SN Computer Science (2024) 5:503
https://doi.org/10.1007/s42979-024-02810-8

ORIGINAL RESEARCH

Which Explanation Should be Selected: A Method Agnostic Model Class Reliance Explanation for Model and Explanation Multiplicity

Abirami Gunasekaran[1] · Pritesh Mistry[1] · Minsi Chen[1]

Received: 28 February 2023 / Accepted: 9 March 2024

Abstract

Feature importance techniques offer valuable insights into machine learning (ML) models by conducting quantitative assessments of the individual contributions of variables to the model's predictive outcomes. This quantification differs across various explanation methods and multiple almost equally accurate models (Rashomon models), creating explanation and model multiplicities. This resulted in a novel framework called method agnostic model class reliance range (MAMCR) for identifying a unified explanation across methods for multiple models. This consensus explanation provides each feature's importance range for a class of models. Using state-of-the-art feature importance methods, experiments on popular machine learning datasets are conducted with a $\varepsilon-$ threshold value of 0.1. The dataset-specific Rashomon set with 200 models, and the prediction accuracy of concerned reference models (m^*) have produced encouraging results in obtaining a consensus model reliance explanation that is consistent across multiple methods. The experiment results ensure whether the prediction accuracy level of models has an impact on the importance range estimation of features. Also, the order of features suggested by MAMCR leads to better performance of models consistently in all the experimented datasets, than the state-of-the-art methods.

Keywords XAI · Model agnostic explanation · Feature importance · Rashomon set · Explanation disagreement · Unified explanation

Introduction

Explainable AI (XAI) is a concept within the machine learning domain that aims to provide human-intelligible explanations behind a predictive model's decisions. That is, to explain how a model produced an output from a given input. Many different types of explanation methods have been proposed and can be found in the literature [5, 15, 32, 41, 52, 62, 73], often ordered into a hierarchical taxonomy. The various levels are not stringent and methods are frequently applied to more than one taxonomy level.

Early attempts at explanations focused solely on using self-explanatory machine learning methods, known as intrinsically interpretable models. These models require no post-hoc analysis for model explanations, and examples include, model classes such as decision trees, regression, and Bayesian-based models [5, 54]. For instance, the interpretation of the coefficients of a linear regression model would be an example of a model-specific interpretability method. Although model transparency is achieved, intrinsically interpretable models sometimes underperform in specific tasks. When more complex, so-called black box methods are desired, interpretability was often insufficient. Black box methods, therefore, require post-hoc analysis for interpretability, i.e. after the model has been trained. Post-hoc analysis can be grouped into several subcategories depending on how they attempt to achieve explainability, for instance, using visualisation methods [1, 27, 36, 65, 74], surrogate models [7, 37, 43, 52, 57], or feature importance methods [8, 11, 15–17, 32, 38, 41].

An explanation method that is tailored for a specific type or class of algorithm application is called model-specific. These methods utilise the internal structure of the model to provide an explanation. These methods [6, 11, 66, 76, 77] are tied with specific model types. For instance, the methods that are interpreting the results of e.g. neural networks

This article is part of the topical collection "Soft Computing in Engineering Applications" guest edited by Kanubhai K. Patel.

✉ Abirami Gunasekaran
a.gunasekaran2@hud.ac.uk

[1] School of Computing and Engineering, University of Huddersfield, Huddersfield HD1 3DH, UK

Published online: 27 April 2024

are model-specific [23, 58]. By definition, the inherently interpretable models' interpretations are always intended for specific models [45]. Providing a deeper insight into the model's decision process with the knowledge of the model's internal workings is the advantage of these model-specific explanation methods. On the other hand, if these methods need to be employed for obtaining the explanation, then the choice of the prediction model to be used is only limited.

Recent developments in the XAI field have focused on model-agnostic methods, which are applicable to any machine learning model. In this study, we look at a particular type of model-agnostic method, known as the feature importance method. These methods can be used with any machine learning model to provide learning behaviour explanations. The learning behaviour essentially represents the important order of the features on which a model makes its predictions [13, 15, 32, 41, 52, 62]. These methods use the input and the predicted output of a model to provide feature importance explanations. Feature importance can be defined as a statistical indicator that quantifies how much a model's output changes with respect to the permutation of one or a set of input variables [73].

A range of successful strategies have been reported in the literature for computing the feature importance values, such as; feature inclusion [32], which introduces features successively into the calculations; feature removal [41], which successively removes features from the calculations; iterative training [38], which retrains the model many times per feature inclusion, or iterative retraining can be avoided [53], by handling the absence of removed features or the inclusion of new features. To do this, several approaches such as, supplementary baseline [64], conditional expectations [63], product of marginal expectations [17], approximation with marginal expectations [41] or replacement with default values [52] can be used.

While many feature importance methods exist, the explanations obtained with one method may not corroborate with that of another method for the same model [21, 35], which can be referred to as *explanation multiplicity*. It is known that many different machine learning models can fit data equally well and produce almost similar accurate predictions, which is referred to as *model multiplicity* [12]. However, the features deemed most important to one model may not be important for another well-performing model [54]. In a scenario involving multiple models and explanations, the selection of a feature importance method for each specific model becomes crucial. When multiple methods are employed, and the resulting explanations present contrasting information, the question arises: which explanation should be trusted? Our research of the literature finds no clear standard framework to help choose an appropriate explanation or method.

Hence, the contributions of the work are as follows:

- Since the explanations based on a single machine learning model using a specific explanation method can be biased over that model/method, a novel framework is proposed to provide a model-agnostic explanation utilising various explanation methods for multiple almost-equally-accurate models. These near-optimal models [48] are called Rashomon set [27].
- Rather than selecting a single predictive model from a set of well-performing models and providing explanations for that model, the proposed method offers an explanation across multiple explanation methods to cover the feature importance of all the well-performing models in the chosen model class.
- When evaluating explanations from different methods across multiple models, the absence of uniqueness in these explanations underscores the importance of a unified explanation that extends across various methods, accurately capturing multiple models with similar performance.
- The proposed framework is designed to achieve method-agnostic model class reliance (MAMCR) range explanations, ensuring unbiased and comprehensive coverage across multiple methods within a specific model class.

This work also addresses the following research questions:

RQ1. Does the quantification of a feature as (un) important depend either on the method or model? While different methods are applied to a model to explain its behaviour, to what extent do the methods disagree with the explanations among them? If they vary, how to select the consensus explanation from the conflicting explanations?

RQ2. Does a model's prediction accuracy affect its reliance range on a feature for its predictions?

The remainder of this paper is structured as follows: Section "Related Work" discusses the existing literature and Section "Method" describes the methodology proposed in this study in detail. Section "Empirical Evaluation" briefly describes the experimental evaluation and Section "Results" reports the result findings of this study. Section "Discussion" summarises the discussion of the work with a note on the limitations of the study. Lastly, we include an Appendix A for the additional data generated from our results.

Related Work

Many strategies for XAI have been developed for providing explanations for black box models. Amongst them, feature importance methods have gained popularity. These methods [13, 15, 32, 38, 41, 62] aim to explain a single model's variable importance values by permuting over the variables.

Explanations can be at the local level [52] for a single instance or at the global level [13] for an entire dataset.

Feature importance methods that provide model agnostic explanations (i.e., irrespective of the model's internal structure) are researched extensively to help understand the prediction behaviour of black-box models. The methodology proposed in this study provides a consensus explanation across multiple explanation methods. In theory, any number of explanation methods can be chosen and processed together, only limited by the computational processing available to the user. To illustrate the methodology, five popular explanation methods are selected, based on their global agnostic explanation compatibility and the ability of the code to reproduce the explanations. The five methods are: LOCO [38], Dalex [8], Skater [13], Shap [41] and Sage [16]. A summary of the methods is provided below.

The **LOCO** (Leave One Covariate Out) method offers global explanations for black-box models. It quantifies the importance of each feature by removing it from the model's input and measures the model performance to quantify the importance of the removed feature. The variation between the model's performance with and without a feature is the feature's contribution towards the model's predictions and therefore its importance. Since the removal of a feature alters the input matrix shape used for training the model, the model has to be retrained every time a feature's importance is to be measured. Here, we employ the LOFO (Leave One Feature Out) method [2], which is a Python implementation of the LOCO explanation method.

The **perturbation based** feature importance methods help to avoid the computational complexity of multiple retraining of separate models. Instead of removing the feature, the actual values of the feature are perturbed. This eliminates the correlation between the removed feature and the target variable and thus, the model performance gets varied. That variation decides the importance of the feature towards the model's predictions. The Dalex, Skater, Shap, and Sage methods belong to the perturbation-based feature importance estimation methods. But they differ in the way that they operationalise the computation, e.g., how many variables are permuted at a time and/or the scoring/loss function used to estimate the model performance deviation.

Dalex method chosen to provide the explanation here explains the model parts globally. But with this method, the local explanation also can be obtained. It permutes one variable at a time to compute the feature's importance value. The importance of the feature is measured based on the loss function (1-AUC) if the trained model does the classification and Root Mean Square Error for regression.

Skater is also a single feature perturbation-based, global feature importance method. It estimates the importance of a model's feature using the mean absolute error for regression models and the cross entropy/f1 score for the classification models.

The **Sage** and **Shap** methods estimate the feature importance by computing the mean average from various feature coalitions. For the feature importance estimation, they use Shapley values [56] which is a distribution concept of the Cooperative game theory approach. The Sage method provides the global explanation whereas the Shap method uses the average of all the local explanations to produce the global interpretation of the model's predictions.

Rashomon Effects

The issue of model multiplicity, where multiple models fit the data equally well but yield different model forms, was initially raised by [12].

No clear criteria are available for choosing the "best" model amongst all those near-optimal models [19]. Moreover, the learning behaviour of the models varies among themselves, i.e. the feature which is important for one model may not be important for another model. Hence, to avoid a biased explanation of a single model, a comprehensive explanation for all the well-performing models is given as a range of explanations [22].

After Fisher, et al., [22], the authors of [19] expanded the Rashomon set concept by defining the cloud of the variable importance (VIC) values for the almost-equally-accurate models and visualising it with the variable importance diagram (VID). The VID informs that the importance of a variable is changed when another variable becomes important.

The non-uniqueness would be reduced by averaging over a group of equally competent models [11]. Based on this idea, the authors of [48] created a collection of 350 nearly optimal logistic regression models on the COMPAS dataset [49] and averaged the feature importance values. They argue that the presented explanation is less biased towards a model class than a single model's explanation. The biased learning of a model is corrected by ensembling the Rashomon set models using the prior domain knowledge [30]. The ratio for the Rashomon set is analysed using a Rashomon elbow [55]. They observe that the model performance does not necessarily vary across different algorithms even though the ratio of Rashomon set models on the dataset is small. However, all these methods provide the solution to the model multiplicity but not to the method explanation multiplicity which is the focus of this work.

Evaluation of Post-Hoc Explanations

The post-hoc methods are required when the trained models become complex black boxes and their prediction behaviours are difficult to interpret. The explanations of these methods are evaluated using various metrics such as fidelity, consistency, and stability [21, 28, 52, 70].

The behaviour of cutting-edge post-hoc explanation approaches has been analysed by numerous researchers using these metrics, and the results show that the methods are susceptible to producing explanations that are unstable, fragile, and easy targets for adversarial attacks [24, 25, 34, 60].

Based on the usability, understandability, plausibility and faithfulness measures, the **most usable** [20], **most understandable** [39, 46] and **most plausible and faithful** [4, 42] explanation among the various obtained explanations from different methods is selected, respectively.

The explanations from different methods are compared in order to assess the quality [70] of the explanation into the **most correct/ best/ effective** [50, 51, 59, 71], or the **most informative** [47] explanation and also the **similarities and differences** [31] between them.

Since the operationalisation of the feature importance estimation varies depending on the approach, so do the explanations. In other words, different method explanations have different preferences for the important attributes. However, how the various approaches interpret the behaviour of the model and their explanation preferences is not taken into account arriving at *'one'* suitable among the many explanations. In the end, the published literature does not offer a consensus explanation across various methodologies which is an additional focus of this work.

