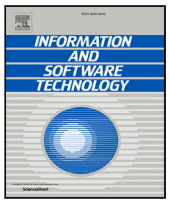
Information and Software Technology 187 (2025) 107865

Contents lists available at ScienceDirect

Information and Software Technology

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



Software Defect Prediction evaluation: New metrics based on the ROC curve

Luigi Lavazza, Sandro Morasca *, Gabriele Rotoloni

Università degli Studi dell'Insubria – Dipartimento di Scienze Teoriche e Applicate, Via O. Rossi, 9, Varese, 21100, Italy

ARTICLE INFO

Keywords:
Fault-proneness models
Binary classifiers
Fault prediction
Accuracy
Performance metrics
ROC curves
Area under the curve (AUC)
Area under the ROC curve (AUROC)
Pearson ϕ
Matthews Correlation Coefficient

ABSTRACT

Context: ROC (Receiver Operating Characteristic) curves are widely used to represent how well fault-proneness models (e.g., probability models) classify software modules as faulty or non-faulty. *AUC*, the Area Under the ROC Curve, is usually used to quantify the overall discriminating power of a fault-proneness model. Alternative indicators proposed, e.g., *RRA* (Ratio of Relevant Areas), consider the area under a portion of a ROC curve. Each point of a ROC curve represents a binary classifier, obtained by setting a specified threshold on the fault-proneness model. Several performance metrics (Precision, Recall, the F-score, etc.) are used to assess a binary classifier.

Objectives: We investigate the relationships linking "under the ROC curve area" indicators such as *AUC* and *RRA* to performance metrics.

Methods: We study these relationships analytically. We introduce iso-PM ROC curves, whose points have the same value $\overline{PM}$ for a given performance metric PM. When evaluating a ROC curve, we identify the iso-PM curve with the same value of *AUC* or *RRA*. Its $\overline{PM}$ can be seen as a property of the ROC curve and fault-proneness model under evaluation.

Results: There is an S-shaped relationship between $\overline{PM}$ and *AUC* for performance metrics that do not depend on the proportion ρ of faulty modules, i.e., dataset balancedness. ϕ (Matthews Correlation Coefficient) depends on ρ: with very imbalanced datasets, *AUC* appears over-optimistic and ϕ over-pessimistic. *RRA* defines the region of interest in terms of ρ, so all performance metrics depend on ρ. *RRA* is related to performance metrics via S-shaped curves.

Conclusion: Our proposal helps gain a better quantitative understanding of the goodness of a ROC curve, especially in practically relevant regions of interest. Also, showing a ROC curve and iso-PM curves provides an intuitive perception of the goodness of a fault-proneness model.

1. Introduction

AUC, i.e., the Area Under a Receiver Operating Characteristic (ROC) Curve is a popular metric for assessing the quality of software defect predictions. Actually, *AUC* accounts for a *family* of binary classifiers, each obtained by (1) setting a specific threshold on fault-proneness or some scoring function chosen and (2) estimating modules faulty if their fault-proneness or score is above the threshold. Each of these classifiers corresponds to a different point of a ROC curve. So, *AUC* provides an *overall* assessment of the discriminating power of a prediction model via the points of a ROC curve.

Instead, performance metrics like Precision, F-score, Accuracy, etc. assess a *specific* classifier. These metrics are defined as functions of a confusion matrix, which contains the numbers of defective and non-defective modules that are correctly and incorrectly estimated (see Section 2.1 and Table 4). So, each point of a ROC curve also corresponds to a confusion matrix.

A few problems arise when using a ROC curve and its *AUC* to evaluate the performance of a fault-proneness model in practical cases.

First, it is not clear how to interpret the values of *AUC*: we know that $AUC = 1$ indicates perfect prediction, and $AUC = 0.5$ indicates that the model has the same performance as random estimation. However, it is definitely not easy to make assessments in intermediate cases, for instance, when deciding whether a model with $AUC = 0.7$ is acceptable.

Second, *AUC* has been criticized because it accounts for the entire ROC curve, including points (hence, thresholds) that are very unlikely to be used in practice. For instance, in a safety-critical application, it is unlikely that practitioners would choose a high value for the probability threshold beyond which a module gets estimated faulty: to stay on the safe side, they would use a low probability threshold. Accordingly, several authors proposed to consider only a part of the area under the curve [1–4].

* Corresponding author.
E-mail address: sandro.morasca@uninsubria.it (S. Morasca).

https://doi.org/10.1016/j.infsof.2025.107865
Received 19 April 2024; Received in revised form 19 December 2024; Accepted 1 August 2025
Available online 9 August 2025

Third, each point of a ROC curve is associated with a confusion matrix, so it also has a specific value for a given performance metric, e.g., every point corresponds to some Accuracy (i.e., the proportion of correctly classified modules) value. So, a ROC curve is associated with a set of Accuracy values, but it is not clear how to summarize or interpret that set of values in a practically useful way.

Finally, the aforementioned problems apply both to the evaluation of a single model and to the comparison of multiple models. For instance, the fact model `m1` has greater *AUC* than model `m2` is sufficient to conclude that model `m1` performs better than `m2`, hence it is preferable to `m2`?

In this paper, we introduce further ways to gain insights into the goodness of a fault-proneness model. The underlying idea is that, since the performance of the single binary classifiers corresponding to the points of a ROC curve can be evaluated via several performance metrics, the goodness of a ROC curve too may be evaluated via these performance metrics. We here extend some initial work [5,6] to explore the relationships that link *AUC* with some of the performance metrics that are most commonly used in Software Engineering in general, and Software Defect Prediction in particular.

Specifically, we propose some ways to associate an entire ROC curve with a single value of a performance metric. To this end, we introduce iso-PM lines, i.e., ROC curves whose points have the same value of the performance metric (PM) of choice (PM can be any performance metric, e.g., Balanced Accuracy, Precision, Gmean, etc.). For instance, iso-*BA* lines are ROC curves whose points have the same value of *BA* (Balanced Accuracy).

By depicting iso-PM lines along with a given ROC curve, it is possible to get a straightforward graphical representation of the performance of the model. For instance, by representing the iso-*BA* lines, the value of *BA* of every point of the ROC curve becomes more evident. When comparing two ROC lines, i.e., two fault proneness models, it is possible to see which one achieves the highest *BA*, etc.

In addition, we introduce a new indicator, $\overline{PM}$ that summarizes the performance of the given ROC curve in terms of the chosen PM.

The ideas described above are also applicable to portions of the area under the curve. There are several ways to select a region of the ROC space: in this paper, we consider the region where performance metrics yield better values than those expected when the classification is performed randomly.

As a result, we provide practitioners with a more comprehensive way of assessing the performance of fault-proneness models and of single classifiers, by exploring the relationships among *AUC* and various performance metrics and showing their effect visually.

The paper is organized as follows. Section 2 provides some background concerning performance[1] measurement. Section 3 highlights the importance of prevalence, i.e. the proportion of positive instances (i.e., defective modules) on performance metrics. Section 4 introduces iso-PM curves and presents there properties, which are then discussed in Section 5. Section 6 illustrates how our proposal can be used in practice. Section 7 provides a brief discussion of the related work. Section 8 draws some conclusions. Mathematical details are provided in the appendices.

Glossary

We use several acronyms throughout the paper, especially in relation to performance metrics. For readers' convenience, their definitions are grouped in Tables 1–3.

[1] In this paper we borrow machine learning terminology, where "performance metrics" indicate the accuracy of predictions.

2. Context

2.1. Prediction performance

The extent to which a binary classifier correctly predicts the faultiness of software modules in a dataset is usually assessed based on a confusion matrix, whose general schema is represented in Fig. 1.

A confusion matrix is a 2×2 matrix that shows how many instances of a test set of n instances are correctly and incorrectly estimated by a binary classifier. As Fig. 1 shows, its cells contain the numbers of instances that are: correctly estimated negative (True Negatives *TN*) and positive (True Positives *TP*); incorrectly estimated negative (False Negatives *FN*) and positive (False Positives *FP*). The column totals *AN* and *AP* denote the numbers of actual negatives and positives. They depend exclusively on the dataset and not on the binary classifier. Instead, the row totals *EN* and *EP* (respectively, the number of estimated negatives and positives) depend on the binary classifier as well.

Based on confusion matrices, many performance metrics have been defined and used. In this paper, we consider some of the most frequently used in Empirical Software Engineering [7–9], listed in Table 4. Besides the definition of each performance metric in terms of confusion matrix cells, Table 4 also indicates the range of each metric and whether increasing values imply better performance ("⇑") or worse performance ("⇓").

From a confusion matrix, one can also derive the proportion of faulty modules, also known as prevalence $\rho = \frac{AP}{n}$. Prevalence is an important characteristic of a dataset, as it does not depend on the binary classifier. As shown in the remainder of the paper, prevalence may influence the values of performance metrics in a substantial way. We do not consider Accuracy ($\frac{TP+TN}{n}$) and F-score (the harmonic mean of *PPV* and *TPR*) because they have been widely criticized as being excessively dependent on ρ [10,11].

2.2. ROC curves and AUC

A ROC curve [12] illustrates the overall diagnostic ability of a binary scoring classifier, in our case a fault-proneness model, i.e., a function that associates a software module with the probability of being faulty. A binary classifier is built by setting a discrimination threshold t and estimating all modules whose score is above t as faulty and the others as non-faulty. Since the value of t is set in a somewhat arbitrary way, many different fault prediction binary classifiers can be obtained. Each of them will be characterized by different values for the pair of performance metrics $\langle FPR, TPR \rangle$.

A ROC curve plots the values of $y = TPR$ against the values of $x = FPR$ obtained on a dataset by using all possible threshold values t on fault-proneness model. A ROC curve is shown in Fig. 2. The possible set of values of pairs $\langle FPR, TPR \rangle$, i.e., the cartesian product $[0, 1] \times [0, 1]$. is called the ROC space.

At any rate, the ROC curve by itself is simply a visual aid, but a performance metric is needed to have an overall evaluation of the fault-proneness model. Researchers and practitioners use the so-called Area Under the Curve (*AUC*) in many domains, including Empirical Software Engineering, to evaluate fault-proneness models [13–17].

AUC is the area below the ROC curve. The longer the ROC curve lingers close to the left side (low *FPR*) and top side (high *TPR*) of the ROC space, the better, and the larger *AUC*. Since the total area of the ROC space is 1, the closer *AUC* is to 1, the better.

Hosmer at al. [18] propose the intervals in Table 5 as guidelines to interpret the values of *AUC* as a measure of how well a fault-proneness model discriminates between faulty and not-faulty modules for all values of threshold t.

Table 1
General terms.

Term	Meaning
BLR	Binary Logistic Regression
CM	Confusion Matrix (it groups TP, FN, FP and FN)
iso-PM	The ROC curve whose points have all the same value of the metric PM (PM can ba any name of a performance metric: e.g., iso-BA indicates the points that have the same value of Balanced Accuracy)
J&M	Jureczko and Madeyski (authors of a dataset used in this paper)
NASA	National Aeronautics and Space Administration (provider of a dataset used in this paper)
PM	Performance Metric
ROC	Receiver Operating Characteristic
RoI	Region of Interest (a portion of the ROC space)

	Actual Neg.	Actual Pos.	
Est. Neg.	TN	FN	$EN = TN + FN$
Est. Pos.	FP	TP	$EP = FP + TP$
	$AN = TN + FP$	$AP = FN + TP$	$n = AN + AP = EN + EP$

Fig. 1. A confusion matrix.

Table 2
Code measures.

Term	Meaning
AMC	Average Method Complexity
CE	Efferent Coupling
NPM	Number of Public Methods
RFC	Response For a Class
WMC	Weighted Method per Class

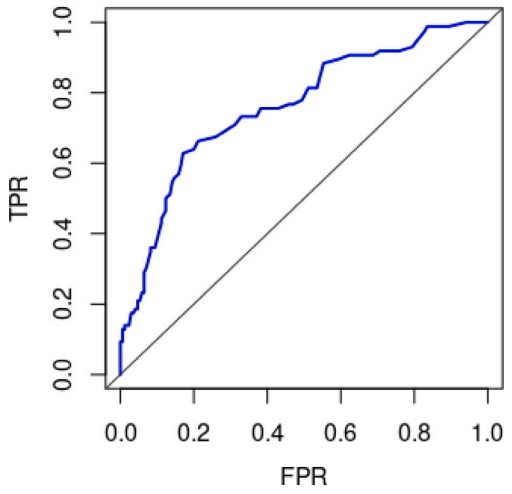


Fig. 2. A ROC curve.

2.3. RRA

In general, a binary classifier should be used only if its performance (measured according to one or more performance metrics of interest) is better than either (Case 1) the performance of a baseline classifier or (Case 2) a given minimum acceptable performance level, expressed via a performance metric of choice.

In Case 1, a typical baseline classifier is a random one, which estimates modules positive with some specified probability p. As such, random classification does not require any knowledge about the modules, and it is extremely inexpensive to implement. Thus, any binary classifier whose performance is not better than the average performance of the random classifier should not be used.

One can theoretically select the specific value of p arbitrarily in the $[0, 1]$ interval. However, the best guess on the probability that a randomly chosen module is positive is $p = \rho$, which is the Maximum Likelihood Estimate of p. Also, ρ is the only value of p such that the expected value of estimated positives obtained after estimating all modules is the number of actual positives.

As for Case 2, one can set a required level for a performance metric of interest and accept only binary classifiers with a better value for that performance metric. For instance, one may set $BA = 0.9$ as the minimum acceptable level of BA for a binary classifier.

Case 2 is applicable when there are context-specific reasons for establishing a minimum performance level. Instead, Case 1 is always applicable.

Morasca and Lavazza [4] proposed restricting the evaluation of the area under the curve to the part of the ROC space in which binary classifiers are deemed acceptable, according to some specified acceptability criteria (both Case 1 and Case 2 being supported). This portion of the ROC space is called Region of Interest (RoI). All other points of the ROC space are excluded from consideration, because their associated binary classifiers are not acceptable and should not be used in practice. Once a RoI is defined, classification performance is evaluated via an indicator named *RRA* (Ratio of Relevant Areas), defined as the proportion of the RoI area that is under the ROC curve, as follows

$$RRA = \frac{area\ of\ the\ RoI\ under\ the\ ROC\ curve}{area\ of\ the\ RoI} \quad (1)$$

RRA excludes those unacceptable points that would only introduce noise in the classification performance evaluation and may lead to selecting suboptimal fault-proneness models, as illustrated by the following example.

Fig. 3 shows the ROC curves associated with two fault-proneness models `m1` and `m2`: these are BLR models, obtained analyzing the `ivy 2.0` dataset (which has $\rho = 0.114$) from the collection by Jureczko and Madeyski [19]. Model `m1` is based on CE (efferent coupling), while `m2` is based on AMC (average method complexity).

The AUC values, $AUC_{m2} = 0.738 > AUC_{m1} = 0.668$, seem to indicate that `m2` is preferable to `m1`. In addition, `m2` is acceptable according to Table 5, while `m1` is not.

However, the evaluation changes when we consider how these two fault-proneness models compare to the random classifier, which has expected values $TPR_{rnd} = FPR_{rnd} = \rho$. Thus, only binary classifiers with $TPR > \rho$ and $FPR < \rho$ certainly behave better than the random classifier and should be taken into account, while the others should be discarded. The horizontal and vertical gray dashed lines in Fig. 3 have equations $TPR = \rho$ and $FPR = \rho$. They delimit the RoI as the upper-left rectangle with vertices $(0, \rho)$, $(0, 1)$, $(\rho, 1)$, and (ρ, ρ), whose points correspond to the classifiers to take into account. It is apparent that in

Table 3
Performance metrics and related entities (see also Table 4).

Term	Meaning
Acc	Accuracy (a performance metric)
AP	The number of actually positive modules (in a dataset)
AN	The number of actually negative modules (in a dataset)
AUC	(alias AUROC) Area Under the ROC Curve
BA	Balanced Accuracy (a performance metric)
D2H	Distance to Heaven (a performance metric)
EN	The number of modules that were estimated (i.e., classified) as negatives
EP	The number of modules that were estimated (i.e., classified) as positives
FM	F-measure, alias F-score (a performance metric)
FN	The number of false negatives, i.e., the number of positive modules incorrectly classified as negative
FP	The number of false positives, i.e., the number of negative modules incorrectly classified as positive
Gmean	The geometric mean of TPR and TNR
GM	The harmonic mean of TPR and TNR
$\overline{PM}$	the value of PM (where PM can be any performance metric, e.g., BA) of the iso-PM ROC curve having the same AUC as a given ROC curve
PPV	Positive Predictive Value, alias Precision (a performance metric)
RRA	Ratio of Relevant Areas (a performance metric)
TN	The number of true negatives, i.e., the number of correctly classified negative modules
TNR	True Negative Rate (a performance metric)
TP	The number of true positives, i.e., the number of correctly classified positive modules
TPR	True Positive Rate (a performance metric)
ϕ	Pearson ϕ, alias Matthews Correlation Coefficient

Table 4
Performance metrics.

Term	Formula	Improve	Range
TPR	$\frac{TP}{AP}$	⇑	[0,1]
TNR	$\frac{TN}{AN}$	⇑	[0,1]
FPR	$\frac{FP}{AN}$	⇓	[0,1]
BA	$\frac{TPR+TNR}{2}$	⇑	[0,1]
$Gmean$	$\sqrt{TPR\ TNR}$	⇑	[0,1]
GM	$2\frac{TPR\ TNR}{TPR+TNR}$	⇑	[0,1]
DtH	$\sqrt{\frac{(1-TPR)^2+(1-TNR)^2}{2}}$	⇓	[0,1]
ϕ	$\frac{TP\ TN-FP\ FN}{\sqrt{EN\ EP\ AN\ AP}}$	⇑	[–1,1]

Table 5
Interpretation of AUC.

