Deep ROC Analysis and AUC as Balanced Average Accuracy to Improve Model Selection, Understanding and Interpretation

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

Abstract—Optimal performance is critical for decision-making tasks from medicine to autonomous driving, however common performance measures may be too general or too specific. For binary classifiers, diagnostic tests or prognosis at a timepoint, measures such as the area under the receiver operating characteristic curve, or the area under the precision recall curve, are too general because they include unrealistic decision thresholds. On the other hand, measures such as accuracy, sensitivity or the F1 score are measures at a single threshold that reflect an individual single probability or predicted risk, rather than a range of individuals or risk. We propose a method in between, deep ROC analysis, that examines groups of probabilities or predicted risks for more insightful analysis. We translate esoteric measures into familiar terms: AUC and the normalized concordant partial AUC are balanced average accuracy (a new finding); the normalized partial AUC is average sensitivity; and the normalized horizontal partial AUC is average specificity. Along with post-test measures, we provide a method that can improve model selection in some cases and provide interpretation and assurance for patients in each risk group. We demonstrate deep ROC analysis in two case studies and provide a toolkit in Python.

Index Terms—Explainable AI, ROC, AUC, C Statistic, Partial AUC, Imbalanced Data

1 Introduction

Two important and common measures of performance for binary diagnostic tests and classifiers (models) are accuracy [1] and the area under the curve [2] (AUC) in a receiver operating characteristic (ROC) plot [3]. Accuracy is a measure at a single operating point or decision threshold on a model's ROC curve, while AUC measures all operating points.

In the medical and health domain a lot of data is imbalanced [4]. For imbalanced data, alternative measures at a single operating point include balanced accuracy [1], the geometric mean (of sensitivity and specificity) [1], [5], the F_1 score [1], [6] and Matthews' Correlation Coefficient [7]. At all operating points, a common alternative is the area under the precision recall curve (AUPRC) a.k.a. average precision (AP) [8]; while less common alternatives include the predictive ROC curve [9], the positive tradeoff (PT) curve [10], the H measure [11] and the area under the cost curve [12].

However, all measures of performance at a single operating point are too specific—they depend on a specific choice of misclassification costs that reflect a single or average patient, and lack information about performance at points nearby where performance may change rapidly [13], [14].

On the other hand, all measures of performance at all operating points, i.e., global measures, are too general. AUC, a global measure, is preferred over accuracy [14] but AUC is criticized because it includes operating points that would not be used in practice [15], [16] and it doesn't provide any information about the distribution of performance along the ROC curve [15].

ROC plots are intended to show the distribution of performance for further analysis [6], [17] but they are visual and do not provide a number of useful quantitative measures—e.g., what is the average sensitivity, AUC and positive predictive value within a group? Precision recall

- *A.M. Carrington is with the Ottawa Hospital Research Institute and the Institute of Clinical Evaluative Sciences, Ottawa, Canada. E-mail: acarrington@ohri.ca*
- *D.G. Manuel is with the Ottawa Hospital Research Institute; the Department of Family Medicine and the School of Epidemiology and Public Health, University of Ottawa; the Institute of Clinical Evaluative Sciences; the Bruyère Research Institute. E-mail: dmanuel@ohri.ca*
- *P. Fieguth is with the Department of Systems Design Engineering, codirects the Vision and Image Processing Lab, and is Associate Dean in the 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 and is an undergraduate student in the Department of Biology, University of Ottawa, Canada. E-mail: ysheikh@ohri.ca*
- *A. Holzinger is with the Alberta Machine Intelligence Institute, University of Alberta, Canada and head of 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 high risk group is most relevant. AUC, as a global measure, obscures all of the group-wise measures. Relative to the AUC, the high risk group has a better balanced average accuracy of 85%, but a significantly lower average sensitivity of 67%. The high risk group has the highest balanced average accuracy among groups—so the result may not improve by optimizing with different hyperparameters. Confidence intervals are omitted for simplicity and post-test measures are discussed in case studies (Sections 5, 6).

ROC horizontal axis (FPR): Probability/risk group:	Global [0,1] All	Left [0,.33] High	Mid [.33,.67] Med	Right [.67,1] Low
Bal Avg Accuracy $= AUC$	0.82			
Group Bal Avg Acc $= \widetilde{cpAUC}$	0.82	0.85	0.81	0.76
Group Avg Sens $= \widetilde{pAUC}$	0.82	0.67	0.84	0.94
Group Avg Spec $= \widetilde{pAUC}x$	0.82	0.93	0.67	0.40
Positive predictive value at a point	0.48 (at threshold=0.5)			

curve (PRC) plots have similar shortcomings.

ROC analysis is typically used to observe the dominance or rank of classifiers overall, to observe where dominance changes when ROC curves cross, or to choose an optimal ROC point or threshold [3], [17].

We posit the need for **deep ROC analysis**—a quantitative analysis of ROC data based on explicitly specified groups of probability or risk (Table 1). In comparison to global measures of performance, or performance at a point, deep ROC analysis can lead to different decisions to select or accept a binary classifier or test.

In our proposed method one may use as many risk groups as needed, only limited by the number of instances (e.g., patients) in the data. The risk groups may be percentiles of predicted risk or probability, intervals in specificity (or its complement, *FPR*), or intervals in sensitivity (*TPR*).

Support for our group-wise approach can be found in a recent systematic review. Wynants *et al.* [18] examined over 100 COVID-19 prediction models and recommended that none of the models be used in practice, in part because of lack of reporting on calibration. Calibration measures performance by groups, similar to our proposed deep ROC analysis, except our method focuses on measures of discrimination, as distinct from calibration [19].

Two key contributions of deep ROC analysis are:

1) **properly measuring AUC in groups with the normalized concordant partial AUC ($\widetilde{\textbf{cpAUC}}$) [20]**
2) **a new interpretation of AUC and $\widetilde{\textbf{cpAUC}}$ as balanced average accuracy**

To compare groups organized left to right in an ROC plot (Table 1) we cannot use the group averages for sensitivity which always increase to the right nor the group averages for specificity which always increase to the left. We need the concept of AUC within a group which is fulfilled by $\widetilde{cpAUC}$. In the Related Work and Background sections that follow we explain why alternatives are improper or insufficient.

Also, current interpretations of AUC are lacking and abstract [16], [21]. If you 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 are as follows.

First, one might receive a comparative explanation: that 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. As the name indicates AUC is the area under the ROC curve which is depicted in an ROC plot. Considering the plot's axes of sensitivity and 1-specificity, the AUC represents how sensitive and specific a classifier or test is at many different operating points along the ROC curve. However this explanation does not tell us what an AUC of 0.8 means precisely: how many errors will the classifier or test commit and in which subgroups?

The second more precise answer is that the AUC can be interpreted as a C statistic: the likelihood that the classifier ranks (scores) a randomly chosen positive patient higher than a randomly chosen negative patient. Therefore, an AUC of 80% means that the classifier is correct 80% of the time in pairwise ranking; and a 2% improvement means that in pairwise ranking, the classifier is correct 2% more often—which seems meaningful at first, but ranking is not decision-making. What is the probability of error for a single patient? What is the probability of error for a subgroup of patients (e.g., those who are predicted with high probability of having the condition)?

Consider, if the 2% improvement only ranks low-risk patients better against each other, or only ranks high-risk patients better against each other, then that improvement may not change our decision-making that distinguishes high from low risk, nor the classifier's output. A classifier is concerned with discriminating those with a condition from those without. Hence, we seek a better interpretation than the C statistic's concept of pairwise-ranking.

Two other interpretations of the AUC are that: AUC equals average sensitivity across all thresholds, and AUC equals average specificity across all thresholds [22]—as observed in the Global column of Table 1.

Hence a classifier with an AUC that is 2% higher, is on average, over all possible thresholds, 2% more sensitive at detecting positives and 2% more specific (i.e., it detects negatives 2% better). These equalities are not true for part of an ROC curve, however, where average sensitivity and average specificity, in general, differ [20].

