular data used in modern data analytics.

Thus, in this work, we evaluate explanations using fully transparent glass-box models. We choose models from which the relevance of a feature to a prediction can be determined, to provide a reference for evaluating the internal fidelity of post hoc explanations. As the internal workings of a glass-box model are apparent, it can be compared against an explanation to determine the fitness of the explainable method. It must be noted that high explanation fidelity for a relatively simple glass box method does not necessarily imply high fidelity for a more complicated black box model. As such, our aim in using this method is not to determine some "best" XAI method. Rather, we aim to more precisely understand the specific workings of each XAI method, how different types of predictive models affect explanation quality, and any patterns of quality that could inform the selection of XAI methods for a given setup and context. Thus, we focus on behaviours and characteristics of models and XAI algorithms. Moreover, given the ease of availability of model-agnostic XAI techniques, we also wish to determine the level of expertise needed to select an XAI technique for use.

In this work, we choose to examine four local, post hoc XAI methods of differing characteristics in order to answer the questions posed:

- **LIME** uses permutation, then a local, linear surrogate model to derive an explanation. Previous works have suggested that LIME's permutation of input x affects explanation stability, particularly as the length of the input increases [46, 50]. In addition to the effect of permutation, in this work we will examine the effects of the choice of local surrogate model.
- While **SHAP** is grounded in game theory, it uses both model-specific and model-agnostic computations to determine feature attribution [27, 28].
- **LINDA-BN** is grounded in statistical modelling and provides a queryable explanation.
- **ACV** uses a global, rather than local, surrogate model to generate explanations. Moreover, ACV provides explanations of multiple types, including feature subsets and feature attribution.

We have chosen datasets and predictive models in such a way as to facilitate answers to these questions. Our experimental setup and choice of algorithms are introduced in 4.

3.2 Evaluation metrics

In this work, the goal of our evaluation is to understand the strengths and limitations of several local, post hoc XAI mechanisms. In particular, we examine these mechanisms with respect to the characteristics of the underlying models and datasets. Thus, in this work, we use a range of datasets, XAI techniques and predictive models for a comprehensive evaluation. We evaluate four XAI techniques, all of which use different explanation-generation mechanisms, with three predictive models of different classes. We also selected datasets by considering a breadth of different characteristics, including the prediction problem, the volume of the data, types of variables present, and the number of features that can be encoded from the data. A detailed description of techniques and datasets used is presented in Sect. 4.

We evaluate three properties when measuring fidelity:

- The correctness of the explanation in identifying the most impactful features;
- The completeness of the explanation in identifying the most impactful features; and
- The correctness of the importance ranking of all features, which is particularly key for feature attribution methods.

The evaluation method for the first two properties use two subsets of features. The basic approach of this method is to compare the subset of features determined to be most relevant by the model ($x_t \subseteq x$) against the features determined to be the most relevant by the explanation ($x_e \subseteq x$). A similar approach was used in [41], where the recall metric was used

to determine fidelity in the form of *completeness* as follows:

$$\text{Recall}(x_t, x_e) = \frac{|x_t \cap x_e|}{|x_t|} \tag{7}$$

Additionally, we also use the precision metric to capture the *correctness* of the explanation as follows:

$$\text{Precision}(x_t, x_e) = \frac{|x_t \cap x_e|}{|x_e|} \tag{8}$$

To evaluate the third property, we extract the importance ranking of features by the model (r_t) and the ranking of features by the explanation (r_e), where r is a reordering of x indicating feature importance. We then use a correlation measure to determine the similarity between r_t and r_e, where a higher similarity would indicate higher *correctness* of the explanation's ranking. We use Kendall's Tau-B to measure rank correlation between r_t and r_e.

As XAI methods, particularly methods relying on feature permutations, often produce unstable explanations [46, 50, 51], we generate five $\phi(x)$ for each x, compute an adjusted or average explanation $\varphi(x)$, and determine x_e and r_e based on $\varphi(x)$.

These metrics are applied at the local level (i.e. when testing individual inputs), but can be averaged out at the dataset level, as in Sect. 5.

Note that, of the chosen predictive methods, only the decision tree model provides a natural subset to be used as x_t. Similarly, only the ACV provides a subset of features as $\phi(x)$. Thus, a heuristic is necessary to identify x_t and x_e for other methods. The procedure for determining this heuristic, as well as the results, are summarised in Sect. 4.4.

4 Design of experiments

4.1 Datasets

We focus our evaluations on tabular data, which has been relatively unexplored in the literature on explanation fidelity. In order to better understand the effect of dataset properties on the results, we chose datasets with specific characteristics, rather than datasets from any particular domain. The datasets all vary in terms of the variable types present, the number of variables present once the dataset is encoded as described below, the volume of the data, and, in the case of the regression datasets, the distribution of the target variable.

A total of 14 open source datasets are used, which include seven *classification* datasets:

1. Adult Income [22]
2. Breast cancer [55]
3. COMPAS [39]
4. Diabetes [2]
5. Iris [15]
6. Mushroom [45]
7. Nursery [40]

and seven *regression* datasets:

1. Bike Rentals [14]
2. Facebook [35, 36]
3. Housing [1]
4. Real Estate [57]
5. Solar Flare [3]
6. Student Scores [10]
7. Wine Quality [11]

A brief profile of each dataset is provided in Tables 1 and 2.

Some pre-processing was necessary for all datasets. All classification datasets were balanced through downsampling, to ensure parity between classes. Where a classification dataset had more than two prediction classes, the target variable was binarized before downsampling. All categorical variables in the datasets were one-hot encoded, and min-max scaling was applied to all numeric variables (excluding the target variable in regression datasets). A train-test split of 70–30 was used, and a sample of the testing set was used to evaluate the XAI methods.

4.2 Predictive models

In this work, we use predictive model types for which extracting x_t is unambiguous. As such, we choose to exclude techniques for which x_t cannot be easily derived, or are unclear. For example, attention mechanisms are an increasingly popular technique used to interpret models without post hoc methods. This is a form of input selection within some neural networks, which allows the model to focus on inputs most relevant to the output [23]. While attention mechanisms are sometimes used as explanation and have been shown to be successful in some domains [53], the validity of attention as explanation is still debated [20, 31, 54]. Thus, we choose not to use it in this work.

We use four different algorithms of three different classes to create predictive models. Firstly, we use a simple CART decision tree to create both classification and regression models, covering all datasets. For all classification datasets, we also create logistic regression and Naïve Bayes models, and for all regression datasets, we create linear regression models. In summary, we have conducted our evaluations using a total of 35 models: 21 classification models and 14 regression models. The accuracy of these models is shown in Tables 3 and 4.

Table 1 A brief profile of the classification datasets used in this work

Dataset	Variable types	Num variables	Training instances	Class balance (%)
Adult Income	Mixed	104	10,977	50.47
Breast Cancer	Continuous	30	296	51.01
COMPAS	Mixed	20	2793	50.13
Diabetes	Continuous	8	375	50.4
Iris	Continuous	4	70	52.86
Mushroom	Discrete	117	5842	50.09
Nursery	Discrete	27	6048	50.39

Table 2 A brief profile of the regression datasets used in this work

Dataset	Variable types	Num variables	Training instances	Target distribution
Bike Rentals	Mixed	62	12,165	Exponential
Facebook	Discrete	49	349	Exponential
Housing	Mixed	23	354	Normal
Real Estate	Continuous	6	289	Normal
Solar Flare	Discrete	32	972	Exponential
Student Scores	Mixed	58	454	Normal
Wine Quality	Continuous	11	3428	Normal

Table 3 Model accuracy on all classification datasets

Dataset	Decision tree	Logistic regression	Naïve Bayes
Adult Income	0.82	0.82	0.81
Breast Cancer	0.88	0.98	0.95
COMPAS	0.71	0.73	0.72
Diabetes	0.69	0.71	0.68
Iris	1.00	1.00	1.00
Mushroom	1.00	1.00	1.00
Nursery	1.00	1.00	1.00

Table 4 Model R-squared error on all regression datasets

Dataset	Decision tree	Linear regression
Bike Rentals	0.88	0.68
Facebook	0.42	0.26
Housing	0.63	0.71
Real Estate	0.55	0.45
Solar Flare	0.14	0.17
Student Scores	0.79	0.90
Wine Quality	0.28	0.29

Given their differences, the extraction of x_t from each type of model varies. We provide a summary of the used methods in Table 5.

