Article:
Pelegrina, G.D. and Siraj, S. orcid.org/0000-0002-7962-9930 (2024) Shapley value-based approaches to explain the quality of predictions by classifiers. IEEE Transactions on Artificial Intelligence, 5 (8). 4217 -4231. ISSN 2691-4581

https://doi.org/10.1109/tai.2024.3365082

Shapley value-based approaches to explain the quality of predictions by classifiers

Guilherme Dean Pelegrina, Sajid Siraj

***Abstract*—The use of algorithm-agnostic approaches for explainable machine learning models is an emerging area of research. When explaining the contribution of individual features towards the predicted outcome, traditionally, the focus remains on explaining the prediction itself, however a little has been done on explaining the quality of prediction of these models, where the quality can be assessed as robustness in predictions when changing the thresholds for classification. In this paper, we propose the use of Shapley values to explain the contribution of each feature towards this robustness, measured in terms of Receiver-operating Characteristics (ROC) curve and the Area under the ROC curve (AUC). With the help of an illustrative example, we demonstrate the proposed idea of explaining the ROC curve, and visualising the uncertainties in these curves. For imbalanced datasets, the use of Precision-Recall Curve (PRC) is considered more appropriate, therefore we also demonstrate how to explain the PRCs with the help of Shapley values. The explanation of robustness can help analysts in a number of ways, for example, it can help in feature selection by identifying the irrelevant features that can be removed to reduce the computational complexity. It can also help in identifying the features having critical contributions or negative contributions towards the quality of predictions.**

***Impact Statement*—Works in machine learning explainability generally focus on feature attribution methods towards local interpretability or performance measures, such as accuracy. Moreover, in these scenarios, the interpretations are based on a single decision threshold. In this paper, we extend the of use of Shapley value-based approach for machine learning interpretability to explain the contribution of features towards the robustness of predictions measure by means of the Receiver-operating Characteristics (ROC) curve and the Area under the ROC curve (AUC). We highlight that our proposal can be useful in feature selection tasks. We also provide all codes supporting this paper in order to ensure reproducibility.**

***Index Terms*—explainable artificial intelligence; machine learning; business analytics.**

I. Introduction

The fields of machine intelligence and decision support tools have grown by leaps and bounds in last two decades [1]–[3]. This growth can be attributed to a significant improvement in the performance of various prediction algorithms (in terms of their accuracy, precision, recall, etc.). However, when considering the practicality of using these prediction algorithms, the majority of them tends to be quite complex. A key contribution in this domain is the recent introduction of algorithm-agnostic explanation approaches to explain the contribution of each feature towards the overall prediction [4]–[7]. It is certainly an important achievement in predictive analytics as most of the models (specially those based on random forests [8], deep neural networks [9], [10] and extreme gradient boosting [11], [12]) were previously treated as black box models and were questioned due to their complex nature. Indeed, besides the performance, aspects such as fairness and explainability (among others) are also important when deciding which machine learning (ML) model to adopt [13]–[16].

One of the widely-used algorithm-agnostic approaches to explain ML models is based on the cooperative game-theoretic concept called the Shapley value [17]. The Shapley value approach is considered as a fair way to divide the payoff in a game among its players. In ML context, it can be used as a feature attribution method, i.e. a measure that indicates how much each feature is contributing in the ML task. Therefore, the main idea is to see the ML problem (classification, prediction, etc.) as a cooperative game such that the features cooperate in order to achieve a specific goal. This goal depends on what one would like to explain, for example, [18] used the Shapley values to explain the coefficient of determination in regression models. As the Shapley value calculation considers all coalitions of regressors, the obtained results were consistent even in scenarios with multicollinearity. [19] associated payoffs in game theory to fairness measures and applied the Shapley values to explain the impact of features when there are disparate results regarding different groups of people (women and men, blacks and whites, etc.). Moreover, in the famous SHAP method proposed by [20], the Shapley values were used to indicate the contribution of each feature in local predictions.

In the field of ML, improving the performance of prediction algorithms has remained the main focus for decades. The performance metrics of accuracy, precision and F-measure are considered almost mandatory when evaluating any prediction algorithm. Although there are recent proposals to estimate the contribution of each feature towards prediction, there is still a need to explain their contributions towards the robustness of these models [21]–[23]. The robustness here refers to the impact on models' performance by changes in the classification threshold. To assess the robustness in prediction models, the Receiver Operating Characteristic (ROC) curve and the area under the ROC curve (AUC) have been widely used in ML [24], which has traditionally been used in the field of

The authors would like to thank the grants #2021/11086-0, #2020/10572-5 and #2020/09838-0, São Paulo Research Foundation (FAPESP), for the financial support. The authors would also like to thank the Department of Management in Leeds University Business School for providing resources to conduct this research.

G. Pelegrina is with the School of Applied Sciences (FCA), University of Campinas (UNICAMP), Limeira, Brazil (e-mail: guidean@unicamp.br). S. Siraj is with the Centre for Decision Research, Leeds University Business School, Leeds, UK, and COMSATS University Islamabad, Wah Campus, Pakistan (e-mail: S.Siraj@leeds.ac.uk).

operational research and signal processing for long time [25]. We contend that the Shapley-values can also be used to explain robustness of the ML models, for example, by explaining the contribution of each feature towards the AUC. Imagine a situation where we achieve a model with the AUC of 0.90. It tells us how robust the prediction model is, however, it does not explain how much each feature has contributed towards this robustness. Also, one may wish to investigate whether the contribution of each feature varies for different specificity, or does it remain consistent regardless of the specificity values.

Considering this gap, we first propose the ShapAUC method, a Shapley-based approach to explain the AUC for any ML model. For this, we assume that features join in coalitions and cooperate to achieve the AUC as a common goal. Based on the AUC calculated for all coalitions of features, we calculate the Shapley values, which indicate the marginal contribution of each feature in the overall model performance. As a second contribution, namely ShapROC, we propose a way that explains the contribution of each feature towards the ROC curve. Shapley value is calculated to provide the marginal contribution of each feature at each point inside the ROC curve. As the use of Precision-Recall Curves (PRCs) is preferred for imbalanced datasets, we also propose to extend the use of Shapley values to decompose PRCs as well as to calculate the contributions towards the area under the PRC (AUPRC). Based on numerical experiments in a real dataset, we show the usefulness of our proposals in explaining the contribution of each feature towards the robustness of a model.

One of the benefits of the proposed approach is to use it in feature selection. As we provide the marginal contributions of features towards robustness, it is possible to identify insignificant features, or more importantly, identifying features having negative contribution. By removing a feature with negative contribution, we may improve the model robustness, and by removing insignificant features, we can reduce the computational complexity of the classifier.

The rest of this paper is organised as follows. Section II discusses the background of our proposals, which lies in the robustness of prediction models and the use of Shapley values as a feature attribution method. In Section III, we present the proposed approaches. Section IV illustrates the use of our proposals in a real dataset. Finally, in Section V, we show our conclusions and future perspectives.

II. Background

This section presents the theoretical background used in our proposals. Firstly, we discuss the key elements when assessing the prediction model robustness based on ROC curve and PRC. Thereafter, we discuss the use of Shapley values as a feature attribution method in ML explainability.

A. Measuring the performance of classifiers

The performance of classifiers can be measured in different ways. A common way of assessing performance is to calculate the accuracy of prediction, which is essentially a ratio of the correct predictions made out of the overall predictions carried out. In practise, this metric might not be very useful in situations where predicting positive outcomes are more important than predicting negative outcomes, or vice versa. Therefore, classification performance usually involves construction of a confusion matrix that shows the number of true positives (TPs), true negatives (TNs), false positives (FPs), and false negatives (FNs). This is illustrated in Figure 1.

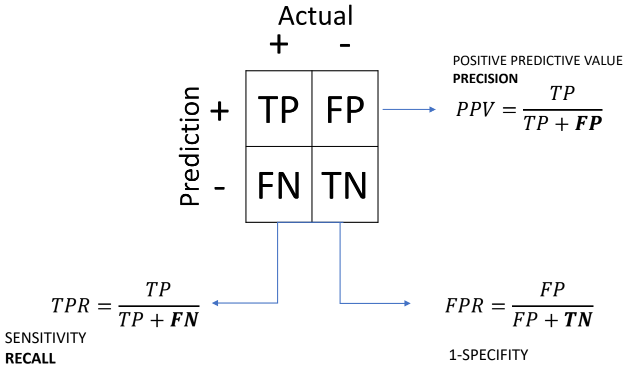

Actual
+
-
Prediction
+
-
TP
FP
FN
TN
POSITIVE PREDICTIVE VALUE
PRECISION
$PPV = \frac{TP}{TP + \boldsymbol{FP}}$
$TPR = \frac{TP}{TP + \boldsymbol{FN}}$
SENSITIVITY
RECALL
$FPR = \frac{FP}{FP + \boldsymbol{TN}}$
1-SPECIFITY

Fig. 1. Confusion matrix showing definitions for performance metrics.

From this confusion matrix, a number of different performance measures can be obtained, for example, the ratio of true positives to actual positives is known as the true positive rate (TPR), also known as Recall or Sensitivity. Similarly, the ratio of true negatives to actual negatives is known as the true negative rate or Specificity. The complementary value of Specificity is known as false positive rate (FPR). Another way of assessing classifier is its ability to detect true positive cases out of the cases that were detected as positives. This is known as the Precision of the classifier. All these metrics capture different aspect of model's performance and are useful, however, for testing robustness of model, the construction of confusion matrix itself has to be investigated and analysed. In ML literature, it is common practise to repeat training and testing several times, and testing the consistency in these performance scores. Another common practise is to assess model's performance by varying different parameters like classification thresholds. We discuss this below in more detail.

1) ROC curve and AUC: Receiver Operating Characteristic (ROC) curve has been widely used in machine learning [24] to assess the robustness in prediction models. In a binary classification problem, the outcome is considered positive when the prediction probability is obtained above a certain threshold. For example, in a fingerprint authentication system, a fingerprint image is scanned and a ML algorithm calculates the probability for it to be a valid fingerprint. If the predicted probability happens to be 0.40, it can be declared unauthorised considering that any value lower than 0.50 is closer to 0 (false) than 1 (true). However, we can relax this requirement by lowering the threshold value from 0.50 to 0.30, in which case, the predicted probability of 0.40 will be declared true (i.e. authorised). This means that lower threshold will have higher risk of authorising the unauthorised fingerprints (false positive), while on the other hand, raising the threshold will have higher risk of rejecting the authorised fingerprints (false negatives). Ideally, a system should be robust enough to detect all true positives regardless of the threshold values. This robustness can be investigated and quantified using ROC curve which is, simply put, a line plot between FPR and TPR values calculated by varying the threshold values for classifying predicted probability. The overall robustness of the algorithm is summarised by calculating the area under this ROC curve (AUC) which is a value between 0 and 1. A value of 1 implies that the algorithm is robust in detecting all true positives no matter what the threshold value is.

B. Precision-recall curve and AUPRC

Although both ROC and AUC have been widely used, their usefulness has been debated for imbalanced classification problems [26], [27]. For example, email spam detection is a classical ML problem where the two classes are highly imbalanced [28]. In this case, people prefer to minimise false positives, and therefore, only interested in the left side of the ROC curve (where FPR is close to 0), however, the calculation of AUC does not prioritise one side over other. For imbalanced classes, the use of Precision-Recall Curve (PRC) is considered more appropriate than the use of ROC and AUC [26], [27]. As the names suggests, the PRC explains the relationship between precision and recall for all threshold values.