What does hold true, is that the average (or balance) of average sensitivity and average specificity, i.e., **balanced average accuracy, is equal to AUC** (Section 7); and for part of an ROC curve, **balanced average accuracy is equal to $\widetilde{\textbf{cpAUC}}$** (Section 8).

Our finding on balanced average accuracy is not to be confused with a previously-known special case. When an ROC curve consists of a single point, S, aside from the peripheral endpoints $(0,0)$ and $(1,1)$, then AUC_S in that special case is equal to balanced accuracy (14) at the point S [23]. This special case occurs for discrete classifiers [3], e.g., a decision rule or decision tree.

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

2 Related Work

Our work seeks to understand and interpret model performance with AUC and related measures in greater detail with a new method and a new interpretation of AUC.

ROC analysis has become a standard tool in the design and evaluation of two-class classification problems [24] with ongoing work [25], [26] and extensions [27]. This is because it allows us to analyze operating points and incorporate costs and priors—which are important issues for many real-world problems where conditions are often non-ideal (non i.i.d.). Analysis of models with ROC plots is a topic with continued growth in the statistical literature [27].

Great advances have been made in machine learning and particularly in deep learning applied to various fields of medicine and smart health with high accuracy [28], [29]. To make such successes even more successful the field of explainable artificial intelligence (xAI) is attracting much interest in the health domain [30], [31].

The xAI community is supporting such efforts in developing methods that provide transparency and traceability for such deep learning approaches which are considered as statistical "black-box" methods. [32]. Recent work on a large-scale nonlinear AUC maximization method (called TSAM) based on triply stochastic gradient descents is relevant for ROC and performance analysis in machine learning generally and for explainable AI specifically [33].

There is literature advising on problems to avoid with measures [6], [16], [34] and there are surveys of available measures [1], [10], [35]. However, there is less literature on how to best use measures together, i.e., overall methods, for greater insight and effectiveness.

Sokolova *et al.* [23] argue that performance measures commonly used in machine learning do not properly address situations where the classes are equally important and several models are compared. They propose three measures: Youden's index[1], likelihood ratios, and discriminant power—but these measures are not popular. Sokolova and Lapalme [36] survey the invariant properties of performance measures and recommend measures for natural language processing.

Steyerberg *et al.* [2] recommend reporting measures of discrimination and calibration, including the C statistic or AUC, and measures from a calibration plot. These measures are popular, and our method supplements them with deeper analysis. For clinical decision-making they also recommend net benefit as a measure of clinical utility. Steyerberg and Vergouwe [37] re-iterate the same categories but narrow down the measures a little further from six to four[2].

Mallett *et al.* [16] discuss various measures of discrimination and clinical utility, including the partial AUC's benefit over AUC, but they do not discuss measuring multiple groups and they do not provide an overall method. Obuchowski and Bullen [24] provide a survey of case studies or applications of AUC and related measures but they do not provide a general method to follow.

Several reporting guidelines have been produced to improve the completeness and transparency of published studies of diagnostic tests and prediction models. For example, STARD asks authors to report their positivity cutoffs, how they were determined and whether they were defined a priori [38]. Similarly, TRIPOD asks authors to define all predictors and the outcome that is predicted by the prediction model, including how and when they were measured [39]. Across all reporting guidelines published to date, none accommodate personalized medicine with classifiers whose thresholds can be tuned or re-calibrated at the point of service specific to the setting, a group, or even a patient. How certain can we be that a test will be suitable for a given population? A GRADE assessment represents our confidence that the true accuracy of a diagnostic test lies above or below a threshold, or in a specified range [40] that depends on prevalence. The range may also consider the cost of a test's direct effects and its downstream health consequences of true and false positives/negatives [40].

Since our proposed method examines performance by groups of risk, or parts of the ROC and AUC data, we discuss precedents for that approach.

Examples of ROC analysis in the literature that applied a group-wise approach include Provost *et al.* who describe the dominant classifier in groups by slope (or skew) where a different classifier dominates in each group [17]. 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 interest.

Bradley [41] 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. Also, while it is scaled to the same range as the AUC or C statistic, it is not shown to have the same or comparable meaning.

Other examples are Carrington *et al.* [20] and Wernly *et al.* [42] who compare classifiers by the partial AUC and the concordant partial AUC in groups by false positive rate—but these measures are not popular or familiar. When those two measures are normalized, however, they are familiar as group average sensitivity and the group's AUC, respectively.

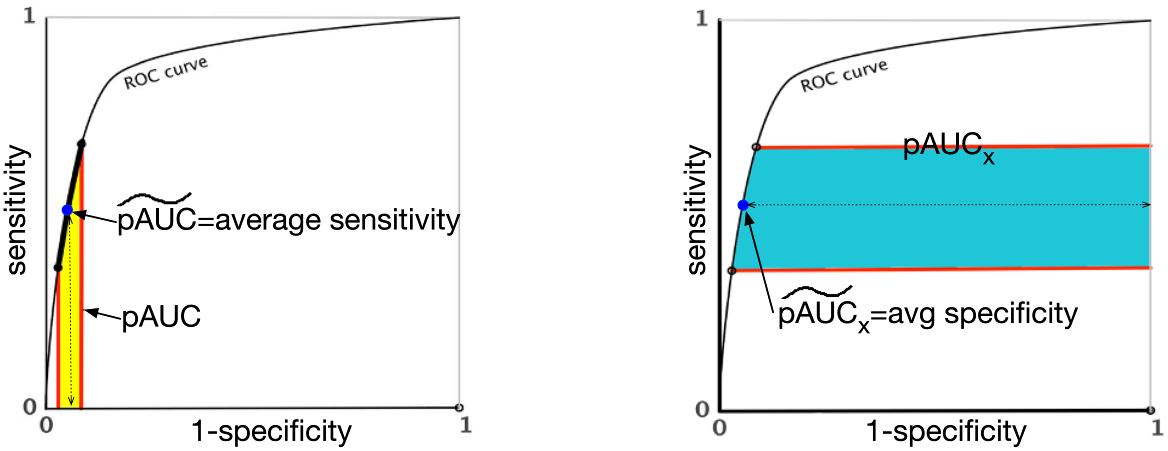
There is also ample literature on performance measures that may be used in a part or group-wise manner without attempting to provide a holistic method [13], [20], [41], [43], [44], [45], [46], [47], [48], [49]—we use and review some of these in the next section.

On interpretations of the AUC, related work includes: the concordant partial AUC as a generalization of the AUC and its relation to the partial C statistic [20], AUC related to utility [11], [12], [50], AUC related to AUPRC [51], [52] and conceptual discussion of requirements for AUC related to utility [53].

Our new interpretations of AUC and our method also support causability for explainable medicine [30], [54], as a step beyond explainable AI [55]. The term causability was coined in reference to the established term usability, and is defined as the measurable extent to which an explanation, from AI, considered by a human expert, achieves a specified level of causal understanding. The quality of explanation can be measured—e.g., with the System Causability Scale [56]—in the same way that usability encompasses measurements for the quality of use.

[1] Youden's index is linearly related to balanced accuracy at the point where the ROC curve intersects the minor diagonal.

[2] Although a fifth measure, the odds ratio, is also discussed.



a) The partial AUC ($pAUC$) is a vertical slice of the area under the ROC curve (AUC) and its normalized value ($\widetilde{pAUC}$) is the average sensitivity (height)

b) The horizontal partial AUC ($pAUCx$) is a horizontal slice of the AUC and its normalized value ($\widetilde{pAUCx}$) is the average specificity (width)

Fig. 1. 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.

3 Background

While Bradley [14] recommended AUC over accuracy, others have since identified issues with the area under the ROC curve (AUC) as a measure of performance [13], [15], [21]—and these criticisms also apply to the C statistic for binary outcomes[3]. The C statistic for binary outcomes [2], [58] we refer to should **not** be confused with Harrell or Uno's C statistics for continuous outcomes [59], [60], [61].