4.2.1 Decision tree

Given that a decision tree model is extremely structured, identifying x_t is relatively simple. When using a decision tree model, we traverse the tree to identify the decision path for input x and take as x_t the unique set of features that fall along the decision path (see Fig. 2). We then calculate the ranking of all features r_t based on each feature's position and frequency on the decision path.

4.2.2 Linear and logistic regression

Given a linear regression model:

$$f(x) = w_0 + w_1 x_1 + \cdots + w_n x_n \tag{9}$$

and a logistic regression model:

$$f(x) = \frac{1}{1 + \exp(-(w_0 + w_1 x_1 + \cdots + w_n x_n))} \tag{10}$$

where w_0 is the intercept and $w \ni w_1 \ldots w_n$ are the coefficients applied to each feature. We consider the weights in the top 5% of the range of $w' \ni |w_1| \ldots |w_n|$ to be the most relevant to the model and take $x_t \ni x_i$ for $i \in n$ if $|w_i| \geq p$ where:

$$p = \max w' - (\max w' - \min w') \times 0.05 \tag{11}$$

We order the features in x by the absolute values of their coefficients $|w_1| \ldots |w_n|$ to determine r_t.

Table 5 A brief summary of the predictive models used in this work

Model	Applied To	Feature extraction	Ranking extraction
Decision tree	All Datasets	Features along the decision path were used as true features	Features were ranked in order and frequency of appearance on the decision path
Logistic regression	All Classification Datasets	Features with coefficients in the top 5% of the range of coefficients	Features were ranked in order of the absolute values of their coefficients
Linear regression	All Regression Datasets	Features with coefficients in the top 5% of the range of coefficients	Features were ranked in order of the absolute values of their coefficients
Naïve Bayes	All Classification Datasets	Features for which the difference in likelihoods given each class were in the top 5% of the range of likelihoods	Features were ranked in order of the absolute values of the difference in likelihoods given each class

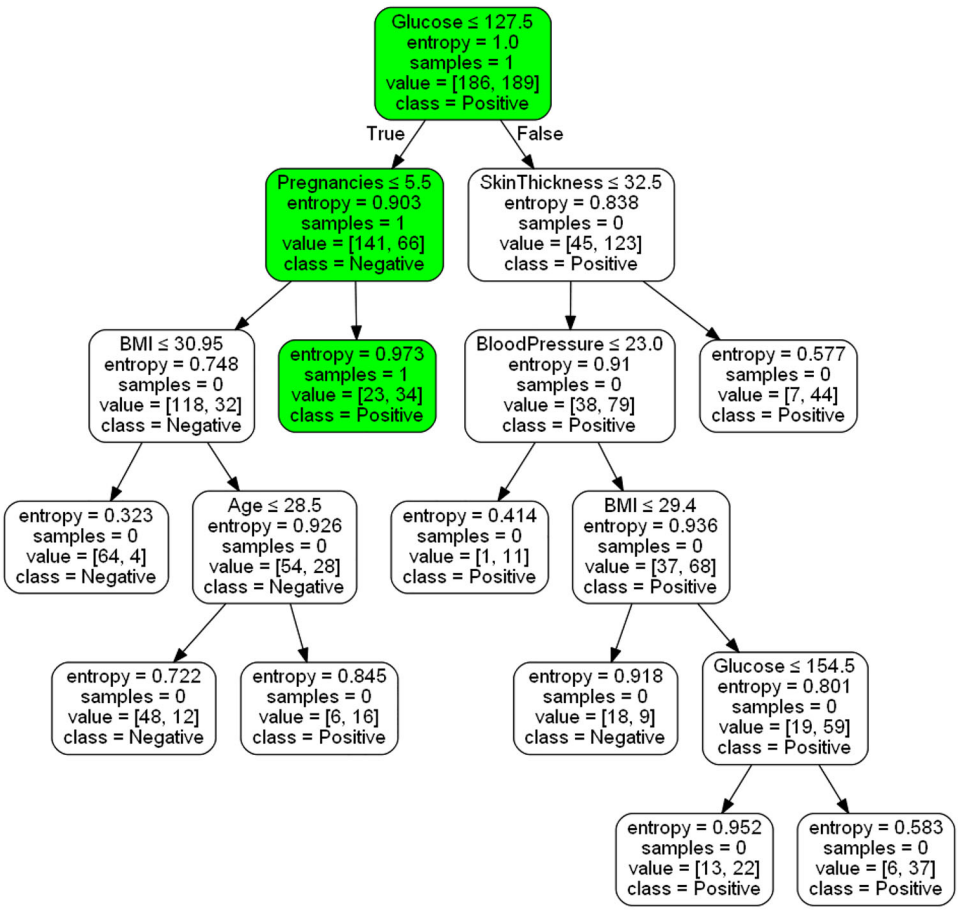


Fig. 2 Prediction of Diabetes, using a decision tree. The highlighted decision path shows all the True Features x_t that are impacting the prediction for a single instance

4.2.3 Naïve Bayes

Since a Naïve Bayes algorithm applies Bayes' theorem with the assumption that all features are conditionally independent from one another, for all $c \in C$ classes in a dataset and given an instance x of length n, $\Pr(c \mid x)$ is calculated as:

$$\Pr(c \mid x) = \Pr(c) \prod_{i=1}^{n} \Pr(x_i \mid c) \tag{12}$$

to determine the most probable c.

As such, we calculate the importance of each feature i given a Naïve Bayes model f and instance x as:

$$w_i = \Pr(x_i \mid f(x)) - \Pr(x_i \mid 1 - f(x)) \quad (13)$$

We calculate $w_1 \ldots w_n \in w'$ and take $x_t \ni x_i$ for $i \in n$ if $|w_i| \geq p$ where p is calculated using Eq. 11. Since we use a Gaussian Naïve Bayes model, the distribution of $Pr(x_i \mid f(x))$ is assumed to be Gaussian and easily calculated using the training set. We order the features in x by the absolute values of w (i.e. $|w_1| \ldots |w_n|$) to determine r_t.

4.3 XAI techniques

what is he

We use and evaluate four XAI methods in this work, each of which use different underlying mechanisms. The format of the provided explanations also varies across the explainable methods. The method by which we extract x_e from each method is described in detail below. A summary of the XAI techniques used, the feature extraction method and the experiments that it is used for is provided in Table 6.

As specified, one of the goals of this work is to understand the effect of explanation mechanisms on explanation fidelity. Thus, we examine XAI techniques with different underlying theoretical principles and explanation-generation procedures. LIME uses linear, local surrogate models to determine feature importance [41], while SHAP is grounded in game theory and differs its approach to determining the contribution of a feature depending on the underlying model [27, 28]. LIME and SHAP are also among the most commonly used local, post hoc techniques for tabular data, and thus, important candidates for evaluation. Although LINDA-BN also uses local surrogate models to generate explanations, unlike LIME, it is grounded in probability theory and uses a Bayesian Network as surrogate model [34]. Moreover, unlike LIME and SHAP, LINDA-BN does not attempt to determine feature importance to the prediction, but examines the conditional dependence between all variables, including the features. Similarly, ACV is also not a feature attribution method, but attempts to identify which subsets of features are necessary for the prediction [5]. Thus, we choose LINDA-BN and ACV in contrast to the more "conventional" LIME and SHAP.

4.3.1 LIME

As a model-agnostic method, LIME was used across all experiments. For a user-defined number of features s, $x_s \subseteq x$, LIME returns weights for each feature in x_s, i.e. $w_1..w_s \in \phi(x)$. For each x, we set $s = n$, generate five explanations, and calculate $|\overline{w_1}| \ldots |\overline{w_s}| \in \varphi(x)$.

For $i \in n$, we take $x_e \ni x_i$ if $\overline{w_i} \geq p$ where:

$$p = \max \varphi(x) - (\max \varphi(x) - \min \varphi(x)) \times 0.05 \quad (14)$$

We order the features in x by the weights in $\varphi(x)$ to extract r_e.

4.3.2 SHAP

Although SHAP is model-agnostic, specific optimisations have also been developed for tree-based and linear and logistic regression models. Thus, we use TreeSHAP when explaining the decision tree models and LinearSHAP when explaining the logistic regression and linear regression models. Model-agnostic computations of SHAP are dependent on the number of features present in the dataset; thus, we use two different SHAP optimisations for the Naïve Bayes models. We use the ExactExplainer to explain the Naïve Bayes classifiers for the Iris and Diabetes datasets, and the PermutationExplainer to explain all other Naïve Bayes models.