C. Explanation in ML models

As the use of ML is getting common, there is an increasing demand (and pressure) for the explainability of these ML models. For example, a bank's customer might ask for reasons why his/her application for credit got refused. Since the introduction of SHapley Additive exPlanations (SHAP) by [20], the use of explainable AI has been introduced in many practical applications like managing financial risk [29], [30], healthcare management [31], [32], studying pandemics [33], and many more. However, so far, explainability has mostly focused on how to explain the contribution of each predictor towards attaining the predicted outcomes. While explaining the predicted outcome is an important area to investigate, it is also important to explain the robustness in predicting these values. For example, if a model is shown to have robustness of 0.85, it is important to explain how different predictors have contributed towards achieving this level of robustness. In this context, the use of Shapley values can give promising results, as it has already been demonstrated to be useful in practical applications involving ML.

D. Shapley values as a feature attribution method

Before defining the Shapley value, let us first introduce the notion of cooperative games [34]. In a cooperative game, there exists a cooperative behaviour among a set of players aiming at achieving a predetermined goal. Several practical situations can be modelled as a cooperative game problem [35]–[37]. For instance, [38]–[40] showed applications in power system expansion planning. In this case, different companies can cooperate in order to reduce power losses or investment cost allocation in power transmission systems. Another example includes modelling supply chain management tasks as a cooperative game problem [41]–[44]. A common goal shared by managers can be the fixed cost paid by shipment orders. Therefore, if they form a coalition and order simultaneously, they could save more money than if they act separately.

Suppose a set $N = \{1, 2, \ldots, n\}$ of n players. Mathematically, one may define a coalition game on N by a characteristic function $\upsilon : \mathcal{P}(N) \rightarrow \mathbb{R}$, where $\mathcal{P}(N)$ is the power set of N, that maps all possible coalitions of players to real numbers, such that[1] $\upsilon(\emptyset) = 0$. One frequently refers to $\upsilon(A)$, where $A \subseteq N$, as the payoff (or the benefit) achieved by the coalition A when cooperating in the game. For example, in the supply chain task mentioned earlier, $\upsilon(A)$ could represent the savings obtained by the coalition of managers A when ordering simultaneously. However, a question that arises in a cooperative game is how to divide the gains obtained by a coalition of players. One of the well-known solutions for such a sharing is called Shapley value [17]. For each player $i \in N$, the associated Shapley value represents the marginal contribution of the player in the game payoff when considering all possible coalitions of players. It can be defined as follows:

$$\phi_i = \sum_{A \subseteq N \backslash \{i\}} \frac{(n - |A| - 1)!\, |A|!}{n!} \Delta_i \upsilon(A), \tag{1}$$

where $\Delta_i \upsilon(A) = \upsilon(A \cup \{i\}) - \upsilon(A)$ and $|A|$ indicates the cardinality of subset A.

An interesting aspect of Shapley value is that it satisfies several desired properties when allocating benefits among players. In this paper, three of them are important (see [45] for another property):

Property 1. Efficiency: The sum of the Shapley values of all players is equal to the payoff of the grand coalition N discounted by the payoff of the empty coalition. As by the definition of a game $\upsilon(\emptyset) = 0$, the gain $\upsilon(N)$ is distributed among the players:

$$\sum_{i=1}^{n} \phi_i = \upsilon(N) - \upsilon(\emptyset) = \upsilon(N). \tag{2}$$

Property 2. Null player: If, for all subset $A \subseteq N$,

$$\upsilon\left(A \cup \{i\}\right) = \upsilon\left(A\right), \tag{3}$$

then $\phi_i = 0$. It means that, if there is no gain when player i joins any coalition (he/she does not contribute in any payoff), he/she will not receive benefits.

Property 3. Symmetry: If two players i and i' are such that

$$\upsilon\left(A \cup \{i\}\right) = \upsilon\left(A \cup \{i'\}\right), \tag{4}$$

for all $A \subset N$ which contains neither i nor i', then $\phi_i = \phi_{i'}$. Therefore, if two players contribute equally when joining all coalitions, they should receive the same amount.

[1] In cooperative game theory, one assumes $\upsilon(\emptyset) = 0$ as there is no gain when there is no player in the coalition. However, when using the Shapley value in machine learning explainability, one may assume a different value for $\upsilon(\emptyset)$. We will further discuss this issue in this paper.

Given these properties, the Shapley value approach is considered to be a fair strategy to divide gains, and the use of such a solution brought attention in the explainable ML research community [18]–[20], [46], [47]. As mentioned in Section I, the main idea is to use it as a feature attribution method. Given a goal that one would like to explain (accuracy, local prediction, fairness measures, etc.), the Shapley values will indicate the contribution of each feature towards this goal. For this purpose, there are some key aspects that must be considered carefully when bringing Eq. (1) to the field of ML:

1) Firstly, one should define $\upsilon(\cdot)$ according to what one would like to explain. For instance, if one aims at analysing the contribution of each feature towards the model's overall accuracy, one has to define $\upsilon(\cdot)$ as the accuracy of the trained model based on the TPs and FPs in the test data.
2) One should be aware of what $\upsilon(\emptyset)$ represents. For example, in the shipment orders example, it is clear that there will be no savings when there is no coalition, so $\upsilon(\emptyset)$ should be zero. However, in a ML scenario, the payoff of the empty set can be a non-zero value, and therefore, calculating $\upsilon(\emptyset)$ might be more complicated in those cases.
3) Finally, another important aspect is the computational complexity of the Shapley values, as it involves retraining the model for all possible coalitions A, and calculating $\upsilon(A)$. This may pose a computational constraint when dealing with a high-dimensional data, as the number of payoffs exponentially increases with the number of features. In order to reduce this effort, one may consider approximation strategies that estimate the Shapley values with less computations [48]–[50].

In the next sections, we explain how to define the aforementioned aspects and how to use the Shapley values to explain the robustness of ML classifiers.

III. Explaining the robustness through Shapley values

As discussed earlier, explaining the predicted outcome is an important area of research but it is also important to explain the robustness in predicting these values. Although there can be different ways to measure robustness, the use of ROC and PR curves are preferred as they span a range of threshold values to classify the predicted outcomes. This makes them independent of threshold values unlike the other measures like accuracy and F-score. Therefore, the ROC and PR curves are also considered preferred approaches for explaining the robustness of ML models. The proposed process for explaining these curves is discussed below.

A. *ShapAUC: Explaining the area under the ROC curve*

Assume a ROC curve obtained after training a ML model. The proposed ShapAUC method provides the contribution of each feature towards the area under this ROC curve. We can safely assume that the random classifier is the baseline for the AUC, which can be achieved even when no feature contributes towards the classifier training. In the case of a random classifier, TPRs are obviously equal to FPRs and therefore, by definition, the AUC will be 0.50. If one includes features in training, the difference between the obtained AUC and the random classifier is then explained by the contribution of such features. For example, if one achieves an AUC of 0.95, the features are contributing to improve the AUC from 0.50 (which could be obtained by a random classifier) to the actual 0.95. In other words, the marginal contribution of all features should sum up to 0.45. Based on this reasoning, we define the payoffs of ShapAUC as follows:

$$\upsilon^{AUC}(A) = AUC_A - 0.50, \tag{5}$$

where AUC_A represents the area under the ROC curve when only the features in A are used in the training step. Note that, if $A = \emptyset$ then $AUC_\emptyset = 0.50$ (the random classifier), and therefore $\upsilon^{AUC}(\emptyset) = AUC_\emptyset - 0.50 = 0$.

Moreover, according to the efficiency property (see Eq. (2)), $\sum_{i=1}^{n} \phi_i = \upsilon^{AUC}(N) - \upsilon^{AUC}(\emptyset) = AUC_N - 0.50$, that is, the sum of the marginal contributions of all features is equal to the difference between the AUC of the grand coalition and the AUC of the random classifier.

After retraining the ML model and calculating the payoffs for all subset of features, we interpret robustness based on the Shapley values calculated by

$$\phi_i^{AUC} = \sum_{A \subseteq N \backslash \{i\}} \frac{(n - |A| - 1)! \, |A|!}{n!} \Delta_i \upsilon^{AUC}(A), \tag{6}$$

where $\Delta_i \upsilon^{AUC}(A) = \upsilon^{AUC}(A \cup \{i\}) - \upsilon^{AUC}(A)$.

The process of ShapAUC is summarised in Figure 2. The procedure involves choosing a subset of features and then calculating the ROC curve for this subset, along with the AUC value. This process is then repeated for all possible subsets of features available in the ML dataset. The contribution of each feature is then estimated with the help of Eq. 6. These contributions can be visualised in a waterfall plot, as shown on the bottom right of Figure 2. The contribution of each feature is provided in a decreasing order (given the absolute values) that cooperates to increase the AUC from the random classifier to the actual value.

B. *ShapROC: Explaining the ROC curve*

In the previous subsection, we propose to explain the area under the ROC curve. It is also possible to explain the curve itself by explaining each point on the curve. The ROC curve is essentially TPR values plotted against the FPR values, so it is possible to formulate each point on the TPR curve as a coalition game. The purpose of ShapROC is to use the Shapley value as a feature attribution method to explain the TPR values in the ROC curve. We explain the process of obtaining ShapROC into three steps, as explained below.

1) Defining the payoffs: Recall that in the case of a random classifier, the TPRs are equal to FPRs. By assuming that the random classifier is the baseline for the TPRs, the difference between a TPR and the associated FPR can be explained by the cooperation of features when they join in a coalition. For example, considering a TPR of 0.85 for a FPR of 0.25, the