Since AUC is an overall measure, McClish and, separately, Thomson and Zucchini [48], [49] proposed the partial AUC ($pAUC$) (Figure 1a), which can be applied to any subset of the false positive rate (1 - specificity). This was a first step toward deep ROC analysis. A later definition of $pAUC$ used a non-parametric fit with fewer assumptions [48]. When the $pAUC$ is normalized by its range for $\Delta x = x_2 - x_1$ for an ROC curve $y = r(x)$ it becomes:

$$\widetilde{pAUC}(x_1, x_2) = \frac{1}{\Delta x} \int_{x_1}^{x_2} r(x)\, dx \tag{1}$$

The name partial AUC is misleading because it does not have all of AUC's characteristics [20].

Mallet *et al.*, who promoted the use of $pAUC$ and compared two tests [16, Fig. 3e,f] with it, suggested "...the tests are equally effective" based on similar $pAUC$ values [16, Pg. 4]. However, the tests had nearly identical sensitivity but starkly different specificity: $78-95\%$ versus $50-60\%$. Mallet *et al.* provide no discussion or rationale for this discrepancy. $pAUC$ considers the width of the range of specificity but not its values.

Soon after, McClish [13] acknowledged that $pAUC$ is flawed because it monotonically increases to the right in an ROC plot—and others also found fault with $pAUC$ [62]. McClish [13] therefore proposed the standardized Partial Area (sPA) which begins with the $pAUC$, subtracts the area under the major diagonal, and then standardizes the result. sPA is intended for comparison to the AUC.

While sPA eliminates or reduces monotonic behaviour, its approach is flawed [20], [63] because it can produce a negative result for an ROC curve that is partly above the major diagonal and partly below it. Such ROC curves occur in real life [22], [27], [63], [64]. Negative values of sPA mean that sPA cannot be interpreted as an AUC or C statistic, because the formulas for the latter are only additive; and other measures we will discuss are interpretable as an AUC or C statistic [20].

If we return our attention to $pAUC$, it does not meet the requirements for an overall measure—but it is useful when properly applied with other measures. The partial AUC [46], when normalized ($\widetilde{pAUC}$), is average sensitivity [20] and therefore has a vertical perspective (Figure 1a).

The $pAUC$ has a horizontal counterpart: the partial area index (PAI) or the normalized horizontal partial AUC ($\widetilde{pAUCx}$) as average specificity [20], [47] (Figure 1b) over the range $\Delta y = y_2 - y_1$ for an ROC curve $x = r^{-1}(y)$:

$$\widetilde{pAUC}_x(y_1, y_2) = \frac{1}{\Delta y} \int_{y_1}^{y_2} 1 - r^{-1}(y) dy \tag{2}$$

We apply both of these measures in our method (Table 1) as well as the next measure.

Carrington *et al.* define the concordant partial AUC ($cpAUC$) and partial C statistic (C_Δ), as fully and properly analogous to the AUC and C statistic [20]—as generalizations, in fact. When normalized, $\widetilde{cpAUC}$ (Figure 2) with $\theta = (x_1, x_2, y_1, y_2)$ is interpreted as the AUC in that part, and can be compared to the AUC or $\widetilde{cpAUC}$ of any other part:

$$\widetilde{cpAUC}(\theta) = \frac{1}{2\Delta x} \int_{x_1}^{x_2} r(x)\, dx + \frac{1}{2\Delta y} \int_{y_1}^{y_2} 1 - r^{-1}(y) dy \tag{3}$$

We use this measure in our method, along with a new finding: that $\widetilde{cpAUC}$ and AUC are balanced average accuracy (Sections 7 and 8). ROC curves that go above and below the

[3] For empirical ROC curves and binary outcomes the AUC and C statistic are equal [19], [34], [57]

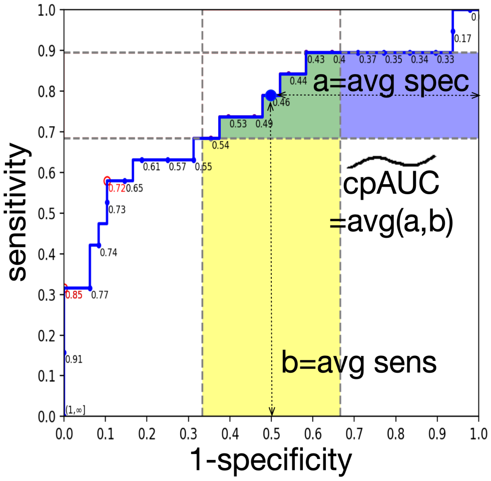


Fig. 2. The normalized concordant partial AUC $\widetilde{cpAUC}$, is illustrated for the middle third of the ROC plot. It has the same meaning and range as the AUC–it is a generalization thereof. It is interpreted as balanced average accuracy. It combines a vertical (yellow) and horizontal (blue) perspective.

major diagonal yield positive values for $\widetilde{cpAUC}$, i.e., they are properly handled.

We note that the averages in the above measures are averages of continuous values, integrals in fact, within a group of predicted risk. They are **not** averages over multiple experiments or cross-validation folds.

4 Method: Deep ROC Analysis

We propose deep ROC analysis for binary classifiers, diagnostic tests, or binary prognosis at a time point to improve or confirm model selection, understanding and explanations. Our method examines measures of discrimination in greater detail, within groups. It may complement calibration measures (if the same groups are used) and it does not include clinical utility (or rewards as utility), which may be evaluated separately. A Python toolkit[4] for the method is provided for general use and limitations of the method are discussed in a later section.

4.1 Design rationale

We have several typical objectives when evaluating model performance:

A. To measure detection of the outcome of interest (positives); and
B. To include, rather than ignore or under-weigh, detection of the other class (negatives). Otherwise too many false positives may occur.
C. To measure pre-test and post-test detection rates.
D. To compare a model against other models.
E. To know whether the model commits more errors in detection for some groups compared to others, especially for the most relevant group(s).

To ensure detection of the outcome, we examine sensitivity (pre-test) to understand what proportion of actual positives will be detected. Pre-test measures are easy to understand and they are "concrete"–i.e., their direct effect on errors is obvious.

To ensure that test results are good we also need post-test measures. Positive predictive value (PPV) also called precision, is a popular [35], [65], [66] and concrete measure, and it is easy to understand: it measures how often a positive test result is correct. However, PPV can be misused [67]: it is only informative for low prevalence; and conversely, negative predictive value (NPV) is only informative for high prevalence. We therefore consider if likelihood ratios [36], [67], [68] are better—and we conclude that it depends on the goal.

If the goal is to evaluate a test in its real-world effect with prevalence [66], [67], or treat patients[5], or explain outcomes and errors [66], then predictive values, PPV or NPV, are recommended.

If the goal is to select a test based on its intrinsic strength [36], [67], [68], without regard for prevalence and the real world effect on results and errors, possibly to account for different settings with different prevalence [68], then likelihood ratios are recommended. Likelihood ratio positive (LR+) measures the detection of true positives relative to false positives, but it is not easy to understand [67] because it deals with odds.

To include detection of the negative class, we use a combined pre-test measure that is (evenly) balanced in its consideration of the positive and negative class, e.g., the AUC, which is balanced average accuracy. In part of an ROC curve, AUC is measured by the normalized concordant partial AUC, which is also balanced average accuracy. For a combined post-test measure, the diagnostic odds ratio is a logical measure associated with the likelihood ratios.

Objectives D and E, that compare performance, require a combined measure, and for that we also use AUC and the normalized concordant partial AUC, as balanced average accuracy.

Accuracy, is not a good alternative to AUC or sensitivity because, for low prevalence, accuracy obscures the outcome (inadequate for both) and for high prevalence it weighs the outcome too much (inadequate as an alternative to AUC).