AUC range	Evaluation
$AUC = 0.5$	Totally random, as good as tossing a coin
$0.5 < AUC < 0.7$	Poor, not much better than a coin toss
$0.7 \leq AUC < 0.8$	Acceptable
$0.8 \leq AUC < 0.9$	Excellent
$0.9 \leq AUC$	Outstanding

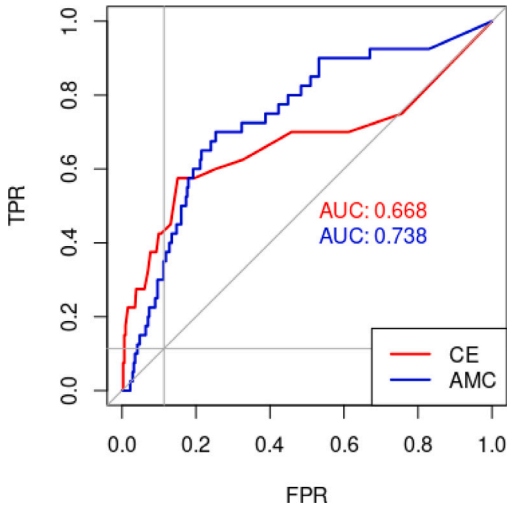


Fig. 3. ROC curves with RoI defined by $FPR < \rho$ and $TPR > \rho$.

the RoI the ROC curve of `m1` is above the ROC curve of `m2`. As a matter of fact, $AUC_{m2} > AUC_{m1}$ only because of the points with $FPR > \rho$, in which both models behave worse than the random classifier.

The values of *RRA*, i.e., the proportions of the area of the RoI under the curves, are $RRA_{m1} = 0.205$ and $RRA_{m2} = 0.073$. Therefore, according to *RRA*, `m1` is definitely preferable.

One can define the RoI in many ways, but in this paper we consider only the region of the ROC space delimited by $TPR > TPR_{rnd} = \rho$ and $FPR > FPR_{rnd} = \rho$.

2.4. The ϕ coefficient

The ϕ coefficient (or mean square contingency coefficient) is a measure of association for two binary variables. Introduced by Pearson [20], and also known as the Yule ϕ coefficient from its introduction by Yule in 1912 [21], this measure is similar to the Pearson correlation coefficient in its interpretation. It is also known as the Matthews Correlation Coefficient (MCC) [22].

The purpose of ϕ, defined in Formula (2), is to assess the strength of the association between estimated and actual values in a confusion matrix

$$\phi = \frac{TP\ \ TN - FP\ \ FN}{\sqrt{EN\ \ EP\ \ AN\ \ AP}} \tag{2}$$

ϕ is in the $[-1, 1]$ range. Specifically, $\phi = 1$ if and only if $FP = FN = 0$, i.e., for perfect predictions. $\phi = 0$ is the expected performance of a perfectly random classifier, which estimates each module positive with probability p and negative with probability $(1 - p)$ for any given p. $\phi = -1$ if and only if $TP = TN = 0$, i.e., for totally perverse classifications, when all actual positives are estimated negatives, and *vice versa*. When $\phi \geq 0$, binary classifiers with higher values of ϕ are preferable.

ϕ is an effect size measure, which quantifies how far the classification performed by a binary classifier is from random classification, which has $\phi = 0$. We generally interpret the positive values of ϕ according to the reference values in Table 6 [23]. One may thus

Table 6
Evaluation [23] of Effect Size as Measured by ϕ (negative values not considered)

ϕ range	Evaluation
$\phi = 0.1$	Small
$\phi = 0.3$	Medium
$\phi = 0.5$	Large

interpret $\phi < 0.2$ as poor, $\phi \in [0.2, 0.4)$ as medium, and $\phi \geq 0.4$ as good.

ϕ is also related to the χ^2 statistic, since $|\phi| = \sqrt{\frac{\chi^2}{n}}$.

3. The impact of ρ on *AUC*, *RRA*, and ϕ

Prevalence ρ is an important characteristic of a dataset and may have an important influence on the performance of a binary classifier. We therefore need to study how it may affect the performance metrics in Table 4 and *AUC* and *RRA*. It is immediate to recognize that, of all the performance metrics in Table 4, ρ only affects ϕ. All other performance metrics are based on *TPR* and *FPR* (or *TNR* = 1- *FPR*), which do not depend on ρ. For instance, *TPR* only depends on cells *TP* and *FN*, so its value does not depend in any way on the values of *FP* and *TN*, nor the total of actual negatives *AN*. Thus, it does not depend on information about the actual negatives, i.e., on the balancedness of the dataset. The same symmetrically applies to *FPR*.

3.1. The impact of ρ on AUC

AUC can be computed based on the knowledge of the points of the ROC curve, i.e., their *FPR* and *TPR* values. Thus, no knowledge of ρ is necessary. While it is clear that the (*FPR*,*TPR*) points of the ROC curve are obtained by applying a binary classifier to a dataset that has a specific ρ, the same ROC curve can correspond to different classifiers and datasets having different ρ.

3.2. The impact of ρ on RRA

RRA, instead, intrinsically depends on ρ. As shown in Fig. 3, the borders of the RoI used in this paper ($TPR = \rho$ and $FPR = \rho$) are determined by ρ. Therefore, the area of the RoI, which is the denominator of Eq. (1), is $\rho(1-\rho)$; the area under the ROC curve in the RoI depends on the points where the ROC curve enters and exits the RoI, and these points depend on ρ as well.

3.3. The impact of ρ on ϕ

The value of ϕ also requires the knowledge of ρ in addition to *TPR* and *FPR*, as shown by Formula (3), which can be mathematically derived from the definitions of *TPR*, *TNR* and ϕ.

$$\phi = \frac{\sqrt{\rho(1-\rho)}(TPR - FPR)}{\sqrt{(\rho TPR + (1-\rho)FPR)(\rho(1-TPR) + (1-\rho)(1-FPR))}} \quad (3)$$

This is expected, since ϕ is an effect size measure, which quantifies how far the classification performed by a binary classifier is from random. So, it must depend on the knowledge of some characteristic of the random classifier, i.e., ρ.

Let us now consider a single point (FPR, TPR) in the ROC space. While it is true that a binary classifier and its confusion matrix are associated with a single point, it is also true that the coordinates (FPR, TPR) do not summarize all of the information in a confusion matrix. Notably, information about prevalence is missing. Thus, a single point (FPR, TPR) can correspond to multiple classifiers. For instance, take the following two confusion matrices, obtained by applying a binary classifier to two datasets, D_a with $\rho_a = 0.8$ and D_b with $\rho_b = 0.9$.

CM_a

$TN_a = 16$	$FN_a = 24$
$FP_a = 4$	$TP_a = 56$

CM_b

$TN_b = 8$	$FN_b = 27$
$FP_b = 2$	$TP_b = 63$

They have the same $FPR = 0.2$ and $TPR = 0.7$ values, so they are both represented by the same point in the ROC space. However, $\phi_a = 0.41$, while $\phi_b = 0.31$. So, each point of a ROC curve by itself provides only partial information that can be used to assess the performance of a binary classifier on a dataset. The same applies to *AUC*, which is computed based on information from a "ρ-unaware" ROC curve. *RRA* and ϕ instead take into account ρ as well.

The impact of ρ can be also appreciated by investigating the possible performances of binary classifiers on balanced and imbalanced datasets. With random classification, it is $TPR_{rnd} = FPR_{rnd} = \rho$ [4]. Let us consider two datasets, D_c and D_d, with $n_c = n_d = 1000$ modules and prevalence $\rho_c = 0.5$ and $\rho_d = 0.01$, respectively. On average, random estimation on D_c results in $TP = TN = FP = FN = 250$. It is relatively easy to improve this performance with a binary classifier, because there are many false positives and false negatives that could be classified better. Instead, when applying random classification to D_d, we expect $TP = 0$, $TN = 980$, $FP = FN = 10$, on average. Thus, it is quite hard to achieve better performance, since there are very few false positives and false negatives to reclassify without disrupting the correct classifications of the other 980 out of 1000 modules. The difficulty of improving performance over random classification when ρ is small is represented well by Formula (3). When ρ is very low, the numerator is very low too and tends to 0 as ρ tends to 0, while the denominator tends to a positive number, so ϕ tends to 0 too.

4. Iso-PM ROC curves

Given a performance metric PM (e.g., ϕ), we define iso-PM curves as follows.

- An iso-PM ROC curve is a ROC curve, hence it is defined in the ROC space, goes through (0,0) and (1,1), and is non-strictly monotonic.
- Except for points (0,0) and (1,1), in which a performance metric PM may not be defined, all the points of an iso-PM curve have $PM = \overline{PM}$.
- Since performance metrics in general have non-zero values when $FPR = 0 \wedge TPR = TPR_0 > 0$ and when $FPR = FPR_0 > 0 \wedge TPR = 0$, an iso-PM ROC curve includes the vertical segment connecting (0,0) with (0, TPR_0) and the horizontal segment connecting (FPR_0,1) with (1,1).

Take a performance metric $PM(FP, TP)$ for which higher values are preferable, like *TPR*, *BA*, *Gmean*, *GM*, and ϕ: $\overline{PM}_1 > \overline{PM}_2$ implies that the iso-PM ROC curve corresponding to $\overline{PM}_1$ is always above the iso-PM ROC curve corresponding to $\overline{PM}_2$, except in points (0,0) and (1,1). As a consequence, $AUC_{\overline{PM}_1} > AUC_{\overline{PM}_2}$ and $RRA_{\overline{PM}_1} > RRA_{\overline{PM}_2}$.

Conversely, take now a performance metric PM for which lower values are preferable, like *FPR* and *DtH*. The lower $\overline{PM}$, the higher the iso-PM ROC curve in the ROC space. As a consequence, higher values of *AUC* and *RRA* are obtained for lower values of $\overline{PM}$.

In either case, we use iso-PM curves to evaluate a ROC curve according any specified performance metric PM, as an alternative to *AUC*. Specifically, given a ROC curve

1. we compute its $AUC = auc$
2. we select the iso-PM curve with $AUC_{\overline{PM}} = auc$
3. we use the value of $PM = \overline{PM}$ as a measure of the goodness of the fault-proneness model that generated the ROC curve.

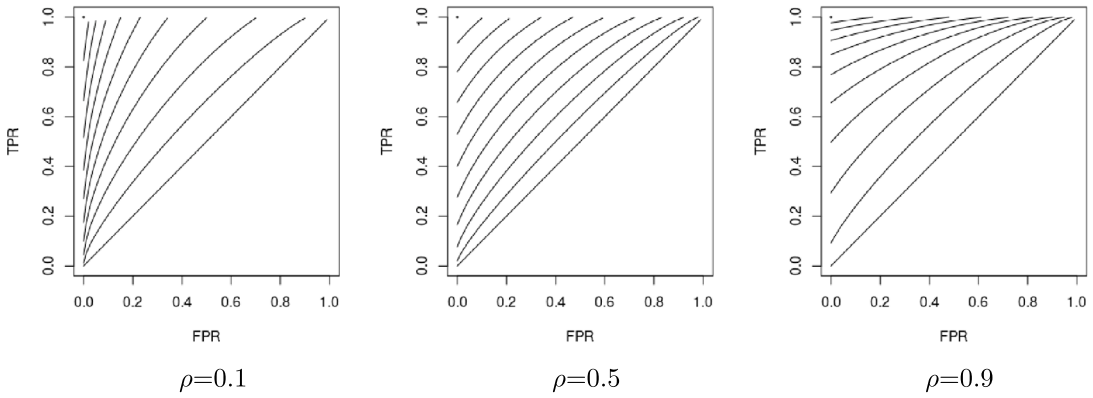


Fig. 4. ROC curves with constant ϕ, for different values of ρ.

We can limit the application of the procedure described above to the RoI: at point (1) we compute the $RRA = rra$ of the curve, at point (2) we look for the iso-PM ROC curve with $RRA = rra$, and (3) we use the value of $PM = \overline{PM}$ of this iso-PM ROC curve to evaluate the original ROC curve.

The following sections describe iso-PM lines for the performance metrics listed in Table 4. We depict and describe the parts of the iso-PM lines that are in the ROC space, since we are not interested in their behavior outside the $FPR \in [0, 1]$ and $TPR \in [0, 1]$ ranges.

4.1. Iso-ϕ ROC curves

For any given value of ρ, the iso-ϕ curves in the ROC space are ellipses that go through points (0,0) and (1,1) [4], of which only the arcs in the ROC space are considered.

Fig. 4 shows the iso-ϕ curves for values of $\overline{\phi}$ multiple of 0.1, when $\rho = 0.1$, $\rho = 0.5$ and $\rho = 0.9$. For the sake of readability, we do not show the vertical and horizontal segments that connect the points (0,0) and (1,1). ϕ depends on ρ, as shown by Formula (3), so the iso-ϕ curves depend on ρ too, as also shown by the differences in the three parts of Fig. 4.

The iso-ϕ curves when $\rho = \hat{\rho}$ are the symmetric, with respect to line $TPR + FPR = 1$, of the iso-ϕ curves obtained for $\rho = 1 - \hat{\rho}$. This phenomenon is visible in Fig. 4, for $\rho = 0.1$ and $\rho = 0.9$.

4.1.1. AUC of Iso-ϕ curves

We solved the equation in Formula (3) for TPR as a function of FPR for any given values of ϕ and ρ. By integrating TPR as a function of FPR, we obtained the formula for AUC for iso-ϕ curves depending on the values of $\overline{\phi}$ and ρ.

As we show in Appendix A.1, by denoting as $H(FPR)$ the integral function of TPR as a function of FPR between 0 and any value FPR,

$$AUC = \frac{\overline{\phi}^2}{\rho + \overline{\phi}^2(1-\rho)} + H\left(\frac{\rho - \overline{\phi}^2\rho}{\rho + \overline{\phi}^2(1-\rho)}\right)$$

To visually illustrate how ρ affects the relationship between AUC and $\overline{\phi}$, Fig. 5 (left) shows how AUC depends on $\overline{\phi}$, for all ρ values that are multiple of 0.01 in the [0.01, 0.99] range. The highest line corresponds to the extreme values of ρ, i.e., 0.01 and 0.99; the lowest line corresponds to $\rho = 0.5$.

Fig. 5 (left) provides a first insight into the relationship between AUC and $\overline{\phi}$. When ρ is not far from 0.5 (i.e., for not very unbalanced datasets), AUC and $\overline{\phi}$ provide coherent indications. For instance, with $\rho = 0.5$, when $\overline{\phi} = 0.3$ (which indicates barely acceptable performance according to Cohen [23]), we have AUC just above 0.7 (the acceptability threshold according to Hosmer at al. [18]). Likewise, when $\overline{\phi}$ is 0.5 (which indicates very good performance according to Table 6), AUC is between 0.8 and 0.9 (indicating excellent performance, according to Table 5). Instead, when ρ is very low or very high, AUC appears much more optimistic than $\overline{\phi}$: when $\rho = 0.1$, $\overline{\phi} = 0.2$, which indicates poor performance, corresponds to AUC well above 0.7, which is usually interpreted as an indication of quite good performance. The closer ρ to zero, the more divergent the indications by AUC and $\overline{\phi}$.

4.1.2. RRA of Iso-ϕ curves

By applying the same procedure described in Section 4.1.1 above, but limited to the RoI, we computed RRA for iso-ϕ curves for various values of $\overline{\phi}_{RoI}$ and ρ.

The interested reader can find the (quite complex) details of the computations in Appendix A.1.

To visually illustrate how ρ affects the relationship between RRA and $\overline{\phi}_{RoI}$, Fig. 5 (right) shows the dependence of RRA on $\overline{\phi}_{RoI}$, for all ρ values that are multiple of 0.01 in the [0.01, 0.99] range.

Fig. 5 (right) shows that, even though the RRA value depends on ρ, RRA appears less influenced by the balancedness of the dataset compared to AUC, especially for high values of $\overline{\phi}_{RoI}$. RRA seems also more aligned with $\overline{\phi}_{RoI}$ than AUC is with $\overline{\phi}$.

4.1.3. Iso-ϕ curves: Examples of usage

When looking at a ROC curves, considering iso-ϕ curves makes it possible to get a better understanding of the performance of a model, as shown by the following example.

Let us consider two fault proneness BLR models, `m1` and `m2`, obtained analyzing the `ant 1.3` dataset from the collection by Jureczko and Madeyski [19]: model `m1` is based on WMC (Weighted Method per Class) and NPM (Number of Public Methods), while `m2` is based on RFC (Response For a Class) and NPM. `m1` achieves $AUC = 0.807$, while `m2` achieves $AUC = 0.825$: therefore, based on AUC values, we should conclude that `m2` is preferable.

However, if the ROC curves of the two models are displayed in a ROC space along with iso-ϕ curves, it is easy to appreciate how well each model performs also in terms of ϕ.

Fig. 6 shows the ROC curves of the mentioned models in ROC spaces where iso-ϕ lines are also shown. The figure also highlights two particular iso-ϕ curves: the dotted red lines are those having $\phi = \overline{\phi}$, while the blue dash-dotted lines are those having $\phi = \overline{\phi}_{RoI}$.

It is easy to see that the ROC curve of model `m1` goes beyond the $\phi = 0.5$ line, while the curve of model `m2` does not. In fact, `m1` reaches $\phi = 0.54$, while `m2` never reaches $\phi = 0.5$. It is also interesting that both models achieve the highest values of ϕ in the RoI delimited by $TPR > \rho$ and $FPR < \rho$ (these boundaries are depicted as dashed lines in Fig. 6), as shown by blue dotted-dashed lines, which are well above the red dotted lines.