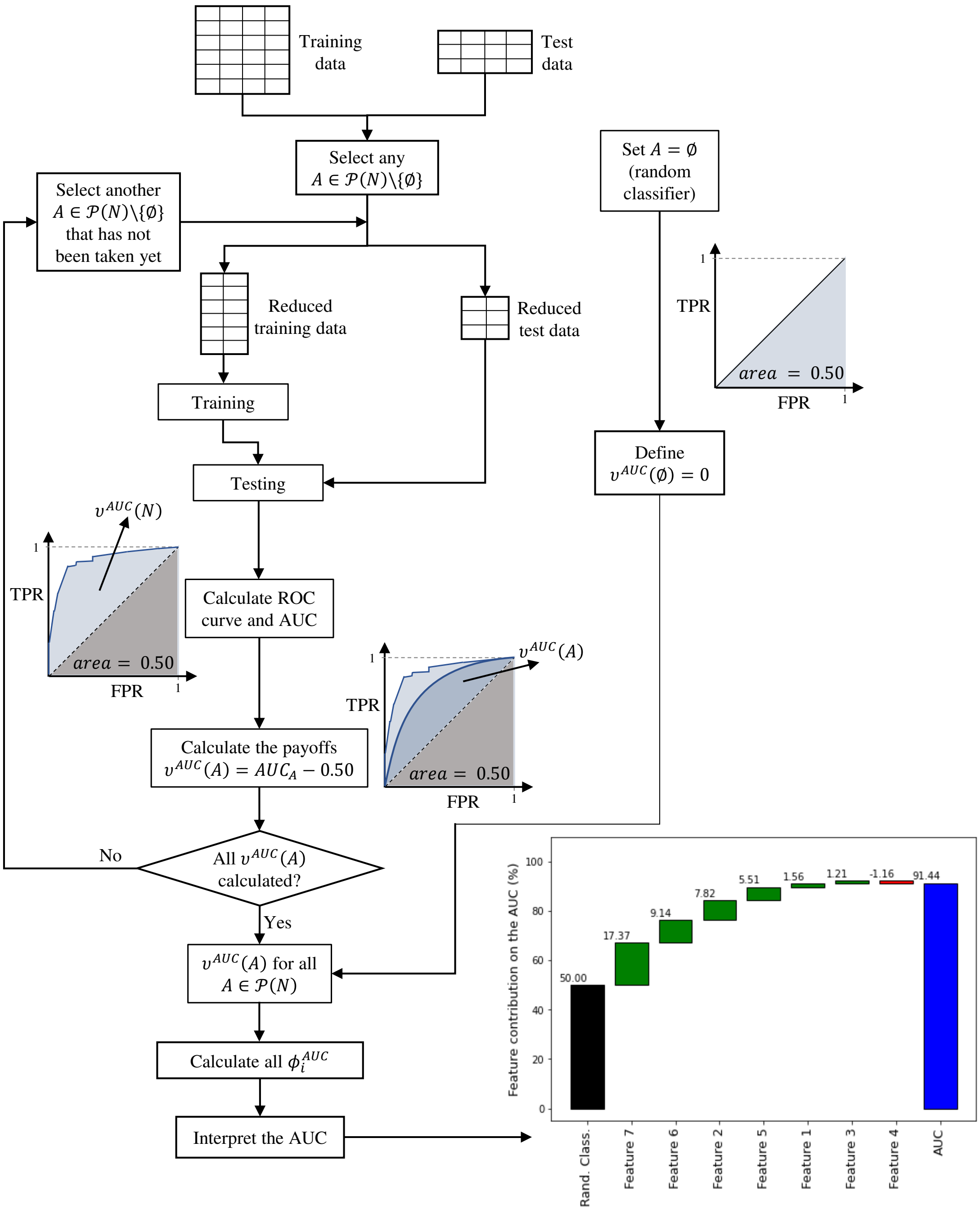

Training data
Test data
Select any $A \in \mathcal{P}(N)\backslash\{\emptyset\}$
Select another $A \in \mathcal{P}(N)\backslash\{\emptyset\}$ that has not been taken yet
Set $A = \emptyset$ (random classifier)
Reduced training data
Reduced test data
TPR
FPR
1
1
$area = 0.50$
Training
Testing
Define $v^{AUC}(\emptyset) = 0$
$v^{AUC}(N)$
Calculate ROC curve and AUC
$v^{AUC}(A)$
Calculate the payoffs $v^{AUC}(A) = AUC_A - 0.50$
No
All $v^{AUC}(A)$ calculated?
Yes
$v^{AUC}(A)$ for all $A \in \mathcal{P}(N)$
Calculate all ϕ_i^{AUC}
Interpret the AUC
Feature contribution on the AUC (%)
0
20
40
60
80
100
50.00
17.37
9.14
7.82
5.51
1.56
1.21
-1.16
91.44
Rand. Class.
Feature 7
Feature 6
Feature 2
Feature 5
Feature 1
Feature 3
Feature 4
AUC

Fig. 2. An overview of the proposed ShapAUC method to evaluate the contribution of each feature towards AUC.

contribution of features is $0.85 - 0.25 = 0.60$. This idea leads to the following definition of payoff:

$$v_{fpr}^{ROC}(A) = tpr_{fpr,A} - fpr, \tag{7}$$

where $0 \leq fpr \leq 1$ and $tpr_{fpr,A}$ is the TPR associated to a given fpr and a coalition A of features.

Note that, for a random classifier, $A = \emptyset$ implies that $tpr_{fpr,\emptyset} = fpr$ and, therefore, $v_{fpr}(\emptyset) = tpr_{fpr,\emptyset} - fpr = fpr - fpr = 0$. In addition, based on the efficiency property, $\sum_{i=1}^{n} \phi_i = v_{fpr}^{ROC}(N) - v_{fpr}^{ROC}(\emptyset) = tpr_{fpr,N} - fpr$. Therefore, by summing up the marginal contribution of each feature, we can explain the net increase in value from fpr to the obtained $tpr_{fpr,N}$.

In Eq. (7), the payoffs $v_{fpr}^{ROC}(A)$ for different coalitions A depend on the fpr values. However, as these values are recalculated for different classification thresholds, different coalitions might end up in generating totally different sets of FPR values, and therefore, making these coalitions incomparable to each other. To address this issue, we have to introduce an additional step for estimating TPR values in the ShapAUC method. We discuss it in the next section.

2) *Estimating the TPR values:*

We estimate the TPRs based on the standard ROC curve (calculated by using all features together). Consider that $(f_k^{ROC}, t_k^{ROC})_{k=1,\cdots,l}$ represents the set of FPR and TPR values used to build the standard ROC curve. Moreover, assume that we intend to explain a specific $tpr'_{fpr',N}$ for a fixed fpr' (e.g. $tpr'_{0.25,N} = 0.85$ for a fixed $fpr' = 0.25$). For each A, as a first step, we find the nearest available FPR values on either side i.e. $f_a^{ROC}, f_b^{ROC} \in \left\{f_1^{ROC}, \ldots, f_l^{ROC}\right\}$ from fpr' such that $f_a^{ROC} \leq fpr' \leq f_b^{ROC}$.

We consider three strategies to estimate the TPR values under analysis, namely optimistic, pessimistic and interpolation strategies. The three strategies are defined below:

- Optimistic strategy: $tpr_{fpr',A} = \max\left(t_a^{ROC}, t_b^{ROC}\right)$.
- Pessimistic strategy: $tpr_{fpr',A} = \min\left(t_a^{ROC}, t_b^{ROC}\right)$.
- Interpolation strategy: $tpr_{fpr',A} = (t_b^{ROC} - t_a^{ROC})(fpr' - f_a^{ROC}) / \left(f_b^{ROC} - f_a^{ROC}\right) + t_a^{ROC}$.

In case of interpolation, if $f_a^{ROC} = f_b^{ROC}$, then we can simply take an average of two values as an estimate: $tpr_{fpr',A} = \left(t_a^{ROC} + t_b^{ROC}\right)/2$.

The use of optimistic strategy will provide higher values of TPRs to calculate payoffs, and therefore, it can be considered as an upper approximation of the contributions. Similarly, the pessimistic strategy will provide the lower approximation of the contributions. The interpolation strategy, on the other hand, might be considered more balanced in a way that it tends to provide a value between the upper and lower approximations. Here, the aim is not to compare these strategies or finding the most appropriate strategy, rather the aim of introducing these strategies is to demonstrate the possibility of estimating the TPR curves in order to compare results from different coalitions.

3) Calculating the Shapley values and visualising the features contribution: After estimating the TPR values, the Shapley values for each slice of the curve can be calculated as follows:

$$\phi_i^{ROC,fpr'} = \sum_{A \subseteq N \setminus \{i\}} \frac{(n - |A| - 1)!\,|A|!}{n!} \Delta_i \upsilon_{fpr'}^{ROC}(A), \quad (8)$$

where $\Delta_i \upsilon_{fpr'}^{ROC}(A) = \upsilon_{fpr'}^{ROC}(A \cup \{i\}) - \upsilon_{fpr'}^{ROC}(A)$.

We can repeat the Shapley value calculations for each FPR in the ROC plot, and therefore, generating a set of curves representing contribution of each feature throughout the curve. This idea can be quite useful for ML analysts to assess the impact of each feature for different FPR values.

Figure 3 illustrates the proposed ShapROC method and the features contribution visualisation towards the TPRs. In the waterfall plot, we can see how features contribute to increase the TPR from the random classifier (e.g. TPR when $fpr = 0.20$) to the actual value (e.g. $t\bar{p}r_{fpr} = 0.91$). In the figure at the bottom right, the contributions of each feature are visible for the whole range of FPR values (varying from 0 to 1).

C. Relation between ShapAUC and ShapROC

The AUC can be approximated by the Riemann integral, that is, by the sum of very small rectangular areas calculated from the TPR and FPR values in the ROC curve. Consider a set of TPR and FPR values $(f_k^{ROC}, t_k^{ROC})_{k=1,\cdots,l}$ such that $0 = f_0^{ROC} < \ldots < f_{k-1}^{ROC} < f_k^{ROC} < \ldots < f_l^{ROC} = 1$ and the difference between any f_k^{ROC} and f_{k-1}^{ROC} is very small. Based on the proposed ShapROC and in the efficiency property, we have that

$$\begin{aligned}
AUC &\approx \sum_{k=2}^{l} \frac{\left(t_k^{ROC} + t_{k-1}^{ROC}\right)\left(f_k^{ROC} - f_{k-1}^{ROC}\right)}{2} \\
&= \sum_{k=2}^{l} \frac{\left(t_k^{ROC} + t_{k-1}^{ROC}\right)\left(f_k^{ROC} - f_{k-1}^{ROC}\right)}{2} - \left(\sum_{k=2}^{l} \frac{\left(f_k^{ROC} + f_{k-1}^{ROC}\right)\left(f_k^{ROC} - f_{k-1}^{ROC}\right)}{2} - 0.50\right) \\
&= 0.50 + \sum_{k=2}^{l} \frac{\left(t_k^{ROC} - f_k^{ROC} + t_{k-1}^{ROC} - f_{k-1}^{ROC}\right)\left(f_k^{ROC} - f_{k-1}^{ROC}\right)}{2} \\
&= 0.50 + \sum_{k=2}^{l} \frac{\left(\left(\sum_{i=1}^{n} \phi_i^{ROC,f_k^{ROC}}\right) + \left(\sum_{i=1}^{n} \phi_i^{ROC,f_{k-1}^{ROC}}\right)\right)\left(f_k^{ROC} - f_{k-1}^{ROC}\right)}{2} \\
&= 0.50 + \sum_{k=2}^{l} \frac{\left(\left(\sum_{i=1}^{n} \phi_i^{ROC,f_k^{ROC}}\left(f_k^{ROC} - f_{k-1}^{ROC}\right)\right) + \left(\sum_{i=1}^{n} \phi_i^{ROC,f_{k-1}^{ROC}}\left(f_k^{ROC} - f_{k-1}^{ROC}\right)\right)\right)}{2} \quad (9) \\
&= 0.50 + \sum_{i=1}^{n} \frac{\left(\left(\sum_{k=2}^{l} \phi_i^{ROC,f_k^{ROC}}\left(f_k^{ROC} - f_{k-1}^{ROC}\right)\right) + \left(\sum_{k=2}^{l} \phi_i^{ROC,f_{k-1}^{ROC}}\left(f_k^{ROC} - f_{k-1}^{ROC}\right)\right)\right)}{2} \\
&= 0.50 + \sum_{i=1}^{n} \frac{\left(\sum_{k=2}^{l}\left(\phi_i^{ROC,f_k^{ROC}}\left(f_k^{ROC} - f_{k-1}^{ROC}\right) + \phi_i^{ROC,f_{k-1}^{ROC}}\left(f_k^{ROC} - f_{k-1}^{ROC}\right)\right)\right)}{2} \\
&= 0.50 + \sum_{i=1}^{n}\sum_{k=2}^{l} \frac{\left(\phi_i^{ROC,f_k^{ROC}} - \phi_i^{ROC,f_{k-1}^{ROC}}\right)\left(f_k^{ROC} - f_{k-1}^{ROC}\right)}{2}.
\end{aligned}$$