4.2 Steps in the Method

Our method has the following steps:

1) Identify the purpose. To evaluate model performance:
   a) in general, or
   b) in general and for specific groups of patients (or instances) by predicted risk or probability
2) Decide whether the group boundaries are by
   a) percentiles of FPR (or its complement, specificity), or
   b) percentiles of TPR, i.e., sensitivity, recall, or
   c) percentiles of the predicted risk or probability
3) Decide how many groups to use and their boundary values. There should be at least 25 patients (preferably 50 or more) in each group.
4) Create a table of average pre-test and post-test measures (e.g., Tables 1 and 2) or plot measures (e.g., Figure 4). These complement an ROC plot (Figure 3).

[4] https://github.com/Big-Life-Lab/partial-AUC-C

[5] Worster *et al.*'s states that predictive values (PPV and NPV) relate to the "probability of disease in an individual patient"

5) Measure how well the model detects positives and negatives, with pre-test measures. Evaluate which models perform best and sufficiently:
   a) in the most relevant group(s), measured by average sensitivity, assuming the outcome is of primary interest
   b) in the most relevant group(s), by "AUC within the group": the concordant partial AUC (balanced average accuracy), as a combined measure, to include negatives (and avoid too many false positives)
   c) in a manner that is even across groups, or that gradually favours relevant group(s), measured by "AUC within the group": the concordant partial AUC (balanced average accuracy)
   d) overall, measured by AUC (balanced average accuracy)
6) Measure how often a test result is correct, with post-test measures, in absolute terms with PPV (or in relative terms with LR+). Evaluate which models perform best and sufficiently:
   a) in the most relevant group(s), measured by average PPV (or LR+), assuming the outcome is of primary interest and the prevalence is low. For high prevalence use NPV (or LR-).
   b) in the most relevant group(s), measured by balanced predictive value (or the odds ratio), as a combined measure, to include negatives.
7) It is highly recommended to produce a calibration plot [18]. If the groups within that plot align to deep ROC analysis then the plot and analysis may be compared directly.

4.3 Calibrated scores

Calibrated scores for models are generally recommended. In binary classification and diagnostic testing, models not only estimate outcomes, they also output classification scores that are used to create ROC curves. Classification scores for some machine learning models are not **probabilistic** by default, yet probabilities are meaningful for interpretation.

By default, a support vector machine produces scores in the range $[-\infty, +\infty]$ or $[a, b], a, b \in \mathbb{R}$ and some neural networks produce scores in the range $[a, b], a, b \in \mathbb{R}$. Calibration turns non-probabilistic scores into probabilities and improves measures of calibration [2], e.g., calibration plots and calibration in the large.

Calibration [69] is an extra stage of processing that uses isotonic regression [70], [71] or Platt's method [72], [73]. It may be built into the model's implemented function or it may be available separately.

Finally, probabilistic or calibrated scores are required for option 2c in our method. Classification scores from logistic regression [74] and naive Bayes [74] are probabilistic (based on model assumptions) but they may not be well calibrated if those assumptions are not correct. Calibration can help in that case.

TABLE 2
The neural network (abbreviated as LSTM) performs consistently well in balanced average accuracy across groups of risk by FPR: [0, 0.33], [0.33, 0.67], [0.67, 1]. Average sensitivity avgSens_θ is always maximal at right, while average specificity avgSpec_θ is always maximal at left, for equally-sized subgroups.

ROC horizontal axis (FPR): Probability/risk group:	Global [0,1] All	Left [0,.33] High	Mid [.33,.67] Med	Right [.67,1] Low
LSTM				
Bal Avg Accuracy = AUC	0.88			
Group Bal Avg Acc = $\widetilde{cpAUC}$	0.88	**0.89**	0.85	0.87
Group Avg Sens = $\widetilde{pAUC}$	0.88	0.76	0.91	**0.97**
Group Avg Spec = $\widetilde{pAUCx}$	0.88	**0.94**	0.57	0.20
Positive predictive value		0.60 at a point (t=0.5)		
Negative predictive value		0.96 at a point (t=0.5)		

TABLE 3
LR performs slightly better than Lactate, but not adequately and SOFA performs poorly in groups of risk by FPR. SOFA performs best in the wrong group.

ROC horizontal axis (FPR): Probability/risk group:	Global [0,1] All	Left [0,.33] High	Mid [.33,.67] Med	Right [.67,1] Low
LR				
Bal Avg Accuracy = AUC	0.82			
Group Bal Avg Acc = $\widetilde{cpAUC}$	0.82	**0.85**	0.81	0.76
Group Avg Sens = $\widetilde{pAUC}$	0.82	0.67	0.84	**0.94**
Group Avg Spec = $\widetilde{pAUCx}$	0.82	**0.93**	0.67	0.40
Positive predictive value		0.48 at a point (t=0.5)		
Negative predictive value		0.95 at a point (t=0.5)		
Lactate				
Bal Avg Accuracy = AUC	0.80			
Group Bal Avg Acc = $\widetilde{cpAUC}$	0.80	**0.81**	0.80	0.80
Group Avg Sens = $\widetilde{pAUC}$	0.80	0.58	0.88	**0.94**
Group Avg Spec = $\widetilde{pAUCx}$	0.80	**0.91**	0.65	0.14
Positive predictive value			-	
Negative predictive value			-	
SOFA				
Bal Avg Accuracy = AUC	0.72			
Group Bal Avg Acc = $\widetilde{cpAUC}$	0.72	0.67	0.72	**0.84**
Group Avg Sens = $\widetilde{pAUC}$	0.72	0.39	0.82	**0.94**
Group Avg Spec = $\widetilde{pAUCx}$	0.72	**0.80**	0.60	0.44
Positive predictive value		0.23 at a point (t=0.5)		
Negative predictive value		0.92 at a point (t=0.5)		

5 Case Study: Mortality prediction based on arterial blood gas analysis of septic patients

Wernly *et al.* [42] provide a useful illustration of the need for partial area measures of discrimination in subgroups. They 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 (ABG) data against several baseline models and clinical scales: Logistic Regression (LR), the SOFA score evaluating functioning of six organs, and against blood lactate levels as a sole predictor. We normalize the partial area measures reported by Wernly *et al.* (Tables 2, 3) for interpretation.

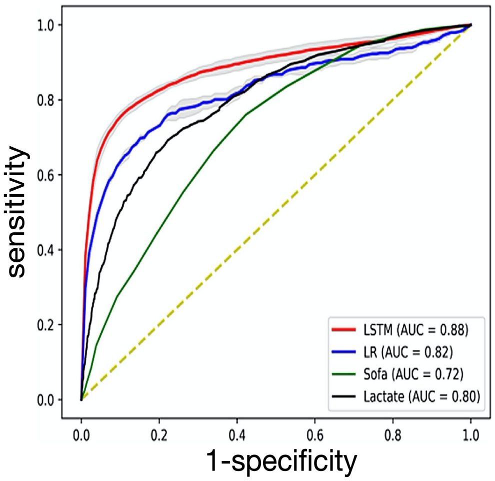


Fig. 3. The ROC plot for the four classifiers: LSTM, LR, Lactate and SOFA. LSTM is almost fully dominant.

First, we examine AUC as an overall measure. SOFA has an AUC, or average balanced accuracy of 0.72 or 72% (Table 3), which is moderately predictive [42], while lactate and Logistic Regression perform well with an AUC of 80% and 82% respectively and LSTM's AUC of 88% (Table 2) is 6% better than others in absolute terms.

In other words, LSTM is 88% accurate on average (on balance) in detecting positives and negatives without regard for class imbalance or prevalence. The AUC of 88% is the average of average sensitivity at 88% and average specificity 88% where these measures are **necessarily equal** for the whole ROC curve, but generally different for a partial ROC curve [20] (Table 2, "whole" column).

In the ROC plot (Figure 3) for FPR<0.35, the curves are above each other (better) in the same order as the AUC values, but for FPR>0.5 Lactate is better than LR, and at FPR>0.63 SOFA is better than LR too.

Wernly *et al.* [42] 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, our split of data into high, medium and low risk thirds by FPR, seems reasonable. Rather than eyeballing averages from the plot, we quantify the concordant partial AUC (or average balanced accuracy) in that region as 0.67, 0.81, 0.85 and 0.89 for SOFA, Lactate, LR and LSTM. This means that in the region that is most relevant: SOFA is 5% worse than what the AUC indicates while LR is 3% better, and both LSTM and Lactate are 1% better.