Previously published work [10] shares some similarity with the work presented in this paper, in analysing the various explanations based on the stability and consistency metrics and combines them into an ensembled explanation across multiple XAI algorithms. However, this work [10] does not account for the problem of model multiplicity.

Method

This section illustrates the methodology that is proposed in [29] for obtaining the method agnostic ensembled explanation of various almost equally accurate machine learning models.

Let $(X,Y) \in \mathbb{R}^{(p+1)}$, where $p > 0$, $X \in \mathbb{R}^{p}$ is the random vector of p input variables and $Y \in \mathbb{R}^{1}$ is the output variable. Let, $\mathcal{ACC}$ be a scoring function for an ML model m and return the prediction accuracy of the model such as $\mathcal{ACC}(m) = \mathcal{ACC}(m;X,Y)$. It is the ratio between the no. of correct predictions and the total predictions made by the model m on the dataset X, compared with the ground truth values Y.

Model building

The process flow of MAMCR is shown in the algorithm 1. There are general preprocessing steps that can be applied to most datasets as a starting point. It includes data cleaning, data scaling and normalisation, data splitting, feature engineering and encoding, and handling imbalanced data. While these are general preprocessing steps, their application and extent may vary depending on the dataset. It's essential to understand the specific characteristics of the dataset and problem domain to determine the most appropriate preprocessing steps. The key is to apply a combination of these techniques in a way that best suits the dataset and the objectives of the analysis.

Then the process proceeds with the modelling of a class of multiple machine learning models on the data. Since the No Free Lunch theorem [75] states that no single machine learning model is considered optimal for all problem domains, multiple machine learning models are fitted to the same dataset to analyse the performances of the models. Thus, the selected set of pre-specified predictive models is referred to as a model class (M) [22].

Algorithm 1 MAMCR

```
Input:
D – Dataset, $\mathcal{R}$ – Rashomon set, $\varepsilon$ - A small positive value
Output:
MAMCR - Method Agnostic MCR Range explanation
begin
    Apply required preprocessing on D
    begin
        M = Train a class of models
        $m^*$ = Maximum Accuracy model in M

        for $l$ = 1 to m models in M do
            if ($\mathcal{ACC}(l) \geq (1\text{-}\varepsilon) * \mathcal{ACC}(m^*)$) then
                Include l into $\mathcal{R}$        ▷ Identifying near optimal models for Rashomon set
            end
        end
        for $j$ = 1 to r model in $\mathcal{R}$ set do
            for $i$ = 1 to n methods do
                Apply method i on j
                Obtain Model Reliance($MR_i$) rank list of j    ▷ Feature Rank Order
            end
            Discover Optimal explanation of j ▷ Model's majority learning using Borda method
            Obtain Optimal_sim$_i^j$ score    ▷ Compare each explanation against optimal explanation using RBO method and get a similarity score
            for $k$=1 to N features do
                Compute Weighted Mean $\theta_{j,k}$
                $MAMCR^K = [MAMCR^{k-}, MAMCR^{k+}]$
                ▷ Model Reliance lower and upper bounds
            end
        end
    end
end
```

M is a model class that consists of m models. Each model takes the input X and converts it to response Y. For classification problems, each model can be assessed in terms of its prediction accuracy. If the model class is built with a set of regression algorithms, then the model performance can be assessed in terms of its R^2 value.

$$Model\ class\ M = m_i \qquad (where\ i = 1, 2, 3 \ldots n) \tag{1}$$

Finding the Rashomon Set Models

From the multiple fitted models of the model class M, the almost equally accurate models form the Rashomon set ($\mathfrak{R}$). A Rashomon set is constructed based on the benchmark model, m^*, and $\varepsilon > 0$, as follows:

$$\mathfrak{R}\ (\varepsilon, m^*, M) = \{m\ \in\ M\ \mid\ \mathcal{ACC}(m) \geq (1-\varepsilon)\ \mathcal{ACC}(m^*)\} \tag{2}$$

Among the multiple trained models, m^* is the one that is selected with the maximum possible accuracy. The ε value is used to indicate a small extended boundary range for discovering the near optimum models from the prediction accuracy of the reference model m^*. The threshold is adjusted by $(1-\varepsilon)$ factor of m^* accuracy and is termed as the *ε-threshold prediction accuracy*. Based on the level of the boundary adjustment for model inclusion, the ε value is set to a small positive integer such as 5% [48] or 1% [61]. Thus, to form the Rashomon set ($\mathfrak{R}$) with the almost-equally-accurate models from the model class M, the models whose prediction accuracy is ≥ the (1-ε) factor of m^* accuracy are chosen.

Obtaining Model Reliance Values and Ranking Lists

The model reliance [22] (or feature importance) indicates how much a model relies on a variable for making its predictions. The model reliance on the variable k (mr^k) is measured by the quantity of change in the model's performance with and without the variable k, where $k = 1, 2, 3$ ...p. The greater the change in the model performance, the greater the contribution that variable has, to the model's predictions. Here, different explanation methods are selected to apply to each of the models in the Rashomon set to obtain their model reliance on p variables. Any global explanation method that returns the explanation in the form of feature importance can be chosen.

$$\begin{aligned} E_{MRR} &= \Big[\big[MRR_1\big], \big[MRR_2\big], \ldots, \big[MRR_r\big]\Big] \\ &= \Big[\big[[e_1^{\mathfrak{R}_1}], [e_2^{\mathfrak{R}_1}], \ldots.., [e_n^{\mathfrak{R}_1}]\big], \big[[e_1^{\mathfrak{R}_2}], [e_2^{\mathfrak{R}_2}], \ldots.., [e_n^{\mathfrak{R}_2}]\big], \ldots \\ &\quad \big[[e_1^{\mathfrak{R}_r}], [e_2^{\mathfrak{R}_r}], \ldots.., [e_n^{\mathfrak{R}_r}]\big]\Big] \end{aligned} \tag{3}$$

The explanation $E_{MRR}[1] = MRR_1 = [e_1^{\mathfrak{R}_1}, e_2^{\mathfrak{R}_1}, \ldots.., e_n^{\mathfrak{R}_1}]$ is the set of model reliance ranking lists obtained for the 1st Rashomon model, $\mathfrak{R}_1$, from n explanation methods.

The $e_n^{\mathfrak{R}_1}$ shows the feature ranking list explanation for the model $\mathfrak{R}_1$, obtained from the n^{th} explanation method. For example, the order can be represented as follows,

$$e_n^{\mathfrak{R}_1} = \left[f_1, f_3, f_4, f_p \ldots f_2\right]$$

Where f_1 is the name of the input feature that has the highest importance value than all other variables $f_2, f_3, f_4, \ldots, f_p$. The model reliance ranking list follows the order $f_1 > f_3 > f_4 > f_p >, \ldots, > f_2$, where variable f_2 has the least importance among the p variables.

Finding the Optimal Explanation and Consistency of Explanations

The methods that operationalise the feature importance computation may not be similar in the computed model reliance values for a model [21]. But, in the order of the features, the methods may exhibit similarity for capturing the model's learning behaviour. As pointed out by [26], if the results of different techniques point to the same conclusion, they very likely reflect the real aspects of the underlying data. Therefore, an optimal explanation reflecting the commonly found feature order among the different methods' explanations of a model should be discovered. This consensus explanation captures the optimal feature order by aggregating all the explanations' feature ranking preferences using the modified Borda Count method [40].

$$\mathcal{O}_j = BORDA(E_{MRR}\ [j]) \quad (where\ j = 1, 2 \ldots\ r,\ r = \mid \mathfrak{R} \mid\ models) \tag{4}$$

The Borda function returns the result as an aggregated model reliance ranking order i.e., $\mathcal{O}_1$ captures the optimal ranking order of the features from the n explanations of the 1st model. Likewise, for each model, a consensus explanation is aggregated from the corresponding model's explanations from n methods. This leads to a total of r number of optimal explanations for the Rashomon set models ($\mathfrak{R}$).

To quantify the consistency of several methods in producing similar explanations to the model, the methods' explanations for the model are compared against the optimal explanation. To find the consistency score, the ranking similarity needs to be compared. Existing statistical methods such as Kendall's τ [33] is considered inadequate for this problem because the ranking lists may not be conjoint. On the other hand, the Rank-Biased Overlap (RBO) [72] could handle the ranking lists even though the lists are incomplete. The RBO similarity between two feature ranking order lists R1 and R2 is calculated using the following equation as per [72].

$$RBO(R1, R2, p) = (1-p)\sum_{d=1}^{\infty} p^{d-1}.A^{d} \qquad where\ A^{d} = \frac{|\,R1_{1:d} \cap R2_{1:d}\,|}{d} \tag{5}$$

The similarity value ranges from 0 to 1, where 0 indicates no similarity between the feature ranking order lists and 1 indicates complete similarity. The p parameter p $(0 < p < 1)$ defines the weight for the top features to be considered. The parameter A_d defines the agreement of overlapping at the depth d. The intersection size of the two feature ranking lists at the specified depth d is the overlap of those 2 lists (For more details, readers are referred to the equations (1-7) in [72]). A similarity score is computed between the model's various explanations and the corresponding optimal explanation. It is referred to as optimal similarity and is calculated as follows,

$$OPTIMAL\_SIM_{i}^{j} = RBO\left(e_{i}^{\mathfrak{R}_j}, \mathcal{O}_j\right) \qquad (where\ i = 1, 2, \ldots n\ explanation\ methods\ and\ j = 1, 2, \ldots r,\ r = |\ \mathfrak{R}\ |\ \ models) \tag{6}$$

The $OPTIMAL\_SIM_{i}^{j}$ defines how much the explanation ($e_{i}^{\mathfrak{R}_j}$) that is obtained from method i for the model $\mathfrak{R}_j$ is similar in complying with the feature order of the optimal explanation $\mathcal{O}_j$ in terms of feature order. The $OPTIMAL\_SIM$ value is computed for all the method explanations of each model. Therefore, n x r similarity scores are obtained.

Computing the Weighted Grand Mean (θ)

Among the various explanations of the Rashomon set models, the optimal similarity scores of the methods are calculated based on the method explanations' compliance with the corresponding model's optimal explanation. This score shows the degree of similarity that the method has, in explaining the model's optimal learning behaviour. Since the different explanation methods produce different feature importance coefficients for each feature, the model has varying levels of reliance on a feature. Therefore, a grand mean (θ) across several methods is to be estimated. For that, a weighted mean [9] is implemented. To weigh the feature importance values that are computed by each method for the model, the optimal similarity score is used. For each feature, the weighted grand mean of the feature importance values based on the methods' optimal similarity score as weight is calculated by,

$$\theta_{j,k} = \frac{\sum_{i=1}^{n} OPTIMAL\_SIM_{i}^{j} * mr_{i}^{k}(\mathfrak{R}_j)}{\sum_{i=1}^{n} OPTIMAL\_SIM_{i}^{j}} \qquad for\ k = 1\ to\ p\ features\ and\ j = 1\ to\ r,\ r = |\ \mathfrak{R}\ |\ \ models \tag{7}$$

Table 1 The dataset details along with the prediction accuracy of the chosen Reference model (m^{*})

Dataset Name	# Predictors	# Data Points	m^{*} Accuracy	ϵ-Threshold Accuracy
COMPAS	06	7214	0.732	0.658
ADULT	14	32561	0.82	0.739
WINE	11	1599	0.828	0.745
HEART	13	1025	0.91	0.819
SONAR	28	208	0.83	0.747

The ϵ value and the number of selected models for the Rashomon set of all the datasets are set to 0.1 and 200 respectively

The grand mean ($\theta_{j,k}$) for the feature k of the model j is calculated by adding the product of the optimal similarity score of the 1 to n methods with its computed feature importance value for the k feature (mr_{1}^{k} to mr_{n}^{k}) and dividing the result with the sum of n methods' weights (i.e., optimal similarity scores of n methods). The grand mean is computed for all the p features for each model of the Rashomon set ($\mathfrak{R}$). Therefore, p x r weighted mean feature importance values are obtained. Based on these computed values, a unified explanation rank order for each of the Rashomon models can be offered by MAMCR.