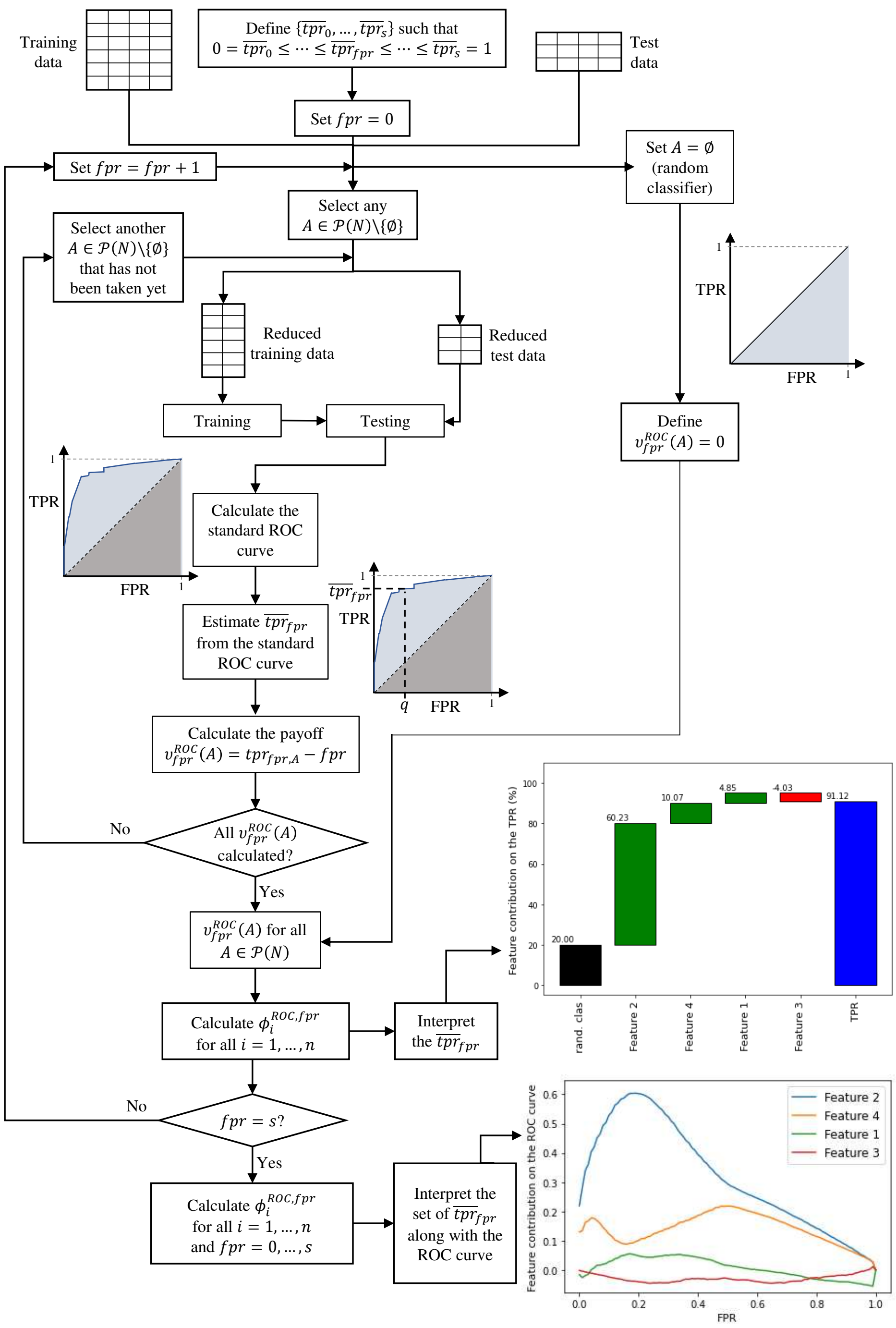

Training data
Define $\{\overline{tpr}_0, \ldots, \overline{tpr}_s\}$ such that $0 = \overline{tpr}_0 \leq \cdots \leq \overline{tpr}_{fpr} \leq \cdots \leq \overline{tpr}_s = 1$
Test data
Set $fpr = 0$
Set $fpr = fpr + 1$
Set $A = \emptyset$ (random classifier)
Select any $A \in \mathcal{P}(N)\backslash\{\emptyset\}$
Select another $A \in \mathcal{P}(N)\backslash\{\emptyset\}$ that has not been taken yet
1
TPR
FPR
1
Reduced training data
Reduced test data
Training
Testing
Define $v_{fpr}^{ROC}(A) = 0$
1
TPR
FPR
1
Calculate the standard ROC curve
1
$\overline{tpr}_{fpr}$
TPR
Estimate $\overline{tpr}_{fpr}$ from the standard ROC curve
q
FPR
1
Calculate the payoff $v_{fpr}^{ROC}(A) = tpr_{fpr,A} - fpr$
No
All $v_{fpr}^{ROC}(A)$ calculated?
Yes
$v_{fpr}^{ROC}(A)$ for all $A \in \mathcal{P}(N)$
Calculate $\phi_i^{ROC,fpr}$ for all $i = 1, \ldots, n$
Interpret the $\overline{tpr}_{fpr}$
No
$fpr = s$?
Yes
Calculate $\phi_i^{ROC,fpr}$ for all $i = 1, \ldots, n$ and $fpr = 0, \ldots, s$
Interpret the set of $\overline{tpr}_{fpr}$ along with the ROC curve
Feature contribution on the TPR (%)
100
80
60
40
20
0
20.00
60.23
10.07
4.85
-4.03
91.12
rand. clas
Feature 2
Feature 4
Feature 1
Feature 3
TPR
Feature contribution on the ROC curve
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.0
0.2
0.4
0.6
0.8
1.0
FPR
Feature 2
Feature 4
Feature 1
Feature 3

Fig. 3. An overview of the proposed ShapROC method to evaluate the contribution of each feature towards TPR values in the ROC curve.

As we proposed in Section III-A, the achieved AUC can be represented as the sum of the random classifier (50%) and the marginal contributions of features. Mathematically, we have that

$$AUC = 0.50 + \sum_{i=1}^{n} \phi_i^{AUC}. \tag{10}$$

In other words, we can say that we may decompose the AUC by the random classifier and the Shapley values ϕ_i^{AUC}, $i = 1, \ldots, n$. Therefore, each ϕ_i^{AUC} represents a "piece of area" from the AUC. By making a parallel between Eq. (9) and Eq. (10), the relation between the proposed ShapAUC and ShapROC approaches is given by the following equation:

$$\phi_i^{AUC} \approx \sum_{k=2}^{l} \frac{\left(\phi_i^{ROC,f_k^{ROC}} - \phi_i^{ROC,f_{k-1}^{ROC}}\right)\left(f_k^{ROC} - f_{k-1}^{ROC}\right)}{2}. \tag{11}$$

Note that the approximation in Eq. (11) is also a sum of small areas. Indeed, as can be visualised in Figure 3 when interpreting the TPR values along with the ROC curve, the area under the contribution of each feature i is an approximation for its contribution in the AUC. If we sum all these areas, we achieve an approximation for the AUC.

D. ShapPRC and ShapAUPRC: Explaining the Precision-recall curve and the area under this curve

As highlighted in Section II-B, the use of Precision-Recall Curve is considered more appropriate than the use of ROC and AUC in scenarios with highly imbalanced datasets. We can also extended the same idea to use Shapley values for explaining the PRC and the AUPRC, termed as ShapPRC and ShapAUPRC, respectively. The proposed ShapAUPRC provides an explanation for the area under the PRC. When assuming a random classifier, the obtained Precision is equal to 0.50 regardless of the Recall values. Therefore, similarly as in the ShapAUC, we have the baseline area of 0.50. We then explain the contributions of features that can potentially improve the AUPRC (from the random classifier). In this case, the payoffs can be defined as follows:

$$v^{AUPRC}(A) = AUPRC_A - 0.50, \tag{12}$$

where $AUPRC_A$ represents the area under the PRC when only the features in A are used in the training step.

The Shapley values for AUPRC can also be estimated using the steps defined for the ShapAUC. The equation for calculating the Shapley values for AUPRC can be defined as below:

$$\phi_i^{AUPRC} = \sum_{A \subseteq N \backslash \{i\}} \frac{(n - |A| - 1)!\,|A|!}{n!} \Delta_i v^{AUPRC}(A), \tag{13}$$

where $\Delta_i v^{AUPRC}(A) = v^{AUPRC}(A \cup \{i\}) - v^{AUPRC}(A)$.

In ShapPRC, we propose to explain the contributions of features towards the Precision values along with the PRC. If one takes a single slice in the PRC, we can also use the ShapPRC to explain the improvement in the Precision value (from the random classifier). The steps are as defined for the ShapROC, with a redefinition of the baseline, payoffs and Shapley values calculation. For all Precision values in the PRC, the baseline remains to be 0.50 regardless of the Recall values. This leads us to the following definition of payoffs:

$$v_{rec}^{PRC}(A) = pre_{rec,A} - 0.50, \tag{14}$$

where $0 \leq rec \leq 1$ and $pre_{rec,A}$ is the Precision value associated with a given Recall value (rec) and a coalition A of features.

When analysing the Precision values along with the PRC, we provide an explanation for each slice in it. So the equation for calculating the Shapley values for each Precision value $pre'_{rec',A}$ (associated with a Recall value rec') can be defined as:

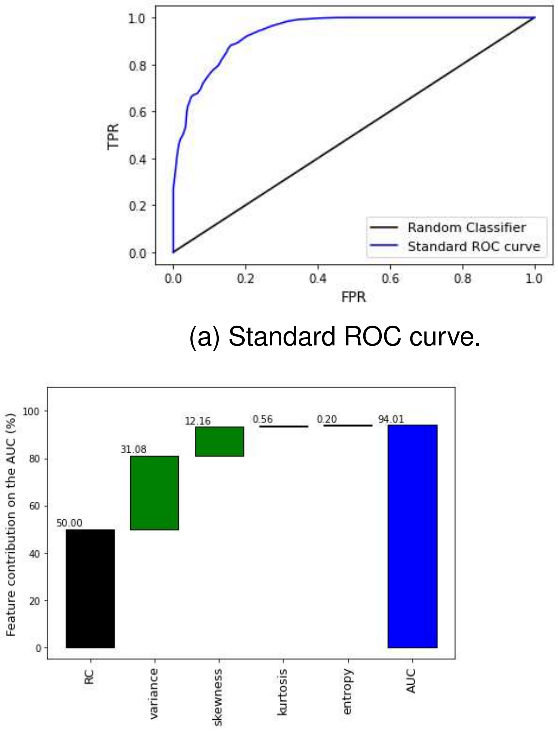
$$\phi_i^{PRC,rec'} = \sum_{A \subseteq N \backslash \{i\}} \frac{(n - |A| - 1)!\,|A|!}{n!} \Delta_i v_{rec'}^{PRC}(A), \tag{15}$$

where $\Delta_i v_{rec'}^{PRC}(A) = v_{rec'}^{PRC}(A \cup \{i\}) - v_{rec'}^{PRC}(A)$.