According to these considerations, whoever tributes more importance to ϕ than to AUC should conclude that the model `m1` based on WMC and NPM is preferable. When considering the RoI, RRA is 0.380 for `m1` and 0.364 for `m2`, while $\overline{\phi}_{RoI}$ is 0.31 for `m1` and 0.29 for `m2`. Therefore, `m1` appears to perform better than `m2` in the RoI.

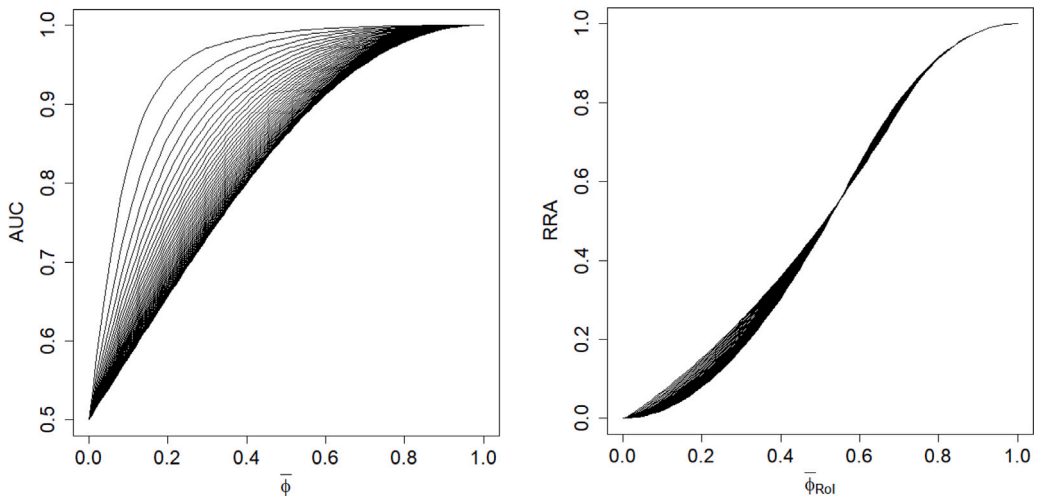


Fig. 5. *AUC* vs. $\overline{\phi}$ (left) and *AUC* vs. $\overline{\phi}_{RoI}$ (right), for various values of ρ.

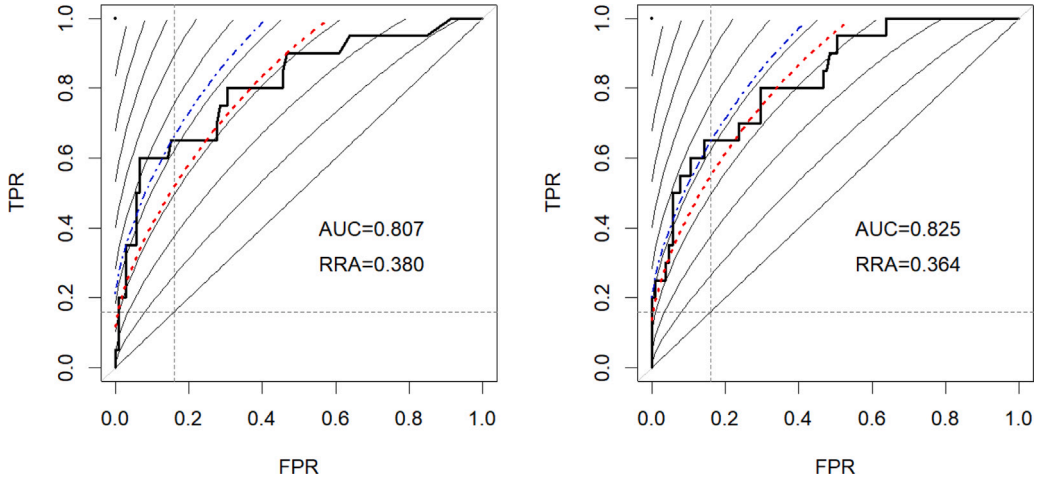


Fig. 6. ROC curves of BLR fault proneness models `m1` (left) and `m2` (right), with iso-ϕ curve.

4.2. Iso-BA ROC curves

In this section and in Sections 4.3–4.5, we deal with performance metrics that are not affected by prevalence ρ.

Since $BA = \frac{TPR+TNR}{2} = \frac{TPR+1-FPR}{2}$, iso-*BA* curves are straight lines, plus the vertical and horizontal segments that connect points (0,0) and (1,1). Fig. 7 (left) shows iso-*BA* curves (omitting the vertical and horizontal segments) for the $\overline{BA}$ values from 0 to 1 that are multiple of 0.1. It can be noticed that when $\overline{BA} = 0$ or $\overline{BA} = 1$, the line consists of a single point: (1,0) and (0,1), respectively.

4.2.1. AUC of Iso-BA curves

The relationship linking *AUC* and $\overline{BA}$ is defined as follows (assuming $AUC \geq 0.5$):

$$\overline{BA} = 1 - \sqrt{\frac{1 - AUC}{2}}$$

or

$$AUC = 1 - 2(1 - \overline{BA})^2$$

It is represented in Fig. 7 (right).

4.2.2. RRA of Iso-BA curves

RRA depends on the RoI area, which in its turn depends on ρ. Hence, the area under the $\overline{BA}$ curve in the RoI is a function of both $\overline{BA}$ and ρ. Its definition is given in Table 7. The details are in Appendix A.2.

Table 7
RRA($\overline{BA}$,ρ).

	$\rho \geq 2\overline{BA} - 1$	$\rho \leq 2\overline{BA} - 1$
$2 - 2\overline{BA} \geq \rho$	$\frac{(2\overline{BA}-1)^2}{2\rho(1-\rho)}$	$\frac{2(2\overline{BA}-1)-\rho}{2(1-\rho)}$
$2 - 2\overline{BA} \leq \rho$	$\frac{4\overline{BA}-3}{2\rho} + \frac{1}{2}$	$1 - \frac{(2-2\overline{BA})^2}{2\rho(1-\rho)}$

The relationship linking $\overline{BA}$ with *RRA* for various values of ρ is shown in Fig. 8. As with ϕ, we get the same curves for $\rho = \hat{\rho}$ and $\rho = 1 - \hat{\rho}$. It can be noted that for any value of ρ it is $RRA = 0.5$ when $\overline{BA} = 0.75$.

Fig. 8 shows that the relationship linking $\overline{BA}$ with *RRA* is close to linear for all values of ρ, except for the extreme values of $\overline{BA}$ and *RRA*. Therefore, $\overline{BA}$ and *RRA* provide essentially equivalent indications in practically relevant situations. The effect of ρ on the relationship appears very limited.

4.2.3. Iso-BA curves: Examples of usage

As an example of usage of iso-*BA* lines, we use again the ROC curves of the models mentioned in Section 4.1.3: Fig. 9 shows the ROC curves of those models, together with iso-*BA* curves. The figure also highlights two specific iso-*BA* curves: the dotted red lines are those having $BA = \overline{BA}$, while the blue dash-dotted lines are those having $BA = \overline{BA}_{RoI}$. Models `m1` and `m2` have $\overline{BA}$ 0.69 and 0.704, respectively;

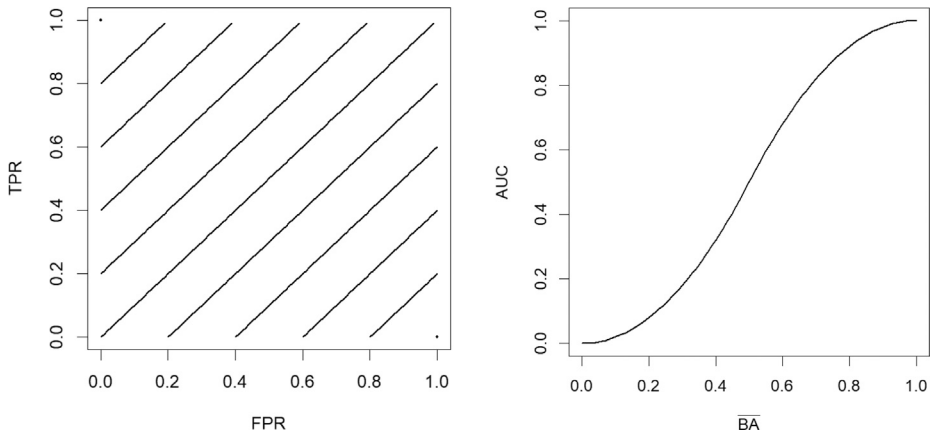


Fig. 7. Iso-*BA* curves (left) and $\overline{BA}$ vs. *AUC* (right).

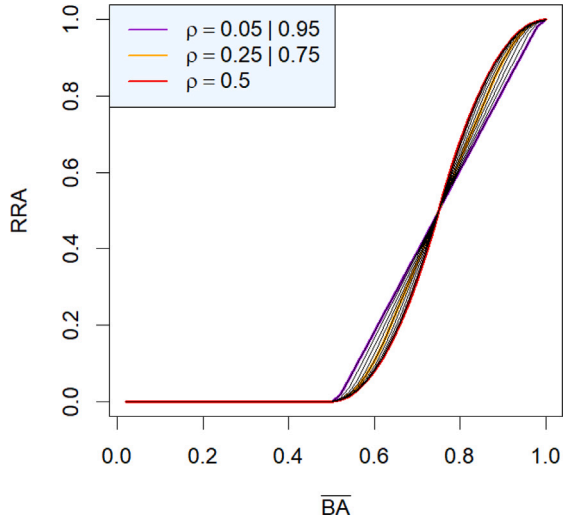


Fig. 8. $\overline{BA}$ vs. *RRA*, for values of ρ that are multiple of 0.05, in the [0.05, 0.95] range.

so m2, which has higher *AUC* and higher $\overline{\phi}$, also has higher $\overline{BA}$, and the red dotted line of model m2 is above the red dotted line of model m1. According to these observations, model m2 should be deemed preferable to m1.

When we consider the RoI, we find that m1 achieves $\overline{BA}_{RoI} = 0.7$, while m2 achieves $\overline{BA}_{RoI} = 0.693$; so, m1 appears better in the RoI according to $\overline{BA}_{RoI}$ just as it appeared better according to *RRA* and $\overline{\phi}_{RoI}$.

Fig. 9 also shows that $\overline{BA}_{RoI}$ is greater than $\overline{BA}$ for model m1, while it is not so for model m2. This is quite interesting, as it makes clear that a performance metric computed in the entire ROC space can be "inflated" by values in regions that correspond to hardly usable thresholds.

4.3. Iso-Gmean curves

Gmean is defined as $Gmean = \sqrt{TPR\,TNR} = \sqrt{TPR\,(1-FPR)}$, hence iso-*Gmean* lines are hyperbolas, plus the usual horizontal and vertical segments at the edges of the ROC space. Fig. 10 (left) shows iso-*Gmean* lines for the $\overline{Gmean}$ values from 0 to 1 that are multiple of 0.1. It can be noticed that when $\overline{Gmean} = 1$, the line consists of the single point (0,1); when $\overline{Gmean} = 0$, the line lays on the x axis.

Table 8
$RRA(\overline{Gmean},\rho)$, where $h = \overline{Gmean}^2$.

	$\rho \geq h$	$\rho < h$
$\rho \leq 1-h$	$1 + h\frac{\log h - \log\rho - \log(1-\rho) - 1}{\rho(1-\rho)}$	$\frac{-h\ \log(1-\rho) - \rho^2}{\rho(1-\rho)}$
$\rho > 1-h$	$\frac{-h\ \log\rho - (1-\rho)^2}{\rho(1-\rho)}$	$1 + \frac{h(1-\log\ h) - 1}{\rho(1-\rho)}$

4.3.1. AUC of Iso-Gmean curves

The relationship linking *AUC* and $\overline{Gmean}$ is defined as follows:

$$AUC = \overline{Gmean}^2(1 - 2\ log(\overline{Gmean}))$$

It is represented in Fig. 10 (right).

4.3.2. RRA of Iso-Gmean curves

The area under the $\overline{Gmean}$ curve in the RoI is a function of both $\overline{Gmean}$ and ρ. Its definition is given in Table 8, where $h = \overline{Gmean}^2$. The details are in Appendix A.3. Note that *RRA* by definition can be greater than zero only if the ROC curve is above the diagonal; accordingly, the formulas in Table 8 assume that such condition holds.

Fig. 11 shows the relationship between *RRA* and $\overline{Gmean}$, for various values of ρ. As with ϕ, we get the same curves for $\rho = \hat{\rho}$ and $\rho = 1-\hat{\rho}$. It can be noted that the line having $\rho = 0.5$ crosses the other ones around point (0.75, 0.55). Above that point, $\overline{Gmean}_{RoI}$ and *RRA* agree that the model is very good; therefore, the interesting part of the curves is below and to the left of (0.75, 0.55).

It can be noted that for any value of ρ, the relation linking *RRA* and $\overline{Gmean}$ is approximately linear, for "interesting" values of *RRA* and $\overline{Gmean}$: in fact, the function relating *RRA* and $\overline{Gmean}$ tends to bend for either low or high values of both metrics, i.e., when both metrics agree that the considered model is either quite poor or very good.

4.3.3. Iso-Gmean curves: Examples of usage

As an example of usage of iso-*Gmean* lines, we use again the ROC curves of the models mentioned in Section 4.1.3: Fig. 12 shows the ROC curves of those models, together with iso-*Gmean* curves. The figure also highlights two particular iso-*Gmean* curves: the dotted red lines are those having $Gmean = \overline{Gmean}$, while the blue dash-dotted lines are those having $Gmean = \overline{Gmean}_{RoI}$. Models m1 and m2 have $\overline{Gmean}$ 0.669 and 0.685, respectively; so m2, which has higher *AUC*, $\overline{\phi}$ and $\overline{BA}$, also has higher $\overline{Gmean}$, and the red dotted line of model m2 is above the red dotted line of model m1.

When we consider the RoI, we find that m1 achieves $\overline{Gmean}_{RoI} = 0.663$, while m2 achieves $\overline{Gmean}_{RoI} = 0.654$; so, m1 appears better in the RoI according to $\overline{Gmean}_{RoI}$ just as it appeared better according to *RRA*, $\overline{\phi}_{RoI}$ and $\overline{BA}_{RoI}$.

It can be observed that $\overline{Gmean}_{RoI}$ is worse than $\overline{Gmean}$ for both m1 and m2. Actually, it is so with all the models we obtained for `ant 1.3`, not only for m1 and m2.

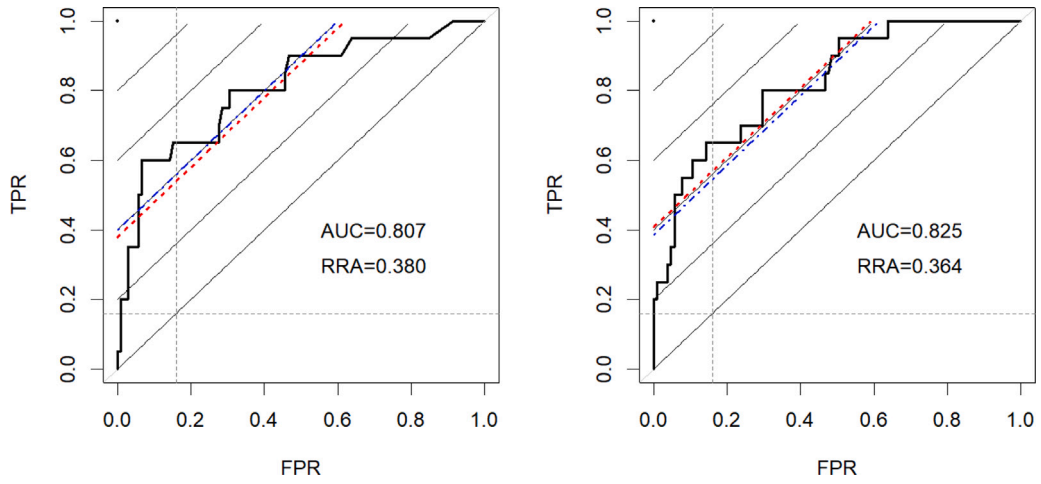


Fig. 9. ROC curves of BLR fault proneness models `m1` (left) and `m2` (right) with iso-*BA* lines.

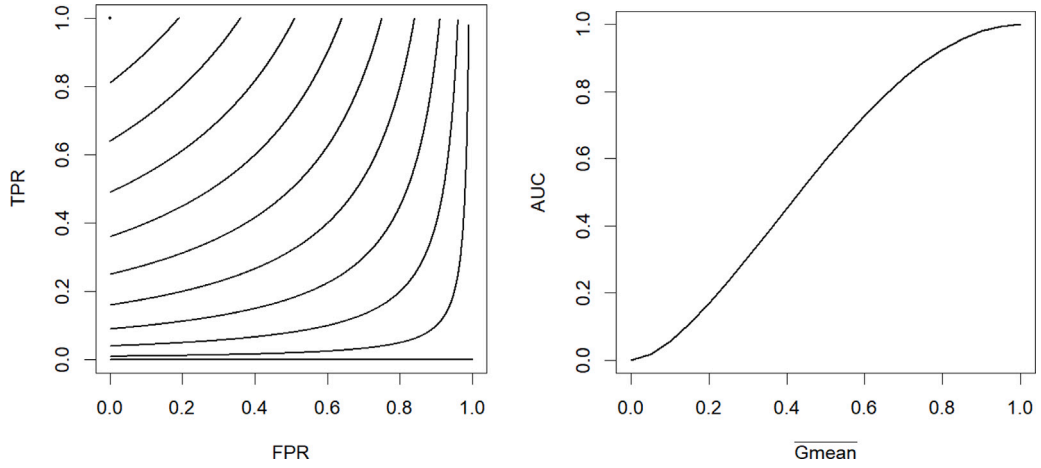


Fig. 10. Iso-*Gmean* curves (left) and $\overline{Gmean}$ vs. *AUC* (right).