If we examine average sensitivity in the high-risk region, the differences between algorithms grow. Between LSTM and LR, an overall difference in AUC (average balanced accuracy) of 6%, and a high-risk difference in the same concept ($\widetilde{cpAUC}$ = average balanced accuracy) of 4%, hides a 9% difference in average sensitivity, which is arguably more important than average specificity and average balanced accuracy for this scenario. That said, it is important to have the complete set of measures—the complete picture and it is helpful to report and compare $\widetilde{cpAUC}$ against the AUC value.

In absolute terms, in the high-risk region, LSTM and LR are 76% and 67% sensitive (on average) while AUC paints a rosier picture. Lactate has 58% average sensitivity, which is not that good, while SOFA is 39% sensitive on average, which is terrible and worse than chance.

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 proposed ABG-LTSM 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 ($\widetilde{cpAUC}$) where it matters least (Table 3), while Lactate performs consistently across all 3 risk groups.

LR and LSTM perform best in the high-risk region (Table 3). Average sensitivity and average specificity, individually cannot be compared across risk groups because they monotonically increase and decrease, respectively, from high-risk to low-risk (left to right).

It is important to repeat that the "partial AUC" is a misnomer. For any data, given equally sized bins, the partial AUC whether normalized ($\widetilde{pAUC}$) or not ($pAUC$) will always have the best value in the rightmost bin. If one interprets it like the AUC, then they will erroneously conclude that LSTM is most accurate overall in the rightmost (low risk) region. The concordant partial AUC [20] ($\widetilde{cpAUC}$) is the proper analogue to AUC.

6 Case study 2: German Breast Cancer Study Group

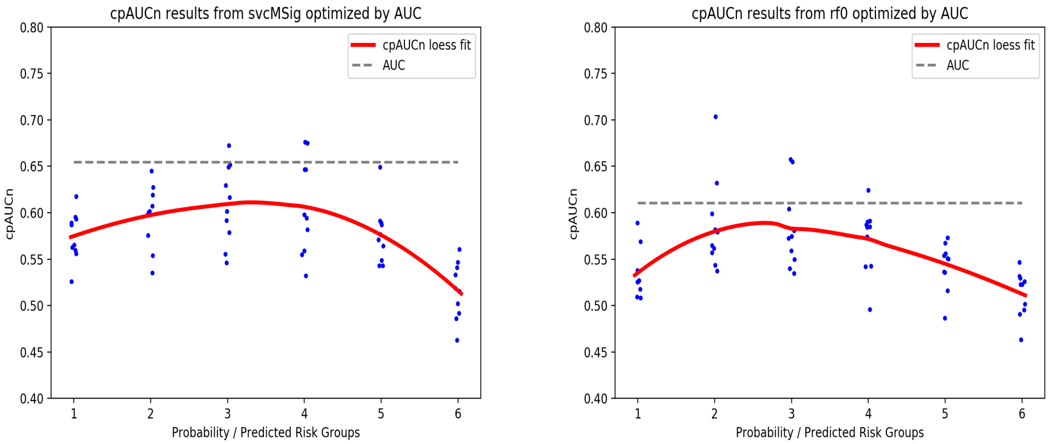
In survival analysis of patients in the German breast cancer study group [75], 33% of patients with positive node primary breast cancer had isolated locoregional recurrence at 2 years after treatment. For this low prevalence situation, the minority of positives are most clinically relevant—i.e., high-risk patients identified by the leftmost part of the ROC plot.

We applied 14 models with many different hyperparameters to predict recurrence: support vector machines (3 kernels), random forests (3 batch sizes), penalized logistic regression (4 loss functions), shallow neural networks (2 activation functions), k-nearest neighbors and decision trees. Based on experiments we focused our analysis on three models: (i) the top performing algorithm by AUC, a support vector machine (SVM) with a Mercer sigmoid kernel [76], (ii) a common statistical algorithm, penalized logistic regression with ridge/L1 loss, and (iii) a random forest model with a small batch size.

We computed and stored results in a 4-dimensional matrix:

- 100 iterations/points in hyperparameter optimization[6]

[6] Because of the efficiency of Bayesian search optimization, we found in experimentation that 100 iterations was sufficient–with only slight gains achieved on occasion by using 200 iterations instead. Alternative methods such as random search or grid search require more iterations.



a) Support Vector Machine performance with a Mercer sigmoid kernel sags by about 3% and 9% in the high risk and low risk groups (left and right) respectively.

b) Random Forests performance with a small batch size sags approximately $5 - 6\%$ in the high and low risk groups.

Fig. 4. For the German Breast cancer Study Group data the performance of all models sags at both extents of 6 risk groups shown along the x-axis. Points from each of 10 folds are jittered for visual clarity. We explain in text the reason why the group measures (red fitted line) are below the overall measure (dashed grey line).

- 10 folds in 2 x 5-fold cross-validation
- 6 groups of probability or predicted risk
- 15 group measures

We found that there was no significant difference in AUC between the best SVM model and the best random forests model when we tested the difference between matched pairs over 10 folds. Similarly between SVM and penalized logistic regression there was no significant difference in AUC. However, in the high risk group, group 1, there was a significant difference between SVM and random forests in both cpAUCn (balanced average accuracy) and pAUCn (average sensitivity), using the same test. And the plot for logistic regression (not shown) dipped in the center and in that group it was significantly different from SVM. Hence, in some situations, deep ROC analysis leads to different decisions for model selection than than standard ROC analysis.

In the plots of performance (Figures 4a and b) there is a noticeable feature: the average of group measures are below the overall measure for all groups in this case study. This is in contrast to the previous case study where the high risk group exceeded overall performance in two cases.

We verified that this behaviour occurs with very small and simple datasets—as was used for this case study. There were 686 samples with 228 positives split into 5 folds. When we use 6 groups for deep ROC analysis, errors in the minority class as a proportion of instances are exaggerated. AUC is balanced average accuracy, so the exaggerated minority class errors weigh as much as majority class errors.

We then tested whether or not using group measures for the objective in Bayesian search optimization of hyperparameters would yield better performance in absolute terms. Instead of optimizing for AUC or AUPRC, we tried optimizing $\widehat{cpAUC}$ in group 1, and also optimizing $\widehat{pAUC}$ in group 1. The results (Table 4) show no significant difference in average performance in any of the four measures.

TABLE 4. Results for a Support Vector Machine with a Mercer Sigmoid kernel. *No maxima are significant.

```
1. Bayesian search maximizing AUC
Max mean_AUC      65.43 +/- 3.28 is at index 64
Max mean_AUPRC    55.60 +/- 7.24 is at index 13
Max mean_cpAUCn1  58.37 +/- 2.62 is at index 34
Max mean_pAUCn1   22.85 +/- 4.51 is at index 34 *
Max mean_avgPPV   46.79 +/- 3.14 is at index 13
Max mean_avgNPV   75.53 +/- 1.50 is at index 70

2. Bayesian search maximizing AUPRC
Max mean_AUC      65.53 +/- 3.84 is at index 0
Max mean_AUPRC    55.37 +/- 7.10 is at index 51
Max mean_cpAUCn1  58.44 +/- 2.77 is at index 86 *
Max mean_pAUCn1   22.67 +/- 4.22 is at index 24
Max mean_avgPPV   46.86 +/- 3.36 is at index 73
Max mean_avgNPV   75.68 +/- 1.64 is at index 3  *

3. Bayesian search maximizing pAUCn.group1
Max mean_AUC      65.74 +/- 3.57 is at index 29 *
Max mean_AUPRC    55.68 +/- 7.21 is at index 62 *
Max mean_cpAUCn1  58.29 +/- 2.61 is at index 67
Max mean_pAUCn1   22.62 +/- 4.47 is at index 49
Max mean_avgPPV   46.89 +/- 3.54 is at index 47 *
Max mean_avgNPV   75.56 +/- 1.60 is at index 87

4. Bayesian search maximizing cpAUCn.group1
Max mean_AUC      65.39 +/- 3.65 is at index 69
Max mean_AUPRC    55.17 +/- 7.00 is at index 0
Max mean_cpAUCn1  58.34 +/- 2.70 is at index 54
Max mean_pAUCn1   22.60 +/- 4.53 is at index 54
Max mean_avgPPV   46.84 +/- 3.44 is at index 15
Max mean_avgNPV   75.59 +/- 1.60 is at index 59
```

Lastly, the difference in values between AUPRC and average PPV pertains to the fact that the former is weighted by change/regions in TPR (or recall as in the PRC plot) while the latter is weighted by change/regions in FPR.