IV. Illustrative Example for Explaining Robustness

We illustrate the proposed approaches[2] based on a real dataset called Banknote Authentication Dataset [51]. This dataset consists of 1372 images that were used to evaluate an authentication procedure for genuine (and forged) banknotes. Wavelet Transforms were applied to these images in order to extract the following (continuous) attributes: *variance*, *skewness*, *kurtosis* and *entropy*.

The ML model was trained using Gaussian Naive Bayes (implementation from scikit-learn), however, the proposed approach can be used with any other ML model. The dataset was split with 80% for training and 20% for testing purpose. The ROC curve obtained is shown in Figure 4a, with AUC value of 94.03%. By applying the ShapAUC approach, we can interpret the contribution of each feature towards this AUC. Figure 4b shows the obtained results (*RC* represents the random classifier). We can see that the highest contribution is assigned to the *variance* (31.08%), while both *kurtosis* and *entropy* have very low contribution to the AUC (0.56% and 0.20%, respectively).



(a) Standard ROC curve.

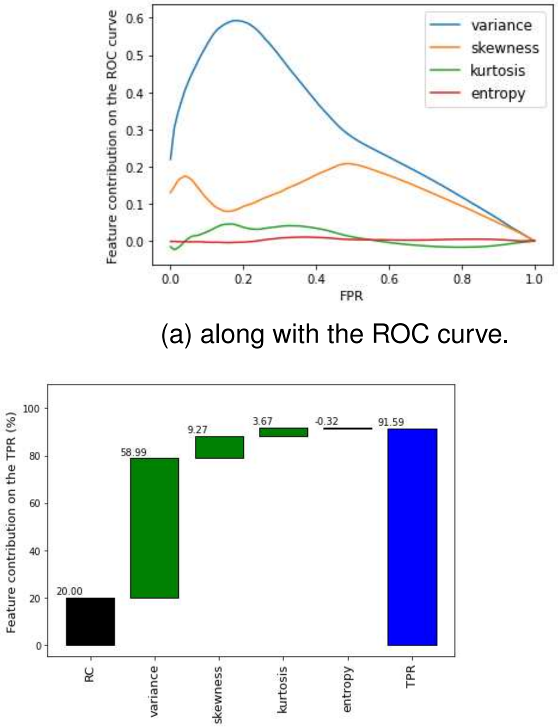
(b) Contributions towards the AUC.

Fig. 4. ROC and AUC for the banknotes dataset.

The results from the ShapROC approach are presented in Figure 5. These results were generated with the interpolation strategy, however, it can be repeated for other strategies as well (see Section IV-A). As seen in Figure 5a, *variance* remains to be the attribute with most contribution towards the TPR values throughout the ROC curve. Both *kurtosis* and *entropy* have practically negligible contribution in the whole range of the FPR values.

[2] All codes can be accessed in https://github.com/shaprob/shaproc.

Figure 5b shows the contributions for a FPR of 20%, i.e. when moving from a TPR of 20% (random classifier) to the actual 91.59%. This waterfall plot can be considered as a slice view of Figure 5a for FPR = 0.2. In this figure, an interesting observations is that the *entropy* feature negatively contributes to the TPR value. In other words, instead of improving the performance, *entropy* is deteriorating the performance of the classifier at this point.



(a) along with the ROC curve.

(b) for a single slice (FPR of 20%).

Fig. 5. Feature contributions towards TPR values.

A. On the use of different strategies for the TPR values estimation

We mentioned in Section III-B that different TPR values estimation strategies could be used in the proposed ShapROC approach. In this section, we discuss the different results that can be achieved by adopting the optimistic, pessimistic or interpolation strategies. Figure 6 presents the estimated ROC curve for these three strategies. As all curves are very similar, we show a magnified version of a selected piece of the ROC curves (to better visualise the differences). It can be seen that the optimistic strategy generates relatively higher values for TPRs and the pessimistic strategy has generated lower values. The interpolation strategy produces values between the other two strategies.

A comparison between the considered strategies is presented in Figure 7. Although there are slight differences among the obtained results, in all cases, both *variance* and *skewness* are the two features with highest contributions towards the AUC and the TPR values (see Figures 7a, 7b and 7c).

One may also note in Figures 7d, 7e and 7f that the shapes of these features do not change a lot with the strategy, although the shapes for *kurtosis* and *entropy* features change slightly. An interesting observation is that *entropy* has a positive overall contribution towards AUC for the optimistic strategy while it has a negative contribution when using the pessimistic strategy. However, as both *kurtosis* and *entropy* have a very low contribution towards robustness, these differences can be considered negligible in terms of explainability.

As mentioned earlier, the aim here is not to propose or evaluate the best strategy, although this can be an area of future work. Without loss of generality, we will consider the interpolation strategy for the onward discussion in this paper.

B. Visualising uncertainties in assessing robustness

A ML model should not be sensitive to a certain subset of the dataset selected for training (and/or testing), and therefore, it is a common practise to create multiple versions of training and testing datasets to assess the performance of ML models. These multiple versions can be created from original dataset by creating different combinations of subsets used for training and testing. The two most common approaches for performing these experiments are K-Fold Cross Validation [52] and Monte-Carlo Cross Validation [53]. Regardless of which approach we take, multiple experiments are involved that end up in multiple performance evaluations like AUC, ROC and PRC. For example, in Figure 8a, we show the ROC curves obtained for the banknotes dataset using Monte-Carlo Cross Validation (using 100 iterations with uniformly distributed sampling). The crisp line in the middle shows the curve generated from expected values, and the shaded area around the curve shows the standard deviation in the values of these curves.

For each iteration, Shapley values can be used to estimate the contribution of each feature towards the AUC. These contributions may vary in each iteration, and therefore, we propose the dispersion in these values as a way of measuring uncertainty. Similarly, Figure 8b explains the contribution of each feature towards the overall AUC where the bar height represents the expected contribution value while the whisker lines show standard deviation (i.e. uncertainty) in the contribution values.

Extending this idea to the whole ROC curve, it is also possible to calculate these contributions of each attribute towards achieving a TPR value (for each FPR value in the ROC curve). This is demonstrated in Figure 9a where the four attributes of banknotes dataset are plotted separately. The figure shows expected contribution values as a solid line in the middle, while the shaded values depict the standard deviation in these contribution values.

As a data analyst, one may need to investigate specific part of this plot, for example, focusing on the FPR value of 0.20, and investigating the contribution of each feature/attribute towards achieving the TPR value for FPR = 0.20. Figure 9b shows such an example which is essentially a sliced view of the ROC curve shown in Figure 8a. To summarise, both the measurement of robustness itself, as well as the uncertainty in measuring robustness can be explained with the help of Shapley values.

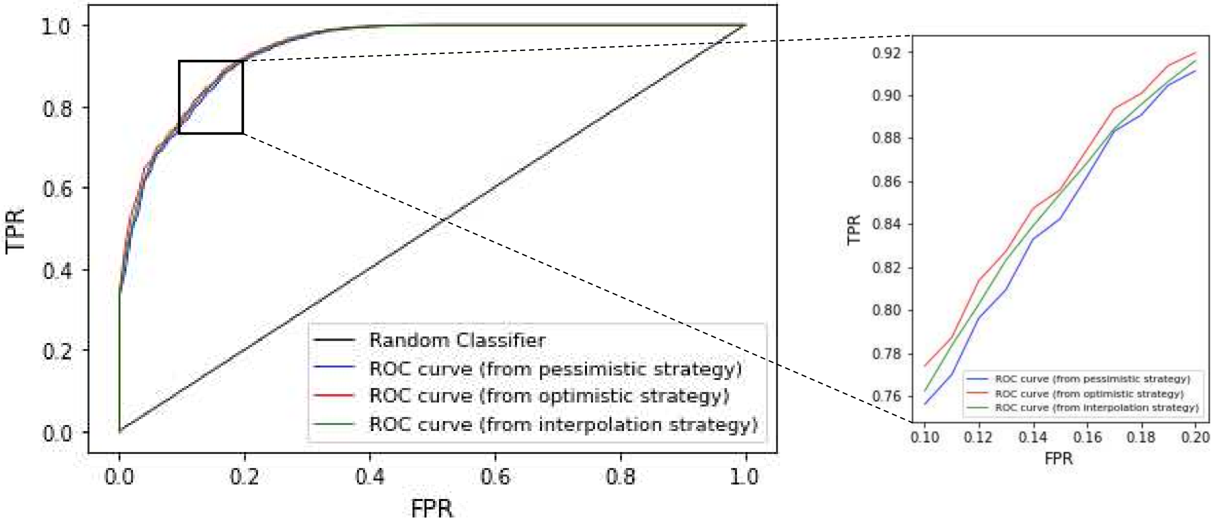


Fig. 6. Estimated ROC curve for different strategies.

C. Analysing imbalanced classification datasets

As mentioned earlier, the use of ROC and AUC has been debated for imbalanced classification problems and the use of Precision-Recall Curve (PRC) is considered more appropriate [54]. We demonstrate the use of ShapPRC with the same illustrative example of banknotes. For demonstration purpose, we sub-sampled the banknotes data to synthetically create an imbalance of 90%-10% (with 10% fake/forged notes). Figure 10a shows the PRC for this derived dataset. The contributions of each feature towards achieving the AUPRC value is shown in Figure 10b. The features of *kurtosis* and *entropy* can be seen to have negative contributions which implies that these features are decreasing the robustness (measured through PRC).

A more detailed picture can be seen in Figure 10c, where *variance* and *skewness* are contributing significantly higher than the other two. Please note that the aim of this experiment is not to compare PRC and ROC curves, as this is out of the scope of this paper. We demonstrate that both curves and their respective areas can be explained with the help of Shapley values.