SHAP provides its explanations as SHAP values, where $\phi(x)$ is single vector of length n for regression models. When explaining classification models, $\phi(x)$ is a vector of SHAP values for each possible class $c \in C$, where each $\phi_c(x)$ is of length n. Though the underlying mechanisms of LIME and SHAP are different, SHAP's explanation is similar in structure to that of LIME, such that $\phi(x) \ni v_1 \ldots v_n$. Therefore, we repeat our feature extraction method for LIME, calculating $|\overline{v_1}| \ldots |\overline{v_n}| \in \varphi(x)$ from five explanations. For $i \in n$, we take $x_e \ni x_i$ if $\overline{v_i} \geq p$ where p is determined by Eq. 14. As with LIME, we order the features based on $\varphi(x)$ to determine r_e.

4.3.3 LINDA-BN

Although LINDA-BN is also a model-agnostic method, as the explanation produced is a probabilistic model, it can currently only be used to explain classifiers. Thus, it was evaluated using only the classification models. The explanation $\phi(x)$ returned by LINDA-BN is a probabilistic graphical model that captures conditional dependencies between variables (including the prediction) and provides a posterior probability for any given prediction class c. That is, $Pr(c)$, given no knowledge about the feature values in x. Given the instance x and $f(x)$ being explained, we can query $\phi(x)$ to determine the impact ($\pi_i \in \pi$) of each feature $x_i \in x$ to $f(x)$ as follows:

$$\pi_i(\phi(x)) = |\Pr(f(x)) - \Pr(f(x) \mid x_i)| \quad (15)$$

In this way, we can create something approximating feature attribution explanations using LINDA-BN. As with

Table 6 A summary of the XAI methods used in this work

XAI method	Explanation-generation mechanism	Used for	Feature extraction	Ranking extraction
LIME	Local, linear surrogate model to determine feature weights	All models	Features with weights in the top 5% of the range of weights	Features ranked in order of absolute value of weights
SHAP	TreeSHAP: Tree traversal to determine contribution of each feature	All decision tree models	Features with contributions in the top 5% of the range of contributions	Features ranked in order of absolute value of contributions
	LinearSHAP: Examining feature coefficients and means to determine contribution	All linear and logistic regression models		
	ExactExplainer and PermutationExplainer: Use of gray codes to determine feature contribution	All Naïve Bayes models		
LINDA-BN	Local Bayesian Network surrogate model to determine conditional dependence between all variables, including the target variable	All classification models	Features with the greatest impact on the target variable in the surrogate model	Features ranked in order of impact on the target variable in the surrogate model
ACV	Global, random forest surrogate model to determine sufficient features for prediction	All models	Smallest set of sufficient features returned as explanation	Features ranked by percentage of occurrence across all sufficient feature sets

LIME and SHAP, we repeat this procedure five times, generating a new Bayesian Network each time and calculating π for each network. When using LINDA-BN, $\varphi(x) \ni \overline{\pi_1} \ldots \overline{\pi_n}$, and we take $x_e \ni x_i$ if $\overline{\pi_i} \geq p$ where p is determined by Eq. 14. We then order the features in x based on $\varphi(x)$ to determine r_e.

4.3.4 ACV

ACV offers explanations in many formats, including the set of all possible sufficient explanations (A-SE), the smallest possible set of sufficient explanations (minimal sufficient explanations, M-SE) and a Local Explanatory Importance (LEI) score for each feature. The LEI is not directly determined by the prediction, but is a measure of how often any given feature is present across all sets in the A-SE. Thus, we take the A-SE as $\phi(x)$, and the smallest set in the A-SE (i.e. the MSE) as $\varphi(x)$. When evaluating ACV, $x_e = \varphi(x)$, and the LEI is used as r_e.

We performed hyperparameter optimisation of ACV's surrogate model to ensure external explanation fidelity.

4.4 Identifying feature subsets

To test the completeness and correctness of the most impactful features, both in the model and as determined by the explanation, subsets of features are required. However, only one predictive method and one XAI method used in this work rely on feature subsets. For all other methods, these subsets must be determined and extracted from x, based on the "weights" attached to each feature.

Thus, a certain percentile of the Top-K features must be chosen as subsets from methods that do not provide subsets. To determine the most appropriate percentile, we conducted tests to identify a percentile from $0.05 \ldots 0.5$ that would provide the most faithful explanations. For each combination of dataset, model and XAI technique, we extracted x_t and x_e using each of the ten percentiles and calculated the $F1$-score to determine the match between the subsets. As with the fidelity evaluation, this resulted in 126 experiments. In 38 experiments, the subsets produced for all percentiles were equally faithful, i.e. the top-K percentile had no impact on the fidelity of the subsets, and these results were set aside. Out of the remaining 88 experiments, 0.05 produced the most faithful subsets in the majority of experiments (see results in Fig. 3). Of these, in 49 experiments, 0.05 produced the most faithful subsets. In another 38, the subsets produced for all percentiles were equally faithful, i.e. the top-K percentile had no impact on the fidelity of the subsets. Thus, 0.05 was chosen to determine x_t and x_e.

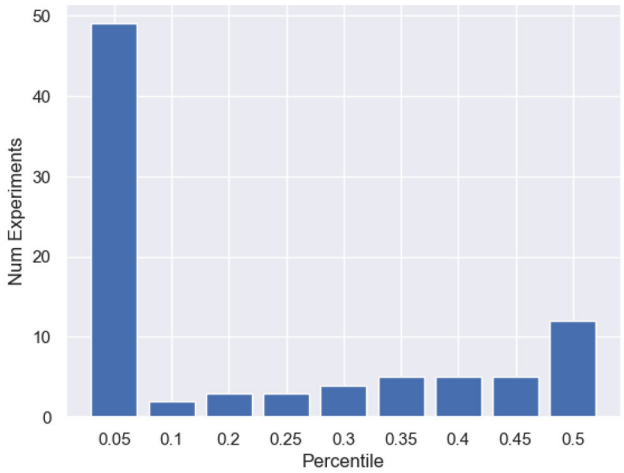


Fig. 3 Distribution of the number of experiments that produces the most faithful explanations, over the percentile used to calculate the top-K features. Taking the top 5% of features as the true and explanation features generally produced the most faithful results in a significant proportion of the 126 experiments conducted

5 Results and analysis

The average results for each combination of model and dataset are presented in Tables 7 and 8.

5.1 Analysis of fidelity results

There are a number of notable observations to be made from these results.

LIME's results are highly inconsistent across the different experimental settings. LIME explanations generally have high precision coupled with low-to-moderate recall scores, indicating that a large number of relevant features are missing from the subset, but also that few extraneous features are included. Feature ranking correlation is also generally moderate. There are no significant differences in LIME performance between classification and regression model. However, model type appears to affect results, and LIME generally performs poorly when explaining Naïve Bayes models. Dataset characteristics also affect LIME explanation quality, and explanations are more faithful for datasets with a larger proportion of categorical features.

SHAP explanations also change significantly under different experimental conditions. Feature subsets derived based on SHAP contributions show high precision in the results, but poor recall and generally low-to-moderate feature rank correlation. SHAP shows no difference in performance between classification and regression datasets, but does show poor performance when explaining logistic and linear regression models, and exceptional precision when explaining decision tree models. Dataset characteristics do not appear to affect SHAP's explanation fidelity.

