A NOVEL PERFORMANCE MEASURE FOR MACHINE LEARNING CLASSIFICATION

Mingxing Gong

Financial Services Analytics, University of Delaware Newark, DE 19716

ABSTRACT

Machine learning models have been widely used in numerous classification problems and performance measures play a critical role in machine learning model development, selection, and evaluation. This paper covers a comprehensive overview of performance measures in machine learning classification. Besides, we proposed a framework to construct a novel evaluation metric that is based on the voting results of three performance measures, each of which has strengths and limitations. The new metric can be proved better than accuracy in terms of consistency and discriminancy.

1. INTRODUCTION

Performance measures play an essential role in machine learning model development, selection, and evaluation. There are a plethora of metrics for classification, such as Accuracy, Recall, Fmeasure, Area Under the ROC Curve (AUC), Mean Squared Error (or known as Brior Score), LogLoss/Entropy, Cohen's Kappa, Matthews Correlation Coefficient (MCC), etc. Generally, the performance measures can be broadly classified into three categories [1]:

- Metrics based on a confusion matrix: accuracy, macro-/micro-averaged accuracy (arithmetic and geometric), precision, recall, F-measure, Matthews correlation coefficient and Cohen's kappa, etc.
- Metrics based on a probabilistic understanding of error, i.e., measuring the deviation from the actual probability: mean absolute error, accuracy ratio, mean squared error (Brier score), LogLoss (cross-entropy), etc.
- Metrics based on discriminatory power: ROC and its variants, AUC, Somers'D, Youden Index, precision-recall curve, Kolmogorov-Smirnov, lift chart, etc.

Confusion-matrix-based metrics are most commonly used, especially accuracy since it is intuitive and straightforward. Most classifiers aim to optimize accuracy. However, the drawbacks of accuracy are also evident. The classifier with accuracy as the objective function is more biased towards majority class. For example, in credit card fraud detection where the data is highly imbalanced, 99.9% of transactions are legitimate, and only 0.1% are fraudulent. A model merely claiming that all transactions are legitimate will reach 99.9% accuracy that almost no other classifiers can beat. However, the model is useless in practice, given no fraudulent transactions could be captured by the model. The drawback of accuracy highlights the importance of recall (or called true positive rate), which measures the effectiveness of the model in detecting fraudulent transactions. Nevertheless, recall is still a biased metric. For example, an extreme case that every transaction is classified as fraudulent would give the highest recall. To reduce false alarms, precision needs to be taken into account to measure the reliability of the prediction. In other words, those metrics alone can only capture part of the information of the model performance. That is why accuracy, recall, precision, or the combination (F-measures) are considered altogether in a balanced manner when dealing with the imbalanced data classification.

Besides, both recall and precision are focus on positive examples and predictions. Neither of them captures how the classifier handles the negative examples. It is challenging to describe all the information of the confusion matrix using one single metric. The Matthews correlation

DOI: 10.5121/ijmit.2021.13101

coefficient(MCC) is one of the measures regarded as a balanced measure, which takes into account both true and false positives and negatives. MCC is commonly used as a reference performance measure on imbalanced data sets in many fields, including bioinformatics, astronomy, etc. [2], where both negatives and positives are very important. Recent study also indicated that MCC is favourable over F1 score and accuracy in binary classification evaluation on imbalanced datasets [3]. It is worth mentioning that all confusion-matrix-based metrics are subject to threshold selection, which is very difficult, especially when misclassification cost and prior probability are unknown.

On the contrary, the probability metrics do not rely on a threshold. The probability metrics measure the deviation or entropy information between the prediction probability p(i; j) and true probability f(i; j) (0 or 1). The three most popular probability metrics are Brier score, cross entropy, and calibration score. The first two measures are minimized when the predicted probability is close to the true probability. The calibration score assesses score reliability. A well-calibrated score indicates that if the classifier prediction probability of positives for a sample of datasets is p, the proportion of positives in the sample is also p. A well-calibrated score does not mean perfect quality of the classifier.