7 AUC IS BALANCED AVERAGE ACCURACY

We show that AUC, or AUC within a part, known as the concordant partial AUC, are interpreted as balanced average accuracy, an average of aggregate measures, which is different from average balanced accuracy (Section 9), an average of point measures. First we provide definitions and notation, and then we demonstrate our claim, which is similar to a proof.

The average of a function $f(z)$ for a continuous domain $z \in \mathcal{Z}$, in the range $\theta_z = [z_1, z_2]$, is the Riemann integral, divided by the size of the range $\Delta z = z_2 - z_1$ as in (4).

$$avg_{\theta_z}\ f(z) = \frac{1}{\Delta z} \int_{z_1}^{z_2} f(z) dz \tag{4}$$

We use x and y in the following equations to represent the axes of an ROC plot, leading to the following typical definitions for an ROC curve and AUC [20], [46], from a vertical perspective:

$$y = r(x) = sens(x) \tag{5}$$

$$AUC = \int_0^1 r(x)\ dx \tag{6}$$

$$= \int_0^1 sens(x)\ dx \tag{7}$$

AUC (6,7) equals average sensitivity [20], [22]—i.e., (6) is in the form of (4) with a normalization factor $1/\Delta z = 1$.

An ROC curve and AUC are also defined as follows [20], [46], from the horizontal perspective:

$$x = r^{-1}(y) = 1 - spec(y) \tag{8}$$

$$AUC = \int_0^1 1 - r^{-1}(y) dy \tag{9}$$

$$= \int_0^1 spec(y) dy \tag{10}$$

AUC (9) equals average specificity [20], [22], [47].

It follows from (6) and (9) that AUC must equal the average of the two (for $\Delta x, \Delta y = 1$):

$$AUC = \frac{1}{2} \int_0^1 r(x) dx + \frac{1}{2} \int_0^1 1 - r^{-1}(y) dy \tag{11}$$

$$= \frac{1}{2} \int_0^1 sens(x) dx + \frac{1}{2} \int_0^1 spec(y) dy \tag{12}$$

$$= avg\ [\ avg_{\Delta x}(sens(x)) + avg_{\Delta y}(spec(y))\ ] \tag{13}$$

We call the above, balanced average accuracy, because it is the balance (average) of average accuracy in each class. We further justify this interpretation as follows.

For AUC, which refers to a whole ROC curve, any weighted average of average sensitivity and average specificity is equal to AUC, because the two parts are equal, but only the simple average generalizes to a partial ROC curve [20], discussed in the next section.

We can see the similarity in form between (12), and balanced accuracy (14), b, at a point w:

$$b(w) = \frac{1}{2} sens(w) + \frac{1}{2} spec(w) \tag{14}$$

$$= avg\ [\ sens(w) + spec(w)\ ] \tag{15}$$

Equations (13) and (14) have similar interpretations—the former, AUC, refers to **average** sensitivity and **average** specificity over the whole ROC curve, while the latter, balanced accuracy, refers to sensitivity and specificity at a point. Hence the name and interpretation.

We further note that Youden's index J at a point w is related to $b(w)$:

$$J(w) = 2b(w) - 1 \tag{16}$$

8 THE NORMALIZED CONCORDANT PARTIAL AUC IS BALANCED AVERAGE ACCURACY

The previous concepts also apply to part of an ROC curve, or one of multiple groups of risk. In such a part or group, AUC is called the normalized concordant partial AUC $\widetilde{cpAUC}$, and we show that it is also Balanced Average Accuracy. $\widetilde{cpAUC}$ is defined as follows [20] for $\Delta x = x_2 - x_1$ and $\Delta y = y_2 - y_1$:

$$\widetilde{cpAUC} = \frac{1}{2\Delta x} \int_{x_1}^{x_2} r(x) dx + \frac{1}{2\Delta y} \int_{y_1}^{y_2} 1 - r^{-1}(y) dy \tag{17}$$

$$= \frac{1}{2\Delta x} \int_{x_1}^{x_2} sens(x) dx + \frac{1}{2\Delta y} \int_{y_1}^{y_2} spec(y) dy \tag{18}$$

$$= avg\ [\ avg_{\Delta x}(sens(x)) + avg_{\Delta y}(spec(y))\ ] \tag{19}$$

We can again see the similarity in form between (18), (14) and (12). It therefore yields the same interpretation: $\widetilde{cpAUC}$ is Balanced Average Accuracy as the balance (or average) of **average** sensitivity and **average** specificity for part of an ROC curve.

9 AUC IS NOT AVERAGE BALANCED ACCURACY

We have shown that AUC is the balance of average accuracy for each class: the average of average sensitivity and average specificity, computed as an integral (or computed discretely at each point[7]). However, AUC is not the average of balanced accuracy at each point. We show this with a simple example (Figs. 6).

It may also help to understand the difference in terms of equations as follows. From balanced accuracy (14) and a continuous average (4) we can express average balanced accuracy for a range $\theta_w = [w_1, w_2]$ where w is a continuous index along the ROC curve:

$$avg_{\theta_w}\ b(w) = \frac{1}{\Delta w} \int_{w_1}^{w_2} b(w)\ dw$$

$$= \frac{1}{\Delta w} \int_{w_1}^{w_2} \frac{1}{2}(sens(w) + spec(w))\ dw \tag{20}$$

Figure 5 shows the vector nature of dw in (20):

$$\Delta w = \Delta sens(w) + \Delta spec(w)$$

$$= |\Delta y| + |\Delta x| \tag{21}$$

$$dw = |dy| + |dx|$$

$$= dy - dx \tag{22}$$

$$k = \frac{1}{2\Delta w} \tag{23}$$

[7] We do not prove the discrete form because it is complex and meticulous, but it is easily seen in every experimental result using our deepROC Python toolkit

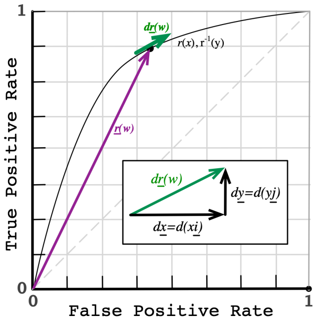


Fig. 5. The vector aspect of balanced accuracy at points w along an ROC curve.

This leads to complexity instead of expressions equal to AUC, such as the following:

$$\begin{aligned} avg_{\theta_w}\ b(x,y;w) &= k\int_{w_1}^{w_2} sens(w)\,[dy-dx] \\ &\quad +k\int_{w_1}^{w_2} spec(w)\,[dy-dx] \\ = k\int_{y_1}^{y_2} sens(y)\,dy &- k\int_{x_1}^{x_2} sens(x)\,dx \\ +k\int_{y_1}^{y_2} spec(y)\,dy &- k\int_{x_1}^{x_2} spec(x)\,dx \end{aligned} \tag{24}$$

Hence, AUC is not average balanced accuracy, but is balanced average accuracy.

10 Limitations

One possible limitation of our method is that the additional information introduces more complexity which could complicate communication of results. Providing guidance to ensure uniform reporting is recommended, wherever possible.

For small datasets that do not meet the preferred threshold of 50 instances or more per group, as in our second case study, we observed that errors bias group performance downward compared to overall measures of performance.