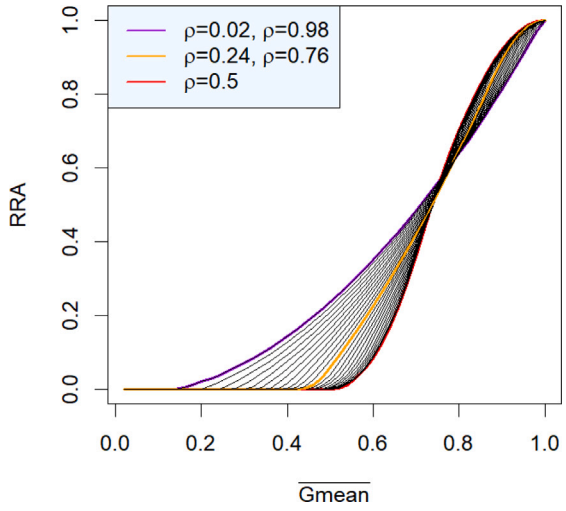


Fig. 11. *RRA* vs. $\overline{Gmean}$, for various values of ρ.

4.4. Iso-GM curves

Since $GM = \frac{2}{\frac{1}{TPR}+\frac{1}{TNR}} = 2\frac{TPR(1-FPR)}{TPR+1-FPR}$, iso-*GM* curves are rectangular hyperbolas, plus the vertical and horizontal segments that connect points (0,0) and (1,1). Fig. 13 (left) shows iso-*GM* curves (omitting the vertical and horizontal segments) for the $\overline{GM}$ values from 0 to 1 that are multiple of 0.1. It can be noticed that when $\overline{GM} = 1$ the line consists of the single point (0,1), while $\overline{GM} = 0$ corresponds to the segment on the x axis.

4.4.1. AUC of Iso-GM curves

The relationship linking *AUC* and $\overline{GM}$ is defined as follows:

$$AUC = \overline{GM} + \frac{\overline{GM}^2}{2}\left(log\left(1-\frac{\overline{GM}}{2}\right) - log\left(\frac{\overline{GM}}{2}\right)\right)$$

It is represented in Fig. 13 (right).

4.4.2. RRAof Iso-GM curves

The area under the $\overline{GM}$ curve in the RoI is a function of both $\overline{GM}$ and ρ. Its definition is given in Table 9. The details are in Appendix A.4. Figure 14 shows the relationship between RRA and $\overline{GM}$, for various

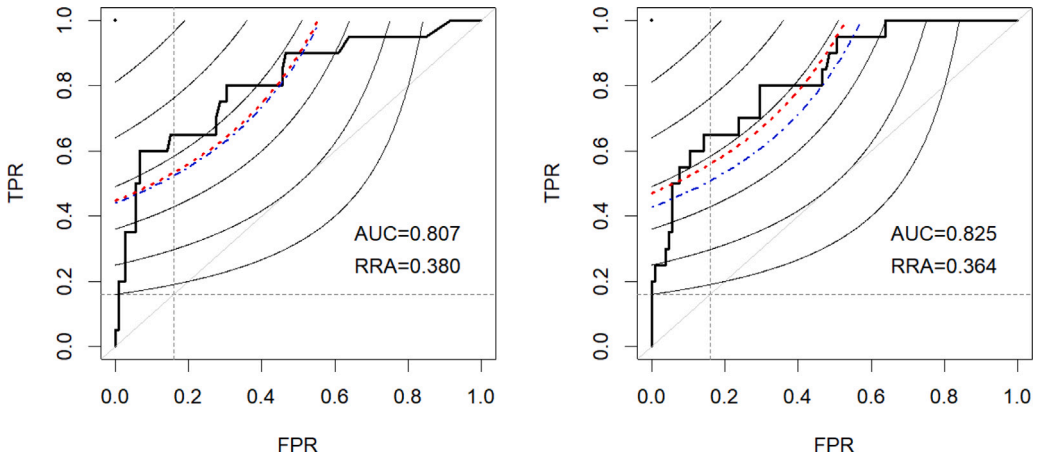


Fig. 12. ROC curves of BLR fault proneness models `m1` (left) and `m2` (right) with iso-*Gmean* lines.

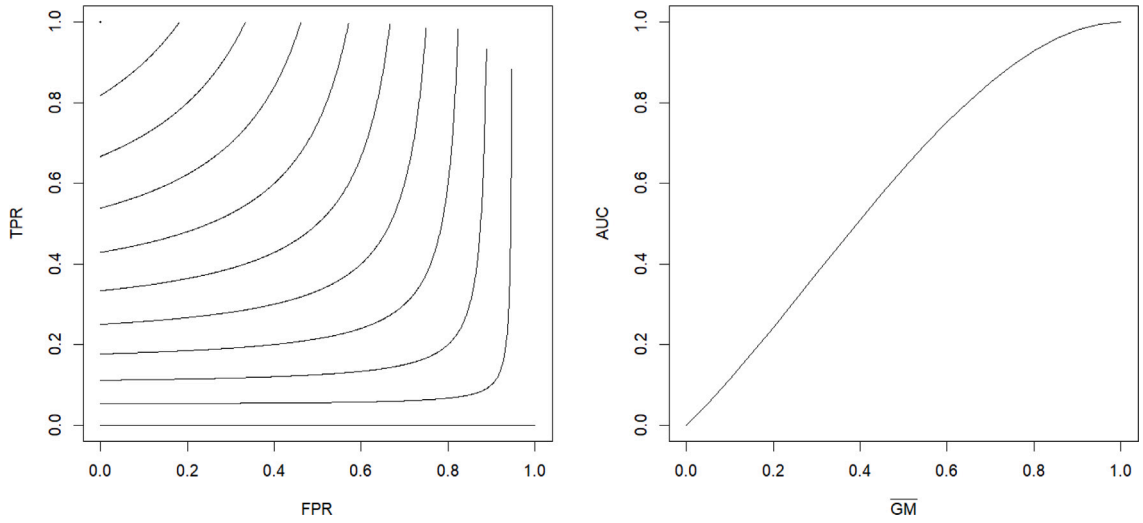


Fig. 13. Iso-*GM* curves (left) and $\overline{GM}$ vs. *AUC* (right).

Table 9
$RRA(\overline{GM},\rho)$, where $h = \frac{\overline{GM}}{2}$.

	$\rho \geq \frac{h}{1-h}$	$\rho < \frac{h}{1-h}$
$\rho \leq \frac{1-2h}{1-h}$	$1 + h\frac{h(2\log h - \log(\rho-h) - \log(1-\rho-h)) - 1}{\rho(1-\rho)}$	$\frac{h\rho + h^2(\log(1-h) - \log(1-\rho-h)) - \rho^2}{\rho(1-\rho)}$
$\rho > \frac{1-2h}{1-h}$	$\frac{-h^2\frac{1-\rho}{1-h} + h^2\log\frac{1-h}{\rho-h} + (1-\rho)\frac{h}{1-h} - (1-\rho)^2}{\rho(1-\rho)}$	$1 + \frac{2h + 2h^2(\log(1-h) - \log h) - 1}{\rho(1-\rho)}$

values of rho. Note that, by definition, *RRA* can be greater than zero only if the ROC curve is above the diagonal; accordingly, the formulas in Table 9 assume that such condition holds.

4.4.3. Iso-GM curves: Examples of usage

As an example of usage of iso-*Gmean* lines, we use again the ROC curves of the models already used in Sections 4.1.3, 4.2.3 and 4.3.3: Fig. 15 shows the ROC curves of those models, along with iso-*GM* curves. The figure also highlights two particular iso-*GM* curves: the dotted red lines are those having $GM = \overline{GM}$, while the blue dash-dotted lines are those having $GM = \overline{GM}_{RoI}$. Models `m1` and `m2` have $\overline{GM}$ 0.653 and 0.671, respectively; so `m2`, which has higher *AUC*, $\overline{\phi}$, $\overline{BA}$ and $\overline{Gmean}$, also has higher $\overline{GM}$, and the red dotted line of model `m2` is above the red dotted line of model `m1`.

When we consider the RoI, we find that `m1` achieves $\overline{GM}_{RoI} = 0.63$, while `m2` achieves $\overline{GM}_{RoI} = 0.617$; so, `m1` appears better in the RoI according to $\overline{GM}_{RoI}$ just as it appeared better according to *RRA*, $\overline{\phi}_{RoI}$, $\overline{BA}_{RoI}$ and $\overline{Gmean}_{RoI}$.

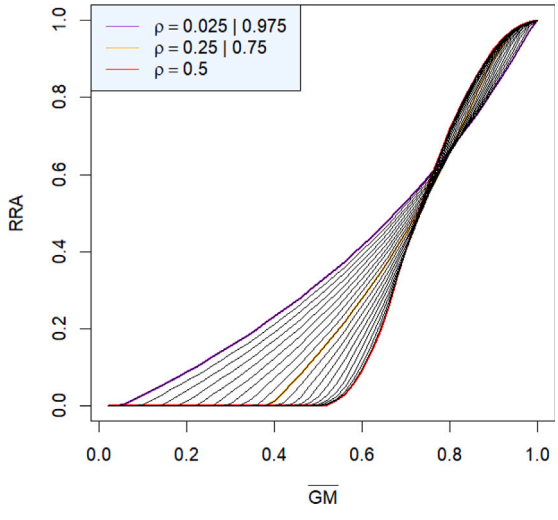


Fig. 14. *RRA* vs. $\overline{GM}$, for various values of ρ.

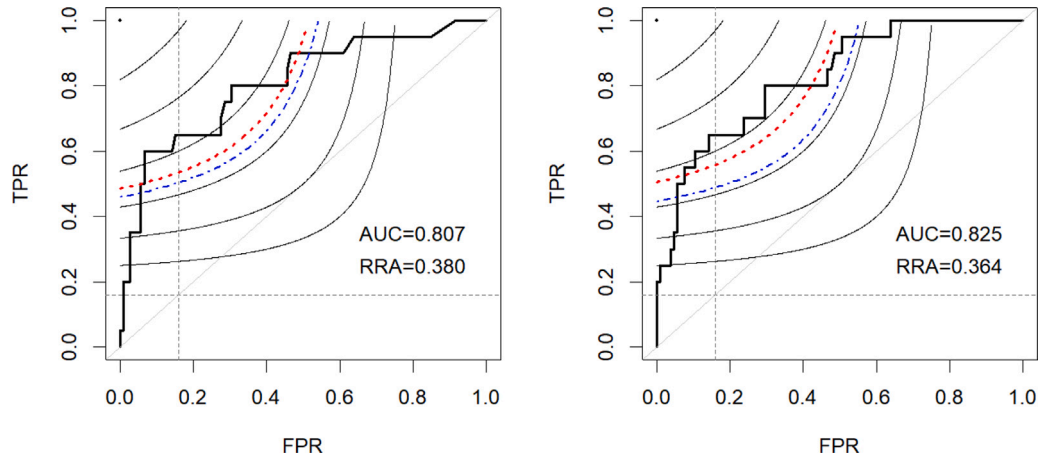


Fig. 15. ROC curves of BLR fault proneness models `m1` (left) and `m2` (right) with iso-*GM* lines.

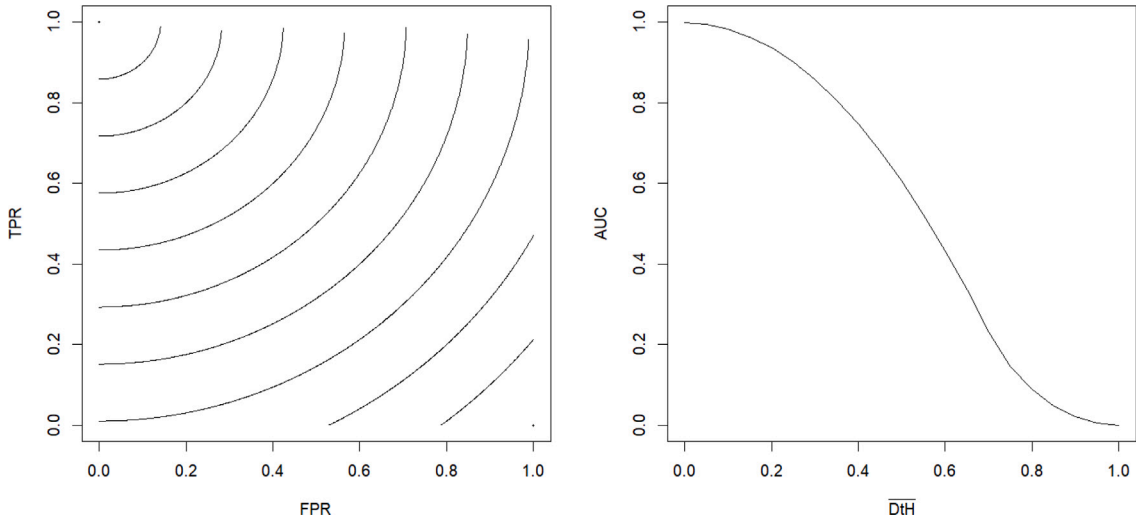


Fig. 16. Iso-*DtH* curves (left) and $\overline{DtH}$ vs. *AUC* (right).

It can be observed that $\overline{GM}_{RoI}$ is worse than $\overline{GM}$ for both `m1` and `m2`. Actually, that is the case with all of the models we obtained for `ant 1.3`, not only for `m1` and `m2`.

4.5. Iso-DTH curves

Since $DtH = \frac{\sqrt{2}}{2}\sqrt{x^2+(1-y)^2}$, iso-*DtH* curves are circumference sections, plus the vertical and horizontal segments that connect points (0,0) and (1,1). Fig. 16 (left) shows iso-*DtH* curves (omitting the vertical and horizontal segments) for the $\overline{DtH}$ values from 0 to 1 that are multiple of 0.1. It can be noticed that when $\overline{DtH} = 0$ or $\overline{DtH} = 1$, the line consists of a single point: (1,0) and (0,1), respectively.

Remember that, differently from the performance metrics discussed in the previous sections, *DtH* indicates better performance with lower values.

4.5.1. AUC of Iso-DtH curves

The relationship linking *AUC* and $\overline{DtH}$ is defined as follows:

$$AUC = 1 - \frac{\pi}{2}\overline{DtH}^2$$

It is represented in Fig. 16 (right).

4.5.2. Iso-DtH curves: relationship with RRA

The area under the $\overline{DtH}$ curve in the RoI is a function of both $\overline{DtH}$ and ρ. Its definition is given in Table 10; the values of the angles α and θ are defined in Table 11 for the different cases. The details are in Appendix A.5. Note that by definition *RRA* can be greater than zero only if the ROC curve is above the diagonal; accordingly, the formulas in Table 10 assume that such condition holds. Note that the plots in Fig. 17 use a normalized version of $\overline{DtH}$ such that $\overline{DtH}_{norm} = \frac{1-\overline{DtH}}{\sqrt{2}}$ (to make the figure easily comparable with the figures representing the relationship linking *RRA* and other *PM*) while the formulas in Table 10 use $\overline{DtH}$.

Table 10
$RRA(\overline{DtH},\rho)$, where $r = \overline{DtH}\sqrt{2}$.

	$\rho \geq 1-r$	$\rho < 1-r$
$r \geq \rho$	$1 - \frac{r\sin\theta}{2\rho} - \frac{r\sin\alpha}{2(1-\rho)} - \frac{(\pi-2\theta-2\alpha)r^2}{4\rho(1-rho)}$	$1 - \frac{r\cos\theta}{2(1-r\sin\theta)} - \frac{r\theta}{2\sin\theta(1-r\sin\theta)}$
$r < \rho$	$1 - \frac{r\cos\theta}{2\rho} - \frac{r^2\theta}{2\rho(1-\rho)}$	$1 - \frac{\pi r^2}{4\rho(1-\rho)}$

Table 11
Values of θ and ϕ, where $r = \overline{DtH}\sqrt{2}$.

	$\rho \geq 1-r$	$\rho < 1-r$
$r \geq \rho$	$\alpha = \arccos(\frac{\rho}{r})$, $\theta = \arccos(\frac{1-\rho}{r})$	$\theta = \arcsin(\frac{\rho}{r})$
$r < \rho$	$\theta = \arcsin(\frac{1-\rho}{r})$	

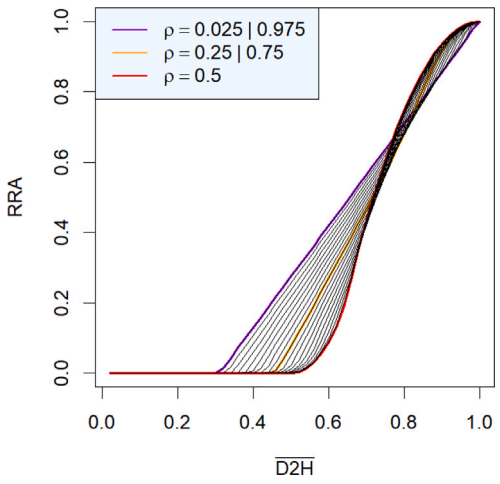


Fig. 17. *RRA* vs. $\overline{DtH}$, for various values of ρ.