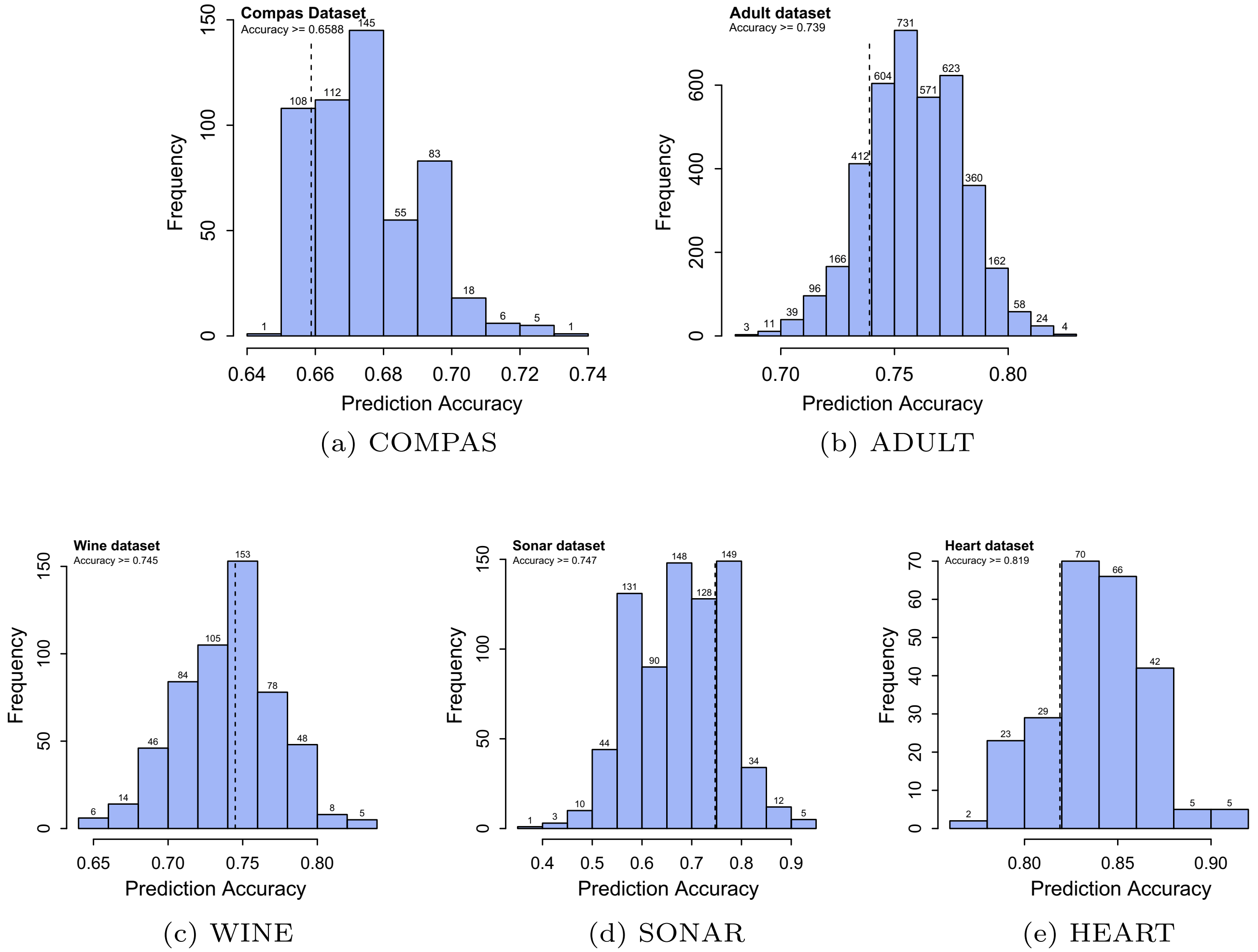
Method Agnostic Model Class Reliance (MAMCR) Explanation

The method agnostic model class reliance explanation of the Rashomon explanation set is given as a comprehensive reliance range for each variable based on the reliance of all the well-performing models under n explanation methods. The model class reliance of all the p variables can be given as a range of lower and upper bounds of weighted feature importance values. The lower and upper bounds of the model class reliance for each variable can be defined as,

$$MAMCR^{k} = \left[MAMCR^{k-}, MAMCR^{k+}\right] \tag{8}$$

$$MAMCR^{k-} = \min_{\theta}\left[\theta_{j,k}\right]_{j=1}^{r} \quad and \quad MAMCR^{k+} = \max_{\theta}\left[\theta_{j,k}\right]_{j=1}^{r} \qquad where\ k = 1\ to\ p\ variables\ and\ r = |\ \mathfrak{R}\ |\ models \tag{9}$$

In the range [$MAMCR^{k-}$, $MAMCR^{k+}$] of variable k, if the $MAMCR^{k+}$ value is low, the variable k is not important for any almost-equally-accurate models in the Rashomon set models $\mathfrak{R}$ whereas if the $MAMCR^{k-}$ is high, then the variable k is most important for every well performing model in $\mathfrak{R}$. Thus, the $MAMCR$ provides a method agnostic variable



(a) COMPAS

(b) ADULT

(c) WINE

(d) SONAR

(e) HEART

Fig. 1 Multiple models that are trained on the given datasets vary in producing accurate predictions. Based on a reference model's (m^*) prediction accuracy, the near-optimal models are selected for the Rashomon set with an extended ε-Threshold which is indicated by the dotted line, and its Accuracy level is represented in the left side of each histogram. Models that produce accurate decisions above the ε -Threshold are considered for the Rashomon set. Each bin in the plot is labelled with the count of trained models that fall in that accuracy level. The Y-Axis holds the count of models and X-Axis carries the prediction accuracy range

importance explanation for all the well-performing models of the Rashomon set ($\mathfrak{R}$).

Empirical Evaluation

The proposed MAMCR framework is illustrated on multiple publicly available real-world datasets using five state-of-the-art, feature importance explainable AI methods, as discussed in the Related Works section. A Logistic Regression model class was employed for this process.

Datasets

The datasets used for the empirical evaluation are briefly discussed herein.

The **COMPAS** (Correction Offender Management Profiling for Alternative Sanctions) dataset[49] was used in the United States Court systems to decide the 2-year recidivism status of previous criminal offenders. The dataset contains 52 features, among them, 12 are date types to denote jail-in and jail-out, offence, and arrest dates. 21 are personal data identifiers such as their first and last names, age, gender, case numbers, and description. The remaining features are mostly numeric values such as the number of days spent in screening, in jail, at COMPAS, etc. The dataset is tailored to use 6 important features by [19, 48] for their explainability analysis and the same is considered in this analysis.

The **Adult** dataset [67], is an extraction of the 1994 American census data. This dataset contains features pertaining to an individual's education, gender, occupation, etc. There are 14 features and 48842 instances with some missing values, therefore an element of preprocessing is required to clean the data. The dataset is mainly used for classification purposes aiming to predict whether an individual earns a salary of greater than 50K or not.

The **Wine** dataset contains information regarding the quality of wine assessed by various chemical properties [14], and is frequently used for classification and regression tasks. From this, the red wine quality dataset [69] is considered in this analysis. The continuous target variable of this dataset is encoded as Good/1 or Bad/0 based on the quality of the wine as suggested in[1] such as if the quality score is >= 6, then, the target is encoded to 1 or else to 0.

The **Heart** dataset [68] contains information pertaining to heart disease in patients. The dataset comes with 76 features although many studies reported in the literature use much fewer. The various attributes contain information on which heart disease can be modelled, for example, age, gender, cholesterol levels, resting blood pressure, etc. The target feature is an integer value ranging from 0, no heart disease presence to 4. When using this dataset for modelling, most studies look at distinguishing presence (1, 2, 3 & 4) from absence (0).

The **Sonar** dataset [44] is a collection of sonar data that can be used to predict the detection of a rock or a mine. It consists of 60 features all of which are of the numerical type. Many of these features are highly correlated but the dataset can be reduced to 28 features by removing the features whose correlation coefficients are above 0.8. The target feature is a binary variable that denotes an “R” for rock and an “M” for mine.

For each dataset, multiple models are trained on various subsets, resulting in varying accuracy levels. The total number of models depends on the level of variation observed in prediction accuracy. This number typically ranges approximately from 240 to 3900 for the given datasets. For example, 3,864 models were trained on the Adult dataset, and 242 models were trained on the Heart dataset.

A dataset-specific reference model with possible maximum prediction accuracy is chosen and the boundary for the near-optimal models' selection is set with the ε value. The trained models whose prediction accuracy values are above the ε*-threshold* are considered for the Rashomon Set. The count of the near-optimal models taken for the model explanation is kept at 200, but not limited to, for each dataset. A summary of how the models are selected is shown in Table 1. The total models trained for each dataset and the distribution of the Rashomon set models' prediction accuracy is shown in Fig. 1. For the COMPAS and Adult datasets, the experiment is elaborated with various ε*-thresholds* and is discussed later in this section.

Results

The variation in the different methods' explanations based on feature importance is illustrated using stacked bar charts. Here we present the results for the COMPAS and Adult datasets in Figs. 2 and 4. Stacked bar charts for the other datasets used in this study can be found in Appendix A (See Figs. 9, 10, 11).

The stacked bar chart of the COMPAS dataset (Fig. 2) demonstrates the variations in the ranking explanation clearly in all 6 features. It is evident that the ranking explanations of the 5 different methods tested vary. There is some agreement amongst the most and least important features, but significant variation for the other features (See Fig. 3).

For instance, the Sage and Dalex methods' explanations rank the Age variable as one of the top 50% important features (rank ≤ 3). But the Skater and Shap methods claim it as least important by ranking it the last ($6th$). Similarly, the Juvenile crime feature is identified by the Sage method as most important by assigning $1st$ or $2nd$ rank, but as least important by the Shap method. As per the Shap method, more than 25% of the models depend on the Current charge variable the most at the rank of 2, whereas all other methods assign majorly the $5th$ or $6th$ rank to the variable.

The method explanations for the Adult income dataset (Fig. 4) vary between several of the variables. The agreed explanation is found with the most and least important features, i.e., Marital status and Race, respectively.

For the Heart dataset also, the ranking variations are found with Sex, Age, CA, Thalach, and Exang variables (Fig. 10). In the Wine dataset explanations, the difference in the ranking explanation occurs with Free sulphur dioxide, PH, Fixed acidity, and Residual sugar variables (Fig. 9). The commonality in the explanations is found with the most and least important features Alcohol and Density, respectively. Similarly in Sonar dataset, where the features are simply labelled as a number, the explanation variations are found in features 28, 31, 41, 44, 9, 51, 55, 56, 57, 59, and 60 (Fig. 11).

This shows that the quantification of feature importance depends on the method that identifies it as the most, moderate, or least important. Based on a method, it can be the most important or least important. For example, the 'Race' feature is identified as the least important feature by the Sage method explanations but as the most important for 50% of the models

[1] https://www.kaggle.com/datasets/uciml/red-wine-quality-cortez-et-al-2009.