Another limitation is that our method pertains to binary classifiers and diagnostic tests, including prognosis at a time point with binary outcomes. Since some authors deride dichotomization, even when and where appropriate for decision-making, we discuss this point in further detail in the following section.

11 Discussion on Continuous Methods versus Predicted Risk and Subgroups in Binary Classification

An opinion piece by Wynants *et al.* [77] argues against thinking or methods that bin, categorize or group data with continuous values. It is opinion because philosophically it is up to the clinician to decide whether or not binning or categorizing helps them make decisions or not. Some of the present authors posit that many detection, diagnosis, prognosis or treatment decision-making problems are categorical in nature (e.g., the patient has strep throat) while other problems in the same categories may be ordinal or continuous (e.g., what dosage to apply).

If one is confident that clinicians and people are diligent and capable of keeping both a category and a number in mind, then there is no concern. Ideally we would all be free to choose our own tools and judge or present evidence in our own way.

It has been argued [77] that categorization loses information, and that is true in terms of information entropy, but not all information is useful. If there is too much unnecessary information, then the signal-to-noise ratio is limited and our understanding and decision-making suffers—i.e., summarizing and categorizing is useful. Summary descriptions are the essence of the word "statistic". Hence, binary classification and subgroups of predicted risk have a role to play.

That said, there are limits to binary and categorical thinking in clinical prediction models–they assume a set of options and tests known a priori, i.e., completeness. However, diagnosing and formulating a therapy for a patient may not be well-defined (explicit) nor complete. Differential diagnosis and treatment, or other decisions, may go beyond any medical protocols and order sets (if/when they exist). Decisions may or may not fall within routine experience and treatment. Diagnosis and decision-making may involve generating, synthesizing and investigating treatment options not previously considered by the clinician; and the dynamic nature of decision-making may involve a clinician's gut feel based on continuous values using Bayesian thinking. In this context, some are concerned that binary classification and categorical/subgroup methods might distract or blind a clinician, regulator, etc.

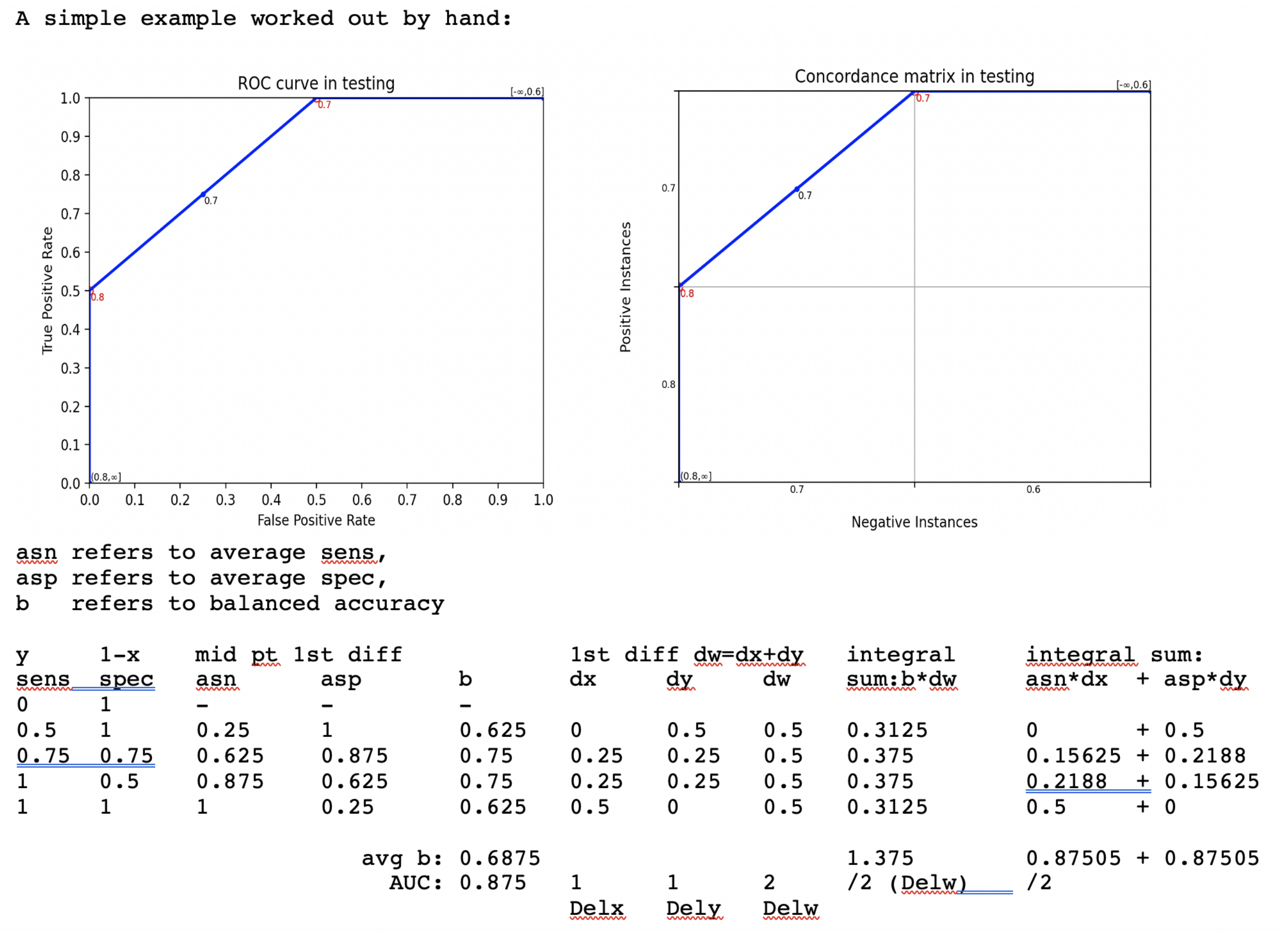
12 Conclusions and Future Work

We have shown that models (or tests) can and do behave differently in different groups of risk—and within those groups their performance that may be better or worse than the average overall. Our method may identify needs for applications using AUC where no deficiencies have been perceived [78], [79].

We have demonstrated that the normalized concordant partial AUC ($\widetilde{cpAUC}$) as balanced average accuracy is useful to interpret a model's performance in each group–it indicates where an algorithm is strong or weak. Our new interpretation of AUC is also helpful since it applies to individuals in contrast to the pairwise interpretation of AUC as a C statistic.

In the first case study, LSTM model aside: our method more clearly differentiates LR versus Lactate in the high risk group that matters most, and it more clearly shows the inadequacy of SOFA in absolute terms. Hence, deep ROC analysis can improve model selection in some cases and it provides an informed view of model performance by groups for assurance.

In the second case study, we observed how the model performs differently in groups with lesser performance in the highest and lowest risk groups for that data. We used Bayesian search optimization with group measures ($\widetilde{cpAUC}_n$ and $\widetilde{pAUC}_n$) as objectives instead of AUC or AUPRC but there was no significant difference in results. Future work could examine if group measures would improve optimization for large datasets.



A simple example worked out by hand:

```
asn refers to average sens,
asp refers to average spec,
b   refers to balanced accuracy

y     1-x    mid pt 1st diff             1st diff dw=dx+dy   integral      integral sum:
sens  spec   asn      asp       b        dx     dy     dw     sum:b*dw      asn*dx   + asp*dy
0     1      -        -         -
0.5   1      0.25     1         0.625    0      0.5    0.5    0.3125        0        + 0.5
0.75  0.75   0.625    0.875     0.75     0.25   0.25   0.5    0.375         0.15625  + 0.2188
1     0.5    0.875    0.625     0.75     0.25   0.25   0.5    0.375         0.2188   + 0.15625
1     1      1        0.25      0.625    0.5    0      0.5    0.3125        0.5      + 0

                         avg b: 0.6875                   1.375         0.87505 + 0.87505
                           AUC: 0.875    1      1      2      /2 (Delw)     /2
                                         Delx   Dely   Delw
```

Fig. 6. A simple example of AUC as balanced average accuracy not average "balanced accuracy".