For metrics based on discriminatory power, ROC is the most popular one [4, 5, 6]. Compared with confusion matrix based metrics, ROC has several advantages. Firstly, ROC does not rely on threshold settings [5]. Secondly, the change of prior class distribution will not impact the ROC curve, since ROC depends on true positive rate and false positive rate, which are calculated against true positives and true negatives independently. Thirdly, ROC analysis does not necessarily have to produce exact probability, and what it has to do is to discriminate positive instances from negative instances [7]. There are several ROC extensions and invariants. To integrate the decision goal with cost information and prior class distribution, ROC convex hull is developed. And under some conditions that only early retrieval information is of any interest, Concentrated ROC (CROC) was proposed to magnify the interested area using an appropriate continuous and monotone function. Besides, ROC curve is closely related to or can be transformed into precision-recall curve, cost curve, lift chart, Kolmogorov–Smirnov (K-S) alike curves, etc. Furthermore, ROC can be extended to multi-class via one vs. one (OVO) or one vs. all (OVA) methodologies. A single summary metric called Area under the ROC curve (AUC) is derived from ROC. Statistical meaning of the AUC measure is the following: AUC of a classifier is equivalent to the probability that the classifier's output probability for a randomly chosen positive instance is higher than a randomly chosen negative instance. A similar measure in multi-class is called volume under ROC space (VUS). There are a lot of discussions on multiclass ROC and VUS [8, 9, 10, 11]. In addition to the three categories of performance measures as mentioned above, the practitioner often assesses the stability of the classifiers' output based on distribution and characteristic analysis. Popularly used metrics include Population Stability Index (PSI) [12], Characteristic Stability Index (CSI), etc. In summary, there is no guarantee that one measure would definitely outperform the other in all situations.

The motivation of this study is that although there are a lot of literature to discuss the evaluation of performance of classifiers in an isolated manner (e.g., [13] wrote an overview of ROC curve; [14] discussed the impact of class imbalance in classification performance metrics based on the binary confusion matrix, etc.), there exists very few comprehensive review in this area. [15] wrote a very general review on evaluation metrics for data classification evaluations. However the review is lack of details and a lot of popular measures are not mentioned. Santafe et al. review the important aspects of the evaluation process in supervised classification in [16], which is more focus on the statistical significance tests of differences of the measures. There are also many literature papers discussing the machine learning performance measures in the specific fields. For example, [17] studied machine learning performance metrics and diagnostic context in radiology; [18] performed a

comparative study of performance metrics of data mining algorithms on medical data; [19] discussed the performance measures in text classification. Motivated by the lack of comprehensive generic review in performance measures, our paper covers a comprehensive review of above-mentioned three categories of measures. In addition, we propose a framework to construct a new and better evaluation metric.

The main contribution of this paper lies in three aspects: it provided a comprehensive and detailed review of performance measures; it highlighted the limitations and common pitfalls of using these performance measures from a practitioner's perspective; it proposed a framework to construct a better evaluation metric. The rest of the paper is structured as follows: Section 2 provides an overview of performance measures as well as their limitations and misuses in practice; Section 3 introduces a solution to compare the performance measures; Section 4 proposes a novel performance measure based on the voting method and Section 5 discusses the experimental results; Section 6 concludes the paper and discusses the potential research opportunities in the future.

2 An overview of performance measures

2.1 Metrics based on a confusion matrix

Confusion-matrix is commonly accepted in evaluating classification tasks. For binary classification, it is a two by two matrix with actual class and predicted class, illustrated as below. For short, TP stands for true positive, FN for false negative, FP for false positive, and TN for true negative. In fraud detection problem setting, we assume positive means fraud and negative means non-fraud. The metrics derived from confusion-matrix are described as follows:

		Actual Class		
		p	**n**	**total**
Predictive class	**p′**	True Positive	False Positive	P′
	n′	False Negative	True Negative	N′
	total	P	N	

Accuracy (Acc) is the most simple and common measure derived from the confusion matrix.

$$ACC = \frac{TP + TN}{P + N} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