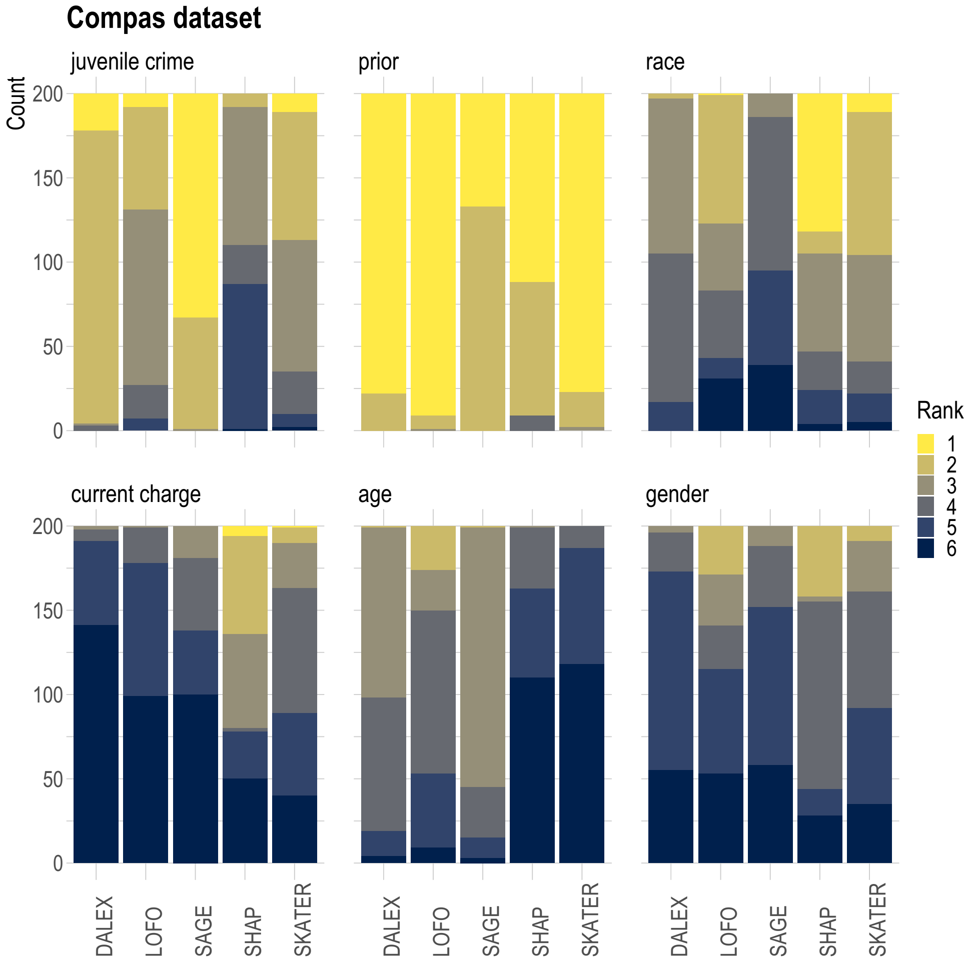
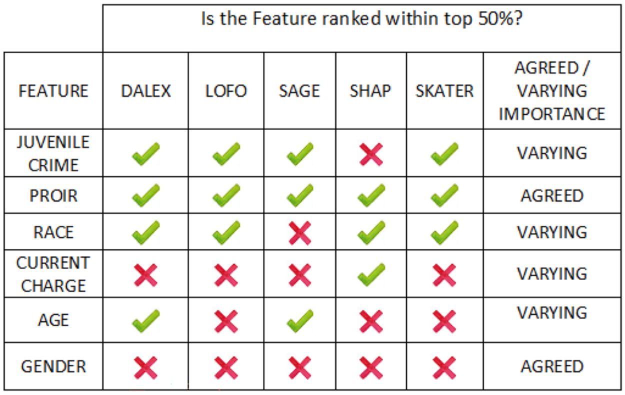


Fig. 2 Ranking explanations obtained for 200 Rashomon models of COMPAS dataset from 5 state-of-the-art methods. The X-axis shows the name of the method from which the ranking explanation was obtained. The Y-axis shows the number of models



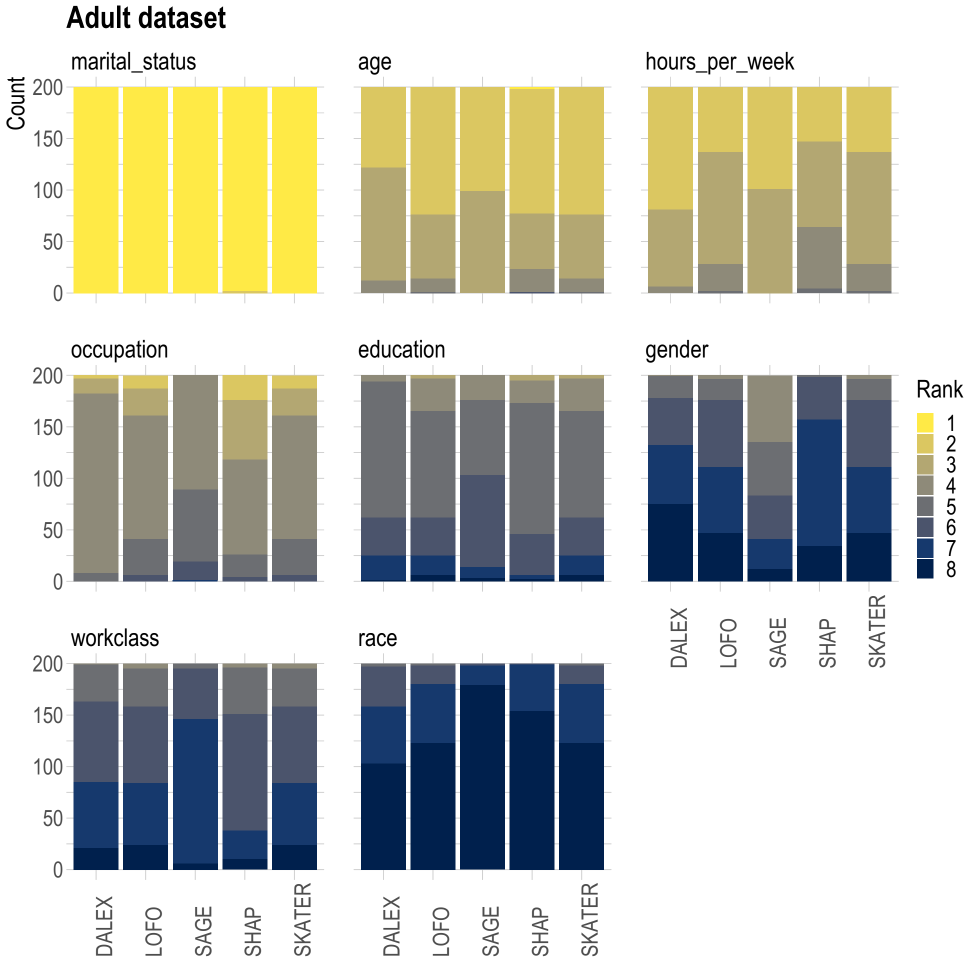
	Is the Feature ranked within top 50%?					
FEATURE	DALEX	LOFO	SAGE	SHAP	SKATER	AGREED / VARYING IMPORTANCE
JUVENILE CRIME	✓	✓	✓	✗	✓	VARYING
PROIR	✓	✓	✓	✓	✓	AGREED
RACE	✓	✓	✗	✓	✓	VARYING
CURRENT CHARGE	✗	✗	✗	✓	✗	VARYING
AGE	✓	✗	✓	✗	✗	VARYING
GENDER	✗	✗	✗	✗	✗	AGREED

Fig. 3 The agreeing or varying explanation status of the 5 methods that each feature (of COMPAS dataset) is in the top 50% ranks. The *varying* status shows the ranking variation between the method explanations and the *agreed* status shows the uniqueness among the explanations

by the Shap method explanations and moderately important for 40% of the models by the Skater and LOFO method explanations. In such a situation, concluding the importance of a feature based on a single method and model is biased towards the chosen method/model. Hence, the MAMCR framework helps to provide a unified explanation for a model class as a method-agnostic explanation range across multiple methods.

To address these conflicting explanations from the multiple tested methods, the MAMCR framework identifies the common ranking order of features using a reference optimal explanation ($\mathcal{O}$) for each Rashomon model using Eq. 4. For each dataset, an optimal explanation is identified from the five explanations for each model of the 200 Rashomon models. Figure 5 shows the ranking distribution of optimal explanations of COMPAS and Adult datasets. For other datasets, it can be found in Appendix A. Since the individual method explanations do not agree among them, the framework discovers the similarity of each explanation of a model by measuring how much it could reflect the ranking order of the corresponding optimal explanation. Based on the commonality with the optimal explanation, it is assigned a similarity score using Eq. 6. The similarity score decides how much each method could contribute its feature importance explanation preferences to the unified explanation. That is, for a method whose explanation similarity is 0.98 and the

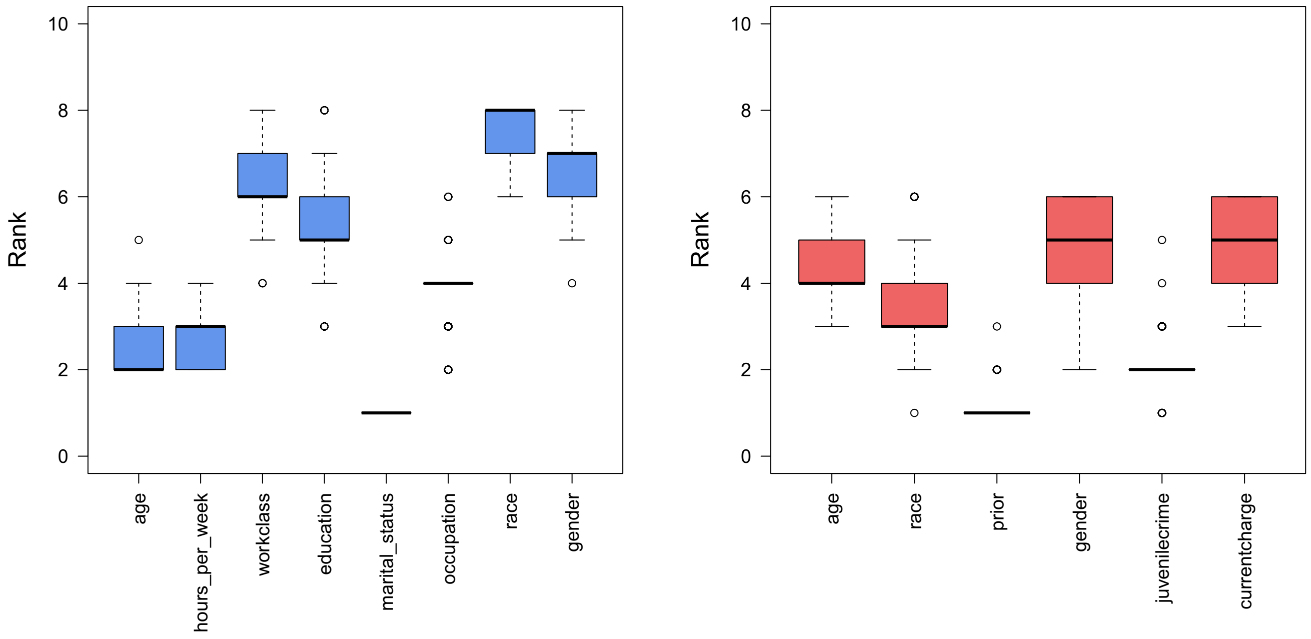
Fig. 4 Ranking explanations obtained for 200 Rashomon models of the Adult dataset from 5 state-of-the-art methods. The X-axis shows the name of the method from which the ranking explanation was obtained. The Y-axis shows the number of models



model reliance value on feature *j* is 0.5, the method can provide 98% of its feature importance value of *j* to the unified explanation, i.e., 0.49. Its explanation preference of feature *j* is considered for 98%. Suppose, the method has a lower similarity such as 0.45 due to its contradicting explanation, the method is allowed to provide only 45% of its explanation preference to the unified explanation, i.e.,0.225. Thus, the similarity score restricts the overestimation of methods and ensures the consistency of the feature importance range bounds. With the optimal similarity scores as the weight for each feature's importance value, the grand mean is computed using Eq. 7. These 200 weighted mean explanations provide the unified explanation order for the Rashomon models and from these values, the minimum and maximum boundary for each feature are determined using Eqs. 8 and 9, respectively. This forms the method agnostic model class reliance range explanation of multiple models across multiple methods.

