Article:
Carrington, A.M., Manuel, D.G., Fieguth, P. et al. (2022) Deep ROC analysis and AUC as balanced average accuracy, for improved classifier selection, audit and explanation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45 (1). pp. 329-341. ISSN: 0162-8828

https://doi.org/10.1109/tpami.2022.3145392

Deep ROC Analysis and AUC as Balanced Average Accuracy, for Improved Classifier Selection, Audit and Explanation

André M. Carrington, Douglas G. Manuel, Paul W. Fieguth, Tim Ramsay, Venet Osmani, Bernhard Wernly, Carol Bennett, Steven Hawken, Olivia Magwood, Yusuf Sheikh, Matthew McInnes and Andreas Holzinger, *Senior Member, IEEE*

Abstract—Optimal performance is desired for decision-making in any field with binary classifiers and diagnostic tests, however common performance measures lack depth in information. The area under the receiver operating characteristic curve (AUC) and the area under the precision recall curve are too general because they evaluate all decision thresholds including unrealistic ones. Conversely, accuracy, sensitivity, specificity, positive predictive value and the F1 score are too specific—they are measured at a single threshold that is optimal for some instances, but not others, which is not equitable. In between both approaches, we propose deep ROC analysis to measure performance in multiple groups of predicted risk (like calibration), or groups of true positive rate or false positive rate. In each group, we measure the group AUC (properly), normalized group AUC, and averages of: sensitivity, specificity, positive and negative predictive value, and likelihood ratio positive and negative. The measurements can be compared between groups, to whole measures, to point measures and between models. We also provide a new interpretation of AUC in whole or part, as balanced average accuracy, relevant to individuals instead of pairs. We evaluate models in three case studies using our method and Python toolkit and confirm its utility.

Index Terms—Performance and Reliability, Performance Analysis and Design Aids, Diagnostic Testing, Artificial Intelligence, ROC, AUC, C Statistic, Explainable AI, Equity, Audit

1 Introduction

Common measures of performance for binary classifiers and binary diagnostic tests are either too general or too specific–they lack depth in information to ensure optimal model selection, equity (in audit) and robustness.

The area under the curve [1] (AUC) in a receiver operating characteristic (ROC) plot [2] is too general because it measures all decision thresholds including unrealistic ones [3], [4], [5], [6], [7], [8], [9] (Figure 1). Whereas measures at a single threshold or ROC point, such as accuracy, sensitivity, specificity, or positive predictive value [10], are too specific—they are optimal for some instances or patients, but not others, reflecting a specific choice of threshold and misclassification costs [8], [11], [12], [13], [14].

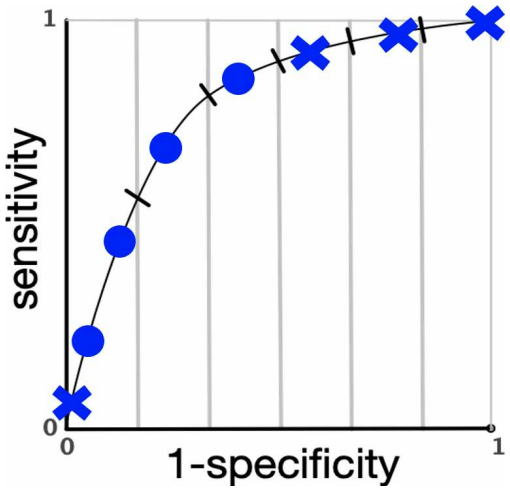


Fig. 1: AUC includes all thresholds, including unrealistic or undesirable ones ('X'). Measures at a single threshold (circle), like sensitivity, are optimal for some patients but not others. Deep ROC avoids both problems with measures for a range of thresholds in each of several groups.

- *A.M. Carrington is with the Ottawa Hospital and Region Imaging Associates and the Department of Systems Design Engineering, University of Waterloo, Canada. E-mail: acarrington@toh.ca*
- *D.G. Manuel is with the Ottawa Hospital Research Institute, the University of Ottawa, the Institute for Clinical Evaluative Sciences and the Bruyère Research Institute. E-mail: dmanuel@ohri.ca*
- *P. Fieguth is with the Department of Systems Design Engineering and Faculty of Engineering, University of Waterloo, Canada. E-mail: paul.fieguth@uwaterloo.ca*
- *V. Osmani is with the e-health group at Fondazione Bruno Kessler Research Institute and the department of Psychology and Cognitive Science at University of Trento, Italy. E-mail: vosmani@fbk.eu*
- *B. Wernly is with the Department of Cardiology, Paracelsus Medical University of Salzburg, Salzburg, Austria. E-mail: b.wernly@salk.at*
- *S. Hawken is with the Ottawa Hospital Research Institute and the University of Ottawa. E-mail: shawken@ohri.ca*
- *M. McInnes is with the Ottawa Hospital Research Institute and the University of Ottawa. E-mail: mmcinnes@toh.on.ca*
- *T. Ramsay is with the Ottawa Hospital Research Institute and the University of Ottawa. E-mail: tramsay@ohri.ca*
- *C. Bennett is with the Ottawa Hospital Research Institute and the Institute of Clinical Evaluative Sciences, Ottawa, Canada. E-mail: cbennett@ohri.ca*
- *O. Magwood is with the Bruyere Research Institute and is a doctoral student at the University of Ottawa, Canada. E-mail: omagwood@bruyere.org*
- *Y. Sheikh is with the Ottawa Hospital Research Institute, Canada. E-mail: ysheikh@ohri.ca*
- *A. Holzinger is with the Alberta Machine Intelligence Institute, University of Alberta, Canada and the Human-Centered AI Lab, Medical University Graz, Austria. E-mail: andreas.holzinger@medunigraz.at.*

- *Corresponding author: Andreas Holzinger.*

TABLE 1: Consider a binary classifier or diagnostic test for data with 30% prevalence. Suppose the group with high predicted risk (or high probability) is most relevant. AUC, as a global measure, obscures all of the group measures (those with subscript i). The high risk group's $AUCn_1$ or balanced average accuracy of 85% is higher than AUC, but its average sensitivity $\overline{se}_1$ is significantly lower at 67%. Average sensitivity always increases to the right, while average specificity ($\overline{sp}_i$) always increases to the left, hence the need for $AUCn_i$ to make comparisons.

FPR:	[0,1]	[0,.33]	[.33,.67]	[.67,1]
Predicted risk:	All	High	Med	Low
Group index i:	0	1	2	3
AUC	82%			
$AUCn_i$	82%	85%	81%	76%
Avg sensitivity $\overline{se}_i$	82%	67%	84%	94%
Avg specificity $\overline{sp}_i$	82%	93%	67%	40%

Measures at a single point (point measures) hide information about how the classifier would perform if the threshold were tuned to different subgroups, different settings where they are used, or personalized to an individual's needs.

Point measures for imbalanced data [15] are also too specific: the F_1 score [10], [16], balanced accuracy [10], the geo-mean [10], [17], and Matthews' Correlation Coefficient [18]. While area measures for imbalanced data are also too general: the area under the precision recall curve (AUPRC) a.k.a. average precision (AP) [19] and the area under the PR Gain curve [20]. Area measures lack information about the distribution of performance over the curve [3].

In between area measures that are too general, and point measures that are too specific, there is a gap in information. ROC plots visually provide information to fill part of the gap, but we propose a better quantitative alternative.

ROC analysis (ROC plots and standard measures) are the standard tool in model selection and evaluation for two-class classification problems [21] with ongoing work [22], [23] and extensions [24].

ROC analysis is one of multiple tools we use to understand and explain models and their test results, as a goal of explainable artificial intelligence (xAI) [25], [26]. In the context of "black-box" methods [27] xAI is important, as well as understanding causality and causability [25], [28].

Whereas ROC plots provide visual information in gaps between measures, we propose **deep ROC analysis** for quantitative measures of a model in multiple parts (partial areas) that span the ROC curve (and plot). We provide an overview of the method in the next section.

Deep ROC analysis supports algorithmic audit to examine how classifiers treat groups of individuals at different levels of predicted risk or probability, e.g., for equity [29] if that is the goal. It also supports robustness since it identifies where a model may fall short in one or more parts or groups, per failure mode effects analysis (FMEA) [29]. We do not cover groups by demographics or other conditions in this paper, although our approach can be extended to that using discrete measures such as the partial C statistic [9].

The contributions of our paper are as follows:

1) We show how to use three pre-test measures effectively in tandem for analysis and audit along with post-test measures.
2) We provide a new interpretation of AUC in part or whole, for individuals, instead of pairs.
3) Our method sometimes identifies a partial area measure as a focus for re-optimization or re-calibration of models.