4.5.3. Iso-DtH curves: Examples of usage

As an example of usage of iso-*DtH* lines, we use again the ROC curves of the models already used in Sections 4.1.3, 4.2.3, 4.3.3 and 4.4.3: Fig. 18 shows the ROC curves of those models, together with iso-*DtH* curves. The figure also highlights two particular iso-*DtH* curves: the dotted red lines are those having $DtH = \overline{DtH}$, while the blue dash-dotted lines are those having $DtH = \overline{DtH}_{RoI}$. Models `m1` and `m2` have $\overline{DtH}$ 0.35 and 0.334, respectively; so `m2`, which has higher *AUC*, $\overline{\phi}$, $\overline{BA}$, $\overline{Gmean}$ and $\overline{GM}$, also has lower $\overline{DtH}$, and the red dotted line of model `m2` is below the red dotted line of model `m1` (remember that lower values of *DtH* indicate better classifications).

When we consider the RoI, we find that `m1` achieves $\overline{DtH}_{RoI} = 0.38$, while `m2` achieves $\overline{DtH}_{RoI} = 0.384$; so, `m1` appears better in the RoI according to $\overline{DtH}_{RoI}$ just as it appeared better according to *RRA*, $\overline{\phi}_{RoI}$, $\overline{BA}_{RoI}$, $\overline{Gmean}_{RoI}$ and $\overline{GM}_{RoI}$.

It can be observed that $\overline{DtH}_{RoI}$ is worse than $\overline{DtH}$ for both `m1` and `m2`. Actually, this is the case for all of the models we obtained for `ant 1.3`, not only for `m1` and `m2`.

4.6. AUC vs. $\overline{PM}$: Summary

The relationship between *AUC* and the various $\overline{PM}$ studied are summarized in Fig. 19, where 1–*DtH* is used instead of *DtH*, to ease the comparison with other performance metrics.

We can notice that, for any value of *AUC*, it is $\overline{BA} > \overline{Gmean} > \overline{GM}$. This is due to the fact that the arithmetic mean is greater than the geometric mean, which in its turn is greater than the harmonic mean. It is also interesting to note that $1-\overline{DtH}$ is always greater than the other three metrics: this fact should be taken into account when using *DtH* for performance evaluation.

5. Discussion

Based on iso-PM curves, we have defined two new metrics—$\overline{PM}$ and $\overline{PM}_{RoI}$—associated with a given ROC curve, i.e., associated with a given fault-proneness model. Given a ROC curve, $\overline{PM}$ and $\overline{PM}_{RoI}$ are obtained as follows:

- An iso-PM ROC curve having the same *AUC* as the given ROC is built: $\overline{PM}$ is the value of PM in all points of the iso-PM curve (with the exception of the points laying on the axes of the ROC space).
- An iso-PM ROC curve having the same *RRA* as the given ROC is built: $\overline{PM}_{RoI}$ is the value of PM in all points of the iso-PM curve.

In this paper we have studied $\overline{PM}$ and $\overline{PM}_{RoI}$ when PM is *BA*, *Gmean*, *GM*, *DtH* or ϕ; however, other performance metrics, like the F-score, for instance, could be used.

Studying iso-PM curves and the relations that link $\overline{PM}$ and $\overline{PM}_{RoI}$ with the *AUC* and *RRA* of a given ROC curve, we highlighted some interesting facts, which can help in the correct interpretation of the performances of fault prediction models.

5.1. When using ϕ, prevalence should be taken into account

As shown in Section 4.1.1, the relationship that links $\overline{\phi}$ and *AUC* depends on prevalence ρ. Accordingly, the indications given by Tables 5 and 6 are not valid in absolute terms. Let us consider the ROC curve shown in Fig. 20, which concerns a dataset having $\rho = 0.01$. The two iso-ϕ lines shown in Fig. 20 have $\overline{\phi} = 0.1$ and $\overline{\phi} = 0.2$, hence the ROC curve has $\overline{\phi}$ in the (0.1, 0.2) range. The two iso-ϕ lines have $AUC = 0.824$ and $AUC = 0.936$. Now, regardless of the actual values of *AUC* and the $\overline{\phi}$ of the ROC curve, we know that the model whose ROC curve is shown in Fig. 20 should be considered excellent (or event outstanding) according to *AUC* and Table 5, while it should be considered poor according to ϕ and Table 6.

Therefore, the indications given by Tables 5 and 6 are not valid in absolute terms. While they agree for not very imbalanced datasets, they tend to disagree for very imbalanced datasets. These disagreements are not unexpected. In fact, Formula (3) shows that ϕ tends to zero with ρ, while *AUC* tends to one when ρ tends to zero, for any non-zero values of *TPR* and *TNR*.

In conclusion, when ρ is very small or very large neither *AUC* nor $\overline{\phi}$ provide reliable indications.

5.2. Some performance metrics are deeply influenced by the portion of the ROC curve outside the RoI

$\overline{PM}$ is a sort of "medium", summary value of PM over the entire ROC curve: in general PM is greater than $\overline{PM}$ in some points of the ROC curve, while it is smaller in other points. Since we are interested in maximizing PM, we should look for the points where PM is greater than $\overline{PM}$. To this end, a threshold on fault proneness has to be chosen carefully. In this section, we propose some observations that can drive the search for optimal PM values, also depending on the considered PM.

By looking at the iso-PM curves shown throughout the paper, it is easy to note that iso-ϕ curves are convex, iso-*BA* curves are straight lines, and iso-*Gmean*, iso-*GM* and iso-*DtH* curves are concave. Now, points that are out of the RoI are likely to achieve a higher values for a PM that has a concave iso-PM curve than for a PM that has a convex iso-PM curve.

For instance, Fig. 6 shows that both `m1` and `m2` achieve the highest values of ϕ in the RoI. Similarly, Fig. 9 shows that both `m1` and `m2` achieve the highest values of *BA* in the RoI, although the maximum *BA* achieved by `m2` out of the RoI is extremely close the value achieved within the RoI. Fig. 12, shows that `m1` achieves the best *Gmean* in the RoI, while `m2` achieves the best *Gmean* outside the RoI. Figs. 15 and 18 show that the maximum values of *GM* and *DtH* are achieved out of the RoI by both `m1` and `m2`.

To check this characteristics of performance metrics, we built several fault proneness models and checked where the best values of the considered performance metrics are, whether in the RoI or outside. We obtained the results shown in Table 12, where we also indicate the number of obtained models for each dataset (the details concerning the construction of the models are in Appendix B).

Table 12 shows that for the 327 fault proneness models that we obtained from the J&M and NASA datasets, the "convex" performance metrics obtained the best value in the RoI around one third of the times, while the best ϕ was in the RoI over 61% of the times.

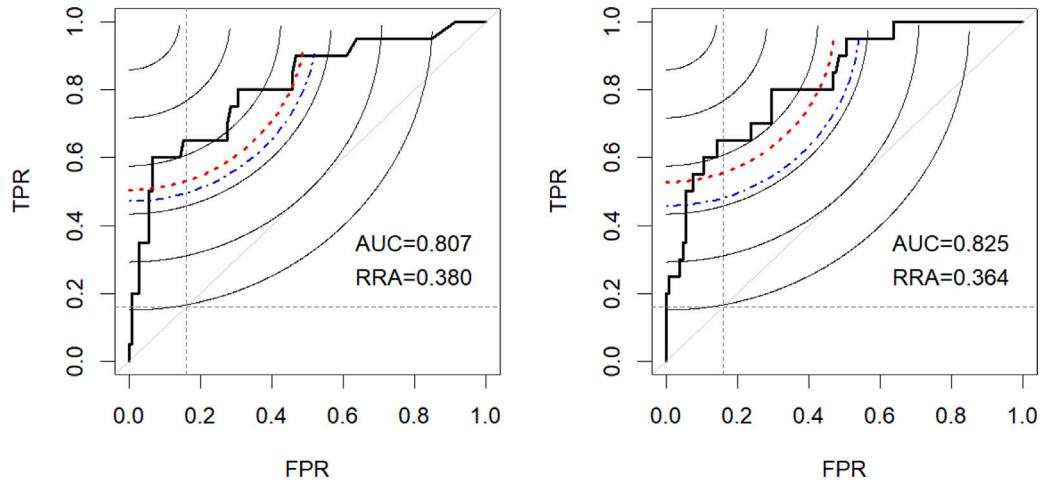


Fig. 18. ROC curves of BLR fault proneness models `m1` (left) and `m2` (right) with iso-*GM* lines.

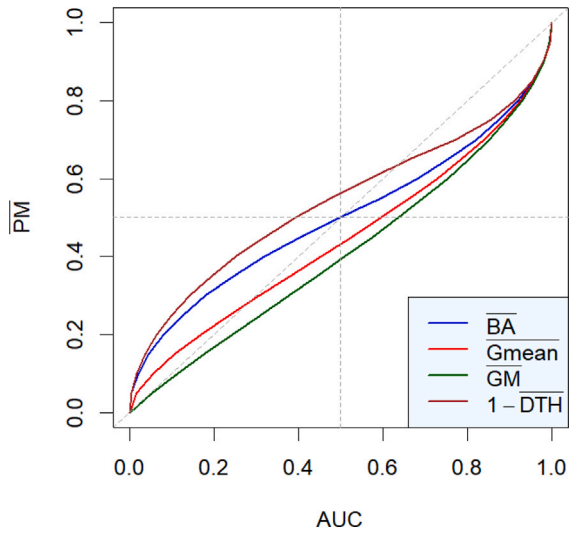


Fig. 19. *AUC* vs. $\overline{PM}$, for various PM.

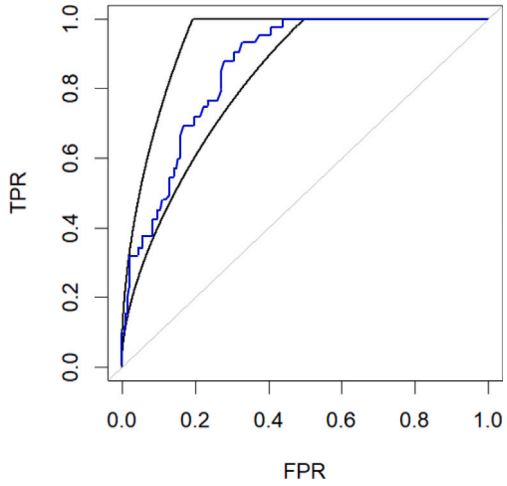


Fig. 20. Iso-ϕ line with $\overline{\phi}$ = 0.1 (lower line) and $\overline{\phi}$ = 0.2 (upper line) when ρ = 0.01.

This finding is interesting because we defined the RoI as the region of the ROC space where both *TPR* and *TNR* (or *FPR*) are better than random. Thus, in the RoI, the considered model performs better (on average) than random fault prediction with respect to both the prediction of actually faulty software modules and actually non-faulty modules. So, Table 12 shows that maximizing performance metrics that have concave iso-PM curves in the ROC space leads quite often to pick "extreme" thresholds that cause either the faulty classification performance (*TPR*) or the non-faulty classification performance (*TNR* or *FPR*) to be worse than random.

Table 12
Number of cases when the best value of a performance metric is in the RoI.

PM	J&M (208)	NASA 1 (69)	NASA 2 (50)	Total (327)
ϕ	139 (66.8%)	27 (39.1%)	34 (68.0%)	200 (61.2%)
BA	103 (49.5%)	6 (8.7%)	9 (18.0%)	118 (36.1%)
Gmean	101 (48.6%)	1 (1.4%)	6 (12.0%)	108 (33.0%)
GM	102 (49.0%)	1 (1.4%)	6 (12.0%)	109 (33.3%)
DtH	101 (48.6%)	1 (1.4%)	6 (12.0%)	108 (33.0%)

Note that picking a faultiness model out of the RoI is not necessarily a bad thing: since ROC curves are monotonic, worse *FPR* and *TNR* correspond to better *TPR*. If a project manager knows that the cost of false negatives is much greater than the cost of false negatives, picking a model that is above the diagonal and to the right of the $FPR = \rho$ line is economically advantageous. However, this kind of procedure voids the usefulness of performance metrics that—like those considered in these paper—aggregate into a single value the performance with respect to both positive and negative modules. In other words, if one needs and is able to reason about performance in terms of cost, an indicator based on cost is for sure more suitable than the performance metrics considered in this paper.

5.3. Performance in the RoI

Suppose that, based on the considerations given in the previous section, we decide to evaluate models based on $\overline{PM}_{RoI}$, that is, among the available fault proneness models, we decide to use the one that yields the highest $\overline{PM}_{RoI}$. Is this a good choice?

We know from the previous section that the best ϕ is usually in the RoI, while the best values of other metrics are most often outside the RoI, but what about the general performance: can we expect that PM values are better (in a statistically significant way) within the RoI than outside?

To this end, we built multiple fault proneness models for each one of the datasets from the J&M and NASA collections, picked the one with the best $\overline{PM}_{RoI}$ and compared the values of PM within he RoI with those outside the RoI using Wilcoxon sign rank test. The results are shown in Table 13.

Table 13
Number of cases when PM values in the RoI are better/worse/equivalent to those outside the RoI.

	J&M	NASA
ϕ	44/3/1	11/1/0
BA	35/10/3	5/6/1
Gmean	34/9/5	4/7/1
GM	34/10/4	4/7/1
DtH	4/35/9	3/8/21

Table 13 shows that if performance is evaluated via ϕ, using the fault proneness model that maximizes $\overline{\phi}_{RoI}$ is a safe choice, since in the vast majority of cases the values of ϕ in the RoI are better than the values of ϕ outside the RoI. The opposite is true for *DtH*, while the performance of *BA*, *Gmean* and *GM* seems to depend on the dataset; in this regard, remember that the NASA datasets have ρ smaller than J&M datasets, on average.

5.4. Concluding remarks

The observations reported above show that it is hardly possible to derive criteria for choosing a performance metric or guidelines concerning the usage of performance metrics that are valid in general.

However, there are well defined relationships among performance metrics and *AUC* and *RRA* that whoever has to evaluate fault proneness models should take into account.

6. Implications and usage

Representing the performance of a defect prediction model via a ROC curve is quite informative, since the ROC curve shows the values of *TPR* and *FPR* for all the possible threshold values that, applied to a fault-proneness model, yield binary defectiveness evaluations.

Knowing the values of *TPR* and *FPR* obtained by applying the model to a test set having *AP* defective and *AN* not defective modules implies knowing *FP* and *FN* (since $FN = AP\ (1 - TPR)$ and $FP = AN\ FPR$). From a practical point of view, these values are very important, because they determine costs: the amount of effort wasted in analyzing non-defective modules is proportional to *FP*, while the costs incurred after software release are proportional[2] to *FN*. Unfortunately, all of the research papers on defect prediction use single-valued performance metrics [9]; the majority of the papers use AUC [9], i.e., a summary of the ROC curve that hides the individual (*TPR*, *FPR*) pairs corresponding to threshold values.

Many researchers, both in software engineering [24] and in other fields [25–27], stated that AUC is "threshold independent". This is not really correct: AUC does not depend on a specific threshold, because it depends on *all* possible threshold values. As other researchers have observed, considering all the thresholds means that even very unlikely threshold values are considered [28,29]. Take for instance a test set with prevalence 10% (i.e., 10% of the modules are defective): in practice, a practitioner could set the threshold at 10%, i.e., modules having fault-proneness greater than 0.1 are considered faulty and subjected to further quality assurance activities. The practitioner could cautiously set the threshold at 5%, to lower the number of false negatives. However, no practitioner would set the threshold at 80% or 90%, since that would imply missing most defects.

Based on these observations, our approach provides several advantages over evaluations based on AUC or performance metrics like F-score, Gmean, ϕ, etc., as described below.

[2] Proportionality holds on average, under the realistic hypothesis that estimated negative modules are released without undergoing further quality improvement activities.

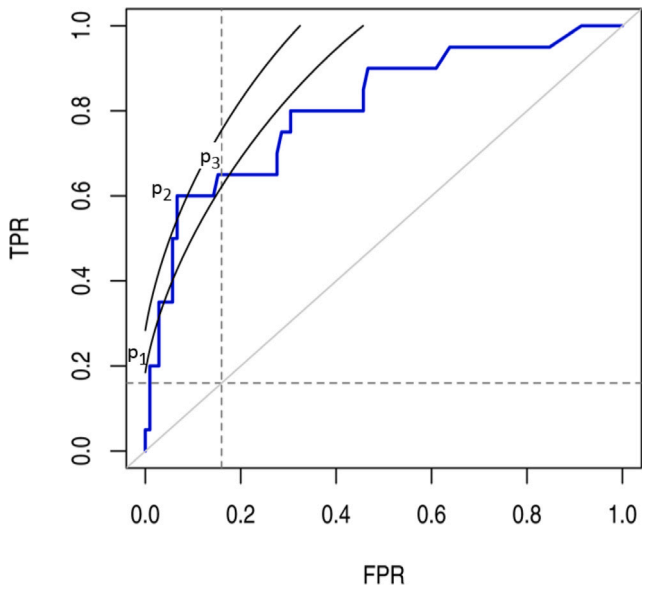


Fig. 21. The ROC curve representing the performance of the BLR model obtained for project `ant 1.3` using WMC and NPM as independent variables. Iso-ϕ lines with $\phi = 0.4$ and $\phi = 0.5$ are also shown. The dashed lines represent $TPR = \rho$ and $FPR = \rho$.

We let developers compute the AUC in a subset of the ROC space, i.e., in a Region of Interest (RoI) where threshold have practically relevant values. In this respect, it is worth noting that in this paper we always defined the RoI as the portion of the ROC space having $TPR > \rho$ and $FPR < \rho$ (i.e., both *TPR* and *FPR* better than random estimation's); nonetheless one can set any border for the RoI. Specifically, one can choose the borders of the RoI so that the portion of the ROC curve in the RoI corresponds to feasible threshold values.