The ranking explanation orders of five state-of-the-art methods and the unified MAMCR ranking explanation order for a model are shown in Fig. 6. The features are included as per the method's rank explanation order and the model performance is determined. The Adult dataset model's performance is reduced when the 2*nd* feature is included as per the Skater's explanation and also when the 4*th* feature is added as per the Sage method's order. However, the MAMCR's unified rank order performs better than the Skater and Sage methods. Though it is found less than the LOFO method's performance, it does not perform poorly like the other methods.

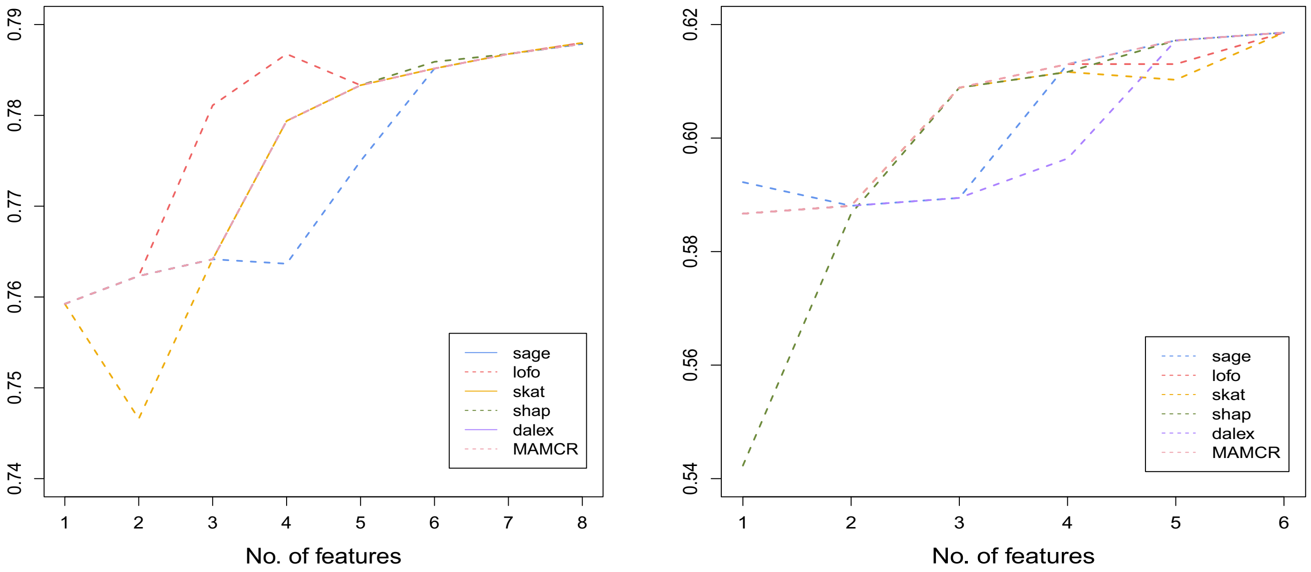
For the COMPAS dataset model, the MAMCR finds the best-unified explanation over the other state-of-the-art methods whereas the other methods' performance is reduced when the features are included in their estimated rank order. For the other three datasets (Wine, Heart, and Sonar), the results can be found in Appendix A. In each dataset, an individual method is identified as showing better performance. However, the same method is not identified as the best in all the datasets. Although the MAMCR method, does not strictly outperform every method for the various datasets used, it does higher on all the datasets consistently. For instance, the LOFO method results in the highest model performance for the Adult dataset but is identified with the lowest performance of all other methods for the Heart dataset when the number of features is equal to 2, 7, and 9, where



(a) ADULT

(b) COMPAS

Fig. 5 The feature importance ranking distribution of 200 optimal explanations of Rashomon sets of two datasets

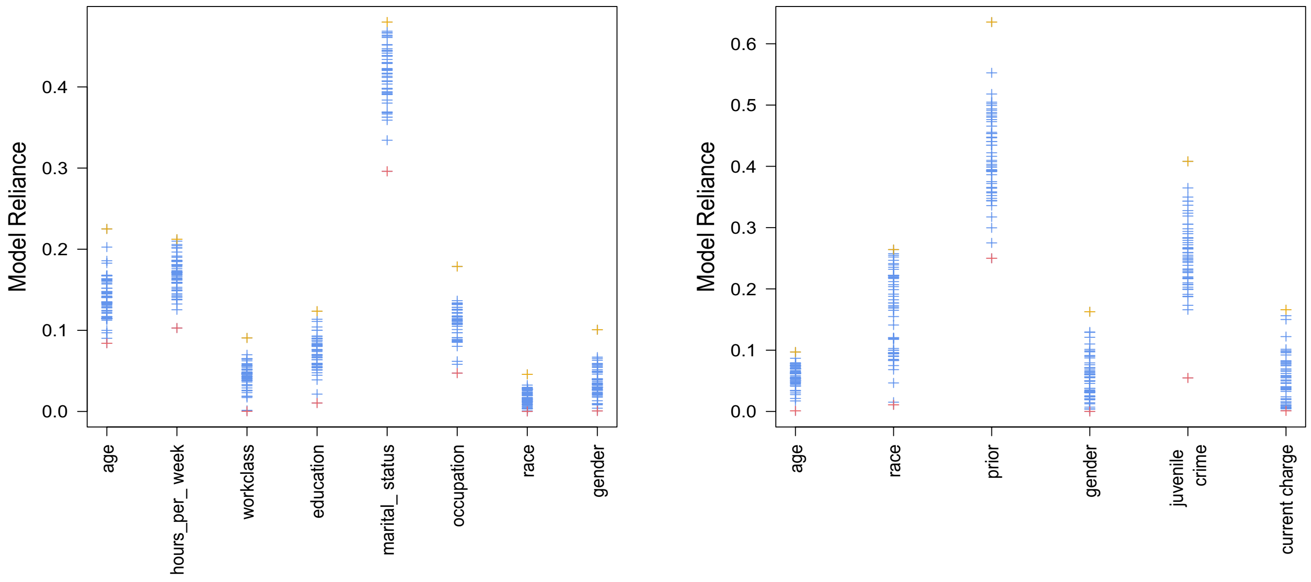


(a) ADULT

(b) COMPAS

Fig. 6 The ranking orders of various state-of-the-art methods are compared with the MAMCR's unified explanation order regarding a model's performance on the given datasets. The X-axis shows the no. of features included for determining the model's performance, while the Y-axis shows the model's prediction accuracy

the MAMCR shows the best model performance followed by the LOFO. Similarly, in Sonar and COMPAS datasets, the Shap method has the least model performance whereas the MAMCR performance proves the best performance even though a minimal number of features are included such as 1,2, or 3. Likewise, the Sage method seems to give a higher performance in the Wine dataset, but it has the lowest performance in the Adult dataset whereas the MAMCR has



(a) ADULT

(b) COMPAS

Fig. 7 The coverage of MAMCR range for randomly trained 50 models on the given datasets. The upper and lower bounds are marked as max and min, respectively. Each dot between the max (yellow) and min (red) bounds represents the importance (Model Reliance) value of a model on the concerned feature

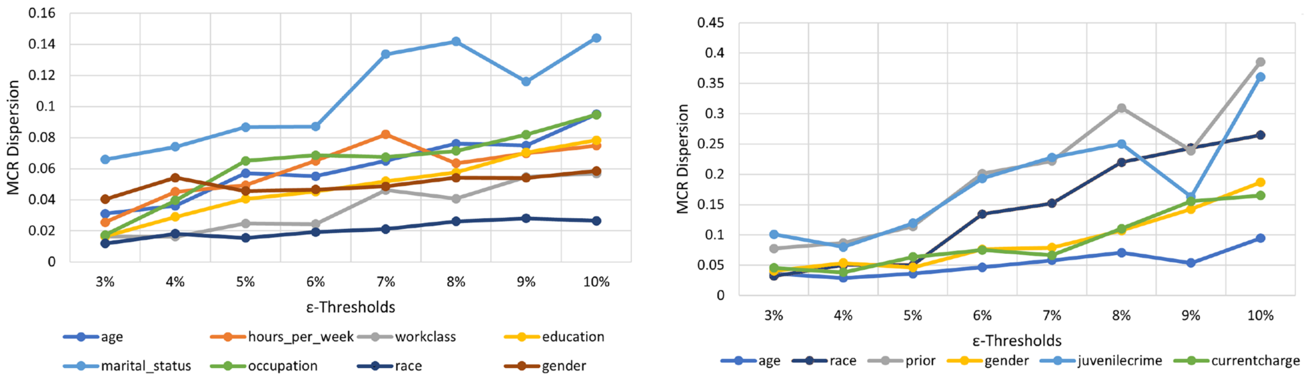
the highest performance in the Adult and closely the best performance with 4 features in the Wine dataset. Thus, the MAMCR demonstrates consistent performance in identifying the efficient unified explanation in all the datasets across model and method explanation multiplicity issues.

To verify the coverage of the MAMCR range for any model of the pre-specified model class that belongs to the corresponding Rashomon set criterion, multiple models are trained on the data and their feature importance is compared against the MAMCR range. For each dataset, 50 logistic regression models are trained with randomly sampled data, and their model reliance values are obtained from the five explanation methods and the unified explanation of each model is compared against the corresponding dataset's MAMCR range. The comparison results are displayed in Fig. 7 for COMPAS and Adult datasets. For other datasets, it could be found in Appendix A. Each model's feature-wise importance value is represented by a filled circle. They are contained within the max and min bounds of the corresponding MAMCR range. If they exceed the bounds, their over or underestimation can be detected.

The Adult and COMPAS datasets are analysed further for the different levels of ε values to understand how the variation in the prediction accuracy of the models affects their reliance on the features to make their prediction. For that, the various ε values have been experimented with from 3% to 10% which includes 1421 Adult dataset models and 3413 COMPAS dataset models. The MAMCR framework is applied to those models and the method agnostic model class reliance ranges are obtained for the above-said ε-thresholds. The reliance range length of the Rashomon models on each feature is computed from the difference between the *maximum* and *minimum* MCR bound values of each feature. This value indicates the dispersion length of the feature importance value that the models have on each feature for the various ε-threshold prediction accuracy levels and is shown in Fig. 8. The length of the importance value increases when the feature is considered the primary predictor by the models in varying prediction accuracy levels to make their predictions otherwise the feature's importance length stays at the same level without much difference. Through this, the importance of the features can be concluded when conflicting feature importance is allocated by different methods.

Discussion

The experiments done on the various datasets warn us that the explanation provided by one approach emphasising a feature as most important may not be projected in the same way by another method. This is illustrated through Figs. 2, 4, 9, 11 and 10. Ning et al.,[48] trained 350 near-optimal logistic regression models on the COMPAS dataset and obtained the feature importance explanations by applying the Sage method. They found the grand mean of the feature importance values and claimed that the 'Race' feature is not an important feature for those Rashomon models. As per their single reference model, 'Race' is identified as an important feature. However,



(a) ADULT (b) COMPAS

Fig. 8 The MCR Dispersion length of Rashomon models on each feature of the datasets depending on various ε-Thresholds. The increasing ε value indicates the extension of the prediction accuracy boundary for the Rashomon models

the aggregated explanation of 350 models does not highlight the 'Race' feature as important.

In our analysis also, the results of the Sage method agree with the conclusions of Ning et al [48]. Because, for more than 90% of the models, 'Race' is not an important feature and is not marked as one of the top 50% ranks (refer to Figs. 2 and 3). But in the case of Shap method's explanations, for more than 40% of the models, the 'Race' feature is the most important feature which disagrees with the analysis of the authors [48]. Other methods like Skater and LOFO also conflict with the claim that the 'Race' feature is not important. From this, it can be observed that for the same set of models, the different explanation methods return disagreeing explanations. Therefore, when a single method is used to explain a model, the resulting explanation tends to be biased towards that particular method. Similarly, when an explanation is generated based on a single model, it tends to exhibit bias towards that specific model. This highlights the importance of considering multiple methods and models to ensure a more comprehensive and unbiased understanding of AI systems.