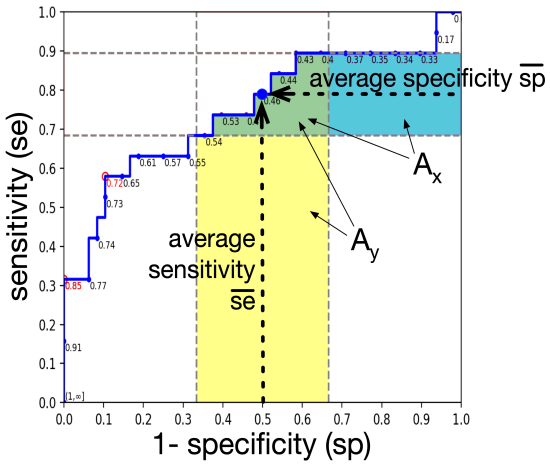


Fig. 2: For part of an ROC curve, the area that represents the model's performance, is the average of the vertical and horizontal areas (A_y and A_x which both include the green area). This is known as the concordant partial AUC, which we denote AUC_i, and it is less than 1. It can be normalized to $[0, 1]$ as $AUCn_i$.

The first contribution is neither obvious nor simple because:

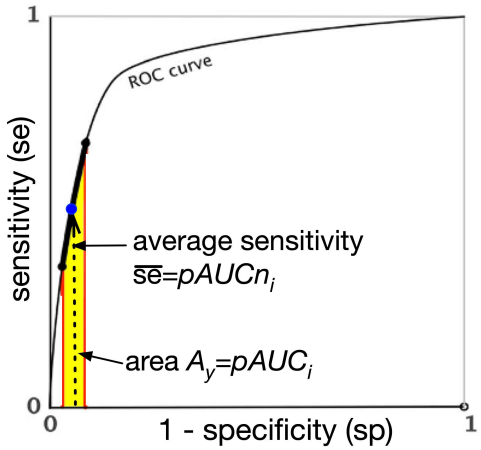
- There is a cacophony (or glut) of performance measures, some of which have flaws.
- Only an avid reader would find and know the three pre-test measures are averages of familiar concepts: sensitivity, specificity and AUC.
- Of the three pre-test measures we use:
  - one was misrepresented [6]
  - another [30] lacked two free boundaries until a recent alternative [9], and
  - the third [9] was normalized implicitly and differently from another explicit measure.
- Several alternative measures (based on the U statistic) do not handle ties in score properly, and ROC plotting functions do not provide all of the implicit points (one for every instance) as necessary to properly compute averages in a part or group. Our python toolkit handles that, and interpolation for any boundaries.

In the sections that follow we review related work, background, our method, two case studies, limitations, conclusions and future work.

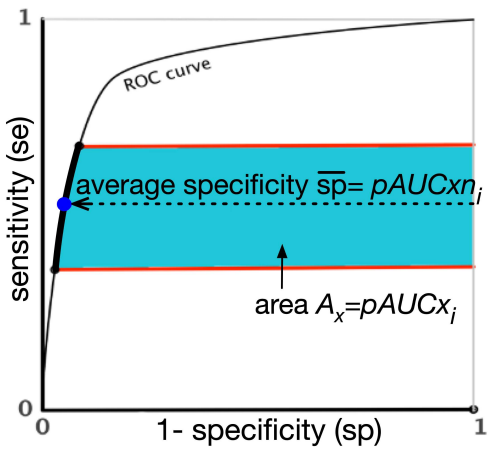
2 Overview of Deep ROC Analysis

We propose a method called deep ROC analysis that measures how well a model discriminates between two classes in multiple groups that span an ROC curve. We can compare measures in one group to another, or to the whole, using 4 pre-test measures and 4 post-test measures.

The pre-test measures in each group are: group AUC (AUC_i) and group normalized AUC ($AUCn_i$) (Figure 2),



a) The partial AUC, $pAUC_i$, is a vertical area, A_y under the ROC curve and its normalized value, $pAUCn_i$, is average sensitivity $\bar{se}$.



b) The horizontal partial AUC $pAUCx$ is a horizontal area, A_x to the right of the ROC curve and its normalized value, $pAUCxn$, is average specificity $\overline{sp}$.

Fig. 3: Two measures used in our method represent average sensitivity and average specificity, but are more commonly known by esoteric labels. Analysis is made complete by balanced average accuracy, as a third measure.

average sensitivity (Figure 3a) and average specificity (Figure 3b). AUC and $AUCn_i$ provide clarity to model comparison when average sensitivity is better but average specificity is worse, or vice-versa—it summarizes both.

The post-test measures are: average positive predictive value, average negative predictive value, average likelihood ratio positive and the average likelihood ratio negative.

One might argue that ROC plots provide in-depth information like deep ROC analysis, but eyeballing measures in groups, and normalizing them, is prone to error. Also, users are prone to mistake partial AUC (the vertical or sensitivity component only) for AUC_i.

In deep ROC analysis, each group is defined as a range of TPR, FPR or percentiles of the threshold[1] (Table 1). Any number of groups may be used, limited only by the number of instances in the data.

Support for group-wise analysis can be found in a recent systematic review by Wynants *et al.* [31] that examines over 100 COVID-19 prediction models and recommends that none of the models be used in practice, in part because of lack of reporting on calibration.

Calibration quantile-quantile (QQ) plots for binary outcomes [32] examine a model or test in groups like deep ROC analysis. However, QQ plots depict net error in the fraction of events—i.e., a combination of false positives and false negatives. Deep ROC analysis reports these separately (for each group) as average sensitivity and average specificity (Figure 3).

Deep ROC analysis uses the generalization of AUC to a partial ROC curve: the normalized concordant partial AUC $AUCn_i$ [9]—where the adjective "concordant" signifies that it equals the normalized partial C (concordance) statistic [9]. No other measures are generalizations of AUC and C with all of the same interpretations [9].

1. Note: when points are not uniquely defined the user must explicitly specify the boundary as a coordinate or they must apply a policy as in [9].

3 Old and New Interpretations of AUC

We show and prove a new interpretation of AUC, for a ROC curve in part or whole: AUC, normalized, is a weighted average that balances average sensitivity and average specificity; and we call this balanced average accuracy (section 9):

$$AUCn_i = \frac{\Delta x}{\Delta x + \Delta y} \cdot \overline{se} + \frac{\Delta y}{\Delta x + \Delta y} \cdot \overline{sp} \tag{1}$$

There is a need for this interpretation because the most common interpretation of AUC is lacking and abstract [4], [5]. Ask someone: what does an AUC of 0.8 or 80% mean? Or what does a 2% improvement in AUC mean? The two most common answers you will receive are as follows.

First, one might receive a geometric and comparative explanation. AUC is the area under the ROC curve which is depicted in an ROC plot. An AUC of 0.5 indicates a classifier (or test) is no better than chance, whereas an AUC of 1.0 means the classifier is perfect at discrimination. This explains AUC as a relative measure but it not tell us what an AUC of 0.8 means in absolute and precise terms: how many errors will the classifier or test commit and for whom?

The second more precise answer is that the AUC can be interpreted as a C statistic: the proportion of patients (or instances) with the outcome event that have a higher predicted score (as they should) than patients without the outcome, for all possible pairs of a positive and negative.

Therefore, an AUC of 80% means that the classifier correctly ranks patients 80% of the time in pairs; and a 2% improvement means that in ranking pairs, the classifier is correct 2% more often. Some thinking in decision-making is pairwise: e.g., how does the current patient or instance compare to another I have seen before? However, this cannot answer important questions in performance measurement. What is the probability of error for a single patient? What is the probability of error for the average patient in a subgroup of patients by predicted risk (e.g., patients at high predicted risk for the condition)? We need more than a pairwise perspective, we need the individual perspective.

Two other less commonly known interpretations of the AUC are that AUC equals average sensitivity across all thresholds, and AUC equals average specificity across all thresholds [33] (Table 1 Global column). This has a concrete interpretation for individuals, across all thresholds. That is, a classifier with an AUC that is 2% higher than another, is on average, over all possible thresholds, 2% more sensitive at detecting positives **and** 2% more specific (i.e., it detects negatives 2% better).

However, this interpretation is not true for part of an ROC curve, where, in general, average sensitivity differs from average specificity (Figure 3) [9] and both differ from normalized AUC.

Carrington *et al.* establish [9] that AUC, for any part of an ROC curve, is the average of the vertical and horizontal areas (7) (Figure 2), and call it the concordant partial AUC. This is useful, but these areas are abstract when it comes to interpretation.

When this measure is normalized and expressed in terms of average sensitivity and average specificity (1), then the interpretation is relatable. It represents a balanced view of detecting positives and negatives. We explain this balance and normalization in a later section (section 9).

Normalization is required to compare performance in one group or part, to the whole, or to any other group (Table 1).

Please note that our newly introduced term, balanced average accuracy, **should not** be confused with balanced accuracy, nor an average thereof. Balanced accuracy only equals AUC in the special case of an ROC curve with one point aside from $(0, 0)$ and $(1, 1)$ [34] which occurs for a discrete classifier [2] such as a decision tree or rule. This special case is sometimes incorrectly described as the result of a "single run" or experiment [34], [35] and sometimes mistaken for AUC in general [36].