Practitioners that are familiar with and trust some performance metrics can use the $\overline{PM}$ we associate with the ROC curve, or the iso-PM lines. Note that the points where a iso-PM line having $PM = K$ crosses the ROC curve indicate the minimum and maximum threshold that guarantee that PM is greater or equal to K. Therefore, one can evaluate the performance of a model by taking into account performance metrics, threshold values and, consequently, the costs due to false positives and false negatives.

For instance, consider Fig. 21, which shows the ROC corresponding to a fault-proneness model for project `ant 1.3`, whose dataset contains the data of $AP = 20$ defective modules and $AN = 105$ not defective modules (hence $\rho = 0.16$). This figure shows a few interesting facts, which can be used to make decisions about the possible usage of the model:

- The part of the ROC curve in the RoI is almost completely above the iso-ϕ curve having $\phi = 0.4$; point `p1` has a value of ϕ just below 0.4.
- Points `p1` and `p3` are the extremes of the portion of ROC curve in the RoI and have coordinates (0.01, 0.2) and (0.15, 0.65), respectively, corresponding to $FP = 1$ and $FN = 16$ and $FP = 16$ and $FN = 7$, respectively. The two points represent the performance of the fault predictions models obtained with threshold 0.51 (`p1`) and 0.17 (`p3`).
- Point `p2` (0.07, 0.6) is the only point where the model achieves $\phi > 0.5$. It corresponds to $FP = 7$ and $FN = 8$, and is obtained when the threshold is set at 0.25.

Based on these values, a practitioner could evaluate if the costs associated with the considered points are acceptable: e.g., setting the threshold at 0.51 (or 0.5) (`p1`) implies that $FN = 16$: releasing 16 faulty modules (i.e., 16/125 = 12.8% of the project's modules) in

production could be not acceptable. Similarly, setting the threshold at 0.17 (p3) implies that $EP = TP + FP = 16 + (20\text{-}7) = 29$ modules require additional quality improvement: these many modules could call for more effort than is available. Point p2 is a trade-off in terms of costs due to false positives and costs due to false negatives and maximizes ϕ: it is a reasonable choice if there are no compelling cost-related constraints.

The proposed approach can be used to draw iso-cost lines, i.e., ROC curves characterized by a constant cost. These are quite interesting for developers, since they indicate where (i.e., for which threshold values) a given type of cost is acceptable and where it is minimum.[3]

7. Related work

Other researchers have represented all points with the same performance in the ROC space. Flach [30] built different lines, which represent all points in the ROC space having the same *PPV*, *Acc*, and *FM*, and different curves representing splitting criteria for Decision Trees, specifically Entropy, Gini-split, and DKM-split. From his research, Flach also noticed that when a metric depends on the class proportion (i.e., prevalence ρ), different isometric curves are possible for different values of ρ. This is the same behavior we experienced with the iso-ϕ curves in Fig. 4.

That research has been used by Hernandez-Orallo et al. to find a point-to-point relation between the ROC space and Cost space (i.e., the graph which correlates cost proportion with classification loss) [31]. In fact, a point in the ROC space corresponds to a line in the Cost space, since the loss computed on a threshold would depend on the cost proportion. The point-to-point relationship is found by intersecting the points of a ROC curve with the isometric predicted positive rate $\frac{EP}{n}$ line. The corresponding set of points in the Cost space can be aggregated and the resulting area under the curve is related to *AUC*.

Isometric functions can also be used to select the best threshold in a ROC curve. For example, [31] considers the use of predicted positive rate for the threshold choice by intersecting the iso-rate line with value ρ with the ROC curve. In another paper, Lachiche and Flach [32] proposed the idea of using iso-accuracy and iso-cost lines to select the best threshold.

In the studies previously mentioned, however, the isometric functions just represent the collection of points in the ROC space having the same value for a specific metric.

AUC has been criticized also outside the software engineering field. For example, Hand shows that *AUC* is a weighted minimum misclassification cost function [28]. However, Hand suggests that practitioners consider other performance metrics, with a more direct correspondence with their objectives. This is because the weights, i.e. the unit costs per misclassification, depend essentially on the binary classifier used, making *AUC* an ill-defined performance metric when interpreted as being related to some kind of cost.

Other researchers have studied the relationship between ϕ and *AUC*. For instance, Chicco and Jurman studied their relationship on tens of thousands of systematically generated ROC curves, considering different values of prevalence ρ [29]. In their study, they find that *AUC* and ϕ are never on opposite trends, but several values of ϕ are possible, given a specific value of *AUC*. They also notice that the range of ϕ values increases as the dataset gets more unbalanced. They also carried out an empirical study, comparing *AUC* and ϕ of a Random Forest model tested on three different datasets, resulting in *AUC* being extremely over-optimistic related to ϕ. In their conclusions, they suggest to stop using *AUC* as a performance metric in favor of ϕ.

Our study differs from that of Chicco and Jurman because of the ϕ used for the comparison. In fact, Chicco and Jurman use the ϕ value obtained with the cut-off threshold $\tau = 0.5$, which appears as an arbitrary choice, and different thresholds may result in different performance. The usage of iso-ϕ curves allows us to study ϕ regardless of a specific threshold.

The conclusions by Chicco and Jurman may not draw a complete picture of the ϕ *AUC* relationship. This becomes even more evident when we consider that in our study *AUC* and ϕ appear to be concordant when working with balanced datasets.

8. Conclusions

In this paper, we investigate the relationship between ROC curves and different performance metrics. We defined iso-PM ROC curves, having the same value on all its points for a specific performance metric. It is possible to compute the Area Under the iso-PM curve and study the relationship between *AUC* and the metric. An alternative to *AUC* that only considers the area with better performance than a random classifier, *RRA*, has also been studied. Since Matthews Correlation Coefficient ϕ and *RRA* depend on the proportion of positive elements in the dataset ρ, they were studied taking prevalence into account. The other metrics that were considered are *BA*, *Gmean*, *GM*, and *DtH*. From the values of *AUC* and *RRA* it is possible to derive the $\overline{PM}$ and $\overline{PM}_{RoI}$, respectively.

Regarding ϕ, the relationship with *AUC* depends on prevalence ρ. When ρ is close to zero, or one, ϕ tends to get closer to zero, while *AUC* tends to get closer to one. Because of this it seems that, when the dataset is very unbalanced, ϕ becomes over-pessimistic, while *AUC* becomes over-optimistic. Because of this, when dealing with very large, or very small values of ρ, both ϕ and *AUC* do not give reliable indications.

When considering all performance metrics, to higher *RRA* and *AUC* values correspond better values of the metrics themselves. If a ROC curve has better *AUC* but worse *RRA* than another curve, it would also have a better $\overline{PM}$ but worse $\overline{PM}_{RoI}$. Despite this, $\overline{Gmean}$, $\overline{GM}$, and $\overline{DtH}$ seem to be better than $\overline{Gmean}_{RoI}$, $\overline{GM}_{RoI}$, and $\overline{DtH}_{RoI}$. This means that they are influenced by the points of the ROC curve with worse performance than a random classifier. This remark has also been studied on real-life datasets, showing how the best PM value is more probable to reside in the RoI for ϕ rather than the other metrics. From the same datasets, we also find that, in general, the Region of Interest contains better values for all metrics, except *DtH*.

This paper provides new insights into commonly used performance metrics and their relationship with the ROC space, *AUC*, and *RRA*.

Future work

Dealing with effort-aware metrics

In this paper, we used "traditional" performance metrics. The reader could wonder if the proposed techniques are applicable also to effort-aware metrics, which are gaining more and more attention. In fact, our proposal is applicable to any model that estimates a module as defective or not defective based on a threshold: by varying the threshold, we get different estimates, each one associated with a specific confusion matrix, hence a specific $\langle TPR, FPR \rangle$ pair, i.e, a point in the ROC space.

As an example, consider the personalized defect prediction (PofB) metric by Chen et al. [33] or its normalized version NPofB [34]. PofBX is the percentage of bugs that a developer can identify by inspecting the most fault-prone modules that cumulatively account for X% of lines of code. Thus, (1) you order the modules in decreasing order of fault-proneness, and (2) you select as estimated positives the highest-ranking modules whose combined size is X% of the total product size. This determines *TP* (the number of defective modules among the selected ones) and *FP* (the number of not defective modules among the selected ones). So, we can compute $TPR = TP/AP$ and $FPR = FP/AN$ for X: by varying X, we obtain a sequence of $\langle TPR, FPR \rangle$ pairs, which defines

[3] As we mention in the Future work section, we plan to address iso-cost lines, also providing a tool that computes and visualizes them.

a ROC curve. With NPofB, the modules are ordered according to their effort proneness divided by their size; the process is then the same as for PofB.

Once we have obtained a ROC curve, we can show it in the ROC space along with iso-PM curves; we can compute AUC, RRA, $\overline{PM}$, etc. as described in the previous sections.

As part of our future work, we plan to apply our techniques to effort-aware metrics, to increase the knowledge of the peculiarities of effort-aware metrics and their relationship with "traditional" performance metrics.

Dealing with cost metrics

We also plan to introduce iso-cost lines, as briefly mentioned in Section 6.

Tool support

We built a tool that automatically computes the metrics described in the paper, and provides several ways of visualizing ROC curves and iso-PM curves. Via the tool, practitioners can use the proposed metrics and visualizations without having to deal with the mathematics reported in this paper. The tool also make explicit the threshold values corresponding to specific points of the ROC curve, noticeably the points where the curve enters and exits the RoI. The tool is publicly available here: https://github.com/BruhZul/roc-and-iso-roc-tool.

CRediT authorship contribution statement

Luigi Lavazza: Writing – review & editing, Writing – original draft, Validation, Supervision, Methodology, Investigation, Formal analysis, Conceptualization. **Sandro Morasca:** Investigation, Formal analysis, Conceptualization, Writing – review & editing, Writing – original draft, Validation, Supervision, Methodology. **Gabriele Rotoloni:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Iso-PM formulas for AUC and RRA

A.1. ϕ

Let us denote *FPR* by x and *TPR* by y for simplicity and $k = \frac{1-\rho}{\rho}$ for mathematical convenience for the time being. As shown in [4], an iso-ϕ curve contains two parts. The first part is an arc of an ellipse with formula

$$(k+\overline{\phi}^2k^2)x^2+2(\overline{\phi}^2k-k)xy+(\overline{\phi}^2+k)y^2-\overline{\phi}^2(k^2+k)x-\overline{\phi}^2(k+1)y=0 \quad \text{(A.1)}$$

for $x \in [0, \frac{1-\overline{\phi}^2}{1+\overline{\phi}^2k}]$, since the ellipse intercepts the horizontal line $y = 1$ when $x = \frac{1-\overline{\phi}^2}{1+\overline{\phi}^2k}$. The second part is the horizontal segment with $y = 1$ between $x = \frac{1-\overline{\phi}^2}{1+\overline{\phi}^2k}$ and $x = 1$. The computation of the area under the second part is trivial, being equal to $1 - \frac{1-\overline{\phi}^2}{1+\overline{\phi}^2k}$.

To compute the area under the first part, we can first represent the arc of the ellipse in explicit form as a function of x by solving Formula (A.1) for y as follows

$$y = \frac{-(2(\overline{\phi}^2k-k)x-\overline{\phi}^2(k+1))}{2(\overline{\phi}^2+k)} + \frac{\sqrt{(2(\overline{\phi}^2k-k)x-\overline{\phi}^2(k+1))^2-4(k+\overline{\phi}^2k^2)(\overline{\phi}^2+k)x^2+4(\overline{\phi}^2+k)(\overline{\phi}^2(k^2+k))x}}{2(\overline{\phi}^2+k)} \quad \text{(A.2)}$$

The integral of this function between 0 and any value of x is here represented as a function, which we here denote as $H(x)$, with the following form

$$H(x) = -\frac{(\overline{\phi}^2k-k)x^2-\overline{\phi}^2(k+1)x}{2(\overline{\phi}^2+k)} + \overline{\phi}(k+1)\sqrt{k}\left((x-\frac{1}{2})\sqrt{D^2-\left(x-\frac{1}{2}\right)^2}+D^2\arcsin\frac{x-\frac{1}{2}}{D}\right) - \overline{\phi}(k+1)\sqrt{k}\left(-\frac{1}{2}\sqrt{D^2-\frac{1}{4}}+D^2\arcsin\frac{-\frac{1}{2}}{D}\right) \quad \text{(A.3)}$$

where, for short, $D = \sqrt{\frac{\overline{\phi}^2+k}{4k}}$.

So, the area under the arc of the ellipse is obtained by computing $H(\frac{1-\overline{\phi}^2}{1+\overline{\phi}^2k})$. Thus, the value of AUC is the sum of $1 - \frac{1-\overline{\phi}^2}{1+\overline{\phi}^2k} = \frac{\overline{\phi}^2}{\rho+\overline{\phi}^2(1-\rho)}$ and $H(\frac{1-\overline{\phi}^2}{1+\overline{\phi}^2k}) = H(\frac{\rho-\overline{\phi}^2\rho}{\rho+\overline{\phi}^2(1-\rho)})$.

To compute the value of RRA, four cases are possible, depending on the mutual positions of the iso-ϕ curve and the borders of the RoI. These mutual positions are described by two predicates, which can be defined in terms of the intersections of the arc of the ellipse with the y-axis and the horizontal line $y = 1$. Specifically, the arc of ellipse intercepts the y axis when $y = \frac{\overline{\phi}^2(k+1)}{\overline{\phi}^2+k}$, and the horizontal line $y = 1$ when $x = \frac{1-\overline{\phi}^2}{1+\overline{\phi}^2k}$.

Thus, whenever $\rho \geq \frac{\overline{\phi}^2(k+1)}{\overline{\phi}^2+k}$, the intercept of the iso-ϕ curve is below or at least on the horizontal $y = \rho$ straight line, and above otherwise.

Whenever $\frac{1-\overline{\phi}^2}{1+\overline{\phi}^2k} \geq \rho$, the value of x at which the iso-ϕ curve intersects the horizontal $y = 1$ straight line is to the right or at least on the vertical $x = \rho$ straight line, and to the left otherwise.

We now discuss all four possible combinations of these two predicates.

Case 1: $\rho \geq \frac{\overline{\phi}^2(k+1)}{\overline{\phi}^2+k}$ and $\frac{1-\overline{\phi}^2}{1+\overline{\phi}^2k} \geq \rho$. These two inequalities hold if and only if $\rho \geq \frac{\overline{\phi}^2}{1-\overline{\phi}^2}$ and $\frac{1-2\overline{\phi}^2}{1-\overline{\phi}^2} \geq \rho$. By summing these two latter inequalities, we have $\frac{1-2\overline{\phi}^2}{1-\overline{\phi}^2} \geq \frac{\overline{\phi}^2}{1-\overline{\phi}^2}$, which implies that this case can occur only if $\overline{\phi}^2 \leq \frac{1}{3}$.

Let us denote by x^* the value of x at which the arc of the eclipse intercepts the horizontal line $y = \rho$ and by y^* the value of y at which the arc of the eclipse intercepts the vertical line $x = \rho$. Both x^* and y^* are functions of ρ. The relevant area under the curve is the area of the portion of the ROC space delimited by the arc of the eclipse, the horizontal segment between (x^*, ρ) and (ρ, ρ), and the vertical segment between (ρ, ρ) and (y^*, ρ). Therefore, the value of $RRA(\overline{\phi}, \rho)$ is

$$RRA(\overline{\phi}, \rho) = \frac{H(\rho) - H(x^*) - (\rho - x^*)\rho}{\rho(1-\rho)} \quad \text{(A.4)}$$

Case 2: $\rho \leq \frac{\overline{\phi}^2(k+1)}{\overline{\phi}^2+k}$ and $\frac{1-\overline{\phi}^2}{1+\overline{\phi}^2k} \geq \rho$. These two inequalities hold if and only if $\overline{\phi}^2 \geq \frac{\rho}{1+\rho}$ and $\frac{1-\rho}{1-2\rho} \geq \overline{\phi}^2$. By summing those two latter inequalities, we find that we must have $\rho \leq \frac{1}{2}$.

The relevant area under the curve is the area of the portion of the ROC space delimited by the arc of the ellipse, the vertical segment between $(0, \rho)$ and $(0, \frac{\overline{\phi}^2(k+1)}{\overline{\phi}^2+k})$, the horizontal segment between $(0, \rho)$ and (ρ, ρ), and the vertical segment between (ρ, ρ) and (ρ, y^*). So,

$$RRA(\overline{\phi}, \rho) = \frac{H(\rho) - \rho^2}{\rho(1-\rho)} \quad \text{(A.5)}$$

Case 3: $\rho \geq \frac{\overline{\phi}^2(k+1)}{\overline{\phi}^2+k}$ and $\frac{1-\overline{\phi}^2}{1+\overline{\phi}^2 k} \leq \rho$. These two inequalities hold if and only if $\overline{\phi}^2 \leq \frac{\rho}{1+\rho}$ and $\frac{1-\rho}{1-2\rho} \leq \overline{\phi}^2$. By summing those two latter inequalities, we find that we must have $\rho \geq \frac{1}{2}$.