This answers the first research question (RQ1) that the quantification of a variable as (un)important depends on both the model and the method that derives the explanation. Hence, through the unified method agnostic explanation, which is an unbiased explanation, the model and method multiplicity problem can be solved.

A detailed analysis with the varying ε-threshold values (refer to Fig. 8) is conducted on the COMPAS and Adult datasets to discover whether the prediction accuracy has an effect on the model's reliance value range. When the ε -threshold value is increased, the prediction accuracy boundary for the Rashomon set consideration is also increased. For the higher prediction accuracy models, the range of the model reliance is dense. For example, the most important features of COMPAS and Adult are 'Juvenile crime' and 'Marital status', respectively and their length of the model reliance is 0.1 and 0.07, respectively. When the ε-threshold value increases, the accuracy level is gradually reduced, and the length of model reliance on those variables is increased to 0.36 and 0.14. When accurate predictions start decreasing, the reliance on important variables gets sparse and thus, the length increases. In the case of the least important features, the sparse length of the model reliance between the varying prediction accuracies is comparatively lower than other variables. For example, the reliance length of the 'Race' feature of Adult and the 'Age' feature of COMPAS datasets do not vary much for any level of ε-threshold models.

The moderately important features that are provided with conflicting explanations by different methods show increasing patterns in their reliance length for the varying ε-threshold values. For example, the Occupation, Age, Work-class, and Hours-per-work features of the Adult dataset and the Race, Gender, and Current charge variables of the COMPAS dataset show such varying length patterns. Through this, we could observe the effect of models' prediction accuracy on the model reliance value range. This answers the second research question (RQ2) such as if the variable is not at all important to any model irrespective of its performance level, the variable's importance value range stays constant. Otherwise, its length varies according to the model's performance.

The 'Race' feature of the COMPAS dataset gets conflicting explanations from multiple methods. From the MAMCR dispersion length explanation (refer to Fig. 7b), it is observed that the 'Race' feature may not seem like an important feature at 3%-threshold but its reliance length is increased when the ε-threshold value increases. Therefore, this feature cannot be concluded as unimportant as the Sage method suggests. Instead, the MAMCR identifies the true importance of the 'Race' feature and confirms that there are some good (equally accurate) models that rely on the 'Race' feature from moderate to a high level for their predictions when the models' prediction accuracy boundary is extended.

Limitations and Future Work

The criteria for considering the Rashomon set is defined by the authors [19, 22] using the *model loss* and the boundary of $\mathfrak{R}$ models is extended up to $(1 + \varepsilon)$ of $\mathcal{L}(m^*)$, whereas we have used *model accuracy* and extended the boundary of $\mathfrak{R}$ models up to $(1 - \varepsilon)$ of $\mathcal{ACC}(m^*)$. However, depending on the ML model and the application domain, other evaluation metrics such as precision, recall, F1 score, etc., can also be used to select the nearby optimal models.

The Borda Count algorithm stands out as a versatile and widely applicable method across various domains and use cases, ranging from voting systems to sports rankings and preference aggregation in recommendation systems. Its simplicity and adaptability render it a popular choice in numerous scenarios. So, it is utilised to aggregate the ranking lists. Other distance-based rank aggregation algorithms such as [3, 18] can also be investigated for future work. However, these algorithms may cater more to specialised applications where factors like similarity, preference relations, or other nuanced considerations hold more importance. Furthermore, it's crucial to consider computational complexity in the aggregation process. While [3] may seem appealing for its ability to capture similarities through cosine distances, its computational demands can escalate rapidly, particularly when handling extensive sets of input lists.

Regarding the scope of explanation, this work unifies the global explanations. It can also be applied to local explanations as well. However, the variations in the explanations may be less in local explanations because the explanation is provided to the model prediction of a single data point. The computed feature importance value can vary but the commonality in identifying the top important feature ranking might be more.

The proposed method can be scaled to large datasets and complex models. Since the considered explanation methods are method-agnostic, they can be plugged into any complex model that deals with tabular datasets. However, the computational complexity depends on the applied methods for obtaining the explanations. Compared to LOFO, Dalex, and Skater methods, the Shap and Sage methods consume more time to produce global explanations.

Appendix A

See Figs. 9, 10, 11, 12, 13, 14, 15 and 16.

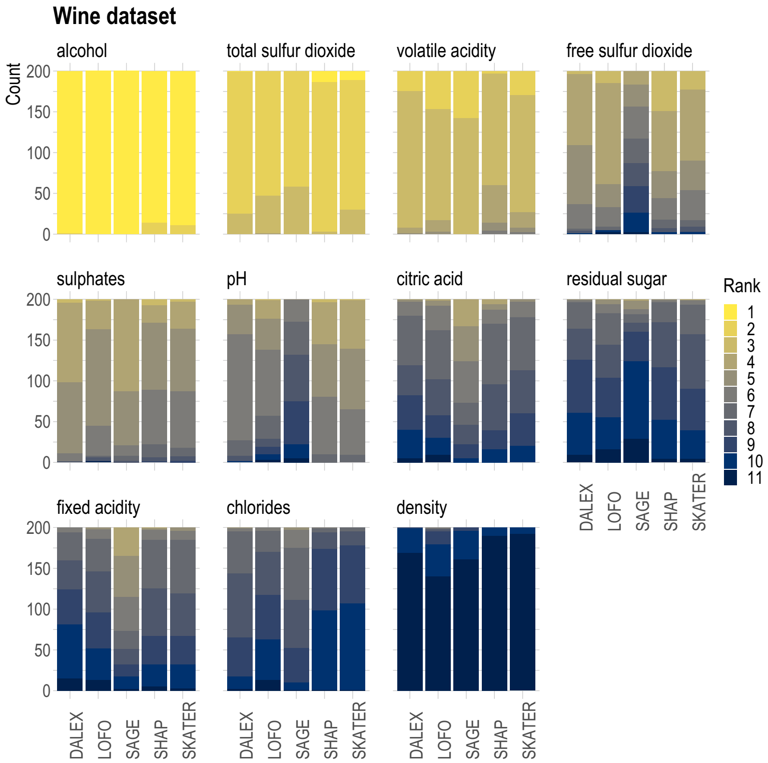


Fig. 9 Ranking explanations obtained for 200 Rashomon models of Wine dataset from 5 state-of-the-art methods. The X-axis shows the name of the method from which the ranking explanation was obtained. The Y-axis shows the number of models

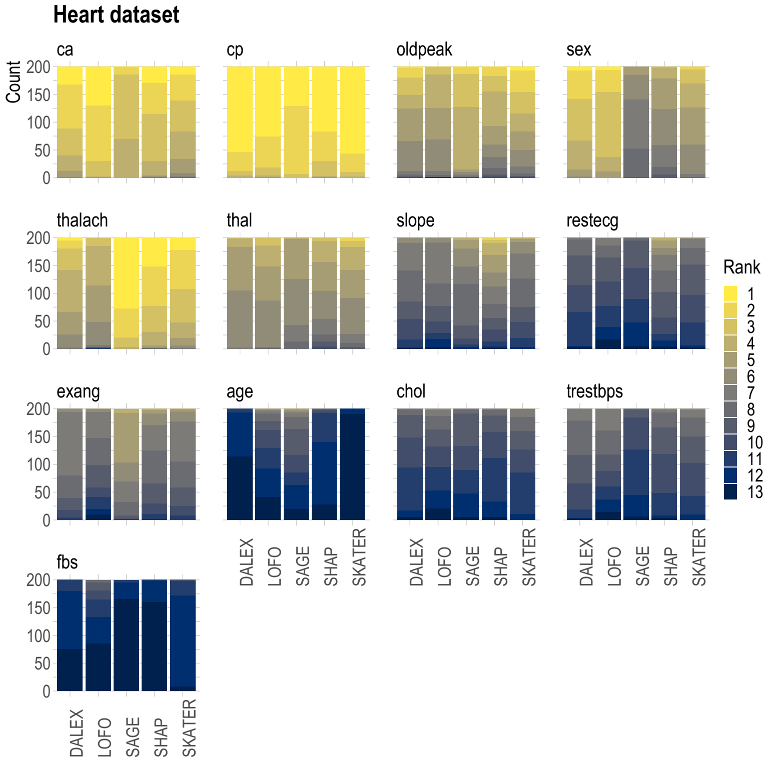


Fig. 10 Ranking explanations obtained for 200 Rashomon models of Heart dataset from 5 state-of-the-art methods. The X-axis shows the name of the method from which the ranking explanation was obtained. The Y-axis shows the number of models

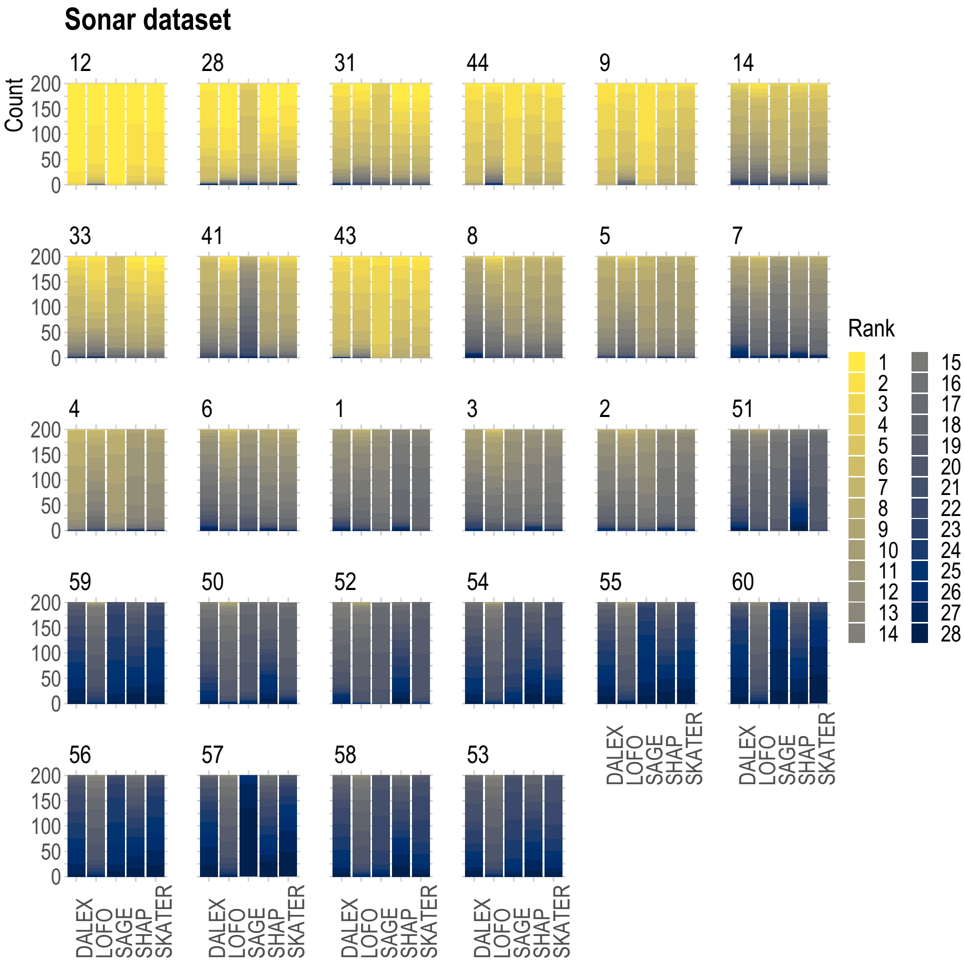


Fig. 11 Ranking explanations obtained for 200 Rashomon models of Sonar dataset from 5 state-of-the-art methods. The X-axis shows the name of the method from which the ranking explanation was obtained. The Y-axis shows the number of models