Table 7 Average precision (Pr), recall (Re) and correlation (Tau) results for XAI methods when explaining the decision tree (DT), logistic regression (LR) and Naïve Bayes (NB) classification models

Dataset	Model	LIME			SHAP			LINDA-BN			ACV		
		Pr	Re	Tau	Pr	Re	Tau	Pr	Re	Tau	Pr	Re	Tau
Adult Income	DT	0.95	0.15	0.24	**1.00**	0.19	**0.59**	0.47	**0.42**	0.07	0.62	0.15	0.43
	LR	**1.00**	**1.00**	**0.75**	0.32	0.34	0.11	0.00	0.46	0.01	0.09	0.09	0.14
	NB	0.15	0.15	0.48	**0.68**	**0.68**	**0.67**	0.03	0.91	0.04	0.22	0.24	0.20
Breast Cancer	DT	**1.00**	**0.43**	0.35	**1.00**	**0.43**	**0.62**	0.09	**0.43**	0.10	0.26	0.25	0.30
	LR	**0.75**	0.24	0.52	0.60	0.17	**0.53**	0.24	**0.57**	0.06	0.38	0.20	0.29
	NB	**0.41**	0.42	0.61	0.40	0.37	**0.69**	0.05	**0.53**	−0.07	0.18	0.20	0.44
COMPAS	DT	0.96	0.26	0.45	**1.00**	0.26	**0.59**	0.33	0.11	0.13	0.70	**0.35**	0.34
	LR	**0.56**	**0.57**	**0.69**	0.50	0.55	0.39	0.01	0.24	−0.04	0.32	0.55	0.18
	NB	0.06	0.05	0.35	**0.82**	**0.82**	**0.52**	0.11	0.55	0.01	0.25	0.38	0.21
Diabetes	DT	0.89	0.30	0.49	**1.00**	0.35	**0.61**	0.49	**0.42**	0.05	0.52	0.35	0.28
	LR	**0.53**	0.54	0.53	0.50	0.51	**0.55**	0.25	**0.66**	0.02	0.36	0.48	0.23
	NB	0.64	0.64	0.63	**0.82**	**0.83**	**0.80**	0.24	0.60	−0.07	0.27	0.42	0.20
Iris	DT	**1.00**	**1.00**	0.71	**1.00**	**1.00**	**1.00**	0.23	0.88	0.03	0.75	**1.00**	0.24
	LR	**0.85**	0.88	**0.83**	0.71	0.81	0.72	0.24	**0.96**	−0.01	0.25	0.50	0.74
	NB	0.81	0.75	0.83	**0.94**	0.94	**0.88**	0.27	**0.96**	−0.01	0.44	0.63	0.58
Mushroom	DT	0.47	0.06	0.20	**0.94**	0.34	**0.54**	0.05	**1.00**	0.00	0.24	0.06	0.15
	LR	**1.00**	**1.00**	**0.86**	0.48	0.58	0.33	0.01	**1.00**	0.00	0.29	0.58	0.11
	NB	0.06	0.20	**0.44**	**0.86**	0.60	0.34	0.02	**0.97**	0.00	0.14	0.11	0.24
Nursery	DT	**1.00**	**1.00**	0.27	**1.00**	**1.00**	**1.00**	0.04	**1.00**	0.00	0.13	0.26	0.18
	LR	**1.00**	**1.00**	0.30	**1.00**	**1.00**	**0.81**	0.04	**1.00**	0.00	0.14	0.27	0.10
	NB	**1.00**	**1.00**	0.78	**1.00**	**1.00**	**0.88**	0.04	**1.00**	0.00	0.14	0.29	0.11

Bold values indicate the best performance for each metric in each row

Table 8 Average precision (Pr), recall (Re) and correlation (Tau) results for XAI methods when explaining the decision tree (DT) and linear regression (LR) regression models

Dataset	Model	LIME			SHAP			ACV		
		Pr	Re	Tau	Pr	Re	Tau	Pr	Re	Tau
Bike Rentals	DT	0.74	**0.07**	0.38	**0.99**	0.06	**0.51**	0.47	0.06	0.22
	LR	**1.00**	**0.88**	**0.78**	**1.00**	0.14	0.63	0.06	0.01	−0.16
Facebook	DT	**1.00**	**0.18**	0.50	0.92	0.17	**0.59**	0.34	0.03	0.37
	LR	**1.00**	**1.00**	**0.74**	**1.00**	0.33	0.59	0.00	0.00	0.17
Housing	DT	0.81	0.11	0.34	**1.00**	**0.14**	**0.55**	0.78	0.12	0.43
	LR	**0.53**	**0.33**	0.44	0.44	0.23	**0.45**	0.18	0.12	0.41
Real Estate	DT	**1.00**	**0.33**	0.48	**1.00**	0.31	**0.56**	0.84	0.27	0.12
	LR	**0.40**	**0.48**	0.44	0.30	0.30	**0.45**	**0.40**	0.43	0.10
Solar Flare	DT	0.89	0.16	0.51	**0.96**	**0.22**	**0.84**	0.21	0.03	0.33
	LR	**1.00**	**1.00**	**0.92**	0.16	0.16	0.51	0.00	0.00	−0.02
Student Results	DT	0.96	0.56	0.19	**1.00**	**0.57**	**0.48**	0.77	0.43	0.43
	LR	**1.00**	**1.00**	**0.91**	0.08	0.08	0.61	0.00	0.00	−0.30
Wine Quality	DT	0.99	**0.33**	0.59	**1.00**	0.31	**0.70**	0.45	0.18	0.10
	LR	**0.43**	**0.52**	0.65	0.34	0.39	**0.68**	0.20	0.22	−0.08

Bold values indicate the best performance for each metric in each row

Table 9 The Wilcoxon signed-rank test was applied to the results of the work on the classification datasets

XAI techniques	Precision			Recall			$F1$-score			Rank correlation		
	T	z	p	T	z	p	T	z	p	T	z	p
LIME & SHAP	64,472	−6.6	0.00	62,995	−4.9	0.00	81,245	−5.4	0.00	542,663	−12.9	0.00
LIME & LINDA-BN	101,122	−28.8	0.00	111,778	−11.5	0.00	124,550	−27.6	0.00	9730	−37.1	0.00
LIME & ACV	73,746	−25.5	0.00	97,228	−15.4	0.00	155,344	−20.9	0.00	198,713	−29	0.00
SHAP & LINDA-BN	47,991	−31.6	0.00	108,833	−8.1	0.00	62,908	−30.8	0.00	3148	−37.4	0.00
SHAP & ACV	43,307	−29.1	0.00	74,658	−19.6	0.00	109,395	−25.4	0.00	46,797	−35.5	0.00
LINDA-BN & ACV	385,472	−11.1	0.00	82,189	−23.1	0.00	392,239	−11.6	0.00	83,384	−29.7	0.00

We find that there is a statistically significant difference in explanation fidelity across all techniques, and when using all metrics

Table 10 The Wilcoxon signed-rank test was applied to the results of the work on the regression datasets

XAI techniques	Precision			Recall			$F1$-score			Rank corrrelation		
	T	z	p	T	z	p	T	z	p	T	z	p
LIME & SHAP	19,874	−9.8	0.00	11,408	−13.5	0.00	16,542	−13.1	0.00	190,176	−17.9	0.00
LIME & ACV	31,400	−22	0.00	34,864	−19.6	0.00	39,536	−20.2	0.00	118,162	−24.6	0.00
SHAP & ACV	13,599	−18.5	0.00	28,498	−13.3	0.00	29,000	−14.4	0.00	36,792	−30	0.00

We find that there is a statistically significant difference in explanation fidelity across all techniques, and when using all metrics

The faithfulness of LINDA-BN's explanations appears to be related to model confidence in predictions. LINDA-BN's explanations generally had correctness, with respect to both correctness of the subset and the correctness of the feature ranking. However, recall scores for LINDA-BN's explanations were generally high. Closer inspection shows that the set x_e returned by LINDA-BN contains numerous features. Thus, simply by volume, this set includes many features that are also present in x_t. This does not appear to be related to model type, or dataset characteristics, but the prediction probability associated with a prediction. This phenomenon was most common when explaining models trained on the Iris, Mushroom and Nursery datasets, all of which produced predictions with high prediction probabilities.

ACV shows little to no consistency in performance, especially when explaining the classification models. ACV explanations generally have low-to-moderate scores for all measures. In particular, ACV generally performs poorly when explaining linear regression models, but otherwise shows little consistency in performance.

We examine these phenomena further in Sect. 6.

5.2 Comparison of XAI techniques

There were notable differences in XAI technique performance for each of the three fidelity criteria examined. We applied the Wilcoxon Signed Rank test with a significance level of $p < 0.05$, to determine whether this difference in performance is statistically significant. Our results Tables 9 and 10) show that there is a significant difference in the performance of the four techniques. Additionally, we also compute the $F1$-score from recall and precision. We find that the differences between XAI techniques on precision and recall do not appear to be a trade-off, as the $F1$-scores are also distinct across the XAI techniques.