The relevant area under the curve is the area of the portion of the ROC space delimited by the arc of the ellipse, the horizontal segment between $(\frac{1-\overline{\phi}^2}{1+\overline{\phi}^2 k}, 1)$ and $(\rho, 1)$, the vertical segment between $(\rho, 1)$ and (ρ, ρ), and the horizontal segment between (x^*, ρ) and (ρ, ρ). So,

$$RRA(\overline{\phi}, \rho) = \frac{H(\frac{1-\overline{\phi}^2}{1+\overline{\phi}^2 k}) - H(x^*) - (\frac{1-\overline{\phi}^2}{1+\overline{\phi}^2 k} - x^*)\rho + (\rho - \frac{1-\overline{\phi}^2}{1+\overline{\phi}^2 k})(1-\rho)}{\rho(1-\rho)} \tag{A.6}$$

Case 4: $\rho \leq \frac{\overline{\phi}^2(k+1)}{\overline{\phi}^2+k}$ and $\frac{1-\overline{\phi}^2}{1+\overline{\phi}^2 k} \leq \rho$. These two inequalities hold if and only if $\rho \leq \frac{\overline{\phi}^2}{1-\overline{\phi}^2}$ and $\frac{1-2\overline{\phi}^2}{1-\overline{\phi}^2} \leq \rho$. This case can occur only if $\overline{\phi}^2 \geq \frac{1}{3}$. The relevant area under the curve is the area of the portion of the ROC space delimited by the arc of the ellipse, the horizontal segment between $(\frac{1-\overline{\phi}^2}{1+\overline{\phi}^2 k}, 1)$ and $(\rho, 1)$, the vertical segment between $(\rho, 1)$ and (ρ, ρ), the horizontal segment between $(0, \rho)$ and (ρ, ρ), and the vertical segment between $(0, \rho)$ and $(0, \frac{\overline{\phi}^2(k+1)}{\overline{\phi}^2+k})$. So,

$$RRA(\overline{\phi}, \rho) = \frac{H(\frac{1-\overline{\phi}^2}{1+\overline{\phi}^2 k}) - \frac{1-\overline{\phi}^2}{1+\overline{\phi}^2 k}\rho + (\rho - \frac{1-\overline{\phi}^2}{1+\overline{\phi}^2 k})(1-\rho)}{\rho(1-\rho)} \tag{A.7}$$

A.2. Balanced accuracy

$$BA = \frac{TPR + 1 - FPR}{2} = \frac{y + 1 - x}{2} \tag{A.8}$$

The iso-BA curve for $BA = b$ is the straight line

$$y = x + 2b - 1 = x + c \tag{A.9}$$

where we write $c = 2b - 1$ for mathematical convenience.

Let us take $c \geq 0$ (which corresponds to $b \geq \frac{1}{2}$, so the straight line is above the bisector.

The value of AUC is equal to 1 minus the area of the triangle of the ROC space above and to the left of the iso-BA curve. The iso-BA curve intersects the y-axis at $(0, c)$ and the horizontal $y = 1$ line at $(1 - c, 1)$. So

$$AUC(c) = 1 - \frac{(1-c)^2}{2} = \frac{1 + 2c - c^2}{2} \tag{A.10}$$

It is immediate to prove that $AUC(c)$ is a monotonically increasing function of c (and therefore of b) and that $AUC(0) = \frac{1}{2}$ and $AUC(1) = 1$.

We now compute $RRA(c, \rho)$. Four cases are possible, depending on the mutual positions of the iso-BA curve and the borders of the RoI. These mutual positions are described by two predicates. Whenever $\rho \geq c$, the intercept of the iso-BA curve is below or at least on the horizontal $y = \rho$ straight line, and above otherwise. Whenever $1 - c \geq \rho$, the value of x at which the iso-BA curve intersects the horizontal $y = 1$ straight line is to the right or at least on the vertical $x = \rho$ straight line, and to the left otherwise. We now discuss all four possible combinations of these two predicates.

Case 1: $\rho \geq c$ and $1 - c \geq \rho$. By summing these two inequalities, we have $1 - c \geq c$, i.e., this case can occur only if $c \leq \frac{1}{2}$. The relevant area under the curve is the area of the right triangle with vertices $(\rho - c, \rho)$, $(\rho, \rho + c)$, and (ρ, ρ). So,

$$RRA(c, \rho) = \frac{(\rho - (\rho - c)) + ((\rho + c) - \rho)}{2\rho(1-\rho)} = \frac{c^2}{2\rho(1-\rho)} \tag{A.11}$$

Case 2: $c \geq \rho$ and $1 - c \geq \rho$. By summing those two inequalities, we have $1 \geq 2\rho$, i.e., this case can occur only if $\rho \leq \frac{1}{2}$. The relevant area under the curve is the area of the trapezoid with vertices $(0, \rho)$, $(0, c)$, $(\rho, \rho + c)$, and (ρ, ρ). So,

$$RRA(c, \rho) = \frac{((c - \rho) + ((\rho + c) - \rho))\rho}{2\rho(1-\rho)} = \frac{2c - \rho}{2(1-\rho)} \tag{A.12}$$

Case 3: $\rho \geq c$ and $\rho \geq 1 - c$. By summing those two inequalities, it can be found that this case can occur only if $\rho \geq \frac{1}{2}$. The relevant area under the curve is the area of the trapezoid with vertices $(\rho - c, \rho)$, $(1 - c, 1)$, $(\rho, 1)$, and (ρ, ρ). So,

$$\begin{aligned} RRA(c, \rho) &= \frac{((\rho - (\rho - c)) + (\rho - (1 - c)))(1-\rho)}{2\rho(1-\rho)} \\ &= \frac{2c + \rho - 1}{2\rho} = \frac{2c - 1}{2\rho} + \frac{1}{2} \end{aligned} \tag{A.13}$$

Case 4: $\rho \leq c$ and $1 - c \leq \rho$. By summing those two inequalities, this case can occur only if $c \geq \frac{1}{2}$. The relevant area under the curve is the area of the RoI minus the right triangle with vertices $(0, c)$, $(1, 1)$, and $(1 - c, 1)$. So,

$$RRA(c, \rho) = \frac{\rho(1-\rho) - \frac{(1-c)^2}{2}}{\rho(1-\rho)} = 1 - \frac{(1-c)^2}{2\rho(1-\rho)} \tag{A.14}$$

A.3. GMean

$$GMean = \sqrt{(1 - FPR)TPR} = \sqrt{(1-x)y} \tag{A.15}$$

The iso-$GMean$ curve for $GMean = \sqrt{h}$ is the rectangular hyperbola

$$y = \frac{h}{1-x} \tag{A.16}$$

The horizontal asymptote of the hyperbola is $y = 0$ and the vertical asymptote is $x = 1$.

The hyperbola intersects the y-axis at $(0, h)$ and the straight line $y = 1$ at $(1 - h, 1)$.

We require that the part of the branch of the hyperbola in the ROC space be not below the bisector. This occurs when $Gmean \geq \frac{1}{2}$.

The value of $AUC(h)$ is

$$AUC(h) = \int_0^{1-h} \frac{h}{1-x} dx + 1 - (1-h) = -h \log h + h = h(1 - \log h) \tag{A.17}$$

$AUC(h)$ is a monotonically increasing function of h, with $AUC(\frac{1}{4}) = \frac{1}{4}(1 - 2\log\frac{1}{2})$ and $AUC(1) = 1$.

We now compute $RRA(h, \rho)$. Just like for Balanced Accuracy, four cases are possible, depending on the mutual positions of the iso-$GMean$ curve and the borders of the RoI. These mutual positions are described by two predicates. Whenever $\rho \geq h$, the intercept of the iso-$GMean$ curve is below or at least on the horizontal $y = \rho$ straight line, and above otherwise. Whenever $1 - h \geq \rho$, the value of x at which the iso-$Gmean$ curve intersects the horizontal $y = 1$ straight line is to the right or at least on the vertical $x = \rho$ straight line, and to the left otherwise. We now discuss all four possible combinations of these two predicates.

Case 1: $\rho \geq h$ and $1 - h \geq \rho$. By summing those two inequalities, this case can occur only if $h \leq \frac{1}{2}$. The horizontal $y = \rho$ straight line intersects the hyperbola at $x = 1 - \frac{h}{\rho}$. The area under the curve between $x = 1 - \frac{h}{\rho}$ and $x = \rho$ is

$$\int_{1-\frac{h}{\rho}}^{\rho} \frac{h}{1-x} dx = h(\log h - \log\rho - \log(1-\rho)) \tag{A.18}$$

The area of the rectangle under the segment between $(1 - \frac{h}{\rho}, \rho)$ and (ρ, ρ) is $\rho^2 - \rho + h$, so

$$RRA(h, \rho) = \frac{h(\log h - \log\rho - \log(1-\rho)) - h + \rho(1-\rho)}{\rho(1-\rho)} \tag{A.19}$$

$$= h\frac{\log h - \log\rho - \log(1-\rho) - 1}{\rho(1-\rho)} + 1 \tag{A.20}$$

Case 2: $h \geq \rho$ and $1-h \geq \rho$. By summing those two inequalities, this case can occur only if $\rho \leq \frac{1}{2}$. The area under the curve between $x = 0$ and $x = \rho$ is

$$\int_0^{\rho} \frac{h}{1-x} dx = -h \log(1-\rho) \tag{A.21}$$

Since the area below the segment between $(0, \rho)$ and (ρ, ρ) is ρ^2, we have

$$RRA(h, \rho) = \frac{-h \log(1-\rho) - \rho^2}{\rho(1-\rho)} \tag{A.22}$$

Case 3: $\rho \geq h$ and $\rho \geq 1-h$. By summing these two inequalities, this case can occur only if $\rho \geq \frac{1}{2}$. The area under the curve between $x = 1 - \frac{h}{\rho}$ and $x = 1-h$ is

$$\int_{1-\frac{h}{\rho}}^{1-h} \frac{h}{1-x} dx = -h \log \rho \tag{A.23}$$

The area in the rectangle with vertices $(1-h, 0)$, $(1-h, 1)$, $(\rho, 1)$, $(\rho, 0)$ is $(1-\rho)(\rho-(1-h))$, so, since the area below the segment between $(1-\frac{h}{\rho}, \rho)$ and (ρ, ρ) is $\rho^2 - \rho + h$, we have

$$RRA(h, \rho) = \frac{-h \log \rho - (1-\rho)^2}{\rho(1-\rho)} \tag{A.24}$$

Case 4: $\rho \leq h$ and $1-h \leq \rho$. By summing those two inequalities, it can be found that this case can occur only if $h \geq \frac{1}{2}$. The relevant area under the curve is $AUC(h)$ minus the parts of the ROC space outside the region of interest, so

$$RRA(h, \rho) = \frac{h(1-\log h) - (1-\rho(1-\rho))}{\rho(1-\rho)} = \frac{h(1-\log h) - 1}{\rho(1-\rho)} + 1 \tag{A.25}$$

A.4. GM

$$GM = \frac{2}{\frac{1}{1-x} + \frac{1}{y}} = \frac{2(1-x)y}{1-x+y} \tag{A.26}$$

The iso-GM curve for $GM = 2h$ is the rectangular hyperbola

$$xy - hx - (1-h)y + h = 0 \tag{A.27}$$

Since $GM \leq 1$, it is $h \leq \frac{1}{2}$. The vertical asymptote of the hyperbola is $x = 1-h$ and the horizontal asymptote is $y = h$, so the center of the hyperbola is point $(1-h, h)$.

Regardless of the value of h, one of the branches of the hyperbola goes through point $(1, 0)$, which is the only point of the ROC space that branch goes through. Since $h \leq \frac{1}{2}$, the other branch goes through several points in the ROC space if $h < \frac{1}{2}$ and only point $(0, 1)$ if $h = \frac{1}{2}$.

This branch of the hyperbola is an "increasing function" of h. More formally, the equation of the hyperbola is

$$y = h \frac{1-x}{1-x-h} \tag{A.28}$$

If $h'' > h'$, we have, for all values of x

$$y'' = h'' \frac{1-x}{1-x-h''} > y' = h' \frac{1-x}{1-x-h'} \tag{A.29}$$

The hyperbola intercepts the y-axis for $y = \frac{h}{1-h}$ and the horizontal straight line $y = 1$ for $x = \frac{1-2h}{1-h}$. It also intercepts the bisector at two points with x equal to $\frac{1}{2} \pm \sqrt{\frac{1}{4} - h}$ if $h < \frac{1}{4}$. If $h > \frac{1}{4}$ the hyperbola is above the bisector and, if $h = \frac{1}{4}$, the hyperbola intercepts the bisector for $x = \frac{1}{2}$.

The value of $AUC(h)$ is

$$AUC(h) = h \int_0^{\frac{1-2h}{1-h}} \frac{1-x}{1-x-h} dx + 1 - \frac{1-2h}{1-h} =$$

$$= h \int_0^{\frac{1-2h}{1-h}} \frac{1-x-h}{1-x-h} dx + h^2 \int_0^{\frac{1-2h}{1-h}} \frac{1}{1-x-h} dx + 1 - \frac{1-2h}{1-h} =$$

$$= 2h + 2h^2(\log(1-h) - \log h) \tag{A.30}$$

$AUC(h)$ is an increasing function of h, with minimum in the degenerate case $h = 0$, in which the hyperbola degenerates into the x- and the x-axis and the vertical straight line $x = 1$, so $AUC(0) = 0$. The maximum is obtained for $h = \frac{1}{2}$, with $AUC(\frac{1}{2}) = 1$.

We now compute $RRA(h, \rho)$. Just like for Balanced Accuracy and the $GMean$, four cases are possible, depending on the mutual positions of the iso-GM curve and the borders of the RoI. These mutual positions are described by two predicates. Whenever $\rho \geq \frac{h}{1-h}$, the intercept of the iso-GM curve is below or at least on the horizontal $y = \rho$ straight line, and above otherwise. Whenever $\frac{1-2h}{1-h} \geq \rho$, the value of x at which the iso-GM curve intersects the horizontal $y = 1$ straight line is to the right or at least on the vertical $x = \rho$ straight line, and to the left otherwise. We now discuss all four possible combinations of these two predicates.

Case 1: $\rho \geq \frac{h}{1-h}$ and $\frac{1-2h}{1-h} \geq \rho$. By summing those two inequalities, this case can occur only if $h \leq \frac{1}{3}$. As a consequence, the hyperbola intersects y-axis at a point with $y \leq \frac{1}{2}$ and the horizontal straight line $y = 1$ at a point with $x \leq \frac{1}{2}$. The horizontal $y = \rho$ straight line intersects the hyperbola at $x = 1 - \frac{h\rho}{\rho-h}$. The area under the curve between $x = 1 - \frac{h\rho}{\rho-h}$ and $x = \rho$ is

$$h \int_{1-\frac{h\rho}{\rho-h}}^{\rho} \frac{1-x}{1-x-h} dx = h \frac{\rho^2 - \rho + h}{\rho - h} + h^2(2\log h - \log(\rho-h) - \log(1-\rho-h)) \tag{A.31}$$

The area of the rectangle under the segment between $(1 - \frac{h\rho}{\rho-h}, \rho)$ and (ρ, ρ) is $\rho \frac{\rho^2-\rho+h}{\rho-h}$, so

$$RRA(h, \rho) = \frac{h^2(2\log h - \log(\rho-h) - \log(1-\rho-h)) - h + \rho(1-\rho)}{\rho(1-\rho)} \tag{A.32}$$

$$= h \frac{h(2\log h - \log(\rho-h) - \log(1-\rho-h)) - 1}{\rho(1-\rho)} + 1 \tag{A.33}$$

Case 2: $\frac{h}{1-h} \geq \rho$ and $\frac{1-2h}{1-h} \geq \rho$. By summing those two inequalities, this case can occur only if $\rho \leq \frac{1}{2}$. The area under the curve between $x = 0$ and $x = \rho$ is

$$h \int_0^{\rho} \frac{1-x}{1-x-h} dx = h\rho + h^2(\log(1-h) - \log(1-\rho-h)) \tag{A.34}$$

Since the area below the segment between $(0, \rho)$ and (ρ, ρ) is ρ^2, we have

$$RRA(h, \rho) = \frac{h\rho + h^2(\log(1-h) - \log(1-\rho-h)) - \rho^2}{\rho(1-\rho)} \tag{A.35}$$

Case 3: $\rho \geq \frac{h}{1-h}$ and $\rho \geq \frac{1-2h}{1-h}$. By summing these two inequalities, this case can occur only if $\rho \geq \frac{1}{2}$. The area under the curve between $x = 1 - \frac{h\rho}{\rho-h}$ and $x = \frac{1-2h}{1-h}$ is

$$h \int_{1-\frac{h\rho}{\rho-h}}^{\frac{1-2h}{1-h}} \frac{1-x}{1-x-h} dx = h^3 \frac{1-\rho}{(\rho-h)(1-h)} + h^2 \log(1-h) - h^2 \log(1-\rho) \tag{A.36}$$