True positive rate (TPR) or recall (also known as sensitivity) is defined as true positive (TP) divided by total number of actual positives ($TP + FN$).

$$TPR = \frac{TP}{P} = \frac{TP}{TP + FN} \tag{2}$$

Specificity or true negative rate (TNR) is calculated as

$$Specificity = \frac{TN}{N} = \frac{TN}{FP + TN} \tag{3}$$

$1 - Specificity$ is called false positive rate (FPR). Sensitivity can be interpreted as accuracy on the actual positive instance, while specificity is characterized as accuracy on the actual negative instances. Similarly, there are also metrics on performance on the predictive positive/negative instances, such as precision and negative predictive value (NPV).

$$Precision = \frac{TP}{P'} = \frac{TP}{TP + FP} \tag{4}$$

$$NPV = \frac{TN}{N'} = \frac{TN}{TN + FN} \tag{5}$$

It is worth noting that none of above-mentioned single metric alone can be sufficient to describe the model performance in fraud detection. For example, similar to accuracy, if we use recall as the performance metric, a model that predicts all transactions are fraudulent can achieve 100% recall rate. However, the model is useless.

The F-measure, a balanced performance metric, considers both recall and precision. The commonly used F_1 measure is the harmonic mean of precision and recall:

$$F_1 = 2 * \frac{Precision * Recall}{Precision + Recall} = \frac{2TP}{2TP + FP + FN} \tag{6}$$

In the F_1 measure, precision is weighted as the same important as recall. The general formula for F_β:

$$F_\beta = (1 + \beta^2) * \frac{Precision * Recall}{\beta^2 * Precision + Recall} \tag{7}$$

which weights β times as much importance to recall as precision. However, as Equation 7 shows that true negatives are neglected in F-measure which focuses more on positive instances and predictions. In addition, we should be very cautious when using F measure to compare model performance in fraud detection, especially considering resampling methods (undersampling legitimate cases or oversampling fraud cases) are often adopted in fraud detection to address imbalance data issues, which may result in different data distributions. Unlike other performance measures (e.g., recall, false positive rate, ROC) that are independent of the data set distribution, F measure will differ significantly on different data sets, as illustrated in Table 1. The model with the same fraud capture rate (recall) and false positive rate exhibits different F_1. Only relying on F_1 score may reach a misleading conclusion that the model overfits or deteriorates.

Table 1: Different F_1 scores for the same model on different data sets

(a) $Recall = 0.5; FPR = 0.091; F_1 = 0.645$

		Actual Class		
		p	**n**	**total**
Predictive class	**p′**	TP 500	FP 50	P′
	n′	FN 500	True 500	N′
	total	P	N	

(b) $Recall = 0.5;\ FPR = 0.091; F_1 = 0.50$

		Actual Class		
		p	**n**	**total**
Predictive class	**p′**	TP 500	FP 500	P′
	n′	FN 500	TN 5000	N′
	total	P	N	

It is challenging to describe all the information of the confusion matrix using one single metric. The Matthews correlation coefficient (MCC) is one of the best such measures, which takes into account both true and false positives and negatives [20]. The MCC can be calculated from confusion matrix:

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{8}$$

MCC varies from $[-1, 1]$, with 1 as perfect prediction, 0 as random prediction and -1 as total disagreement between the actual and prediction.

Cohen's kappa [21] is another measure to correct the predictive accuracy due to chance. It is calculated as:

$$k = \frac{p_0 - p_c}{1 - p_c} \tag{9}$$

p_0 is the observed accuracy, represented by $\frac{TP+TN}{TP+TN+FP+FN}$ from confusion matrix, while p_c is expected accuracy, calculated as

$$\frac{(TN + FP) * (TN + FN) + (FN + TP) * (FP + TP)}{(P + N)^2} \tag{10}$$

It is important to note that the aforementioned confusion-matrix based measures are subject to thresholds and prior class distribution. The different selection of thresholds will result in changes of confusion matrix. As long as the predicted probability is greater than the threshold, the example would be labeled as positive and vice versa. It does not matter how close the predicted probability is to the actual probability. For example, there are two classifiers for a new instance, which actually belongs to the positive class. One predicts the likelihood of positive event is 0.51 and the second classifier predicts the probability of positive event is 0.99. If the threshold is 0.50, the confusion-matrix based measures would be exactly the same for these two classifiers, although intuitively the second classifier is much better.