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
C_Δ:	The partial C statistic
FNR:	False negative rate
FPR:	False positive rate, or 1-specificity
$pAUC$:	Partial area under the ROC curve (i.e., vertical)
$\widetilde{pAUC}$:	Normalized partial area under the ROC curve
$cpAUC$:	Concordant partial area under the ROC curve
$\widetilde{cpAUC}$:	Normalized concordant partial area under the ROC curve
$pAUCx$:	Horizontal partial area under the curve
$\widetilde{pAUCx}$:	Normalized horizontal partial area under the ROC curve
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] 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.

[2] 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.

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

[4] 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.

[5] 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.

[6] 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.

[7] 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

[8] 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.

[9] S.-Y. Shiu and C. Gatsonis, "The predictive receiver operating characteristic curve for the joint assessment of the positive and negative predictive values," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 366, no. 1874, pp. 2313–2333, 2008.

[10] C. O'Reilly and T. Nielsen, "Revisiting the ROC curve for diagnostic applications with an unbalanced class distribution," *2013 8th International Workshop on Systems, Signal Processing and Their Applications, WoSSPA 2013*, pp. 413–420, 2013.

[11] D. J. Hand, "Measuring classifier performance: A coherent alternative to the area under the ROC curve," *Machine Learning*, vol. 77, no. 1, pp. 103–123, 2009.

[12] P. A. Flach, J. Hernández-Orallo, and C. F. Ramirez, "A coherent interpretation of auc as a measure of aggregated classification performance," in *ICML*, 2011.

[13] 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

[14] 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.

[15] 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.

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

[17] F. Provost, T. Fawcett, and R. Kohavi, "The case against accuracy estimation for comparing classifiers," *Proceedings of the 15ˆ{th} International Conference on Machine Learning*, no. January 2013, pp. 445–553, 1998.

[18] 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.

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

[20] 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.

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

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

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] J. Xing, K. Li, W. Hu, C. Yuan, and H. Ling, "Diagnosing deep learning models for high accuracy age estimation from a single image," *Pattern Recognition*, vol. 66, no. 6, pp. 106–116, 2017.

[29] H. Zuo, H. Fan, E. Blasch, and H. Ling, "Combining convolutional and recurrent neural networks for human skin detection," *IEEE Signal Processing Letters*, vol. 24, no. 3, pp. 289–293, 2017.

[30] 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.

[31] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (xai): Toward medical xai," *IEEE Transactions on Neural Networks and Learning Systems*, vol. in print, 2021.

[32] 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.

[33] Z. Dang, X. Li, B. Gu, C. Deng, and H. Huang, "Large-scale nonlinear auc maximization via triply stochastic gradients," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[34] 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.

[35] D. M. W. Powers, "Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation," Flinders University, Tech. Rep. December, 2007.

[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] 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.

[38] 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.

[39] 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.

[40] 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.

[41] A. P. Bradley, "Half-auc for the evaluation of sensitive or specific classifiers," *Pattern Recognition Letters*, vol. 38, pp. 93–98, 2014.

[42] 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.

[43] 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.

[44] 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.

[45] M. S. Pepe, *The statistical evaluation of medical tests for classification and prediction*. Oxford University Press, 2003.

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

[47] 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.

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

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

[50] J. Hernández-Orallo, P. Flach, and C. Ferri, "A Unified View of Performance Metrics: Translating Threshold Choice into Expected Classification Loss," *Journal of Machine Learning Research*, vol. 13, pp. 2813–2869, 2012. [Online]. Available: http://www.jmlr.org/papers/volume13/hernandez-orallo12a/hernandez-orallo12a.pdf

[51] W. Su, Y. Yuan, and M. Zhu, "A relationship between the average precision and the area under the roc curve," in *Proceedings of the 2015 International Conference on The Theory of Information Retrieval (ICTIR'15)*, J. Allan and B. Croft, Eds. ACM SIGIR, 2015.

[52] J. Davis and M. Goadrich, "The relationship between Precision-Recall and ROC curves," *Proceedings of the 23rd international conference on Machine learning - ICML '06*, pp. 233–240, 2006. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1143844.1143874

[53] N. H. Shah, A. Milstein, and S. C. Bagley, "Making machine learning models clinically useful," *JAMA*, 2019.

[54] 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.

[55] 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.

[56] A. Holzinger, A. Carrington, and H. Mueller, "Measuring the quality of explanations: The system causability scale (scs). comparing human and machine explanations," *KI - Kuenstliche Intelligenz (German Journal of Artificial intelligence), Special Issue on Interactive Machine Learning, Edited by Kristian Kersting, TU Darmstadt*, vol. 34, no. 2, pp. 193–198, 2020.

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

[58] M. J. Pencina and R. B. D'Agostino, "Evaluating discrimination of risk prediction models: The C statistic," *JAMA - Journal of the American Medical Association*, vol. 314, no. 10, pp. 1063–1064, 2015.

[59] C. Guo, Y. So, and W. Jang, "Evaluating Predictive Accuracy of Survival Models with PROC PHREG," *Proceedings of the SAS Global Forum 2017 Conference*, pp. 1–16, 2017. [Online]. Available: https://pdfs.semanticscholar.org/0f63/7c13f7eac0dbbeb1a691da46197593fa131b.pdf

[60] H. Uno, T. Cai, M. J. Pencina, R. B. D'Agostino, and L. Wei, "On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data," *Statistics in medicine*, vol. 30, no. 10, pp. 1105–1117, 2011.

[61] F. Harrell Jr., R. Califf, D. Pryor, K. Lee, and R. Rosati, "Evaluating the yield of medical tests," *Journal of the American Medical Association*, vol. 247, no. 18, pp. 2543–2546, 1982. [Online]. Available: http://dx.doi.org/10.1001/jama.1982.03320430047030

[62] 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.

[63] 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.

[64] 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.

[65] B. Ozenne, F. Subtil, and D. Maucort-Boulch, "The precision–recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases," *Journal of clinical epidemiology*, vol. 68, no. 8, pp. 855–859, 2015.

[66] D. G. Altman and J. M. Bland, "Diagnostic tests 2: predictive values," *BMJ*, vol. 309, no. July, p. 16104, 1994.

[67] A. Worster, G. Innes, and R. B. Abu-laban, "Diagnostic testing: an emergency medicine perspective," *Canadian Journal of Emergency Medicine*, vol. 4, no. 5, 2002.

[68] J. J. Deeks and D. G. Altman, "Diagnostic tests 4: likelihood ratios," *Bmj*, vol. 329, no. 7458, pp. 168–169, 2004.

[69] 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.

[70] 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.

[71] 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.

[72] 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.

[73] 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.

[74] C. M. Bishop, *Pattern recognition and machine learning*. springer, 2006.

[75] 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.

[76] A. M. Carrington, P. W. Fieguth, and H. H. Chen, "A new mercer sigmoid kernel for clinical data classification," in *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2014, pp. 6397–6401.

[77] L. Wynants, M. van Smeden, D. J. McLernon, D. Timmerman, E. W. Steyerberg, and B. Van Calster, "Three myths about risk thresholds for prediction models," *BMC medicine*, vol. 17, no. 1, p. 192, 2019.

[78] A. Borji, "Saliency prediction in the deep learning era: Successes and limitations," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 2, pp. 679–700, 2021.

[79] W. Wang, Q. Lai, H. Fu, J. Shen, H. Ling, and R. Yang, "Salient object detection in the deep learning era: An in-depth survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. in print, 2021.

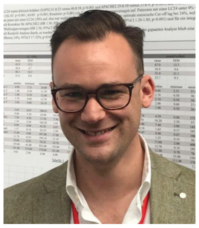
André M. Carrington is a Post Doctoral Research Fellow at the Ottawa Health Research Institute. André received his Ph.D. in Systems Design Engineering and Masters in Mathematics (Computer Science) from the University of Waterloo.

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.

Matt McInnes is a radiologist with 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.

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.