The area below $y = \rho$ between $x = 1 - \frac{h\rho}{\rho-h}$ and $x = \frac{1-2h}{1-h}$ is

$$\rho h^2 \frac{1-\rho}{(\rho-h)(1-h)} \tag{A.37}$$

So, the area above $y = \rho$ between $x = 1 - \frac{h\rho}{\rho-h}$ and $x = \frac{1-2h}{1-h}$ is

$$h^3 \frac{1-\rho}{(\rho-h)(1-h)} - \rho h^2 \frac{1-\rho}{(\rho-h)(1-h)} + h^2 \log(1-h) - h^2 \log(1-\rho) \tag{A.38}$$

$$= -h^2 \frac{1-\rho}{1-h} + h^2 \log \frac{1-h}{\rho-h} \tag{A.39}$$

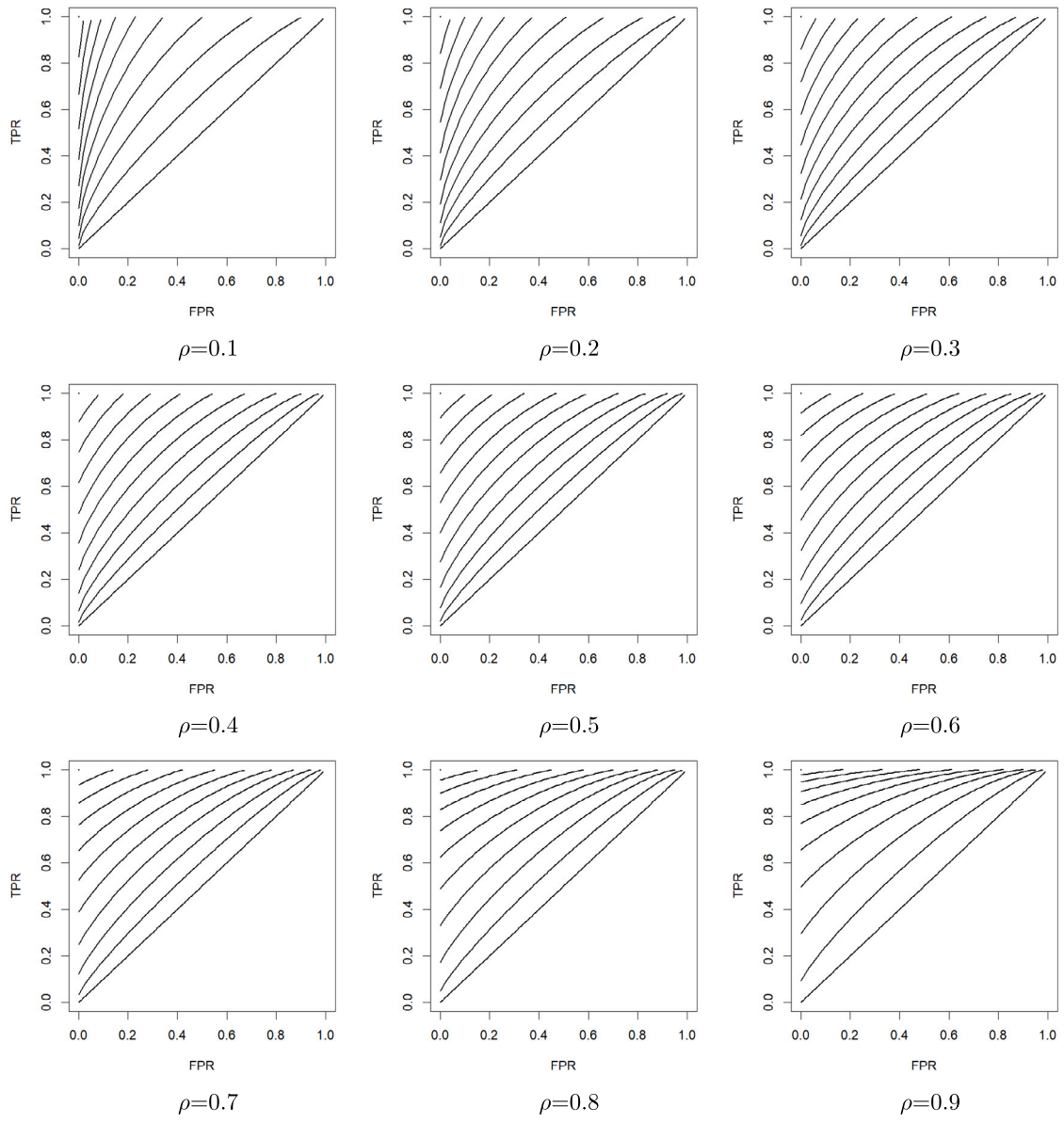


Fig. C.22. Iso-ϕ curves, for different values of ρ.

The area above $y = \rho$ between $x = \frac{1-2h}{1-h}$ and $x = \rho$ is

$$(1-\rho)\frac{h}{1-h} - (1-\rho)^2 \tag{A.40}$$

So, the value of RRA is

$$RRA(h,\rho) = \frac{-h^2\frac{1-\rho}{1-h} + h^2\log\frac{1-h}{\rho-h} + (1-\rho)\frac{h}{1-h} - (1-\rho)^2}{\rho(1-\rho)} \tag{A.41}$$

Case 4: $\rho \leq \frac{h}{1-h}$ and $\frac{1-2h}{1-h} \leq \rho$. By summing those two inequalities, this case can occur only if $h \geq \frac{1}{3}$. As a consequence, the hyperbola intersects y-axis at a point with $y \geq \frac{1}{2}$ and the horizontal straight line $y = 1$ at a point with $x \geq \frac{1}{2}$. The relevant area under the curve is $AUC(h)$ minus the parts of the ROC space outside the region of interest, so

$$RRA(h,\rho) = \frac{2h + 2h^2(\log(1-h) - \log h) - \rho^2 - 1 + \rho}{\rho(1-\rho)} \tag{A.42}$$

$$= \frac{2h + 2h^2(\log(1-h) - \log h) - 1}{\rho(1-\rho)} + 1 \tag{A.43}$$

A.5. DtH

$$DtH(r) = \sqrt{\frac{x^2 + (1-y)^2}{2}} \tag{A.44}$$

What needs to be maximized is actually $1 - DtH(r)$.

The iso-DtH curve for $DtH\sqrt{2} = r$ is the arc of circumference

$$y = 1 - \sqrt{r^2 - x^2} \tag{A.45}$$

for x between $x = 0$ and $x = r$.

We require that the iso-DtH curve be not below the bisector. This occurs when $r \leq \frac{\sqrt{2}}{2}$.

The value of $AUC(f)$ is

$$AUC(r) = 1 - \frac{\pi}{4}r^2 \tag{A.46}$$

$AUC(r)$ is a monotonically decreasing function of r, with $AUC(0) = 1$ and $AUC(\frac{1}{2}) = 1 - \frac{\pi}{8}$

We now compute $RRA(r,\rho)$. Just like for Balanced Accuracy and the GMean, four cases are possible, depending on the mutual positions of

the iso-DtH curve and the borders of the RoI. These mutual positions are described by two predicates. Whenever $\rho \geq 1 - r$, the intercept of the iso-DtH curve is below or at least on the horizontal $y = \rho$ straight line, and above otherwise. Whenever $r \geq \rho$, the value of x at which the iso-DtH curve intersects the horizontal $y = 1$ straight line is to the right or at least on the vertical $x = \rho$ straight line, and to the left otherwise. We now discuss all four possible combinations of these two predicates.

Case 1: $\rho \geq 1 - r$ and $r \geq \rho$. By summing those two inequalities, this case can occur only if $r \geq \frac{1}{2}$. Let α be the angle between the horizontal straight line $y = 1$ and the segment between $(0, 1)$ and the intersection between the circumference and the vertical straight line $x = \rho$. So, $\rho = r\cos\alpha$. Let θ be the angle between the vertical straight line $x = 0$ and the segment between $(0, 1)$ and the intersection between the circumference and the horizontal straight line $y = \rho$. So, $1 - \rho = r\cos\theta$. The relevant area under the curve is

$$\rho(1-\rho) - \frac{(1-\rho)r\sin\theta}{2} - \frac{\rho r\sin\alpha}{2} - (\frac{\pi}{2} - \theta - \alpha)\frac{r^2}{2} \tag{A.47}$$

So

$$RRA(r,\rho) = 1 - \frac{1}{2}\frac{r\sin\theta}{\rho} - \frac{1}{2}\frac{r\sin\alpha}{1-\rho} - (\frac{\pi}{2} - \theta - \alpha)\frac{r^2}{2}\frac{1}{\rho(1-\rho)} \tag{A.48}$$

Case 2: $1 - r \geq \rho$ and $r \geq \rho$. By summing those two inequalities, this case can occur only if $\rho \leq \frac{1}{2}$. Let θ denote the angle between the y-axis and the segment connecting point $(0, 1)$ to the intersection point of the circumference and the vertical straight line $x = \rho$. We can also denote $\rho = r\sin(\theta)$. The relevant area under the curve between $x = 0$ and $x = \rho$ is

$$\rho - \frac{\rho r\cos\theta}{2} - \frac{r^2\theta}{2} - \rho^2 \tag{A.49}$$

So

$$RRA(g,\rho) = 1 - \frac{1}{2}\frac{r\cos\theta}{1-\rho} - \frac{r^2}{2}\frac{\theta}{\rho(1-\rho)} \tag{A.50}$$

$$= 1 - \frac{1}{2}\frac{\sqrt{r^2-\rho^2}}{1-\rho} - \frac{r^2}{2}\frac{\arcsin\frac{\rho}{r}}{\rho(1-\rho)} \tag{A.51}$$

$$= 1 - \frac{r}{2}\frac{\cos\theta}{1-r\sin\theta} - \frac{r}{2}\frac{\theta}{\sin\theta(1-r\sin\theta)} \tag{A.52}$$

Case 3: $\rho \geq 1 - r$ and $\rho \geq r$. By summing these two inequalities, this case can occur only if $\rho \geq \frac{1}{2}$. The relevant area under the curve is

$$(1-\rho)r - \frac{(1-\rho)r\cos\theta}{2} - \theta\frac{r^2}{2} + (\rho - r)(1-\rho) =$$

$$= -\frac{r}{2}(1-\rho)\cos\theta - \theta\frac{r^2}{2} + \rho(1-\rho) \tag{A.53}$$

So,

$$RRA(r,\rho) = -\frac{r}{2}\frac{\cos\theta}{\rho} - \frac{r^2}{2}\theta\frac{1}{\rho(1-\rho)} + 1 \tag{A.54}$$

Case 4: $\rho \leq 1 - r$ and $r \leq \rho$. By summing those two inequalities, it can be found that this case can occur only if $r \leq \frac{1}{2}$. The relevant area under the curve is the area of the region of interest minus the are of the quarter circle with radius r, so

$$RRA(r,\rho) = 1 - \frac{\pi}{4}\frac{r^2}{\rho(1-\rho)} \tag{A.55}$$

Appendix B. Models

The models used in Section 5 were derived as follows.

All models were derived via Binary Logistic Regression (BLR). Only models that passed a few tests (including sign test, Wald test and Likelihood Ratio Test) were selected. Even though the relationships studied in this paper are valid for any model, it seemed reasonable to limit the examples reported to "good enough" models.

The models based on the J&M datasets used as independent variables any combination of RFC (response for a class), WMC (weighted methods per class), CBO (coupling between objects) and NPM (number of public methods). The models based on the NASA datasets were built using two subsets of independent variables: 1) any combination of Num. operands, Num. operators, Num. unique operands and Num. unique operators (these are the models named NASA1 in Section 5.2), and (2) Branch count, Cyclomatic complexity, Essential complexity and LOC executable (NASA2 in Section 5.2).

In Section 5.3, for each NASA dataset we selected the best model from the union of the NASA1 and NASA2 models.

Appendix C. Additional graphs

Iso-ϕ curves, for multiple values of ρ are shown in Fig. C.22.

Data availability

The datasets we used are well-known ones and in the public domain.

References

[1] D. McClish, Analyzing a portion of the ROC curve, Med. Decis. Mak. Int. J. Soc. Med. Decis. Mak. 9 (1989) 190–195, http://dx.doi.org/10.1177/0272989X8900900307.

[2] Y. Jiang, C.E. Metz, R.M. Nishikawa, A receiver operating characteristic partial area index for highly sensitive diagnostic tests, Radiol. 201 (3) (1996) 745–750.

[3] H. Yang, K. Lu, X. Lyu, F. Hu, Two-way partial AUC and its properties, Stat. Methods Med. Res. 28 (1) (2019) 184–195.

[4] S. Morasca, L. Lavazza, On the assessment of software defect prediction models via ROC curves, Empir. Softw. Eng. 25 (5) (2020) 3977–4019.

[5] L. Lavazza, S. Morasca, G. Rotoloni, On the reliability of the area under the ROC curve in empirical software engineering, in: Proceedings of the 24th International Conference on Evaluation and Assessment in Software Engineering (EASE). Association for Computing Machinery, ACM, 2023.

[6] L. Lavazza, S. Morasca, G. Rotoloni, An experience in the evaluation of fault prediction, in: International Conference on Product-Focused Software Process Improvement, Springer, 2023, pp. 323–338.

[7] H. Alsolai, M. Roper, A systematic literature review of machine learning techniques for software maintainability prediction, Inf. Softw. Technol. 119 (2020) 106214.

[8] C. Watson, N. Cooper, D.N. Palacio, K. Moran, D. Poshyvanyk, A systematic literature review on the use of deep learning in software engineering research, ACM Trans. Softw. Eng. Methodol. (TOSEM) 31 (2) (2022) 1–58.

[9] R. Moussa, F. Sarro, On the use of evaluation measures for defect prediction studies, in: Proceedings of the 31st ACM SIGSOFT International Symposium on Software Testing and Analysis, ISSTA, ACM, 2022.

[10] D. Chicco, G. Jurman, The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation, BMC Genomics 21 (1) (2020) 1–13.

[11] L. Lavazza, S. Morasca, Common problems with the usage of F-measure and accuracy metrics in medical research, IEEE Access (2023).

[12] T. Fawcett, An introduction to ROC analysis, Pattern Recogn. Lett. 27 (8) (2006) 861–874, http://dx.doi.org/10.1016/j.patrec.2005.10.010.

[13] E. Arisholm, L.C. Briand, M. Fuglerud, Data mining techniques for building fault-proneness models in telecom java software, in: Software Reliability, 2007. ISSRE'07. The 18th IEEE International Symposium on, IEEE, 2007, pp. 215–224.

[14] S. Beecham, T. Hall, D. Bowes, D. Gray, S. Counsell, S. Black, A Systematic Review of Fault Prediction Approaches Used in Software Engineering, Tech. Rep., Technical Report Lero-TR-2010-04, Lero, 2010.

[15] C. Catal, Performance evaluation metrics for software fault prediction studies, Acta Polytech. Hung. 9 (4) (2012) 193–206.

[16] C. Catal, B. Diri, A systematic review of software fault prediction studies, Expert Syst. Appl. 36 (4) (2009) 7346–7354.

[17] Y. Singh, A. Kaur, R. Malhotra, Empirical validation of object-oriented metrics for predicting fault proneness models, Softw. Qual. J. 18 (1) (2010) 3.

[18] D.W. Hosmer Jr., S. Lemeshow, R.X. Sturdivant, Applied Logistic Regression, John Wiley & Sons, 2013.

[19] M. Jureczko, L. Madeyski, Towards identifying software project clusters with regard to defect prediction, in: Proceedings of the 6th International Conference on Predictive Models in Software Engineering, 2010, pp. 1–10.

[20] H. Cramér, Mathematical Methods of Statistics, Princeton University Press, 1946.

[21] G.U. Yule, On the methods of measuring association between two attributes, J. R. Stat. Soc. 75 (6) (1912) 579–652.

[22] B.W. Matthews, Comparison of the predicted and observed secondary structure of T4 phage lysozyme, Biochim. Biophys. Acta (BBA) Protein Struct. 405 (2) (1975) 442–451.
[23] J. Cohen, Statistical Power Analysis for the Behavioral Sciences Lawrence Earlbaum Associates, Routledge, New York, NY, USA, 1988, pp. 20–26.
[24] J. Jiarpakdee, C. Tantithamthavorn, A.E. Hassan, The impact of correlated metrics on the interpretation of defect models, IEEE Trans. Softw. Eng. 47 (2) (2019) 320–331.
[25] A.P. Bradley, The use of the area under the ROC curve in the evaluation of machine learning algorithms, Pattern Recognit. 30 (7) (1997) 1145–1159.
[26] C. O'Reilly, T. Nielsen, Revisiting the ROC curve for diagnostic applications with an unbalanced class distribution, in: 2013 8th International Workshop on Systems, Signal Processing and their Applications, WoSSPA, IEEE, 2013, pp. 413–420.
[27] A. Jiménez-Valverde, Threshold-dependence as a desirable attribute for discrimination assessment: Implications for the evaluation of species distribution models, Biodivers. Conserv. 23 (2014) 369–385.
[28] D.J. Hand, Measuring classifier performance: A coherent alternative to the area under the ROC curve, Mach. Learn. 77 (1) (2009) 103–123, http://dx.doi.org/10.1007/s10994-009-5119-5.
[29] D. Chicco, G. Jurman, The matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification, BioData Min. 16 (1) (2023) 1–23.
[30] P.A. Flach, The geometry of ROC space: Understanding machine learning metrics through ROC isometrics, in: Machine Learning, Proceedings of the Twentieth International Conference (ICML 2003), August 21-24, 2003, Washington, DC, USA, 2003, pp. 194–201, URL http://www.aaai.org/Library/ICML/2003/icml03-028.php.
[31] J. Hernández-Orallo, P. Flach, C. Ferri, ROC curves in cost space, Mach. Learn. 93 (1) (2013) 71–91, http://dx.doi.org/10.1007/s10994-013-5328-9.
[32] N. Lachiche, P.A. Flach, Improving accuracy and cost of two-class and multi-class probabilistic classifiers using ROC curves, in: Proceedings of the 20th International Conference on Machine Learning, ICML-03, 2003, pp. 416–423.
[33] H. Chen, W. Liu, D. Gao, X. Peng, W. Zhao, Personalized defect prediction for individual source files, Comput. Sci. 44 (4) (2017) 90–95, http://dx.doi.org/10.11896/j.issn.1002-137X.2017.04.020.
[34] J. Çarka, M. Esposito, D. Falessi, On effort-aware metrics for defect prediction, Empir. Softw. Eng. 27 (6) (2022) 152.