Various factors appear to underlie these differences. LIME's and SHAP's performance differs significantly given the underlying model, particularly when the precision metric is used (i.e. in *correctness* of the subset of most important features). While SHAP performs best for the decision tree model, it must be noted that LIME's performance is comparable. In reverse, LIME's explanations for the logistic and linear regression models are more faithful than that of SHAP. Similarly, SHAP clearly outperforms LIME when explaining Naïve Bayes models. This is noteworthy, as it suggests that the type of predictive model used should be taken into consideration when choosing a post hoc XAI technique. Though LIME and SHAP are model-agnostic methods, the characteristics of the underlying model affect explanation fidelity when using these techniques. LINDA-BN and ACV generally produce less precise explanations than LIME and SHAP.

LINDA-BN performed best with respect to the completeness of the subsets. The low precision scores that generally pair with high recall scores, and close examination of LINDA-BN's explanations suggest that this is likely because LINDA-BN's feature "weights" are often similar. This is especially true when the model is confident regarding a prediction. The other XAI methods generally produce less complete explanations than LINDA-BN. However, the consistently low correlation between r_t and r_e suggests that

LINDA-BN simply does not perform adequately as a feature attribution technique. Thus, while LINDA-BN generally has higher recall than the other methods, this cannot be said to be an advantage of LINDA-BN.

Feature ranking correlation across all XAI methods was poor. These results suggest that none of the XAI methods could fully capture the true importance of all features to the model. This is particularly noteworthy when considering LIME and SHAP as they were intended to generate feature attribution explanations, which is closely tied to feature ranking. ACV's low feature ranking correlation suggests that the LEI scores, which were used to determine r_e, are not fully aligned with the importance of the model. It must be noted that the LEI is suggested by the creators of ACV to show the necessity of each feature to the prediction [5], and thus, it may not fully align with its importance to the model.

Overall, we note significant differences in the performance of the four XAI methods. ACV was the most consistent method, though it generally performed poorly across all measures, indicating that it could not fully capture the features most important to the model. LINDA-BN's explanation fidelity is more consistent across model types than that of LIME or SHAP, but generally does not provide precise explanations. While LIME and SHAP generally provide precise explanations, the explanations they offer may not be complete. Moreover, the characteristics of the underlying model and data appear to affect the faithfulness of LIME and SHAP.

6 Discussion

There are two key insights to be derived from these results. Firstly, the explanation mechanism itself affects the explanation quality. This is particularly noticeable when using the different optimisations of SHAP. Though they all share a similar theoretical grounding, the implementation of the explanation mechanism greatly affects explanation quality. Thus, it is key to note that choice of XAI method should consider the potential implications of the explanation mechanism, not simply its grounding. Moreover, as noted in section 5.2, it is apparent that there are significant differences between the fidelity of different post hoc techniques. This suggests that, though a number of model-agnostic techniques are open source and easily accessible, *machine learning expertise may be needed to choose the most appropriate XAI method for the task*. It may also be more appropriate to use model-specific XAI techniques, rather than the model-agnostic techniques applied in this work.

Secondly, *explanation quality is heavily impacted by all aspects of the technical setting, including data and the chosen predictive model*. The strengths and weaknesses of the different XAI mechanisms, as well as their causes, are explored in Sects. 6.2, 6.3, 6.4 and 6.5. While these methods are all model-agnostic, their performance across different predictive model types is highly inconsistent.

These findings indicate clear implications for the use of XAI in real-life settings. In this work, the class of the model used is known and, the models themselves are fully transparent. In cases where the predictive model is considered to be intellectual property, the category of the predictive model may be unknown to end users [43]. Given the clear impacts of model type on explanation quality, it will likely become difficult to choose the most appropriate XAI method for the task without knowledge of model type. This then becomes a consideration not only for the choice of an XAI method but also for the choice of the predictive model.

6.1 Correctness versus completeness

There is a clear distinction between the precision and recall scores for almost all XAI techniques evaluated. In particular, precision results for LIME and SHAP are generally higher than recall. That is, while the features most highly weighted by LIME and SHAP do not necessarily include all relevant features (i.e. the explanation is not complete), in general, all features weighted highly by the explanation are relevant to the model (i.e. explanation correctness is high). Thus, a user applying LIME and SHAP explanations can be confident that highly weighted features are truly relevant to the model, but must also be aware that other factors could have impacted on the model prediction.

LINDA-BN generally has low precision, often paired with high recall. This is a characteristic of the permutation method of LINDA-BN. Because LINDA-BN was intended to determine the conditional dependence between variables within a small neighbourhood, the variance ϵ of permutations is generally low. Thus, in a neighbourhood where small changes in feature values do not necessarily affect the model's prediction, the Bayesian Network returned by LINDA-BN reflects this. When x_e is extracted from the network using our method, $\Pr(f(x) \mid x_i)$ is the same or similar for a large number of features, and a large set of features is returned as x_e. Such a result indicates that a user can have strong confidence in the original model's prediction [34]. However, a faithful feature attribution explanation cannot be extracted from LINDA-BN for such a prediction.

All XAI techniques generally show a low-to-moderate correlation between r_t and r_e. This is particularly noteworthy when using LIME and SHAP, which explicitly attempt to explain the importance of each feature to the prediction. Combined with the findings regarding precision and recall, we can come to the conclusion that while a user of LIME and SHAP can be confident that highly ranked features are relevant, they must be aware that the ranking of features as a whole may not be correct. Moreover, neither can they assume that they have identified all relevant features.

It must also be noted that LINDA-BN and ACV were not intended to provide feature attribution explanations. Their internal explanation-generation mechanisms reflect this, and thus, it is not surprising that both show low correlation scores. We explore this further in Sects. 6.4 and 6.5.

6.2 LIME

6.2.1 Impact of surrogate model

Notably, LIME performed poorly when explaining Naïve Bayes models. It is important to note that Gaussian Naïve Bayes was used in this work, which assumes that the relationship between features and the target variable has a Gaussian, i.e. nonlinear, distribution. We demonstrate this using ICE plots in Fig. 4.

This issue stems from the choice of surrogate model for LIME. LIME uses a Ridge Regression model as a surrogate to mimic the computations of the original model, returning the coefficients of the surrogate model as $\phi(x)$. When fitting nonlinear relationships between variables, Linear Regression and associated methods increase the coefficients of less important features in order to capture complex interaction and cancellation effects [28]. Thus, the coefficients of the model become more nuanced and can no longer be interpreted purely as feature importance. Therefore, when using such models as a surrogate for a predictive model, if the original model assumes a nonlinear relationship between variables, the coefficients of the surrogate model can no longer be interpreted simply as importance. One exception to this rule is in cases where the dataset contains all or mostly categorical features. When explaining Naïve Bayes, LIME's precision was generally poorest for the Breast Cancer and Diabetes datasets, which have only continuous variables. LIME permutes categorical features and continuous features differently, resulting in smaller variation in feature values, and thus in the prediction, when permuting categorical variables. And so, LIME is able to compute their importance more accurately.

6.2.2 Impact of permutation

As noted, LIME's permutation procedure differs for categorical and continuous features. Therefore, LIME explanations are more sound for datasets with a greater proportion of categorical features. This is especially true when explaining Linear Regression models, for which LIME performs best given that both the predictive model and the surrogate model are of the same class. When explaining Linear Regression models, LIME's precision falls below 0.5 for only two datasets, both of which have only continuous features (Real Estate and Wine Quality). The models trained on the Facebook and Solar Flare datasets, which have all categorical features, and the Student Score dataset, which has a high proportion of categorical features, have both perfect precision and perfect recall.

We also note that the length of the input does not affect the result. LIME's explanation stability is known to be related to input length, due to the nature of LIME's permutation method [46, 50]. It is reasonable to expect that a decrease in explanation consistency would affect the correctness of the explanation. However, when considering LIME's fidelity results, particularly explanation correctness, such a relationship is absent. This could suggest that our strategy of taking an average explanation from multiple explanations reduced the impacts of instability. The datasets with the longest inputs in this study are also composed primarily of categorical variables, due to the encoding method. It is also possible that the impact of the permutation method was reduced because of the large proportion of categorical variables, for which there is smaller variability in permutations.

Overall, our results suggest that the fidelity of LIME explanations are determined by the types of variables used, as well as the linearity of the relationship between variables and model prediction.