2.2 Metrics based on probability deviation

Measures based on probability are threshold-free metrics that assess the deviation from the true probability. There are several probability-based measures including MAE (mean absolute error) which shows how much the predictions deviate from the actual outcome; Brier score,or called mean square error, a quadratic version of MAE which penalizes greater on the deviation; Cross entropy, derived from information theory; calibration score, which is also known as Expected vs. Actual (EvA) analysis. Below are the formulas for these common probability-based metrics.
The mean absolute error (MAE) is defined as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - p_i|$$

The Brier score is defined as:

$$Brier\_score = \frac{1}{n}\sum_{i=1}^{n}(y_i - p_i)^2$$

The cross entropy is defined as:

$$cross\_entropy = \sum_{i=1}^{n} -y_i log(p_i)$$

Where p_i is the predicted probability and y_i is the ground truth probability (0 or 1 for binary classification). Brier score and cross entropy are the most commonly used metrics, which are widely used as the objective function in machine learning algorithm. Generally, cross entropy is preferred in classification problems and brier score is preferred in regression models. In practice, brier score and cross entropy are more considered as objective function/loss function, instead of performance measures. The reason is the most classification models are used for ranking-order purposes and the absolute accuracy of the model is not of significant concern.

2.3 Metrics based on discriminatory power

Two-class ROC

ROC (Receiver Operations Characteristics) is one of the most widely used performance metrics, especially in binary classification problems. A ROC curve example is shown below (Figure 1). The horizontal axis is represented by false positive rate (FPR), and the vertical axis is true positive rate (TPR). The diagonal connecting (0,0) and (1,1) indicates the situation in which the model does not provide any separation power of bad from good over random sampling. Classifiers with ROC curves away from the diagonal and located on the upper-left side are considered excellent in separation power. The maximum vertical distance between ROC curve and diagonal is known as Youden's index, which may be used as a criterion for selecting operating cut-off point.

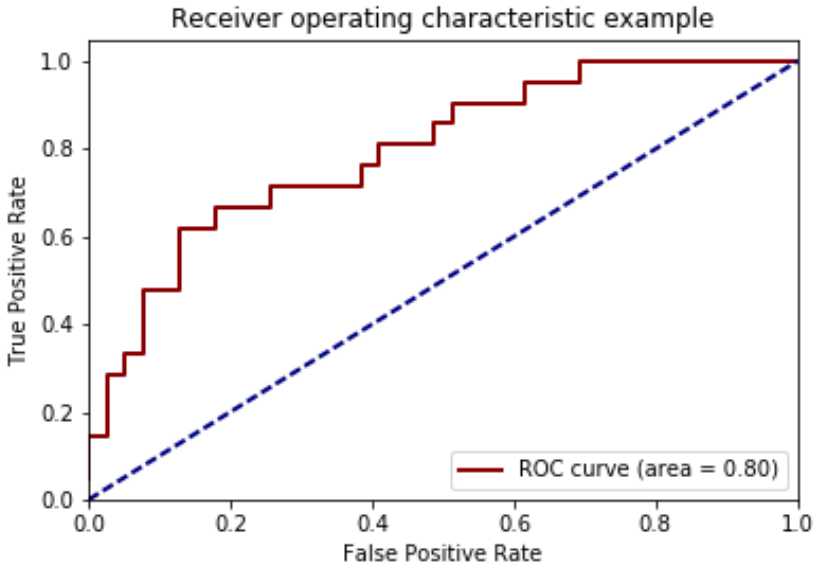

Figure 1: A ROC curve example *Source: scikit-learn simulation [22]*

ROC is considered as a trade-off between true positive rate and false positive rate. One useful feature of the ROC curve is that they remain unchanged when altering class distribution [21]. Since ROC is based on TPR and FPR, which are calculated on actual positives and actual negatives respectively, ROC curves are consequently independent of class distribution. Another useful feature of ROC curve is that the curves exhaust all possible thresholds, which contain more information than static threshold dependent metrics, such as precision, accuracy, and recall. Therefore ROC curve is deemed as a useful measure for overall ranking and separation power for model evaluation.