In the sections that follow we discuss related work, background, our method, two case studies, limitations, conclusions and future work.

4 Related Work

ROC analysis is commonly used to select prediction models by computing and comparing the AUC. ROC curves are also plotted to interpret the AUC—e.g., is one model's ROC curve better than another in all parts/groups or only some?

Instead of visual ROC comparison, deep ROC analysis quantifies that comparison in parts/groups and quantifies part/group performance for a single ROC curve. To achieve this, we use the three measures from the literature (one recent) which are readily and inherently meaningful: AUC, average sensitivity and average specificity **in a part or group**.

Historically, measures have not been normalized to be interpreted as AUC, average sensitivity or average specificity—and some measures have been misunderstood. Hence, we review the history of AUC-related work and alternatives.

4.1 Measures we use

Two decades ago Bradley [11] recommended AUC [33] over accuracy as an overall measure of performance. However, numerous authors have identified issues with the AUC [3], [5], [8], [37]. For example, the AUC, as an overall measure, includes unrealistic or unused thresholds [3], [5] and it lacks information about the distribution of errors along the ROC curve [3]. These criticisms also apply to the C statistic (for binary outcomes) which is equal to the AUC of an empirical ROC curve [32], [38], [39].

Some criticisms of AUC expect it to fulfill calibration [3] and clinical utility [3], [5] too. However, standard advice from Steyerberg *et al.* [1] recommends reporting measures of discrimination and calibration (separately), including the C statistic or AUC, and a calibration plot. This approach is common in medical literature and our method supplements it with deeper analysis. They also recommend measuring net benefit as a measure of clinical utility in a separate category, for clinical decision-making.

Instead of an overall measure, McClish [6] and, separately, Thomson and Zucchini [7] proposed the partial AUC ($pAUC$) (Figure 3a) in parametric form[2]. We show the non-parametric form (2) [40] but our discussion pertains to both forms. For a range of FPR $[x_2, x_1]$ and an ROC curve $y = r(x)$, where y represents TPR, partial AUC computes an area A_y:

$$pAUC(x_1, x_2) = A_y = \int_{x_1}^{x_2} r(x)\, dx \tag{2}$$

This looks like a sensible generalization of AUC:

$$AUC = \int_0^1 r(x)\, dx \tag{3}$$

The partial AUC was a first step toward more in-depth analysis. However, the name "partial AUC" is misleading because, when normalized (signified by "n") to the range $[0, 1]$ by its maximum possible area $(1 \cdot \Delta x) = x_2 - x_1$, it is **average sensitivity** $\overline{se}$ (Figure 3a) [9] without any component of specificity:

$$pAUCn(x_1, x_2) = \frac{A_y}{\Delta x} = \overline{se} \tag{4}$$

In contrast, AUC summarizes and represents sensitivity and specificity [33]. The formula for AUC itself (3) is an over-simplification that works because the (whole) AUC is a special case where the horizontal and vertical areas are the same, so the horizontal perspective is redundant [9].

Mallet *et al.* used partial AUC to mistakenly claim two "tests are equally effective" [4, Fig. 3e,f], however, that is clearly false when the range of sensitivity was the same, but one test had markedly better specificity ($78 - 95\%$) than the other ($50 - 60\%$).

Shortly thereafter, McClish acknowledged that $pAUC$ is flawed as a measure of AUC in a part [8] because, when standardized, it monotonically increases to the right in an ROC plot—and others also found fault with it [41]. Hence, McClish proposed the standardized Partial Area (sPA) [8] which begins with the $pAUC$, subtracts the area under the major diagonal, and then standardizes the result.

Unfortunately, sPA is also flawed [9], [42]. It produces **negative** results for ROC curves that are partly above the major diagonal and partly below it, which occur in real life [24], [33], [42], [43] with **positive** AUC values.

Hence, to briefly summarize, the partial AUC, $pAUC$ measures sensitivity but not AUC. When normalized, it is **average sensitivity** [9], [30].

2. The parametric form assumes binormality in ROC data.

Its counterpart measures specificity. Jiang *et al.* [30] define a partial area index, PAI, which measures **average specificity** $\overline{se} = PAI$ based on the area to the right of an ROC curve. PAI assumes a fixed boundary at the top-right of an ROC plot and is parametric with a binormality assumption. For our purposes, we need both boundaries to be selected rather than fixed, hence we use a similar but non-parametric form (without assumptions) by Carrington *et al.* [9] as follows.

The horizontal partial AUC (Figure 3b), for a range of TPR $[y_1, y_2]$ in a ROC curve $x = r^{-1}(y)$, is normalized by its maximum possible area $(1 \cdot \Delta y) = y_2 - y_1$:

$$pAUCxn(y_1, y_2) = \frac{1}{\Delta y} \int_{y_1}^{y_2} 1 - r^{-1}(y) dy \tag{5}$$

$$= \frac{A_x}{\Delta y} = \overline{sp} \tag{6}$$

Next, to measure AUC in part of an ROC curve, as McClish sought, we refer to Carrington *et al.*'s concordant partial AUC [9] which we denote AUC_i (Figure 2) as a proper generalization of AUC to a part or group labelled i with vertical and horizontal perspectives. It is defined for the range $\theta_i = (x_1, x_2, y_1, y_2)$ as follows:

$$AUC_i(\theta_i) = \frac{1}{2} \int_{x_1}^{x_2} r(x)\ dx + \frac{1}{2} \int_{y_1}^{y_2} 1 - r^{-1}(y) dy \tag{7}$$

Carrington *et al.* derived its meaning from how the partial C statistic is computed, which can be visualized in the concordance matrix [9]. They normalize it to $[0, 1]$ [9, Table 3] consistent with the following:

$$AUCn_i(\theta_i) = \frac{AUC_i}{\frac{1}{2}(\Delta x + \Delta y)} \tag{8}$$

The equation above (8) is equal to an earlier equation we referenced (1), just in a different form.

Our method uses three familiar concepts of sensitivity, specificity and AUC. We measure the average sensitivity in a group of predicted risk (or thresholds), average specificity in a group, and normalized AUC in a group (Table 1)—along with other (post-test) measures.

4.2 Measures we do not use

There are several alternatives that we do not apply because they do not have familiar interpretations with inherent and well-established value, like the measures we use.

Bradley [44] provides an alternative, the half-AUC, to examine the area in an ROC plot in two parts, separated by the minor diagonal which extends from the top left to the bottom right, and where sensitivity and specificity are separately emphasized in each part. This approach is sensible, but limits analysis to two groups with fixed bounds. While it is scaled to the same range as the AUC or C statistic, it is not shown to have the same meaning. Empirically, its values are close to AUC_i with two groups, which does have established meaning.

Wu *et al.*'s novel partial area index [45] learns costs from human decision-makers to formulate a measurement baseline—a neat concept although its methodology needs more detail.

Yang *et al.*'s two-way AUC [46] examines a portion of AUC limited by cutoffs in sensitivity and false positive rate. It was introduced to choose between two models, not analyze the performance of one. Two-way AUC looks at an ROC part or group in isolation, not compared to the whole—where the whole is required to properly evaluate C statistics, AUC, fairness and equity. Carrington *et al.* describe the same concept, local concordance, as an interim step they discard as flawed and insufficient [9]. Also, Yang *et al.* modify and reference a U statistic that does not handle ties in scores.

Kallus and Zhou [47] define the cross AUC (xAUC) as a way to examine fairness issues in a model, by examining how events in group a rank against non-events in group b, and how events in group b rank against non-events in group a (and other combinations). There are 4 xAUC measures (see balanced xAUC [47]) for each pair of groups: relative measures, whereas AUC and other measures that we use are absolute (and comparative). For 4 groups, there are 6 pairs multiplied by 4 measures (24 in total). Kallus and Zhou also define a within-group AUC, but that is flawed like local concordance [9] (discussed previously).

In contrast to xAUC, Carrington *et al.*'s [9] partial C statistic measures the events in group a against all non-events, and the non-events in group a against all events—together as one measure. The concordant partial AUC (AUC_i) is equal in value but it is defined in a continuous instead of discrete form.

Also, in Kallus and Zhou [47]: groups are defined by instances instead of boundaries in TPR and/or FPR; and ties in score are not properly handled.

Narasimhan *et al.* [48] describe pair-wise measures for fairness between and in groups—mostly relative measures, but they also include measures for average sensitivity and average specificity. They do not measure AUC (or C) in a group—and their measures have the same shortcomings as Kallus and Zhou.

Hand and Till [49] generalize AUC to the multi-class case instead of the partial case.

Examples of ROC analysis in the literature that applied a group-wise approach like our method include Provost *et al.* who describe the dominant classifier in groups by slope (or skew) where a different classifier dominates in each group [50]. Dominance ensures better performance by a variety of common measures: accuracy, sensitivity, specificity, balanced accuracy, positive predictive value, etc. However, the question arises: how much better is the performance? Provost *et al.* do not quantify the difference, but they show confidence intervals toward that goal.