6.3 SHAP

It is important to note that SHAP values are not defined as the importance of a feature to the model. Rather, SHAP values indicate the contribution of that feature in moving the actual prediction away from the expected value of the prediction, i.e. the average value of the target variable in the training set. In our evaluations, we compared SHAP values against the importance of the feature to the model. In the case of decision trees, we found that SHAP values aligned closely with the model regarding feature importance, likely because position in the tree matters both to the prediction and TreeSHAP's calculation of SHAP values [28]. On the other hand, LinearSHAP assumes that the model treats features in the dataset as conditionally independent, and calculates SHAP values as for a given feature $x_i \in x$ as [27]:

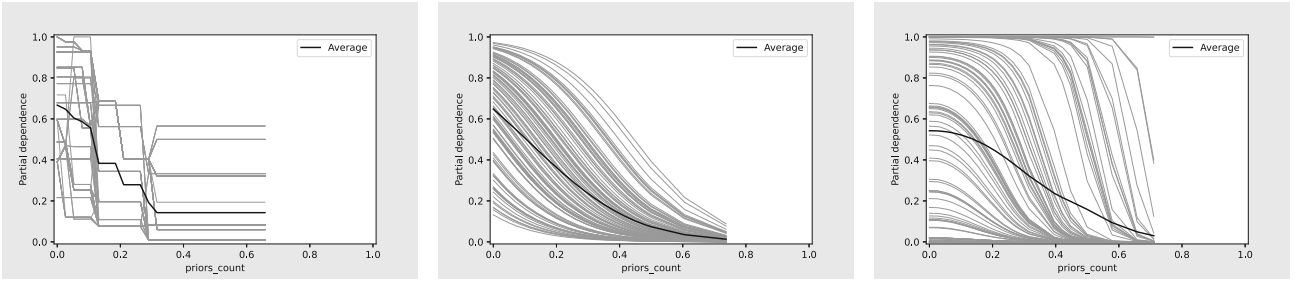
$$\phi_i(f, x) = w_i(x_i - E[x_i]) \tag{16}$$

Therefore, SHAP values, when produced by LinearSHAP, do not fully align with model importance.

Given this, we suggest that SHAP may be more useful for supporting end user decision-making, rather than data engineers or data scientists investigating the quality of a model.

6.4 LINDA-BN

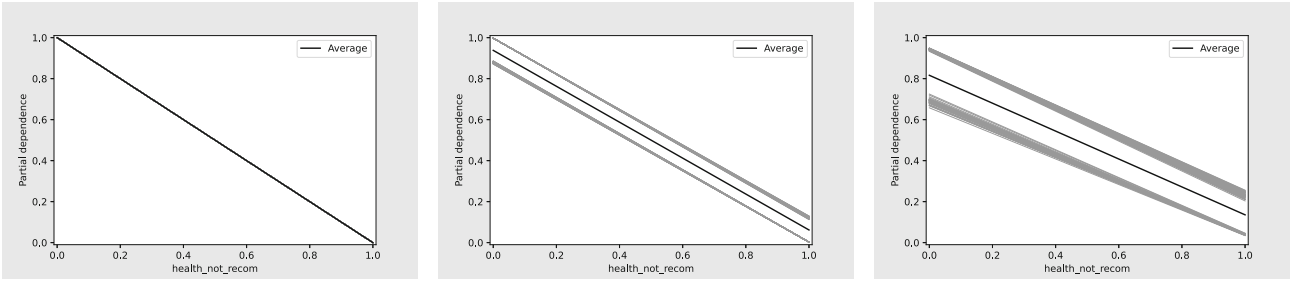
LINDA-BN was designed to aid in understanding the dependencies between variables and determining the truthfulness of a prediction. Thus, rather than returning a static explanation, it returns a Bayesian Network describing the relation-



(a) COMPAS - Decision Tree
LIME Precision: 0.96
SHAP Precision: 1.00

(b) COMPAS - Logistic Regression
LIME Precision: 0.56
SHAP Precision: 0.50

(c) COMPAS - Naïve Bayes
LIME Precision: 0.06
SHAP Precision: 0.82



(d) Nursery - Decision Tree
LIME Precision: 1.00
SHAP Precision: 1.00

(e) Nursery - Logistic Regression
LIME Precision: 1.00
SHAP Precision: 1.00

(f) Nursery - Naïve Bayes
LIME Precision: 1.00
SHAP Precision: 1.00

Fig. 4 ICE plots showing the relationship between the model prediction and a single feature for the COMPAS (**a**–**c**) and the Nursery (**d**–**f**) datasets. As the nonlinearity of the relationship between variables increases, the fidelity of the explanation decreases

ships between features within a small neighbourhood. This network can be further queried and explored to generate more insights into a prediction [34]. Generalising this explanation to a simple feature attribution explanation over-simplifies the information that this Bayesian Network provides, and so should not be the sole purpose of LINDA-BN.

However, in some cases, approximating the impact of each feature on the target variable does provide useful insights, even when the results suggest that the explanation is not sound. For example, as previously noted, the feature attribution scores extracted from the network returned by LINDA-BN could be used to determine the certainty of an explanation. Therefore, we suggest that the greatest strength of LINDA-BN is not in attempting to extract a single type of explanation from the Bayesian Network, but in querying and exploring the relationships between the variables using various methods, including the one presented in this work.

It must be noted, however, that understanding LINDA-BN's explanations requires statistical knowledge and an understanding of Bayesian Networks. While a technical expert may have this knowledge, a domain expert in another field may not. Thus, LINDA-BN's use for an end-user must also be evaluated prior to implementation.

6.5 ACV

There are several notable insights that can be derived from the described results. Firstly, extensive hyperparameter optimisation was conducted when training the surrogate model for ACV, resulting in relatively high surrogate model accuracy with respect to the original model's predictions. However, our results suggest that this high external fidelity does not translate into high internal fidelity, particularly at the local level.

Secondly, in cases where no feature subset meets π, no A-SE explanations are returned by ACV. As we use the A-SE to derive both the M-SE (which we take as $\phi(x)$) and the LEI (which is used to determine r_e), in cases where no explanation meets the π-level, we get poor scores for all three measures.

Finally, it is important to note that, although we used ACV as a post hoc explanation method in this work, ACV can function as predictive model in its own right. If used in this way, ACV would no longer have to mimic a distinct and separate model and can provide faithful explanations for its own behaviour. So, while ACV may underperform as a local, post hoc method in comparison with LIME and SHAP, its application as a self-explaining predictive model can eliminate the need for post hoc explanations altogether, which is generally more desirable [43].

Table 11 A brief summary of the strengths and weaknesses of the four XAI methods tested

	Strengths	Weaknesses	Should be used for
LIME	Generally precise in identifying features	Generally, does not identify all important features	Model types with relatively distinct relationships between features and prediction
	Performs well for decision tree and linear and logistic regression models	Does not correctly rank features based on importance	Datasets with mostly categorical variables
	More accurate for datasets with higher proportion of categorical variables	Cannot accurately work with Naïve Bayes model - relationship between prediction and features are too nonlinear	Contexts in which not all relevant variables may be important (for example, not in medical decision-making)
		More continuous variables result in poorer explanation correctness	
		Explanation quality is subject to dataset characteristics, and, in some cases, model characteristics	
SHAP	Generally precise in identifying features	Generally, does not identify all important features	Not for investigations of model quality or model fairness - performance may be inconsistent across model types
	Performs well for decision tree and Naïve Bayes models	Does not correctly rank features based on importance	End user decision-making
		Cannot accurately work with linear and logistic regression models	Contexts in which not all relevant variables may be important (for example, not in medical decision-making)
		Explanation quality is subject to model type and SHAP implementation for model type	
LINDA-BN	Fidelity is relatively consistent across model types	Cannot always distinguish between most important and least important features	Not feature attribution
		Explanations cannot be taken as accurate feature attribution/feature ranking	Model debugging and determining model confidence with no ground truth (i.e. in context)
ACV	Can show feature necessity to prediction (i.e. features without which predictions cannot be accurately made)	Cannot always distinguish between most important and least important features	A self-explaining model - does not function well post hoc
		Explanations cannot be taken as accurate feature attribution/feature ranking	

6.6 Summary of insights

Based on the insights identified in this section, we consolidate the strengths and weaknesses of the XAI techniques in Table 11.

These strengths and weaknesses can assist a data scientist or a practitioner in choosing an XAI technique to use. While these consolidated findings are not recommendations or guidelines for use, to the best of our knowledge, this is the first known attempt at moving towards such guidelines. Future work performing benchmarks of XAI performance can be useful in producing more concrete recommendations for XAI use in practice and in data science.