D. Using robustness analysis for feature engineering

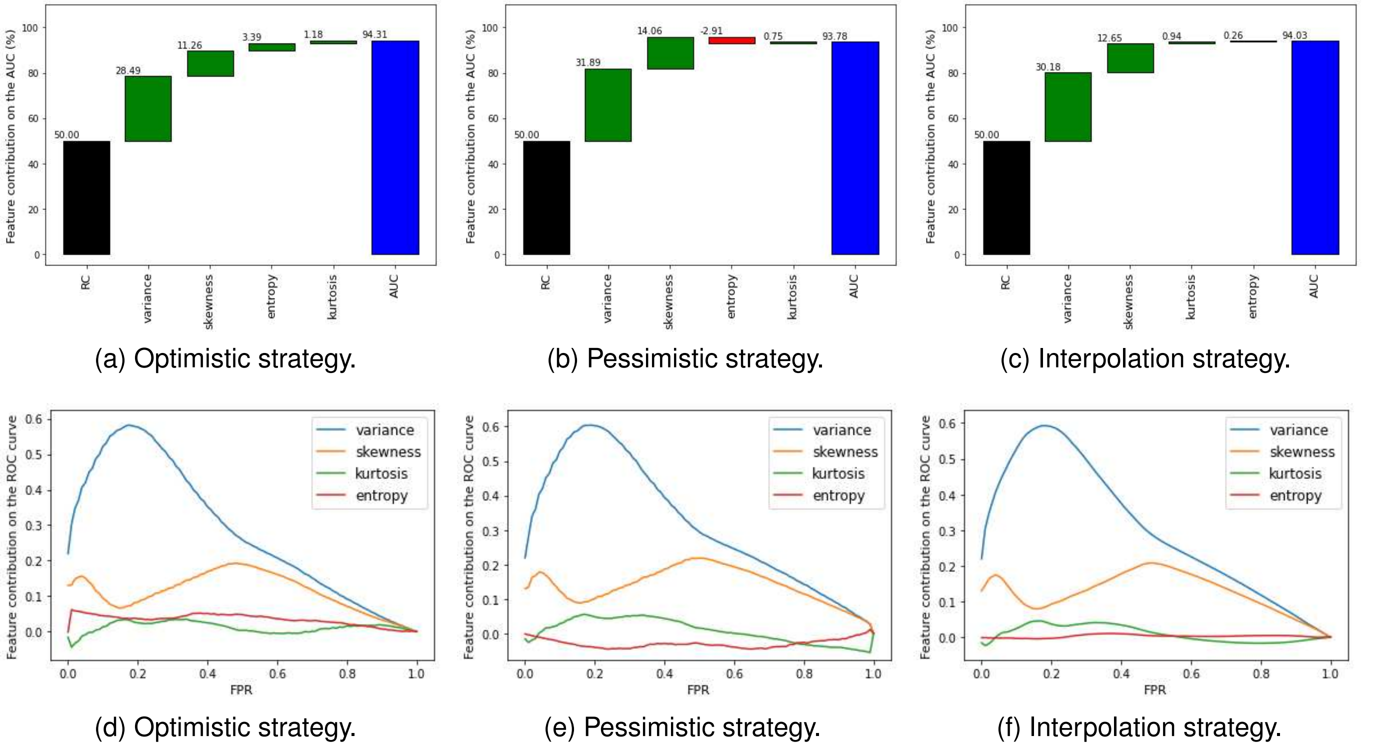
The gain or loss in the robustness by removing a specific feature is not necessarily equivalent to its marginal contribution. Instead, this is given by the difference between the payoffs with and without such a feature. However, as the Shapley value provides the marginal contribution when all possible coalitions are considered, it also serves as an indicative whether the presence (or absence) of a feature impacts the model robustness. Features whose contributions are insignificant could arguably be removed without affecting the robustness of the model.

Recall Figure 8b for the illustrative example, it showed that both *kurtosis* and *entropy* have practically no contributions towards AUC while *variance* has the highest contribution. By removing the *kurtosis* and *entropy* features, a slight improvement can be achieved on the model robustness. This is shown in Figure 11a where the AUC increases slightly from 94.03% to 95.19%. However, if a more useful feature, like *variance*, is removed from the dataset, the AUC may decrease significantly. This is demonstrated in Figure 11b for the illustrative example where the AUC decrease from 94.03% to 73.98%, which is a detrimental change in robustness that might end up in making the model practically useless.

Also, to investigate the impact of having duplicate feature, we create a novel feature which is a copy of *variance*. Figure 12 shows the contributions of this duplicate variable towards the AUC and the ROC curve. In Figure 12a, we see that both *variance* and the novel feature (represented by *dupl. variance*) contributes equally towards the AUC (16.86%). This is attested in Figure 12b where the contributions of the duplicate variance are identical to the original variance curve. The figure uses vertical and horizontal markers on these two curves to highlight their overlap. As the novel feature is a duplicate of *variance*, the payoffs $\upsilon\left(A \cup \{variance\}\right) = \upsilon\left(A \cup \{\textit{dupl. variance}\}\right)$ and, therefore, $\phi_{variance} = \phi_{\textit{dupl. variance}}$ (see the Symmetry property in Section II-D).

Note that the use of PRCs is preferred for assessing robustness in case of having imbalanced dataset. Recall Figure 10b where *kurtosis* and *entropy* both contributed negatively towards the AUPRC. Therefore, an obvious recommendation would be to remove these two features from the model. Figure 13b shows the results after removing these two features. Clearly, the performance of the classifier has improved from 74.67% to 86.52%. This is a significant improvement in a sense that the model has improved by more than 15%.

This demonstrates the usefulness of our proposed approach in selecting or removing features, which is a critical step in machine learning applications. Appendix-I further illustrates the use of our proposed approach on the datasets for Red Wine Quality [55], Rice [56], and Pima Indians Diabetes [57] as well.



(a) Optimistic strategy.

(b) Pessimistic strategy.

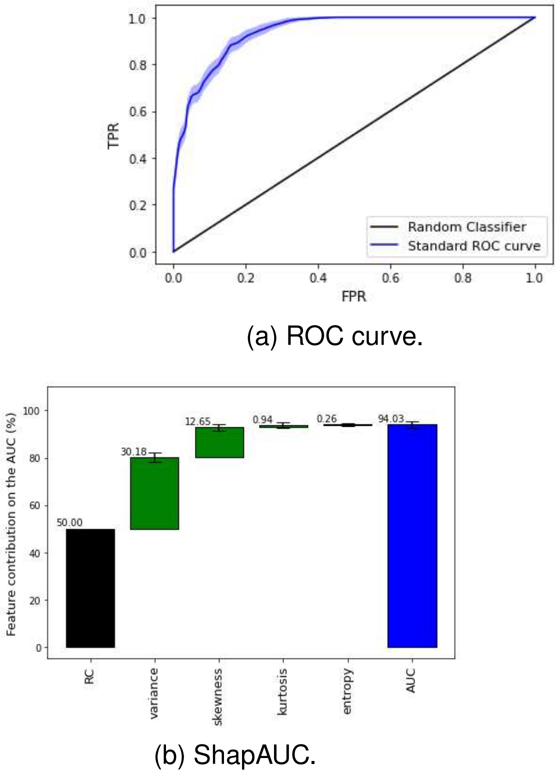
(c) Interpolation strategy.

(d) Optimistic strategy.

(e) Pessimistic strategy.

(f) Interpolation strategy.

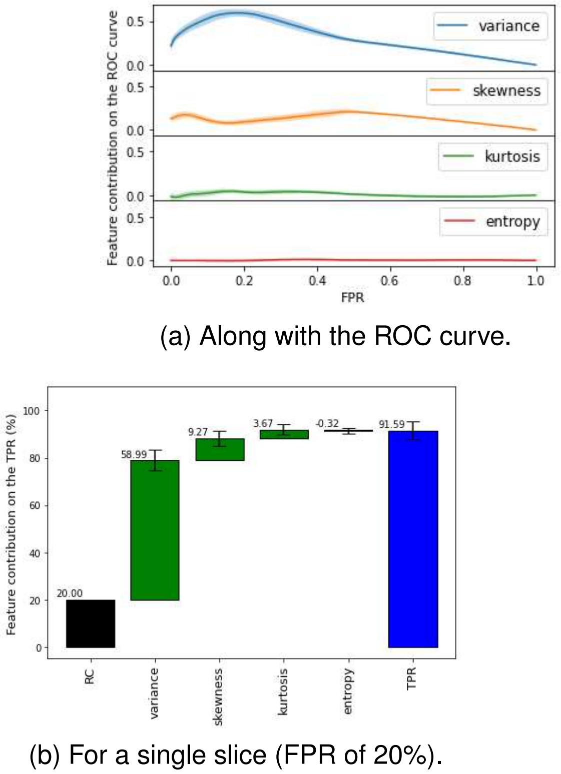
Fig. 7. Comparison between the optimistic, pessimistic and interpolation strategies. Plots in the left: contributions towards the AUC. Plots in the right: contributions towards TPR values along with the ROC curve.



(a) ROC curve.

(b) ShapAUC.

Fig. 8. Uncertainties in ROC curve and ShapAUC.



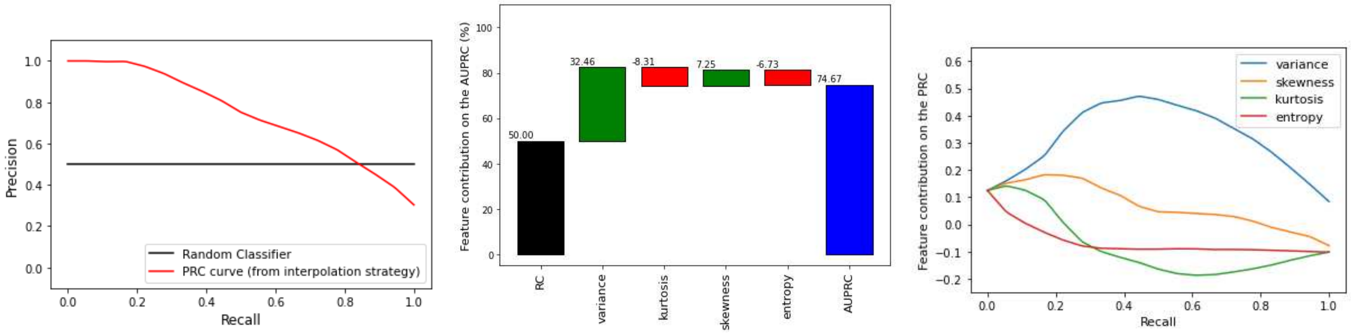
(a) Along with the ROC curve.

(b) For a single slice (FPR of 20%).

Fig. 9. Uncertainties in ShapROC.

V. Conclusions

We proposed techniques that can be used to explain the contribution of each feature towards the robustness of ML

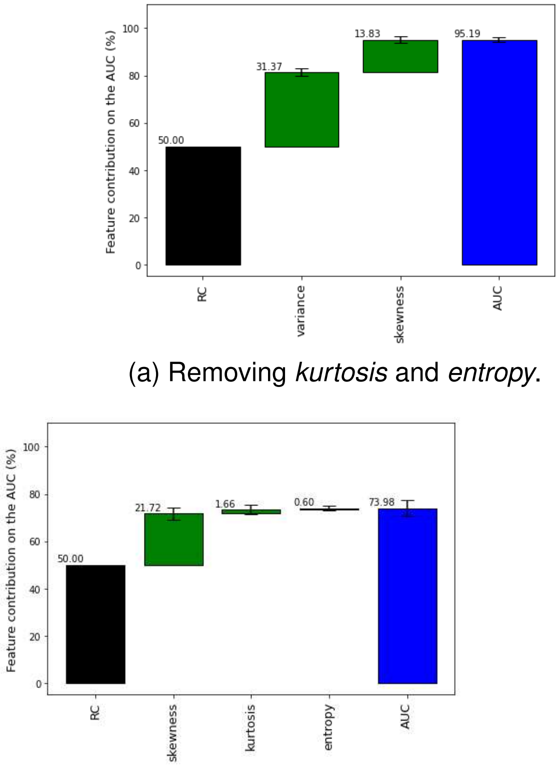


(a) PRC.

(b) ShapAUPRC.

(c) ShapPRC.

Fig. 10. Application of ShapPRC and ShapAUPRC in imbalanced dataset.



(a) Removing *kurtosis* and *entropy*.

(b) Removing *variance*.

Fig. 11. Application of ShapAUC as a feature selection method.

models. For explaining the area under the ROC curve, we propose to use 0.50 as the baseline value as any random classifier can achieve this value without help from any useful feature. Then, we propose to estimate the contribution of features towards adding robustness with the help of Shapley values. We also propose to explain each point at the ROC curve and therefore creating a decomposition of an overall curve into a set of individual curves (for each feature). As the use of PRC is considered more appropriate for imbalanced datasets, we also extended the idea to use Shapley values for explaining the PRC and the AUPRC. Explaining the robustness of classifiers can help analysts in auditing various features in their models and to revise their performance tuning parameters accordingly.