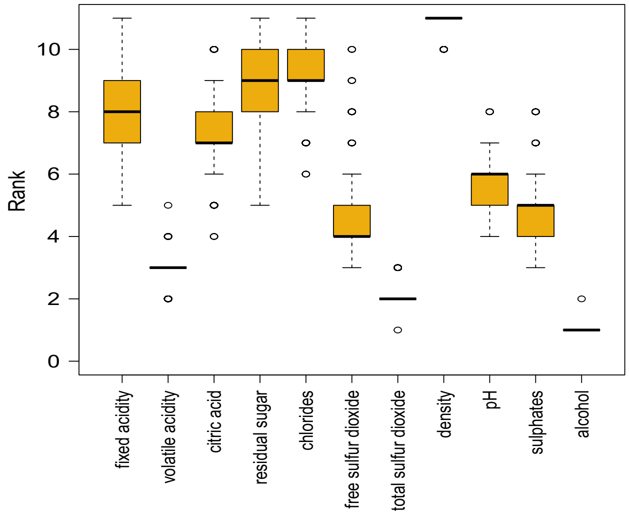


Fig. 12 The feature importance ranking distribution of 200 optimal explanations of Rashomon set of Wine dataset. Each box plot represents the ranking dispersion of each feature

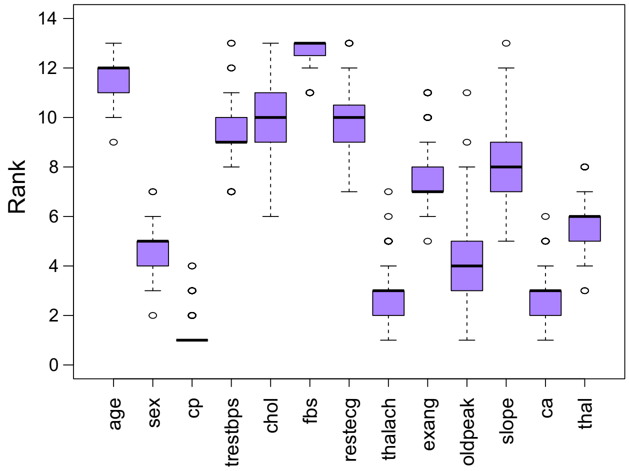


Fig. 13 The feature importance ranking distribution of 200 optimal explanations of Rashomon set of Heart dataset. Each box plot represents the ranking dispersion of each feature

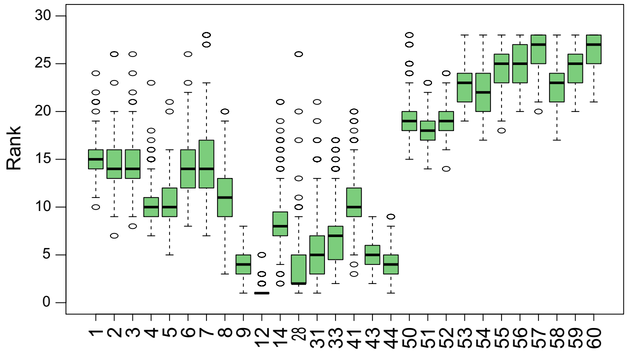
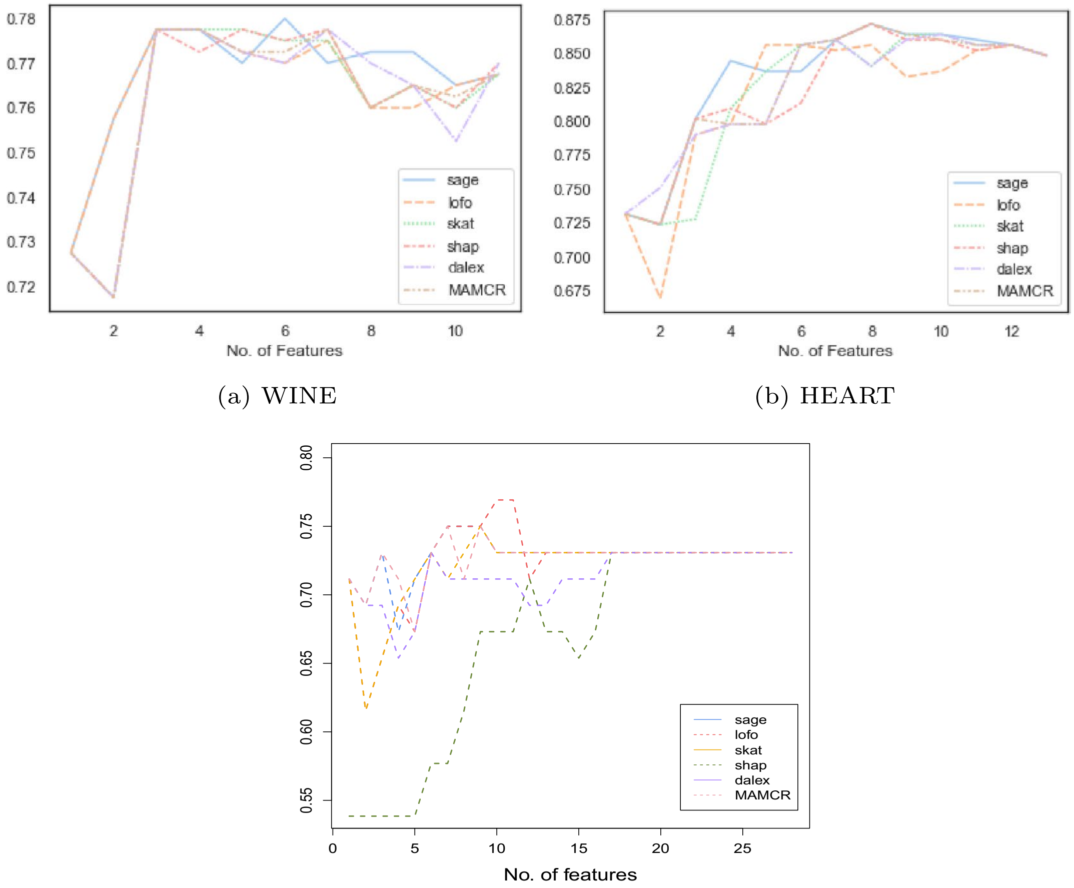


Fig. 14 The feature importance ranking distribution of 200 optimal explanations of Rashomon set of Sonar dataset. Each box plot represents the ranking dispersion of each feature

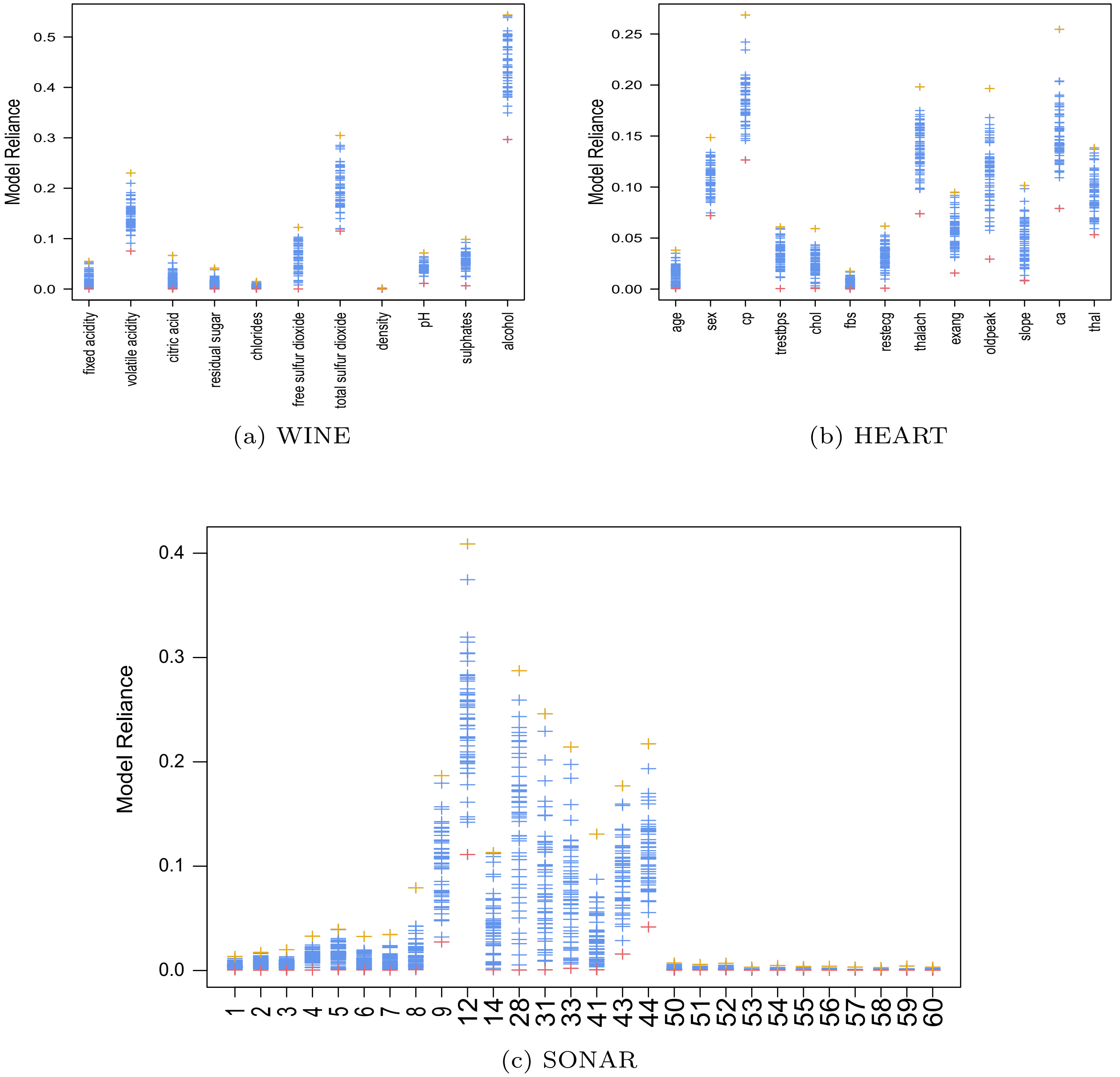


(a) WINE

(b) HEART

(c) SONAR

Fig. 15 The ranking order of various state-of-the-art methods are compared with the MAMCR's unified explanation order regarding a model's performance on the given datasets. The X-axis shows the number of features included for determining the model's performance, while the Y-axis shows the model's prediction accuracy



(a) WINE

(b) HEART

(c) SONAR

Fig. 16 The coverage of MAMCR range for randomly trained 50 models on the given datasets. The upper and lower bounds are marked as max and min, respectively. Each dot between the max (yellow) and min (red) bounds represents the importance (Model Reliance) value of a model on the concerned feature

Author Contributions Conceptualization, A.G., M.C. and P.M.; methodology, A.G.; formal analysis and investigation, A.G.; writing—original draft preparation, A.G., and P.M.; writing—review and editing, A.G., P.M. and M.C.; visualization, A.G. and P.M. All authors have read and agreed to the published version of the manuscript.

Data availability The data that support the findings of this study are openly available in Propublica Datastore at https://www.propublica.org/datastore [49], Kaggle Datasets at https://www.kaggle.com/datasets [44] and UCI ML Repository at https://archive.ics.uci.edu/ [67–69].

Declarations

Conflict of interest All authors declare that they have no conflicts of interest.