Carrington *et al.* [9] and Wernly *et al.* [51] compare classifiers by the partial AUC and the concordant partial AUC in groups, but without normalization which eases comparison and interpretation.

4.3 Average precision as an alternative to AUC

Deeper analysis of ROC plots often arises in the context of imbalanced data, and common alternatives employed in that case are the F_1 score [10], [16] at a point, which is too specific, and the area under the precision recall (PR) curve (AUPRC) a.k.a. average precision (AP) [19], which is too general.

Like AUC, the trouble with AP or AUPRC is that it includes all thresholds–and there are no partial measures

defined for them, but they focus on precision or positive predictive value (PPV), which is relevant to low prevalence data. Ideally, one should be able to compute both pre and post test measures, as in our method. Our method allows average PPV to be computed in subgroups and in intervals defined in several different ways including intervals that correspond to calibration plots.

A further criticism of F_1 and AP or AUPRC is that they vary with the prevalence of data in hidden (implicit) ways, i.e., their biases are not detectable. However, Flach and Kull fix that issue in AUPRC with PR gain curves, while Powers [52] provides versions of the F_1 score and precision (in AUPRC) that can explicitly account for the skew (which includes prevalence).

4.4 Multiple measures in tandem

Since Deep ROC analysis uses multiple measures in tandem, we review similar work.

Steyerberg *et al.* [1] discuss five categories of performance measures to report: overall performance, discrimination, calibration, reclassification and clinical usefulness—a good reference although Hilden aptly criticizes reclassification.

Steyerberg and Vergouwe [53] discuss three categories and only four measures in total.

Sokolova *et al.* [34] claim that common performance measures are insufficient when two classes are equally important or compare algorithms. They propose using: Youden's index, likelihood ratios, and discriminant power.

Sokolova and Lapalme [36] survey the invariant properties of performance measures and recommend measures for natural language processing.

Mallett *et al.* [4] discuss various measures of discrimination and clinical utility, while Obuchowski and Bullen [21] provide a survey of case studies or applications of AUC and related measures.

Several reporting guidelines for diagnostic tests and prediction models have been produced with guidance on measures to report for completeness and transparency. For example, STARD asks authors to report positivity cut-offs, how they were determined and whether they were defined a priori [54]. TRIPOD asks authors to define all predictors and the outcome that is predicted by the prediction model, including how and when they were measured [55].

A GRADE assessment reports our confidence that the true accuracy of a diagnostic test lies above or below a threshold, or in a specified range [56] that depends on prevalence, and optionally, the cost of a test's effects.

5 Method: Deep ROC Analysis

We provide an explanation and justification of deep ROC analysis (DRA) in section 2. To apply DRA, one performs the following steps (e.g., using our Python toolkit[3] for reference).

1) Ensure that the models you wish to analyze, use probabilities or calibrated scores (discussed further below).
2) Decide if there is a region in an ROC plot or group of persons in your data by predicted risk, that is of greatest interest or concern. For example, in health care: those at greatest risk, or eligible for treatment based on risk, or applicable to a specific/different clinical setting, or the model's weak spot in prediction, or the user/clinician's weak spot in prediction, or the region (at right) applicable to ruling-in or screening, or the region at left, applicable to ruling-out or diagnosis, or the region of Bayes error where positives and negatives may be hardest to distinguish (the middle).
3) Based on your business or clinical needs in step two, decide how you will define groups for analysis:
   a) as groups in FPR (or its complement, specificity)
   b) as groups in TPR, i.e., sensitivity or recall
   c) as groups in percentiles of predicted risk or probability (possibly to match a calibration quantile-quantile plot)
4) Based on your answers in steps two and three, decide how many groups of predicted risk/probability to use and their boundary values. Consider examples in case studies with three and six group if that helps. There should be at least 30 patients in each group—the minimum number of samples for a normal distribution, to report means and confidence intervals.
5) Create a table of average pre-test and post-test measures as in Table 2.
6) Assess if one model is better than another using standard ROC analysis. For example, which model has the best AUC (note any statistical significance per DeLong's method). Is the ROC curve of the best model dominant (higher) throughout the plot or only in some regions? Add any other measures you use or prefer to the table (e.g., AUPRC, se, sp, F_1).
7) Assess if one model is better than another in the group(s) of interest, or in each group, using deep ROC analysis. That is, for group i, is AUCn_i better for one model? If so, is it better in both $\overline{se}_i$ and $\overline{sp}_i$? Is it better in $\overline{PPV}_i$ and $\overline{NPV}_i$? Based on the answers from step two, you may favour positive measures ($\overline{se}_i$ and $\overline{PPV}_i$) over negative measures ($\overline{sp}_i$ and $\overline{NPV}_i$), or vice-versa.
8) Within a model, assess which groups according to AUCn_i perform better, the same or worse than the overall AUC and compared to each other. Note statistical significance according to confidence intervals [57].
   - Observe the values underlying each value of AUCn_i in case there is anything unusual: average sensitivity ($\overline{se}_i$) and average specificity ($\overline{sp}_i$). The usual behaviour of groups from left to right (in FPR) or bottom to top (in TPR) are monotonically increasing $\overline{se}_i$ and monotonically decreasing $\overline{sp}_i$.

We elaborate on the calibrated scores mentioned in step 1. Calibrated scores are needed to meaningful interpret and compare model results as predicted risk or probabilities. Some machine learning models do not produce calibrated scores by default.

In binary classification and diagnostic testing, models not only estimate binary outcomes, they also output classification scores that are used to create ROC curves. Statistical models such as logistic regression or naive Bayes produce scores which are probabilities in the range $[0, 1]$ [58], however machine learning models may produce scores in the range $[-\infty, +\infty]$ or $[a, b], a, b \in \mathbb{R}$ like some support vector machines and some neural networks respectively, while others may score in $[0, 1]$.

3. https://github.com/Big-Life-Lab/deepROC

TABLE 2: Performance of Adult Income models with confidence intervals, in 3 even groups by false positive rate (FPR) or non-events, for high, medium and low predicted income.

FPR Pred. Risk	[0,1] All	[0,.33] High	[.33,.67] Medium	[.67,1] Low
AUC	86.6% ±0.5	**Penalized Logistic Regression**		
$AUCn_i$	86.6% ±0.5	85.1% ±0.5	86.2% ±1.0	91.8% ±0.8
$\overline{se}_i$	86.6% ±0.5	67.2% ±1.0	93.9% ±0.4	98.6% ±0.2
$\overline{sp}_i$	86.6% ±0.5	91.9% ±0.2	54.1% ±0.5	18.1% ±0.7
$\overline{PPV}_i$	55.4% ±0.7	69.6% ±1.0	38.0% ±0.1	27.3% ±0.0
$\overline{NPV}_i$	90.3% ±0.1	86.1% ±0.2	96.4% ±0.3	97.5% ±0.3
AUC	86.3% ±0.5	**Random Forests**		
$AUCn_i$	86.3% ±0.5	85.2% ±0.5	85.2% ±0.6	90.8% ±1.1
$\overline{se}_i$	86.3% ±0.5	67.0% ±1.0	93.4% ±0.5	98.4% ±0.2
$\overline{sp}_i$	86.3% ±0.5	92.1% ±0.3	53.8% ±0.4	17.1% ±0.3
$\overline{PPV}_i$	55.3% ±0.7	69.7% ±1.0	37.9% ±0.1	27.3% ±0.1
$\overline{NPV}_i$	90.2% ±0.2	86.1% ±0.1	96.1% ±0.3	97.1% ±0.4
AUC	86.6% ±0.4	**Extreme Gradient Boosting**		
$AUCn_i$	86.6% ±0.4	85.5% ±0.4	85.6% ±0.9	91.9% ±0.6
$\overline{se}_i$	86.6% ±0.4	67.7% ±0.8	93.6% ±0.4	98.6% ±0.2
$\overline{sp}_i$	86.6% ±0.4	92.1% ±0.3	53.5% ±0.7	18.6% ±1.1
$\overline{PPV}_i$	56.0% ±0.6	70.7% ±0.9	37.9% ±0.1	27.4% ±0.1
$\overline{NPV}_i$	90.3% ±0.1	86.2% ±0.1	96.2% ±0.2	97.6% ±0.4

Calibration [59] is a process that has two purposes, either: (a) to turn non-probabilistic scores into probabilities; or (b) to change a model's output distribution to improve measures of calibration [1] such as calibration in the large.

Calibration [59] is an extra stage of processing that uses isotonic regression [60], [61] or Platt's method [62], [63]. It may be built into a model, or it made be performed as a separate post-processing step.

6 Case Study 1: US Adult Income

The Adult data set contains information on 48,842 adults from the United States Census Bureau regarding demographics, work and capital gains/losses which are used to predict personal income as either >50k or <50k.