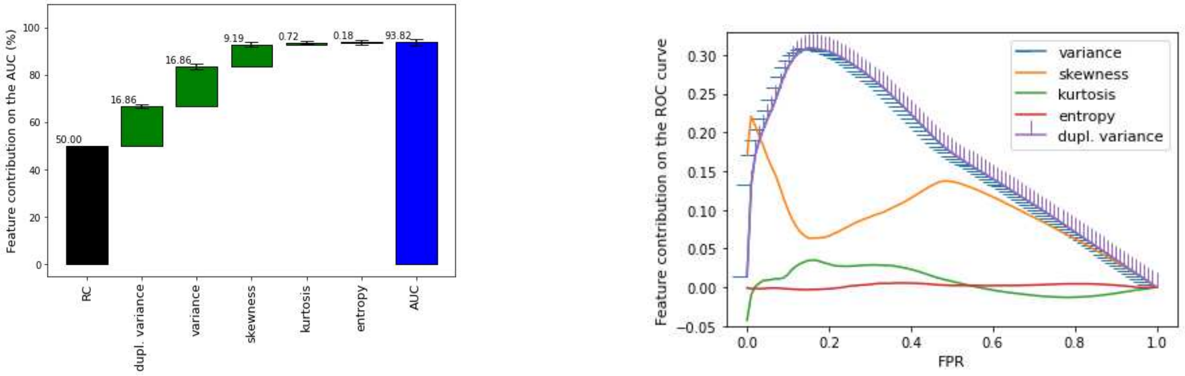
We demonstrate the use of our proposed approaches in feature selection. Based on the estimated Shapley values, it is possible to spot a feature that contributes negatively, and therefore, can be removed from the model. Also, this can help us identify features that should not be removed from the model due to their critical contributions towards robustness. In addition, it is also possible to identify a feature having insignificant contribution to the model's robustness, and therefore, removing such feature might help increasing the overall computational efficiency.

A. *Limitations and future work*

As mentioned in Section III-B, to explain a ROC curve, we generate multiple curves where each curve represents one of the coalitions among the players (i.e. features). Comparing these curves is not a straightforward task due to the fact that each curve has a different set of FPR/TPR values which does not necessarily align with other curves. For this, we proposed the three possible strategies for estimating these values: optimistic, pessimistic and interpolation strategies. However, one may argue that these strategies are sub-optimal, and better strategies are possible. Therefore, we consider this an area of further research.

The visualisation of ShapROC (as shown in Figure 5a) can assist data analysts in assessing the contribution of each feature across a range of FPR values. However, as we move on this curve from left to right, the contribution of random classifier dominates the contribution of features. This is visible in the figure where all individual curves are approaching zero, and therefore, the relative importance of these features cannot be inspected visually. An analyst might be interested in assessing these contributions of features in a relative sense, and therefore, a normalised version of this plot might be more useful in such case. We show an example of normalisation in Figure 14 which might be more useful when comparing the relative contributions of features towards achieving the TPR values. The same can be applied to ShapPRC plots as well. We consider this another area of future work that can help data analysts and researchers.

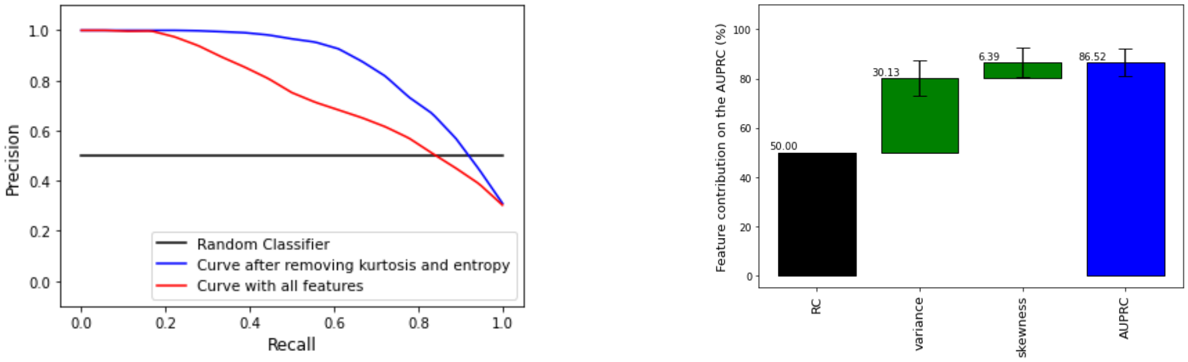
We demonstrated the use of Shapley values to explain the contribution of each feature towards the robustness of classifiers. However, this idea can be extended further to assess



(a) Contributions towards AUC.

(b) Contributions along with the ROC curve.

Fig. 12. Evaluating the inclusion of a duplicate *variance*.



(a) PRCs.

(b) AUPRCs.

Fig. 13. Application of ShapAUPRC as a feature selection method.

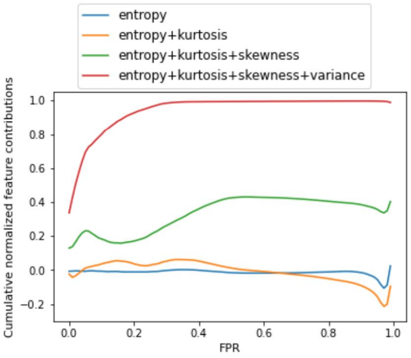


Fig. 14. Example of relative feature contributions visualisation.

the interaction among the features. In some cases, it might be important to estimate the contribution of some combinations of features instead of treating them as standalone/independent features. We consider this another important area of future work with practical implications.

Finally, it is possible to further investigate datasets from different application domains using the proposed approaches, which is considered to be another area of future research.

Appendix

The following three Figures demonstrate the contribution of each feature towards the AUC for the Red Wine Quality [55], Rice [56], and Pima Indians Diabetes [57] datasets. By removing features *pH* and *free sulfur dioxide* in the Wine dataset, the overall AUC have improved from 86.95% to 87.15%. For the Rice dataset, by removing feature *Extent*, the overall AUC practically remained the same (from 95.27% to 95.29%). In the Diabetes dataset, by removing feature *Glucose*, which has the highest contribution towards AUC, the robustness decreased from 79.82% to 72.59%.

References

[1] V. Kumar and M. L. Garg, "Predictive analytics: A review of trends and techniques," *International Journal of Computer Applications*, vol. 182, no. 1, pp. 31–37, 2018.

[2] T. P. Liang and Y. H. Liu, "Research landscape of business intelligence and big data analytics: A bibliometrics study," *Expert Systems with Applications*, vol. 111, pp. 2–10, 2018.

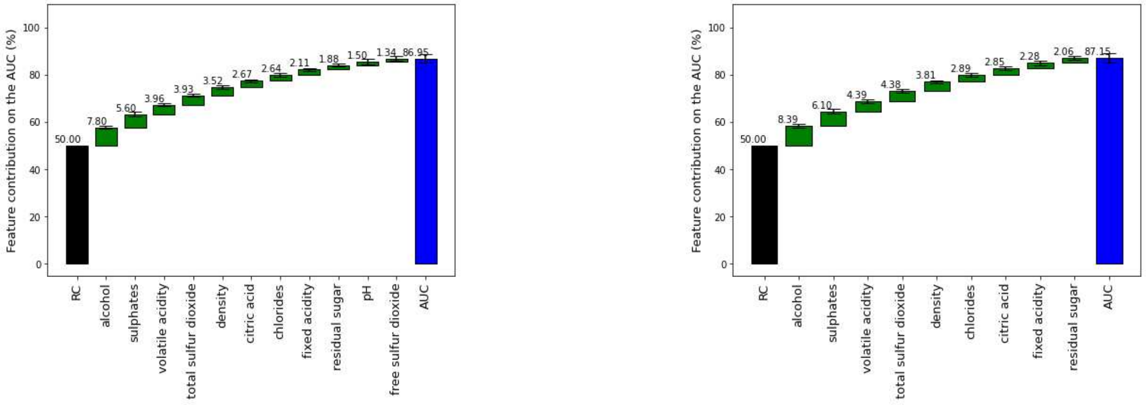
[3] X. Zhang, F. T. Chan, C. Yan, and I. Bose, "Towards risk-aware artificial intelligence and machine learning systems: An overview," *Decision Support Systems*, vol. 159, p. 113800, 2022.

[4] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM Computing Surveys*, vol. 51, no. 5, pp. 1–42, 2018.

[5] C. Molnar, *Interpretable machine learning*, 2021. [Online]. Available: https://christophm.github.io/interpretable-ml-book/

[6] N. Burkart and M. F. Huber, "A survey on the explainability of supervised machine learning," *Journal of Artificial Intelligence Research*, vol. 70, pp. 245–317, 2021.

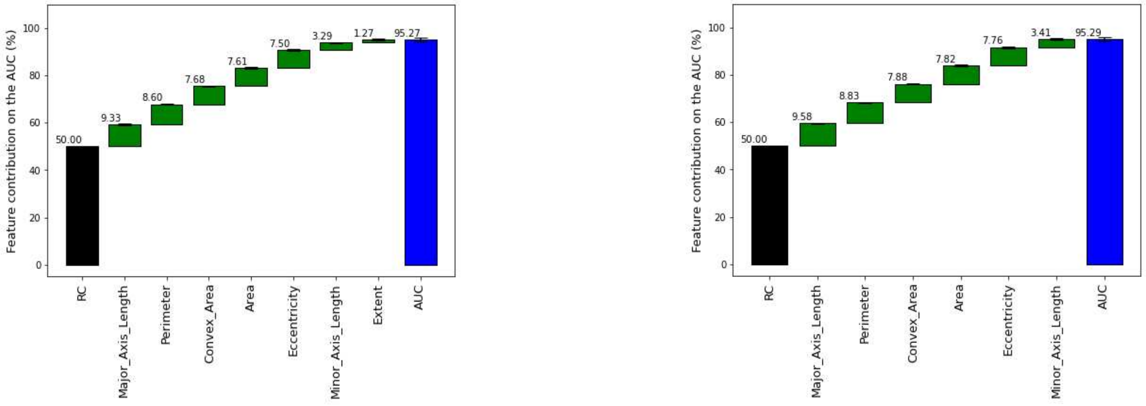
[7] E. M. Kenny, C. Ford, M. Quinn, and M. T. Keane, "Explaining black-box classifiers using post-hoc explanations-by-example: The effect of explanations and error-rates in XAI user studies," *Artificial Intelligence*, vol. 294, p. 103459, 2021.



(a) With all features.

(b) Removing *pH* and *free sulfur dioxide*.

Fig. 15. Application of ShapAUC as a feature selection method - Red Wine Quality dataset.



(a) With all features.

(b) Removing *Extent*.

Fig. 16. Application of ShapAUC as a feature selection method - Rice dataset.

[8] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[9] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[10] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT Press, 2016.

[11] T. Chen and T. He, "Xgboost - extreme gradient boosting," *R package version 0.4-2 1.4*, pp. 1–4, 2015.

[12] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[13] J. Kleinberg, S. Mullainathan, and M. Raghavan, "Inherent trade-offs in the fair determination of risk scores," in *8th Innovations in Theoretical Computer Science Conference (ITCS 2017). Leibniz International Proceedings in Informatics (LIPIcs)*, C. H. Papadimitriou, Ed., vol. 67. Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2017, pp. 43:1–23.