References

1. Adadi Amina, Berrada Mohammed. Peeking inside the black-box: a survey on explainable artificial intelligence (xai). IEEE Access. 2018;6:52138–60.
2. Ahmet E. LOFO (Leave One Feature Out) Importance. 2019.
3. Akritidis L, Fevgas A, Bozanis P, Manolopoulos Y. An unsupervised distance-based model for weighted rank aggregation with list pruning. Expert Syst Appl. 2022;202: 117435.
4. Alonso Jose M, Javier T-A, Alberto B. Experimental study on generating multi-modal explanations of black-box classifiers in terms of gray-box classifiers. In: 2020 IEEE International Conference on fuzzy systems (FUZZ-IEEE), 2020; p. 1–8. IEEE.
5. Arrieta AB, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, García S, Gil-López S, Molina D, Benjamins R, et al. Explainable artificial intelligence (xai): concepts, taxonomies, opportunities and challenges toward responsible ai. Inform Fusion. 2020;58:82–115.
6. Artelt A, Hammer B. Efficient computation of counterfactual explanations of lvq models. arXiv preprint arXiv:1908.00735, 2019.
7. Bastani O, Kim C, Bastani H. Interpretability via model extraction. arXiv preprint arXiv:1706.09773, 2017.
8. Biecek P. Dalex: explainers for complex predictive models in r. J Mach Learn Res. 2018;19(1):3245–9.
9. Bland JM, Kerry SM. The Weighted comparison of means. Bmj. 1998;316(7125):129.
10. Bobek S, Bałaga P, Nalepa Grzegorz J. Towards model-agnostic ensemble explanations. In: International Conference on computational science, Springer, p. 39–51. 2021.
11. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32.
12. Breiman L. Statistical modeling: the two cultures (with comments and a rejoinder by the author). Stat Sci. 2001;16(3):199–231.
13. Choudhary P, Kramer A, and datascience.com team. datascienceinc/Skater: Enable Interpretability via Rule Extraction(BRL), 2018.
14. Cortez P, Cerdeira A, Almeida F, Matos T, Reis J. Modeling wine preferences by data mining from physicochemical properties. Decis Support Syst. 2009;47(4):547–53 (**Smart Business Networks: Concepts and Empirical Evidence**).
15. Covert I, Lundberg S, Lee S-I. Feature removal is a unifying principle for model explanation methods. arXiv preprint arXiv:2011.03623, 2020.
16. Covert I, Lundberg SM, Lee S-I. Understanding global feature contributions with additive importance measures. Adv Neural Inf Process Syst. 2020;33:17212–23.
17. Datta A, Sen S, Zick Y. Algorithmic transparency via quantitative input influence: theory and experiments with learning systems. In: 2016 IEEE symposium on security and privacy (SP), 2016; p. 598–617. IEEE.
18. Desarkar MS, Sarkar S, Mitra P. Preference relations based unsupervised rank aggregation for metasearch. Expert Syst Appl. 2016;49:86–98.
19. Dong J, Rudin C. Exploring the cloud of variable importance for the set of all good models. Nat Mach Intell. 2020;2(12):810–24.
20. Duell JA. A comparative approach to explainable artificial intelligence methods in application to high-dimensional electronic health records: Examining the usability of xai. arXiv preprint arXiv:2103.04951, 2021.
21. Fan M, Wei W, Xie X, Liu Y, Guan X, Liu T. Can we trust your explanations? sanity checks for interpreters in android malware analysis. IEEE Trans Inf Forensics Secur. 2020;16:838–53.
22. Fisher A, Rudin C, Dominici F. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. J Mach Learn Res. 2019;20(177):1–81.
23. Fong R, Patrick M, Vedaldi A. Understanding deep networks via extremal perturbations and smooth masks. In: Proceedings of the IEEE/CVF International Conference on computer vision, 2019; p. 2950–2958.
24. Garreau D, von Luxburg U. Looking deeper into tabular lime. arXiv preprint arXiv:2008.11092, 2020.
25. Ghorbani A, Abid A, Zou J. Interpretation of neural networks is fragile. In: Proceedings of the AAAI Conference on artificial intelligence. 2019;33:3681–8.
26. Gifi A. Nonlinear multivariate analysis. Wiley-Blackwell; 1990.
27. Goldstein A, Kapelner A, Bleich J, Pitkin E. Peeking inside the black box: visualizing statistical learning with plots of individual conditional expectation. J Comput Graph Stat. 2015;24(1):44–65.
28. Guidotti R, Ruggieri S. Assessing the stability of interpretable models. arXiv preprint arXiv:1810.09352, 2018.
29. Gunasekaran A, Chen M, Hill R, McCabe K. Method agnostic model class reliance (mamcr) explanation of multiple machine learning models. In: International Conference on soft computing and its engineering applications, Springer, 2022; p. 56–71.
30. Hamamoto M, Egi M. Model-agnostic ensemble-based explanation correction leveraging rashomon effect. In: 2021 IEEE Symposium Series on Computational Intelligence (SSCI), 2021; p. 01–08. IEEE.
31. Hazwani IA, Schmid J, Sachdeva M, Bernard J. A design space for explainable ranking and ranking models. arXiv preprint arXiv:2205.15305, 2022.
32. Horel E, Giesecke K. Computationally efficient feature significance and importance for machine learning models. arXiv preprint arXiv:1905.09849, 2019.
33. Kendall MG. Rank correlation methods. 1948.
34. Kindermans P-J, Hooker S, Adebayo J, Alber M, Schütt KT, Dähne S, Erhan D, Kim B. The (un) reliability of saliency methods. In: Explainable AI: interpreting, explaining and visualizing deep learning. Springer; 2019, p. 267–280.
35. Kobylińska K, Orłowski T, Adamek M, Biecek P. Explainable machine learning for lung cancer screening models. Appl Sci. 2022;12(4):1926.
36. Krause J, Perer A, Ng K. Interacting with predictions: visual inspection of black-box machine learning models. In: Proceedings of the 2016 CHI Conference on human factors in computing systems, 2016; p. 5686–5697.
37. Le F, Srivatsa M, Reddy KK, Roy K. Using graphical models as explanations in deep neural networks. In: 2019 IEEE 16th International Conference on mobile ad hoc and sensor systems (MASS), 2019; p. 283–289. IEEE.
38. Lei J, G'Sell M, Rinaldo A, Tibshirani RJ, Wasserman L. Distribution-free predictive inference for regression. J Am Stat Assoc. 2018;113(523):1094–111.
39. Lenders D, et al. Getting the best of both worlds? combining local and global methods to make ai explainable. 2020.
40. Lin S. Rank aggregation methods. Wiley Interdiscipl Rev omput Stat. 2010;2(5):555–70.
41. Lundberg SM Lee S-In. A unified approach to interpreting model predictions. In: Advances in neural information processing systems. 2017;30.

42. Luo CF, Bhambhoria R, Dahan S, Zhu X. Evaluating explanation correctness in legal decision making. 2022.
43. Luštrek M, Gams M, Martinčić-Ipšić S, et al. What makes classification trees comprehensible? Expert Syst Appl. 2016;62:333–46.
44. Mayur D. SONAR Mine Dataset, 2022.
45. Molnar C. Interpretable machine learning. Lulu. com, 2020.
46. Nayebi A, Tipirneni S, Foreman B, Reddy CK, Subbian V. An empirical comparison of explainable artificial intelligence methods for clinical data: a case study on traumatic brain injury. arXiv preprint arXiv:2208.06717, 2022.
47. Nguyen TT, Nguyen TL, Ifrim G. A model-agnostic approach to quantifying the informativeness of explanation methods for time series classification. In: International Workshop on Advanced Analytics and Learning on Temporal Data. Springer, 2020; p. 77–94.
48. Ning Y, Eng Hock Ong M, Chakraborty B, Goldstein BA, Ting DSW, Vaughan R, Liu N. Shapley variable importance cloud for interpretable machine learning. Patterns. 2022;3(4):100452.
49. ProPublica Data store. COMPAS Recidivism Risk Score Data and Analysis of Broward County of Florida, 2016.
50. Pruthi D, Bansal R, Dhingra B, Soares LB, Collins M, Lipton ZC, Neubig G, Cohen WW. Evaluating explanations: How much do explanations from the teacher aid students? Trans Asoc Comput Linguist. 2022;10:359–75.
51. Ratul QEA, Serra E, Cuzzocrea A. Evaluating attribution methods in machine learning interpretability. In: 2021 IEEE International Conference on Big Data (Big Data), 2021; p. 5239–5245. IEEE.
52. 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; p. 1135–1144.
53. Robnik-Šikonja M, Kononenko I. Explaining classifications for individual instances. IEEE Trans Knowl Data Eng. 2008;20(5):589–600.
54. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell. 2019;1(5):206–15.
55. Semenova L, Rudin C, Parr R. A study in Rashomon curves and volumes: A new perspective on generalization and model simplicity in machine learning. arXiv preprint arXiv: 1908.01755, 2019.
56. Shapley LS. A value for n-person games. Contrib Theory Games. 1953;2:307–17.
57. Shi S, Zhang X, Fan W. A modified perturbed sampling method for local interpretable model-agnostic explanation. arXiv preprint arXiv:2002.07434, 2020.
58. Shrikumar A, Greenside P, Kundaje A. Learning important features through propagating activation differences. In: International Conference on machine learning, PMLR, 2017; p. 3145–3153.
59. Silva W, Fernandes K, Cardoso JS. How to produce complementary explanations using an ensemble model. In: 2019 International Joint Conference on Neural Networks (IJCNN), 2019; p. 1–8. IEEE.
60. Slack D, Hilgard S, Jia E, Singh S, Lakkaraju H. Fooling lime and shap: adversarial attacks on post hoc explanation methods. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 2020; p. 180–186.
61. Smith G, Mansilla R, Goulding J. Model class reliance for random forests. Adv Neural Inf Process Syst. 2020;33:22305–15.
62. Staniak M, Biecek P. Explanations of model predictions with live and breakdown packages. arXiv preprint arXiv:1804.01955, 2018.
63. Štrumbelj E, Kononenko I. Explaining prediction models and individual predictions with feature contributions. Knowl Inf Syst. 2014;41(3):647–65.
64. Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. In: International Conference on machine learning, PMLR, 2017; p. 3319–3328.
65. Tamagnini P, Krause J, Dasgupta A, Bertini E. Interpreting black-box classifiers using instance-level visual explanations. In: Proceedings of the 2nd Workshop on human-in-the-loop data analytics, 2017; p. 1–6.
66. Thiagarajan JJ, Kailkhura B, Sattigeri P, Ramamurthy KN. Treeview: Peeking into deep neural networks via feature-space partitioning. arXiv preprint arXiv:1611.07429, 2016.
67. UCI Machine Learning. Adult Quality Dataset, 1996.
68. UCI Machine Learning. Heart Disease Dataset, 1998.
69. UCI Machine Learning. Wine Quality Dataset, 2009.
70. Velmurugan M, Ouyang C, Moreira C, Sindhgatta R. Evaluating explainable methods for predictive process analytics: a functionally-grounded approach. arXiv preprint arXiv: 2012.04218, 2020.
71. Warnecke A, Arp D, Wressnegger C, Rieck K. Evaluating explanation methods for deep learning in security. In: 2020 IEEE European symposium on security and privacy (EuroS &P), IEEE, 2020; p. 158–174.
72. Webber W, Moffat A, Zobel J. A similarity measure for indefinite rankings. ACM Trans Inform Syst (TOIS). 2010;28(4):1–38.
73. Wei P, Zhenzhou L, Song J. Variable importance analysis: a comprehensive review. Reliabil Eng Syst Saf. 2015;142:399–432.
74. Wexler J, Pushkarna M, Bolukbasi T, Wattenberg M, Viégas F, Wilson J. The what-if tool: interactive probing of machine learning models. IEEE Trans Visual Comput Graph. 2019;26(1):56–65.
75. Wolpert DH. The supervised learning no-free-lunch theorems. Soft computing and industry, 2002; p. 25–42.
76. Yang SX, Tian YJ, Zhang CH. Rule extraction from support vector machines and its applications. In: 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, volume 3, IEEE, 2011; p. 221–224.
77. Zhou Z-H, Jiang Y, Chen S-F. Extracting symbolic rules from trained neural network ensembles. AI Commun. 2003;16(1):3–15.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.