We observe the following from deep ROC analysis (DRA) on the 10-fold cross-validation results (Table 2) with penalized logistic regression (LR), random forests (RF) and extreme gradient boosting (XGB), using groups by FPR:

- Two models, LR and XGB, have the same mean AUC, with a tighter confidence interval for XGB. DRA shows that XGB is slightly better for high predicted income in all group measures, while LR is better for medium predicted income.
- RF has a worse average PPV than other models for low predicted income, with statistical significance.
- The rightmost group has a higher $AUCn_3$ because each part of the ROC curve from left to right has a diminishing height (Figure 4) and diminishing contribution in specificity compared to near perfect sensitivity.

It may be more natural to examine groups by sensitivity (TPR) or events (Table 3), where we observe the following:

- The high risk group (at left) has a higher $AUCn_1$ because each part of the ROC curve from top to bottom has a diminishing width and diminishing contribution in average sensitivity.
- The medium risk group performs significantly worse than the overall AUC.

TABLE 3: If events (actual positives) are of interest, then analyze 3 even groups of sensitivity (TPR) or events, for high, medium and low predicted income.

TPR Pred. Risk	[0,1] All	[0,.33] High	[.33,.67] Medium	[.67,1] Low
AUC	86.6% ±0.4	**Extreme Gradient Boosting**		
$AUCn_i$	86.6% ±0.4	94.4% ±0.6	83.8% ±0.5	85.4% ±0.4
$\overline{se}_i$	86.6% ±0.4	22.8% ±0.7	53.6% ±0.5	92.4% ±0.3
$\overline{sp}_i$	86.6% ±0.4	99.3% ±0.1	93.5% ±0.4	67.1% ±0.9
$\overline{PPV}_i$	56.0% ±0.6	89.3% ±1.7	71.9% ±1.2	40.3% ±0.3
$\overline{NPV}_i$	90.3% ±0.1	79.2% ±0.0	85.9% ±0.1	95.2% ±0.2

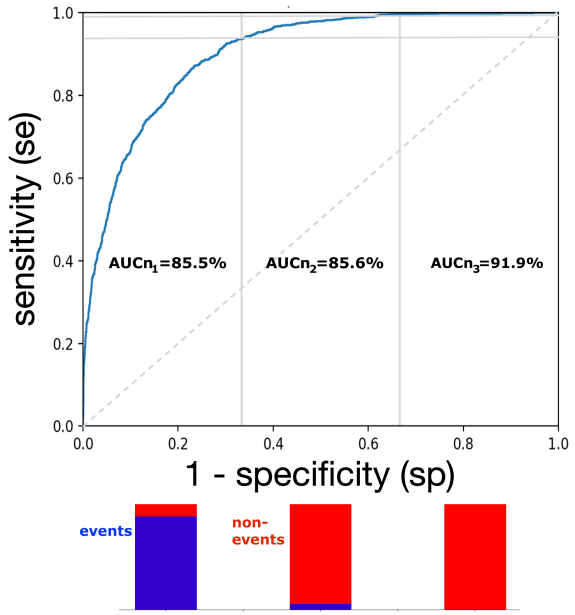


Fig. 4: This ROC plot from an Extreme Gradient Boosting (XGB) machine on the Adult income data, shows that for even groups of FPR or specificity, $AUCn_i$ tends to be larger at right, where the change in height of the curve in the group becomes vanishingly small. The near-perfect performance in the 3rd group's vertical dominates in contribution. $AUCn_i$ is as unaffected by the class ratio, as AUC is.

7 Case Study 2: Mortality Prediction Based on Arterial Blood Gas Analysis of Septic Patients

Wernly *et al.* [51] compare four different machine learning and clinical algorithms to predict the 32.4% of septic patients who would pass away within the next 96 hours in a multi-center ICU observational study. They evaluate a recurrent neural network using long-short term memory (LSTM) on arterial blood gas data against several baseline models and clinical scales: Logistic Regression (LR), the SOFA score evaluating functioning of six organs, and against blood

TABLE 4: The neural network using long-short term memory (LSTM) performs consistently well in AUC_i across groups of high, medium and low predicted risk, defined by FPR. Average sensitivity is always maximal at right, while average specificity is always maximal at left.

FPR Pred. Risk	[0,1] All	[0,.33] High	[.33,.67] Medium	[.67,1] Low
AUC	88%		**LSTM**	
$AUCn_i$	88%	**89%**	85%	87%
$\overline{se}_i$	88%	76%	91%	97%
$\overline{sp}_i$	88%	94%	57%	20%
PPV			60% at t=0.5	
NPV			96% at t=0.5	

TABLE 5: Logistic regression (LR) performs slightly better than Lactate as a predictor, but not adequately and SOFA performs poorly in groups of risk by FPR. SOFA performs best in the wrong group.

FPR Pred. Risk	[0,1] All	[0,.33] High	[.33,.67] Medium	[.67,1] Low
AUC	82%		**Logistic Regression**	
$AUCn_i$	82%	**85%**	81%	76%
$\overline{se}_i$	82%	67%	84%	94%
$\overline{sp}_i$	82%	93%	67%	40%
PPV			48% at t=0.5	
NPV			95% at t=0.5	
AUC	80%		**Lactate**	
$AUCn_i$	80%	**81%**	80%	80%
$\overline{se}_i$	80%	58%	88%	94%
$\overline{sp}_i$	80%	91%	65%	14%
PPV			-	
NPV			-	
AUC	72%		**SOFA**	
$AUCn_i$	72%	67%	72%	**84%**
$\overline{se}_i$	72%	39%	82%	94%
$\overline{sp}_i$	72%	80%	60%	44%
PPV			23% at t=0.5	
NPV			92% at t=0.5	

lactate levels as a sole predictor. We normalize the partial area measures reported by Wernly *et al.* (Tables 4, 5) for interpretation.

First, we examine AUC as an overall measure. SOFA has an AUC, or average balanced accuracy of 72% (Table 5), which is moderately predictive [51], while Lactate and Logistic Regression perform well with an AUC of 80% and 82% respectively and LSTM's AUC of 88% is best (Table 4) at 6% above the others.

Since AUC is equal to average sensitivity and average specificity, this means that, across all thresholds, LSTM is on average 88% sensitive and 88% specific. For parts of an ROC curve however, $AUCn_i$, average sensitivity and average specificity differ (Table 4) as expected [9].

In the ROC plot (Figure 5) for FPR<0.35, the curves are above each other (better) in the same order as AUC, but for FPR>0.5 Lactate is better than LR, and at FPR>0.63 SOFA is better than LR too.

Wernly *et al.* [51] indicate that high-risk patients are the most clinically relevant: predicting patients with "poor prognosis" and predicting with "high accuracy, with low false-positive rates". Hence, that brings focus to the high risk group among the three groups shown.

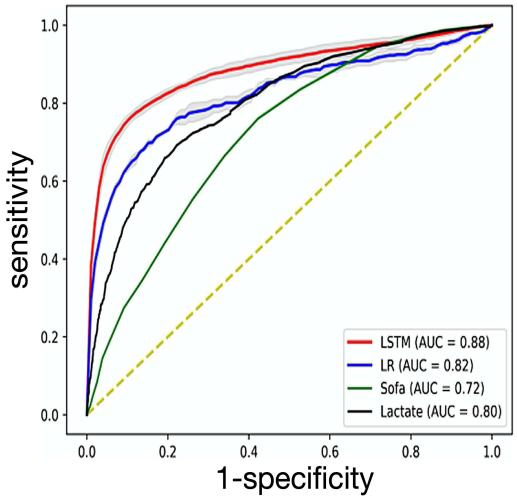


Fig. 5: The ROC plot for the four classifiers: SOFA, Lactate, Logistic Regression (LR) and LSTM. LSTM is best (dominant) in most regions.

In the high risk group $AUCn_i$ is 67%, 81%, 85% and 89% for SOFA, Lactate, Logistic Regression and LSTM, respectively. SOFA is 5% worse in that group, than what the overall AUC indicates, while LR is 3% better, and LSTM and Lactate are 1% better. This analysis confirms that LSTM not only performs best overall, but also in that particular group, in quantitative terms beyond eyeballing the ROC plot or choosing a single point in each group.

If we examine average sensitivity in the high-risk region, the differences between LSTM and Logistic Regression grows from 4% in $AUCn_i$ to a 9% difference in average sensitivity—which is more important than average specificity in this scenario. Lactate has 58% average sensitivity in the high risk group, which is hidden if one only looks at AUC as a summary measure, and SOFA is only 39% sensitive there on average.

The poor sensitivity of SOFA is striking, but it makes sense. That is, in high-risk patients, there will be a lot of morbidity or organ dysfunction which SOFA identifies, e.g., if creatinine rises from 1.0 to 2.0 mg/dL. However, a rise in creatinine from 3.0 to 6.0 mg/dL might not reflect the same importance; and the same concept applies to bilirubin, coagulation, etc. This underscores the merit of risk stratification tools with higher granularity, as in the LTSM that Wernly et al. propose rather than SOFA.