[14] M. Choraś, M. Pawlicki, D. Puchalski, and R. Kozik, "Machine learning - The results are not the only thing that matters! What about security, explainability and fairness?" in *International Conference on Computational Science (ICCS 2020). Lecture Notes in Computer Science*, vol. 12140. Springer, Cham, 2020, pp. 615–628.

[15] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial Intelligence*, vol. 267, pp. 1–38, 2019.

[16] M. Bücker, G. Szepannek, A. Gosiewska, and P. Biecek, "Transparency, auditability, and explainability of machine learning models in credit scoring," *Journal of the Operational Research Society*, vol. 73, no. 1, pp. 70–90, 2022.

[17] L. S. Shapley, "A value for n-person games," in *Annals of mathematics studies: Vol. 28. Contributions to the theory of games, Vol. II*, W. Kuhn and A. W. Tucker, Eds. Princeton: Princeton University Press, 1953, pp. 307–317.

[18] S. Lipovetsky and M. Conklin, "Analysis of regression in game theory approach," *Applied Stochastic Models in Business and Industry*, vol. 17, no. 4, pp. 319–330, 2001.

[19] T. Begley, T. Schwedes, C. Frye, and I. Feige, "Explainability for fair machine learning," *arXiv preprint:2010.07389*, 2020.

[20] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[21] R. Wang and K. Tang, "Feature selection for maximizing the area under the roc curve," in *2009 IEEE International Conference on Data Mining Workshops (ICDM)*. IEEE, 2009, pp. 400–405.

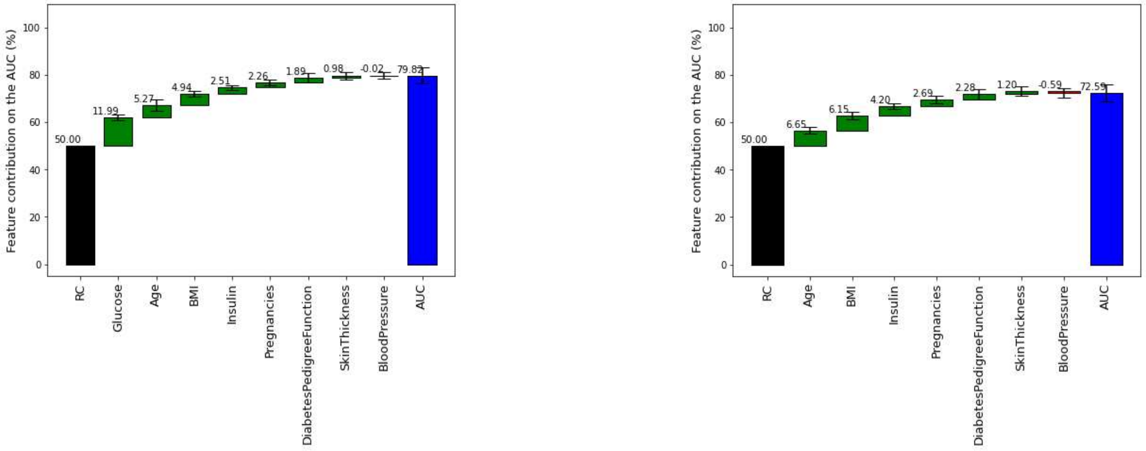
[22] A. J. Serrano, E. Soria, J. D. Martín, R. Magdalena, and J. Gómez, "Feature selection using roc curves on classification problems," in *The 2010 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2010, pp. 1–6.

[23] P. Xu, X. Liu, D. Hadley, S. Huang, J. Krischer, and C. Beam, "Feature selection using bootstrapped roc curves," *Journal of Proteomics & Bioinformatics*, vol. S9, pp. 1–10, 2014.

[24] A. P. Bradley, "The use of the area under the roc curve in the evaluation of machine learning algorithms," *Pattern recognition*, vol. 30, no. 7, pp. 1145–1159, 1997.

[25] C. W. Therrien, *Decision estimation and classification: an introduction to pattern recognition and related topics*. John Wiley & Sons, Inc., 1989.

[26] L. A. Jeni, J. F. Cohn, and F. De La Torre, "Facing imbalanced data: Recommendations for the use of performance metrics," in *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII 2013)*. IEEE, 2013, pp. 245–251.



(a) With all features.

(b) Removing *Glucose*.

Fig. 17. Application of ShapAUC as a feature selection method - Pima Indians Diabetes dataset.

[27] T. Saito and M. Rehmsmeier, “The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets,” *PLoS ONE*, vol. 10, no. 3, p. e0118432, 2015.

[28] J. Alqatawna, H. Faris, K. Jaradat, M. Al-Zewairi, and O. Adwan, “Improving knowledge based spam detection methods: The effect of malicious related features in imbalance data distribution,” *International Journal of Communications, Network and System Sciences*, vol. 8, pp. 118–129, 2015.

[29] N. Bussmann, P. Giudici, D. Marinelli, and J. Papenbrock, “Explainable ai in fintech risk management,” *Frontiers in Artificial Intelligence*, vol. 3, pp. 1–5, 2020.

[30] E. Dumitrescu, S. Hué, C. Hurlin, and S. Tokpavi, “Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects,” *European Journal of Operational Research*, vol. 297, no. 3, pp. 1178–1192, 2022.

[31] S. M. Lundberg, B. Nair, M. S. Vavilala, M. Horibe, M. J. Eisses, T. Adams, D. E. Liston, D. K.-W. Low, S.-F. Newman, J. Kim *et al.*, “Explainable machine-learning predictions for the prevention of hypoxaemia during surgery,” *Nature biomedical engineering*, vol. 2, no. 10, pp. 749–760, 2018.

[32] F. Weng, J. Zhu, C. Yang, W. Gao, and H. Zhang, “Analysis of financial pressure impacts on the health care industry with an explainable machine learning method: China versus the usa,” *Expert Systems with Applications*, vol. 210, p. 118482, 2022.

[33] B. Davazdahemami, H. M. Zolbanin, and D. Delen, “An explanatory machine learning framework for studying pandemics: The case of covid-19 emergency department readmissions,” *Decision Support Systems*, vol. 161, p. 113730, 2022, data Analytics and Decision-Making Systems: Implications of the Global Outbreaks.

[34] B. Peleg and P. Sudhölter, *Introduction to the theory of cooperative games*, 2nd ed. Springer Science & Business Media, 2007.

[35] L. Z. Wang, L. Fang, and K. W. Hipel, “Water resources allocation: A cooperative game theoretic approach,” *Journal of Environmental Informatics*, vol. 2, no. 2, pp. 11–22, 2003.

[36] I. Curiel, *Cooperative game theory and applications: Cooperative games arising from combinatorial optimization problems*. Springer Science & Business Media, 2013, no. Vol. 16.

[37] F. Bistaffa, A. Farinelli, G. Chalkiadakis, and S. D. Ramchurn, “A cooperative game-theoretic approach to the social ridesharing problem,” *Artificial Intelligence*, vol. 246, pp. 86–117, 2017.

[38] M. Kristiansen, M. Korpås, and H. G. Svendsen, “A generic framework for power system flexibility analysis using cooperative game theory,” *Applied Energy*, vol. 212, pp. 223–232, 2018.

[39] J. He, Y. Li, H. Li, H. Tong, Z. Yuan, X. Yang, and W. Huang, “Application of Game Theory in Integrated Energy System Systems: A Review,” *IEEE Access*, vol. 8, pp. 93 380–93 397, 2020.

[40] A. Churkin, J. Bialek, D. Pozo, E. Sauma, and N. Korgin, “Review of cooperative game theory applications in power system expansion planning,” *Renewable and Sustainable Energy Reviews*, vol. 145, p. 111056, 2021.

[41] A. Meca, I. García-Jurado, and P. Borm, “Cooperation and competition in inventory games,” *Mathematical Methods of Operations Research*, vol. 57, pp. 481–493, 2003.

[42] G. P. Cachon and S. Netessine, “Game theory in supply chain analysis,” *INFORMS Tutorials in Operations Research. Models, methods, and applications for innovative decision making*, pp. 200–233, 2006.

[43] M. G. Fiestras-Janeiro, I. García-Jurado, A. Meca, and M. A. Mosquera, “Cooperative game theory and inventory management,” *European Journal of Operational Research*, vol. 210, pp. 459–466, 2011.

[44] X.-X. Zheng, Z. Liu, K. W. Li, J. Huang, and J. Chen, “Cooperative game approaches to coordinating a three-echelon closed-loop supply chain with fairness concerns,” *International Journal of Production Economics*, vol. 212, pp. 92–110, 2019.

[45] H. P. Young, “Monotonic solutions of cooperative games,” *International Journal of Game Theory*, vol. 14, pp. 65–72, 1985.

[46] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, “From local explanations to global understanding with explainable AI for trees,” *Nature Machine Intelligence*, vol. 2, no. 1, pp. 56–67, 2020.

[47] K. Aas, M. Jullum, and A. Løland, “Explaining individual predictions when features are dependent: More accurate approximations to Shapley values,” *Artificial Intelligence*, vol. 298, p. 103502, 2021.

[48] J. Castro, D. Gómez, and J. Tejada, “Polynomial calculation of the Shapley value based on sampling,” *Computers and Operations Research*, vol. 36, no. 5, pp. 1726–1730, 2009.

[49] E. Štrumbelj and I. Kononenko, “Explaining prediction models and individual predictions with feature contributions,” *Knowledge and Information Systems*, vol. 41, pp. 647–665, 2014.

[50] C. Condevaux, S. Harispe, and S. Mussard, “Fair and Efficient Alternatives to Shapley-based Attribution Methods,” in *Machine Learning and Knowledge Discovery in Databases, ECML PKDD 2022, LNCS*, M. R. Amini, S. Canu, A. Fischer, T. Guns, P. Kralj Novak, and G. Tsoumakas, Eds., vol. 13713. Springer, Cham, 2023, pp. 309–324.

[51] V. Lohweg, E. Gillich, and J. Schaede, “New concepts in banknote authentication,” p. 0, Aug 2009.

[52] P. Refaeilzadeh, L. Tang, and H. Liu, “Cross-validation,” *Encyclopedia of database systems*, vol. 5, pp. 532–538, 2009.

[53] Q.-S. Xu and Y.-Z. Liang, “Monte carlo cross validation,” *Chemometrics and Intelligent Laboratory Systems*, vol. 56, no. 1, pp. 1–11, 2001.

[54] J. Cook and V. Ramadas, “When to consult precision-recall curves,” *The Stata Journal*, vol. 20, no. 1, pp. 131–148, 2020.

[55] P. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis, “Modeling wine preferences by data mining from physicochemical properties,” *Decision Support Systems*, vol. 47, no. 4, pp. 547–553, 2009.

[56] I. Cinar and M. Koklu, “Classification of rice varieties using artificial intelligence methods,” *International Journal of Intelligent Systems and Applications in Engineering*, vol. 7, pp. 188–194, 2019.

[57] J. W. Smith, J. E. Everhart, W. C. Dickson, W. C. Knowler, and R. S. Johannes, “Using the ADAP learning algorithm to forecast the onset of diabetes mellitus,” in *Proceedings of the Annual Symposium on Computer Applications in Medical Care*. American Medical Informatics Association, 1988, pp. 261–265.