6.7 Limitations and future work

There are a number of limitations associated with our work. Firstly, we have evaluated XAI methods using only tabular data. While evaluating explanation fidelity for this form of data is relatively under-explored in the literature, it is unclear how our findings in this work may be generalised for other dataset types. In particular, Naïve Bayes models are commonly used in text classification. It is currently unclear how well our findings regarding the fidelity of linear explanation-generation mechanisms would apply for Naïve Bayes models trained on text data, rather than tabular data.

Secondly, in this work, we evaluate explanations by using fully transparent predictive models. Thus, we focus on the behaviours of the models and XAI techniques applied, rather than the classes of the models. It is key to note that high explanation fidelity for the relatively simple models we have used in this work does not necessarily translate into high explanation fidelity for other, more complicated predictive models. However, the findings of this work can provide guidance in selecting an appropriate XAI method, if considering the model or dataset behaviours, rather than model class. For example, if a model is known to assume a nonlinear relationship between features and target variable, such as with Naïve Bayes models, linear surrogate models may not be able to accurately capture those relationships. It is noted in [6] that an adversarial party could manipulate choice of explainable technique or explanation generation parameters in order to produce an explanation of benefit to them—a method which may be difficult for an external auditor or examiner to detect. However, with a clearer understanding of how explanation generation mechanisms interact with predictive techniques, such manipulations can be detected.

However, this work and the fidelity evaluation method used within could serve as the first step in developing a fidelity evaluation method for other contexts. As noted earlier, current fidelity evaluation methods at the local level performed poorly when applied to more complex tabular data than those used in this work. Future work could use the findings of this work in order to develop a new fidelity evaluation method compatible with the more complex time series tabular data, and which can examine explanations of black box predictive models. Our findings in this work regarding the behaviour of XAI techniques under certain situations can be used to validate such a method. We also suggest that glass box methods can be used as a proxy for black box models in developing a new evaluation method.

Similarly, we must also note that the evaluation methods outlined in this work necessitates that explanations be provided or can be reduced to some form of feature attribution or ranking. Though we found that LINDA-BN and ACV generally showed poor fidelity, this is because we simplified or focused on one aspect of the explanation, rather than assessing the explanation as a whole. Therefore, we suggest that future work should also explore methods to evaluate the fidelity of other forms of explanations. This includes not only non-static, queryable explanations, such as LINDA-BN's explanations, but also counterfactuals, among others.

Furthermore, we suggest that the findings of this work could be used to motivate and inform future XAI development. In this work, we have outlined the strengths and weaknesses of each technique and identified potential causes for these. Future XAI development could use the insights presented in this work to develop XAI techniques that address these weaknesses.

Finally, it is important to note that the findings in this work are only the first step in determining the suitability of an XAI technique for a specific situation. When evaluating the XAI techniques used, we consider only the technical context of the explanation, such as dataset characteristics and the type of model used. This fails to consider how an explanation might be understood or perceived by a user. For example, past user studies have demonstrated that feature attribution explanations are often not helpful for users performing specific tasks [42, 52, 58], instead suggesting the rule-based or counterfactual-based explanations are most useful [9, 42, 52]. Moreover, some explainable methods, such as LINDA-BN, produce an explanation that a user must have a certain level of statistical knowledge to accurately understand, and so may not be useful for an end user. Thus, the use of cognitive metrics, used to gauge explanation quality with respect to a human user [26], also becomes necessary given a context.

7 Conclusion

Given the increasing use of explainable AI methods to understand the decision-making of opaque predictive models, it becomes necessary to understand how *well* an explainable method can interpret any given model. However, evaluation methods to achieve this still remain an open question in the field of XAI, particularly for tabular data. In this work, we examine and evaluate the explanations of four model-agnostic, open-source XAI methods, using fully transparent "glass box" models trained on fourteen open-source datasets. We examined the various dataset characteristics and model behaviours that affected the faithfulness of an explanation to the underlying model.

Overall, our results suggest that model-agnostic XAI methods show significant differences in explanation quality under different technical contexts. This is likely to necessitate technical expert involvement and close examination of techniques when deploying XAI techniques. Our key findings are as follows:

- the explanation mechanism, i.e. the engineering of the XAI method, has a strong effect on explanation quality,

which may once again require a technical expert to examine and assess whether a given mechanism is suitable for the technical context; and
- there is no one "best", most faithful XAI method, even for a single dataset or model type—all of the methods show significant differences in performance across datasets and models.

In addition to these insights, we also specifically highlight some characteristics of all four methods that affect explanation quality. The insights we present in this paper are simply the first step in understanding the strengths and weaknesses of XAI methods in certain contexts. More extensive experiments can help in understanding the suitability of XAI techniques when used with different datasets and model types, and, importantly, for different user groups.

Acknowledgements Computational resources and services used in this work were provided by HPC and Research Support Group, Queensland University of Technology, Brisbane, Australia. We thank the Australian Research Training Program for the first author's scholarship.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions

Declarations

Conflict of interest The authors have no competing interests to declare that are relevant to the content of this article.

References

1. (1978) Housing. UCI Machine Learning Repository, https://archive.ics.uci.edu/ml/machine-learning-databases/housing/
2. (1988) Pima indians diabetes. UCI Machine Learning Repository, https://archive.ics.uci.edu/ml/datasets/pima+indians+diabetes
3. (1989) Solar Flare. UCI Machine Learning Repository, https://archive.ics.uci.edu/ml/datasets/solar+flare
4. Agarwal, C., Krishna, S., Saxena, E., et al.: OpenXAI: Towards a Transparent Evaluation of Model Explanations, (2022) arXiv:2206.11104
5. Amoukou, S.I., Brunel, N.J.B.: Consistent Sufficient Explanations and Minimal Local Rules for explaining regression and classification models, (2021) https://arxiv.org/abs/2111.04658
6. Bordt, S., Finck, M., Raidl, E., et al.: Post-hoc explanations fail to achieve their purpose in adversarial contexts. In: 2022 ACM Conference on Fairness, Accountability, and Transparency, (2022) https://doi.org/10.1145/3531146.3533153
7. Borisov, V., Leemann, T., Seßler, K., et al.: Deep Neural Networks and Tabular Data: A Survey, (2021) https://arxiv.org/abs/2110.01889
8. Carvalho, D.V., Pereira, E.M., Cardoso, J.S.: Machine learning interpretability: a survey on methods and metrics. Electronics (2019). https://doi.org/10.3390/electronics8080832
9. Chou, Y.L., Moreira, C., Bruza, P., et al.: Counterfactuals and causability in explainable artificial intelligence: theory, algorithms, and applications. Inf. Fus. (2022). https://doi.org/10.1016/j.inffus.2021.11.003
10. Cortez, P.: Student Performance. UCI Machine Learning Repository, (2014) https://archive.ics.uci.edu/ml/datasets/Student+Performance
11. Cortez, P., Cerdeira, A., Almeida, F., et al.: Wine Quality. UCI Machine Learning Repository, (2009) https://archive-beta.ics.uci.edu/ml/datasets/wine+quality
12. Doshi-Velez, F., Kim, B.: Towards A Rigorous Science of Interpretable Machine Learning, (2017) https://arxiv.org/abs/1702.08608
13. Du, M., Liu, N., Yang, F., et al.: On attribution of recurrent neural network predictions via additive decomposition. In: The World Wide Web Conference - WWW'19, (2019) https://doi.org/10.1145/3308558.3313545
14. Fanaee-T, H.: Bike Sharing Dataset. UCI Machine Learning Repository, (2013) https://archive-beta.ics.uci.edu/ml/datasets/bike+sharing+dataset
15. Fisher, R.: Iris. UCI Machine Learning Repository, (1988) https://archive.ics.uci.edu/ml/datasets/iris
16. Fong, RC., Vedaldi, A.: Interpretable explanations of black boxes by meaningful perturbation. In: 2017 IEEE International Conference on Computer Vision (ICCV), (2017) https://doi.org/10.1109/iccv.2017.371
17. Goethals, S., Martens, D., Evgeniou, T.: The non-linear nature of the cost of comprehensibility. J. Big Data (2022). https://doi.org/10.1186/s40537-022-00579-2
18. Guidotti, R., Monreale, A., Ruggieri, S., et al.: A survey of methods for explaining black box models. ACM Comput. Surv. (2018). https://doi.org/10.1145/3236009
19. Guidotti, R., Monreale, A., Giannotti, F., et al.: Factual and counterfactual explanations for black box decision making. IEEE Intell. Syst. **34**(6), 14–23 (2019). https://doi.org/10.1109/mis.2019.2957223
20. Jain, S., Wallace, B.C.: Attention is not explanation. In: Proceedings of the 2019 Conference of the North, (2019) https://doi.org/10.18653/v1/D19-1002
21. Kenny, E.M., Ford, C., Quinn, M., et al.: Explaining black-box classifiers using post-hoc explanations-by-example: the effect of explanations and error-rates in XAI user studies. Artif. Intell. **294**(103), 459 (2021). https://doi.org/10.1016/j.artint.2021.103459
22. Kohavi, R., Becker, B.: Adult. UCI Machine Learning Repository, (1996) https://archive.ics.uci.edu/ml/datasets/Adult
23. Konstantinov, A.V., Utkin, L.V.: Attention-like feature explanation for tabular data. Int. J. Data Sci. Anal. (2022). https://doi.org/10.1007/s41060-022-00351-y
24. Koska, C., Filipović, A.: Blackbox AI: State regulation or corporate responsibility? Digitale Welt (2019). https://doi.org/10.1007/s42354-019-0208-5
25. Lakkaraju, H., Bach, S.H., Leskovec, J.: Interpretable decision sets. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, (2016) https://doi.org/10.1145/2939672.2939874