Scores such as SOFA or qSOFA or Lactate concentrations were developed to "rule in" high-risk patients. However, the approach by Wernly *et al.* is different, they want to "rule out" patients who are very unlikely to benefit from further critical care. SOFA performs best ($AUCn_3$) where it matters least (Table 5), while Lactate performs consistently across all three risk groups.

8 Case Study 3: German Breast Cancer Study Group

In survival analysis of patients in the German breast cancer study group [64], 33% of patients with positive node primary breast cancer had isolated locoregional recurrence at two

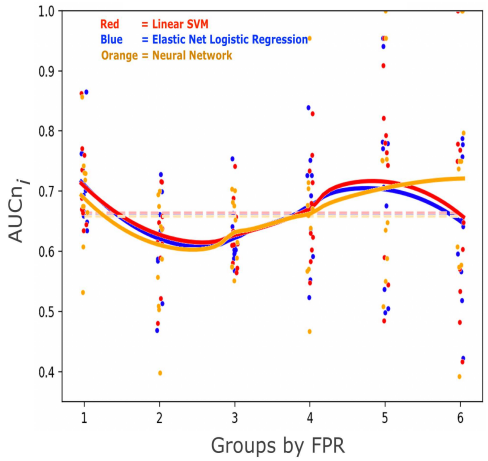


Fig. 6: For the German Breast cancer Study Group the AUC (dashed lines) are shown relative to the loess fitted values of AUCn$_i$ across 6 groups by FPR. Points from each of 10 cross-validation folds are jittered for visual clarity.

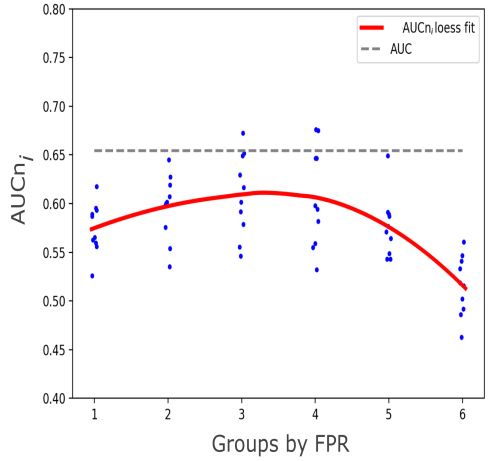


Fig. 7: Local normalization has undesirable effects: values less than AUC and sagging at left and right.

years after treatment. For this low prevalence situation, the minority of positives are most clinically relevant—i.e., high-risk patients identified by the leftmost part of the ROC plot.

We applied four models to this data set: penalized logistic regression (LR), a linear support vector machine (LSVM), a non-linear support vector machine with a Gaussian RBF kernel (SVM), and a neural network (NN) with Rectified Linear Units. Because the data are small, shallow learning algorithms were used.

We performed 10-fold cross-validation using a random search of the hyperparameter space.

We perform deep ROC analysis using 6 groups by FPR. Rather than reporting results in tabular form, we plot AUCn$_i$ (Table 6).

The plot reveals that a linear support vector machine (red), elastic net logistic regression (blue) and neural network (orange) perform about the same overall with AUC values of 66.4%, 66.1% and 65.8% respectively (dashed line) but the AUCn$_i$ in each group fit with loess curves, differ. Hence the need for deep ROC analysis to quantify inequities in performance between groups of predicted risk. Applying isotonic regression for calibration may reduce those inequities.

Additional plots for the other measures, average sensitivity, average specificity, average PPV and average NPV, were also produced but are not shown for brevity.

9 AUC is Balanced Average Accuracy

Carrington *et al.* [9] generalize AUC with the concordant partial AUC, which we denote AUCn$_i$. For any contiguous part or group of an ROC curve it meets the following properties of AUC:

1) converges to and equals AUC for bounds at (0,0) & (1,1)
2) adds up to AUC for a covering but non-overlapping set of ROC curve parts or groups
3) is explicitly related to the ROC curve's average sensitivity and average specificity
4) is interpretable as a C statistic and equal to it
5) is interpretable as meaningful area(s) "under" the ROC curve
6) can be computed for any empirical or continuously defined ROC curve
7) properly handles ties in scores
8) is non-negative
9) can be normalized to [0,1] for fair comparison with other (normalized) AUC measures
10) ignores, or has the effect of ignoring the class ratio, in the same manner as AUC

The last property is the balance in balanced average accuracy. We explain that as follows.

AUC equals average sensitivity $\overline{se}$ related to vertical area; and equals average specificity $\overline{sp}$ related to horizontal area [33]. $\overline{se}$ and $\overline{sp}$ are given the same weight thereby ignoring imbalance in the class ratio.

For part of an ROC curve with a small vertical area in comparison to a larger horizontal area (Figure 3), we must give proportional weight to those areas in order to continue ignoring the class ratio in the same way as AUC (i.e., for a balanced view). Hence, the proportional weights in:

$$AUCn_i = \frac{\Delta x}{\Delta x + \Delta y} \cdot \overline{se} + \frac{\Delta y}{\Delta x + \Delta y} \cdot \overline{sp} \tag{9}$$

which can be compared to any other AUCn$_i$ or AUC. We refer to this as **balanced average accuracy**. This is the normalization, Carrington *et al.* applied in a table [9, Table 3] for the concordant partial AUC, while showing an alternative normalization in formulas for the partial C statistic. That is, equation (9) above is a global normalization, whereas the alternative is to normalize "locally" in each of the two terms separately (two signified by "nn"). We do **not recommend** the following local normalization (Figure 7):

$$AUCnn_i = \frac{1}{2} \cdot \frac{A_y}{\Delta x} + \frac{1}{2} \cdot \frac{A_x}{\Delta y} \tag{10}$$

$$= \frac{1}{2}\overline{se} + \frac{1}{2}\overline{sp} \tag{11}$$

The local normalization is affected by the class distribution in groups and has group values almost always less than AUC with values that demonstrably sag lower than AUC at

the left and right extremes. It fails to meet the last criterion for proper balance. The normalized partial C statistics in Carrington *et al.* [9, Eq.4,7] should be updated accordingly.

We note that any measure derived from AUC that attempts to generalize it, can easily meet AUC's first property for convergence and equality. So showing that equality is superfluous. The critical thinking is in understanding the full suite of AUC's properties and applying them.

Also, measuring part of an ROC curve has two requirements that append to the previous list:

11) can be computed for any unique bounds at any point continuously along the ROC curve
12) can be computed for any unique bounds at any point continuously along the ROC curve

The partial AUC and horizontal partial AUC meet all of AUC's properties except the third and fourth. Hence, they are not proper generalizations of AUC (from the whole to a part).

10 Limitations

One possible limitation of our method is that the additional information introduces more complexity which could complicate communication of results.

Another limitation of our method is that there is always the chance that there may not be any clear choice between models, with or without statistical significance. Additional analyses may be prudent, such as: decision curve analysis [14].

A third limitation is that we recommend at least 30 patients in each group (in the test set) since that satisfies the sample size for a normal distribution when reporting means and confidence intervals. For small data this may limit the number of groups.

11 Conclusions

We selected and interpreted three key pre-test measures from the literature and showed how to use them in tandem as familiar concepts, but newly applied to groups or parts of an ROC curve or plot. We call this method deep ROC analysis.

We have shown that models (or tests) can and do behave differently in different risk groups—with better or worse performance than what AUC indicates.

We provided a new interpretation of AUC, in whole or part, that permits a new and pragmatic interpretation for individual patients or instances, not just pairs.

In the first case study (adult income) our method largely confirms the similarity in overall AUC, however it also allows one to choose XGB for slightly better performance in the high risk group or choose logistic regression for slightly better performance in the medium risk group.

In the second case study, our method indicates that logistic regression is not just better than Lactate overall (by 2%) but it is particularly better in the high risk group which matters most (with 5% and 9% better AUCn_1 and $\overline{se}$). It also further describes SOFA's inadequacy.

In the third case study deep ROC analysis shows that penalized logistic regression and a linear support vector machine perform similarly, but in a non-uniform manner across risk groups; while a neural network performed differently. Depending on one's needs, the choice of model may differ based on these details.

In summary, deep ROC analysis provides in-depth information by groups that will sometimes improve model evaluation, selection, explanation and audit over standard ROC analysis. We recognize that other analysis (of utility) may still be needed and performed separately.

Our approach could support tuning a model's threshold to a particular clinical setting and risk group, or choosing thresholds at the point of service for more personalized medicine.

Our method also indicates group measures to consider as objectives in loss functions for optimization or re-optimization of models.

12 Future Work

Future work may include computing the odds ratio at each point and taking its average in the group; or comparing results with AUPRC and the F_1 score. Examining how DeLong's method pertains to part of an ROC curve would also be helpful.

Recent work permits neural networks to be optimized for proxy measures of AUC, instead of cross-entropy. Hence, future work may consider proxy measures for AUCn_i and linear combinations thereof.

Further investigation is also warranted to understand any advantages or disadvantages of measuring average PPV and average NPV by traversing the ROC curve uniformly, as we do, instead of traversing the precision recall curve by units of recall.

List of abbreviations

AI:	Artificial intelligence
AUC:	Area under the ROC curve
AUPRC:	Area under the precision recall curve
C:	The C statistic for binary outcomes, but not Harrell or Uno's C statistic
FNR:	False negative rate
FPR:	False positive rate, or 1-specificity
$pAUC_i$:	Partial area under the ROC curve (i.e., vertical)
$pAUCn_i$:	Partial area under the ROC curve (normalized)
AUC_i:	Concordant partial area under the ROC curve
$AUCn_i$:	Concordant partial area under the ROC curve (normalized) or $cpAUCn_i$
$pAUCx_i$:	Horizontal partial area under the curve
$pAUCxn_i$:	Horizontal partial area under the ROC curve (normalized)
LR:	Logistic Regression
LSTM:	Long Short-Term Memory
PAI:	Partial area index
PRC:	Precision recall curve
ROC:	Receiver operating characteristic
SOFA:	Sequential organ failure assessment
sPA:	Standardized partial area
TNR:	True negative rate, or specificity, or selectivity
TPR:	True positive rate, or sensitivity, or recall
xAI:	Explainable artificial intelligence

AVAILABILITY OF CODE AND DATA

The Python code that produced the measurement numbers, plots and tables, is available at:

`https://github.com/Big-Life-Lab/deepROC`

`http://deepROC.org`

The German Breast Cancer data is available at:

`https://biostat.app.vumc.org/wiki/Main/DataSets`

ACKNOWLEDGEMENTS

Parts of this work has received funding by the Austrian Science Fund (FWF), Project: P-32554 "A reference model for explainable Artificial Intelligence in the medical domain".

REFERENCES

[1] E. W. Steyerberg, M. W. Kattan, M. Gonen, N. Obuchowski, M. J. Pencina, A. J. Vickers, T. Gerds, and N. R. Cook, "Assessing the Performance of Prediction Models: a Framework for Some Traditional and Novel Measures," *Epidemiology*, vol. 21, no. 1, pp. 128–138, 2009.

[2] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, pp. 861–874, 2006.

[3] J. M. Lobo, A. Jiménez-valverde, and R. Real, "AUC: a misleading measure of the performance of predictive distribution models," *Global Ecology and Biogeography*, no. 17, pp. 145–151, 2008.

[4] S. Mallett, S. Halligan, M. Thompson, G. S. Collins, and D. G. Altman, "Interpreting diagnostic accuracy studies for patient care," *Bmj*, vol. 345, 2012.

[5] K. Wagstaff, "Machine learning that matters," *arXiv preprint arXiv:1206.4656*, 2012.

[6] D. K. McClish, "Analyzing a Portion of the ROC Curve," *Medical decision making*, pp. 190–195, 1989.

[7] M. Thomson and W. Zucchini, "On the statistical analysis of ROC curves," *Statistics in Medicine*, vol. 8, pp. 1277–1290, 1989.

[8] D. K. McClish, "Evaluation of the Accuracy of Medical Tests in a Region around the Optimal Point," *Academic Radiology*, vol. 19, no. 12, pp. 1484–1490, 2012. [Online]. Available: http://dx.doi.org/10.1016/j.acra.2012.09.004

[9] A. M. Carrington, P. W. Fieguth, H. Qazi, A. Holzinger, H. H. Chen, F. Mayr, and D. G. Manuel, "A new concordant partial auc and partial c statistic for imbalanced data in the evaluation of machine learning algorithms," *Springer/Nature BMC Medical Informatics and Decision Making*, vol. 20, no. 1, pp. 1–12, 2020.

[10] G. Santafe, I. Inza, and J. A. Lozano, "Dealing with the evaluation of supervised classification algorithms," *Artificial Intelligence Review*, vol. 44, no. 4, pp. 467–508, 2015.

[11] A. P. Bradley, "The use of the area under the {ROC} curve in the evaluation of machine learning algorithms," *Pattern Recognition*, vol. 30, pp. 1145–1159, 1997.

[12] C. E. Metz, "Basic principles of roc analysis," in *Seminars in nuclear medicine*, vol. 8, no. 4. Elsevier, 1978, pp. 283–298.

[13] S. Rosset, "Model selection via the auc," in *Proceedings of the twenty-first international conference on Machine learning*, 2004, p. 89.

[14] A. J. Vickers and E. B. Elkin, "Decision curve analysis: a novel method for evaluating prediction models," *Medical Decision Making*, vol. 26, no. 6, pp. 565–574, 2006.

[15] R. Dubey, J. Zhou, Y. Wang, P. M. Thompson, and J. Ye, "Analysis of sampling techniques for imbalanced data: An n= 648 adni study," *NeuroImage*, vol. 87, pp. 220–241, 2014.

[16] P. Flach, "Performance Evaluation in Machine Learning: The Good, the Bad, the Ugly, and the Way Forward," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 9808–9814, 2019.

[17] M. Wu and J. Ye, "A small sphere and large margin approach for novelty detection using training data with outliers," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 11, pp. 2088–2092, 2009.

[18] Q. Zhu, "On the performance of Matthews correlation coefficient (MCC) for imbalanced dataset," *Pattern Recognition Letters*, vol. 136, pp. 71–80, 2020. [Online]. Available: https://doi.org/10.1016/j.patrec.2020.03.030

[19] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," *PLoS ONE*, vol. 10, no. 3, pp. 1–21, 2015.

[20] P. A. Flach and M. Kull, "Precision-recall-gain curves: Pr analysis done right." in *NIPS*, vol. 15, 2015.

[21] N. A. Obuchowski and J. A. Bullen, "Receiver operating characteristic (roc) curves: review of methods with applications in diagnostic medicine," *Physics in Medicine & Biology*, vol. 63, no. 7, p. 07TR01, 2018.

[22] J.-H. Xue and P. Hall, "Why does rebalancing class-unbalanced data improve auc for linear discriminant analysis?" *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 5, pp. 1109–1112, 2014.

[23] T. C. Landgrebe and R. P. Duin, "Efficient multiclass roc approximation by decomposition via confusion matrix perturbation analysis," *IEEE transactions on pattern analysis and machine intelligence*, vol. 30, no. 5, pp. 810–822, 2008.

[24] S. Pérez-Fernández, P. Martínez-Camblor, P. Filzmoser, and N. Corral, "nsroc: An r package for non-standard roc curve analysis," *The R Journal*, vol. 10, no. 2, pp. 55–77, 2018.

[25] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Mueller, "Causability and explainability of artificial intelligence in medicine," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 9, no. 4, pp. 1–13, 2019.

[26] A. Carrington, P. Fieguth, and H. Chen, "Measures of model interpretability for model selection," in *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*. Springer, 2018, pp. 329–349.

[27] A. Holzinger, P. Kieseberg, E. Weippl, and A. M. Tjoa, "Current advances, trends and challenges of machine learning and knowledge extraction: From machine learning to explainable ai," in *Springer Lecture Notes in Computer Science LNCS 11015*. Cham: Springer, 2018, pp. 1–8.

[28] A. Holzinger, B. Malle, A. Saranti, and B. Pfeifer, "Towards multi-modal causability with graph neural networks enabling information fusion for explainable ai," *Information Fusion*, vol. 71, no. 7, pp. 28–37, 2021.

[29] I. D. Raji, A. Smart, R. N. White, M. Mitchell, T. Gebru, B. Hutchinson, J. Smith-Loud, D. Theron, and P. Barnes, "Closing the ai accountability gap: Defining an end-to-end framework for internal algorithmic auditing," in *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 2020, pp. 33–44.

[30] Y. Jiang, C. E. Metz, and R. M. Nishikawa, "A receiver operating characteristic partial area index for highly sensitive diagnostic tests." *Radiology*, vol. 201, no. 3, pp. 745–750, 2014.

[31] L. Wynants, B. Van Calster, G. S. Collins, R. D. Riley, G. Heinze, E. Schuit, M. M. Bonten, D. L. Dahly, J. A. Damen, T. P. Debray *et al.*, "Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal," *bmj*, vol. 369, 2020.

[32] E. W. Steyerberg, *Clinical Prediction Models*. Springer, 2009.

[33] X.-H. Zhou, D. K. McClish, and N. A. Obuchowski, *Statistical methods in diagnostic medicine*. John Wiley and Sons, 2002.

[34] M. Sokolova, N. Japkowicz, and S. Szpakowicz, "Beyond accuracy, f-score and roc: a family of discriminant measures for performance evaluation," in *Australasian joint conference on artificial intelligence*. Springer, 2006, pp. 1015–1021.