26. Li, X.H., Cao, C.C., Shi, Y., et al.: A survey of data-driven and knowledge-aware eXplainable AI. IEEE Trans. Knowl. Data Eng. (2020). https://doi.org/10.1109/tkde.2020.2983930
27. Lundberg, S.M., Lee, S.I.: A Unified approach to interpreting model predictions. In: Proceedings of the 2017 Neural Information Processing Systems Conference, (2017) https://doi.org/10.5555/3295222.3295230
28. Lundberg, S.M., Erion, G., Chen, H., et al.: From local explanations to global understanding with explainable AI for trees. Nat. Mach. Intell. (2020). https://doi.org/10.1038/s42256-019-0138-9
29. Maksymiuk, S., Gosiewska, A., Biecek, P.: Landscape of R Packages for eXplainable Artificial Intelligence, (2020) https://arxiv.org/abs/2009.13248
30. Markus, A.F., Kors, J.A., Rijnbeek, P.R.: The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. J. Biomed. Inform. (2021). https://doi.org/10.1016/j.jbi.2020.103655
31. Meng, C., Trinh, L., Xu, N., et al.: Interpretability and fairness evaluation of deep learning models on MIMIC-IV dataset. Sci. Rep. (2022). https://doi.org/10.1038/s41598-022-11012-2
32. Messalas, A., Kanellopoulos, Y., Makris, C.: Model-agnostic interpretability with shapley values. In: 2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA), (2019) https://doi.org/10.1109/iisa.2019.8900669
33. Ming, Y., Qu, H., Bertini, E.: RuleMatrix: visualizing and understanding classifiers with rules. IEEE Trans. Vis. Comput. Graph. (2019). https://doi.org/10.1109/tvcg.2018.2864812
34. Moreira, C., Chou, Y.L., Velmurugan, M., et al.: LINDA-BN: an interpretable probabilistic approach for demystifying black-box predictive models. Decis. Support Syst. (2021). https://doi.org/10.1016/j.dss.2021.113561
35. Moro, S., Rita, P., Vala, B.:Facebook metrics. UCI Machine Learning Repository, (2016a) https://archive-beta.ics.uci.edu/ml/datasets/Facebook+metrics
36. Moro, S., Rita, P., Vala, B.: Predicting social media performance metrics and evaluation of the impact on brand building: a data mining approach. J. Bus. Res. (2016). https://doi.org/10.1016/j.jbusres.2016.02.010
37. Naretto, F., Bodria, F., Giannotti, F., et al.: Benchmark analysis of black box local explanation methods. In: Proceedings of the 3rd Italian Workshop on Explainable Artificial Intelligence co-located with 21th International Conference of the Italian Association for Artificial Intelligence(AIxIA 2022), (2022) https://ceur-ws.org/Vol-3277/paper5.pdf
38. Nguyen, D.: Comparing automatic and human evaluation of local explanations for text classification. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1 (Long Papers), (2018) https://doi.org/10.18653/v1/N18-1097
39. ProPublica: compas-analysis. GitHub (2016). https://github.com/propublica/compas-analysis
40. Rajkovic, V.: Nursery. UCI Machine Learning Repository, (1997) https://archive.ics.uci.edu/ml/datasets/nursery
41. Ribeiro, MT., Singh, S., Guestrin, C.: 'Why Should I Trust You?': Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, (2016) https://doi.org/10.1145/2939672.2939778
42. . Ribeiro, MT., Singh, S., Guestrin, C.: Anchors: high-precision model-agnostic explanations. In: Proceeding of the 32nd AAAI Conference on Artificial Intelligence, (2018) https://doi.org/10.1609/aaai.v32i1.11491
43. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat. Mach. Intell. (2019). https://doi.org/10.1038/s42256-019-0048-x
44. Samek, W., Binder, A., Montavon, G., et al.: Evaluating the visualization of what a deep neural network has learned. IEEE Trans. Neural Netw. Learn. Syst. (2017). https://doi.org/10.1109/tnnls.2016.2599820
45. Schlimmer, J.: Mushroom. UCI Machine Learning Repository, (1987) https://archive.ics.uci.edu/ml/datasets/Mushroom
46. Shankaranarayana, SM., Runje, D.: ALIME: autoencoder based approach for local interpretability. In: Intelligent Data Engineering and Automated Learning—-IDEAL, (2019) https://doi.org/10.1007/978-3-030-33607-3_49
47. Shin, D.: The effects of explainability and causability on perception, trust, and acceptance: implications for explainable AI. Int. J. Human–Comput. Stud. **146**(102), 551 (2021). https://doi.org/10.1016/j.ijhcs.2020.102551
48. Shwartz-Ziv, R., Armon, A.: Tabular data: deep learning is not all you need. Inf. Fus. **81**, 84–90 (2022). https://doi.org/10.1016/j.inffus.2021.11.011
49. Velmurugan, M., Ouyang, C., Moreira, C., et al.: Evaluating fidelity of explainable methods for predictive process analytics. In: CAiSE Forum 2021: Intelligent Information Systems, (2021a) https://doi.org/10.1007/978-3-030-79108-7_8
50. Velmurugan, M., Ouyang, C., Moreira, C., et al.: Evaluating stability of post-hoc explanations for business process predictions. In: ICSOC2021: Service-Oriented Computing, (2021b) https://doi.org/10.1007/978-3-030-91431-8_4
51. Visani, G., Bagli, E., Chesani, F., et al.: Statistical stability indices for LIME: obtaining reliable explanations for machine learning models. J. Oper. Res. Soc. (2021). https://doi.org/10.1080/01605682.2020.1865846
52. Weerts, HJP., van Ipenburg, W., Pechenizkiy, M.: A Human-Grounded Evaluation of SHAP for Alert Processing, (2019) https://arxiv.org/abs/1907.03324
53. Wickramanayake, B., He, Z., Ouyang, C., et al.: Building interpretable models for business process prediction using shared and specialised attention mechanisms. Knowl.-Based Syst. (2022). https://doi.org/10.1016/j.knosys.2022.108773
54. Wiegreffe, S., Pinter, Y.: Attention is not not explanation. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), (2019) https://doi.org/10.18653/v1/d19-1002
55. Wolberg, WH., Street, N., Mangasarian, OL.: Breast Cancer Wisconsin (Diagnostic) Data Set. UCI Machine Learning Repository, (1995) https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)
56. Yang, F., Du, M., Hu, X.: Evaluating Explanation Without Ground Truth in Interpretable Machine Learning, (2019) https://arxiv.org/abs/1907.06831
57. Yeh, IC.: Real Estate Valuation Data Set. UCI Machine Learning Repository, (2018) https://archive.ics.uci.edu/ml/datasets/real+estate+valuation+data+set
58. Zhang, Y., Liao, Q.V., Bellamy, R.K.E.: Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, (2020) https://doi.org/10.1145/3351095.3372852
59. Zhou, J., Gandomi, A.H., Chen, F., et al.: Evaluating the quality of machine learning explanations: a survey on methods and metrics. Electronics **10**, 593 (2021). https://doi.org/10.3390/electronics10050593