[35] V. García, R. A. Mollineda, and J. S. Sánchez, "Index of balanced accuracy: A performance measure for skewed class distributions," in *Iberian conference on pattern recognition and image analysis*. Springer, 2009, pp. 441–448.

[36] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information processing & management*, vol. 45, no. 4, pp. 427–437, 2009.

[37] J. Hilden, "The area under the roc curve and its competitors," *Medical Decision Making*, vol. 11, no. 2, pp. 95–101, 1991.

[38] A. J. Vickers and A. M. Cronin, "Everything you always wanted to know about evaluating prediction models (but were too afraid to ask)," *Urology*, vol. 76, no. 6, pp. 1298–1301, 2010.

[39] N. R. Cook, "Statistical evaluation of prognostic versus diagnostic models: Beyond the ROC curve," *Clinical Chemistry*, vol. 54, no. 1, pp. 17–23, 2008.

[40] L. E. Dodd and M. S. Pepe, "Partial AUC estimation and regression," *Biometrics*, vol. 59, no. 3, pp. 614–623, 2003.

[41] H. Ma, A. I. Bandos, H. E. Rockette, and D. Gur, "On use of partial area under the roc curve for evaluation of diagnostic performance," *Statistics in medicine*, vol. 32, no. 20, pp. 3449–3458, 2013.

[42] J.-M. Vivo, M. Franco, and D. Vicari, "Rethinking an roc partial area index for evaluating the classification performance at a high specificity range," *Advances in Data Analysis and Classification*, vol. 12, no. 3, pp. 683–704, 2018.
[43] C. E. Metz and H. B. Kronman, "Statistical significance tests for binormal roc curves," *Journal of Mathematical Psychology*, vol. 22, no. 3, pp. 218–243, 1980.
[44] A. P. Bradley, "Half-auc for the evaluation of sensitive or specific classifiers," *Pattern Recognition Letters*, vol. 38, pp. 93–98, 2014.
[45] T. Wu, H. Huang, G. Du, and Y. Sun, "A novel partial area index of receiver operating characteristic (ROC) curve," *Medical Imaging 2008: Image Perception, Observer Performance, and Technology Assessment*, vol. 6917, no. 69170, p. 69170B, 2008.
[46] H. Yang, K. Lu, X. Lyu, and F. Hu, "Two-way partial auc and its properties," *Statistical methods in medical research*, vol. 28, no. 1, pp. 184–195, 2019.
[47] N. Kallus and A. Zhou, "The fairness of risk scores beyond classification: Bipartite ranking and the xauc metric," *Advances in neural information processing systems*, vol. 32, pp. 3438–3448, 2019.
[48] H. Narasimhan, A. Cotter, M. Gupta, and S. Wang, "Pairwise fairness for ranking and regression," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, 2020, pp. 5248–5255.
[49] D. J. Hand and R. J. Till, "A simple generalisation of the area under the roc curve for multiple class classification problems," *Machine learning*, vol. 45, no. 2, pp. 171–186, 2001.
[50] F. Provost, T. Fawcett, and R. Kohavi, "The case against accuracy estimation for comparing classifiers," *Proceedings of the 15^{th} International Conference on Machine Learning*, no. January 2013, pp. 445–553, 1998.
[51] B. Wernly, B. Mamandipoor, P. Baldia, C. Jung, and V. Osmani, "Machine learning predicts mortality in septic patients using only routinely available abg variables: a multi-centre evaluation," *International Journal of Medical Informatics*, p. 104312, 2020.
[52] D. M. W. Powers, "Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation," Flinders University, Tech. Rep. December, 2007.
[53] E. W. Steyerberg and Y. Vergouwe, "Towards better clinical prediction models: Seven steps for development and an ABCD for validation," *European Heart Journal*, vol. 35, no. 29, pp. 1925–1931, 2014.
[54] J. F. Cohen, D. A. Korevaar, D. G. Altman, D. E. Bruns, C. A. Gatsonis, L. Hooft, L. Irwig, D. Levine, J. B. Reitsma, H. C. De Vet *et al.*, "Stard 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration," *BMJ open*, vol. 6, no. 11, 2016.
[55] G. S. Collins, J. B. Reitsma, D. G. Altman, and K. G. Moons, "Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (tripod) the tripod statement," *Circulation*, vol. 131, no. 2, pp. 211–219, 2015.
[56] M. Hultcrantz, R. A. Mustafa, M. M. Leeflang, V. Lavergne, K. Estrada-Orozco, M. T. Ansari, A. Izcovich, J. Singh, L. Y. Chong, A. Rutjes *et al.*, "Defining ranges for certainty ratings of diagnostic accuracy: a grade concept paper," *Journal of clinical epidemiology*, vol. 117, pp. 138–148, 2020.
[57] P. C. Austin and E. W. Steyerberg, "Interpreting the concordance statistic of a logistic regression model: relation to the variance and odds ratio of a continuous explanatory variable," *BMC medical research methodology*, vol. 12, no. 1, p. 82, 2012.
[58] C. M. Bishop, *Pattern recognition and machine learning*. springer, 2006.
[59] A. Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," in *Proceedings of the 22nd international conference on Machine learning*, 2005, pp. 625–632.
[60] R. L. Dykstra and T. Robertson, "An algorithm for isotonic regression for two or more independent variables," *The Annals of Statistics*, pp. 708–716, 1982.
[61] P. Mair, K. Hornik, and J. de Leeuw, "Isotone optimization in r: pool-adjacent-violators algorithm (pava) and active set methods," *Journal of statistical software*, vol. 32, no. 5, pp. 1–24, 2009.
[62] J. Platt *et al.*, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Advances in large margin classifiers*, vol. 10, no. 3, pp. 61–74, 1999.
[63] H.-T. Lin, C.-J. Lin, and R. C. Weng, "A note on platt's probabilistic outputs for support vector machines," *Machine learning*, vol. 68, no. 3, pp. 267–276, 2007.
[64] M. Schumacher, "Rauschecker for the german breast cancer study group, randomized 2× 2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive lbreast cancer patients," *Journal of Clinical Oncology*, vol. 12, pp. 2086–2093, 1994.

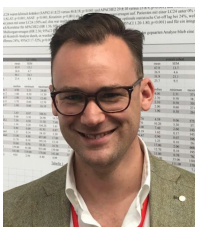
André M. Carrington is an Associate Scientist at the Ottawa Hospital and Region Imaging Associates. André received his Ph.D. in Systems Design Engineering focused on Medical AI and his Masters in Mathematics (Computer Science) from the University of Waterloo.

Douglas G. Manuel is a Medical Doctor with a Masters in Epidemiology and Royal College specialization in Public Health and Preventive Medicine. He is a Clinician Scientist at the Ottawa Hospital Research Institute and the Bruyère Research Institute and a Professor in the Departments of Family Medicine and the School of Epidemiology, Public Health and Preventive Medicine at the University of Ottawa.

Paul W. Fieguth is a Professor in Systems Design Engineering and Associate Dean in the Faculty of Engineering at the University of Waterloo and co-Director of the Vision & Image Processing group. Paul received his Ph.D. in electrical engineering from the Massachusetts Institute of Technology.

Tim Ramsay is the head of the Ottawa Methods Centre at the Ottawa Hospital Research Institute and a Professor with the University of Ottawa, Canada.

Venet Osmani is a Senior Researcher in the eHealth Group at the Fondazione Bruno Kessler Research Institute and a Professor in the Department of Psychology and Cognitive Science, University of Trento, Italy.



Bernhard Wernly practices internal medicine in the Department of Cardiology at the Paracelsus Medical University of Salzburg, Salzburg, Austria.

Carol Bennett is a research associate at the Ottawa Hospital Research Institute and the Institute for Clinical Evaluative Sciences, Ottawa, Canada.

Steve Hawken is the head of Big Data initiatives in the Ottawa Methods Centre at the Ottawa Hospital Research Institute and a Professor with the University of Ottawa, Canada.

Olivia Magwood is a research associate at Bruyère Research Institute and a doctoral student at the University of Ottawa. She has a Master's degree in public health.

Yusuf Sheikh is a part-time researcher at the Ottawa Hospital Research Institute. Yusuf is completing his Bachelor's degree in Biomedical Science at the University of Ottawa, Canada.

Matthew McInnes is a radiologist with the Ottawa Hospital Research Institute and a Professor with the University of Ottawa, Canada.

Andreas Holzinger (M'00) is Visiting Professor for explainable AI at the University of Alberta, Canada since 2019 and head of the Human-Centered AI Lab at the Medical University Graz, Austria. He received his PhD in cognitive science from Graz University and his second PhD in computer science from Graz University of Technology. Andreas promotes a synergistic approach to put the human-in-control of AI to align it with human values, privacy, security and safety. For his pioneer work in this area he was elected ordinary member of the Academia Europaea, the European Academy of Sciences, and is full member of the European Lab for Learning and Intelligent Systems.