Unfortunately, while a ROC graph is a valuable visualization technique, it may not help select the classifiers in certain situations. Only when one classifier clearly dominates another over the entire operating space can it be declared better. For example, Figure 2 shows that we can claim that classifier 0 outperforms other classifiers since the curve is always above other curves. However, it is a little challenging to tell whether classifier 4 is better than classifier 2 or not.

Thus, A single summary metric called Area under the ROC curve (AUC) is derived from ROC. Statistical meaning of the AUC measure is the following: AUC of a classifier is equivalent to the probability that the classifier will evaluate a randomly chosen positive instance as a higher probability than a randomly chosen negative instance. Let $\{p_1, ..., p_m\}$ be the predicted probabilities of the m positives to belong to the positive class sorted by descending order. And $\{n_1, ..., n_k\}$ be the predicted probabilities of the k negatives to belong to the positive class sorted by descending order. AUC can be defined as

$$AUC = \frac{1}{mk} \sum_{i=1}^{m} \sum_{j=1}^{k} 1_{p_i > n_j} \tag{11}$$

Where $1_{p_i > n_j}$ is an indicator function, equal to 1 when $p_i > n_j$.

AUC is useful to rank the examples by the estimated probability and compare the performance

of the classifier. Huang and Ling theoretically and empirically show that AUC is more preferred than accuracy [23]. However, AUC as a single summary metric has its drawbacks. For example, the AUC for classifier 1 is 0.55, and the AUC for classifier 2 is 0.65. Classifier 2 performs much better than classifier 1 in terms of AUC. However, considering a situation like a marketing campaign, where people may only be interested in the very top-ranked examples, classifier 1 may be a better choice. And also, AUC is purely a ranked order metric, and it ignores the probability values [23].

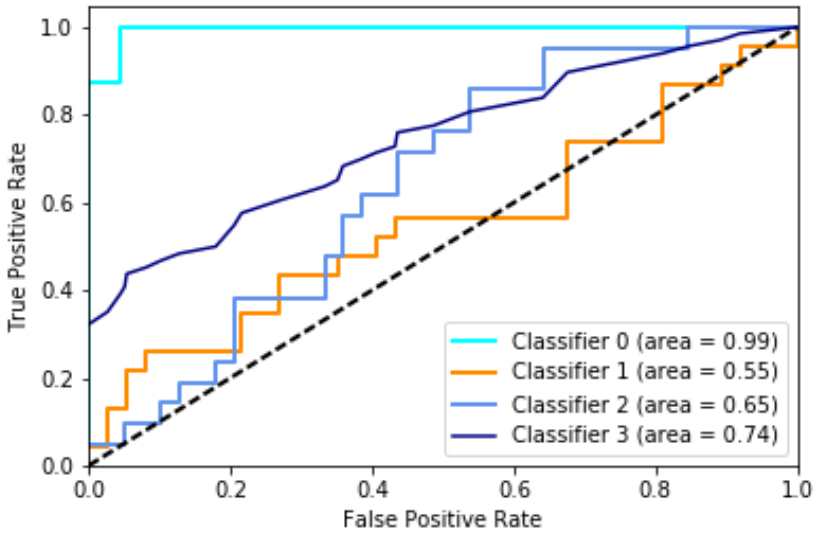

Figure 2: ROC comparisons *Source: scikit-learn simulation [22]*

There are many variants of ROC and AUC. Gini coefficient, also called Somer's D, is one of them, which is widely used in the evaluation of credit score modeling in banking industry. Gini can be converted from AUC approximately, using the formula: $Gini = (AUC - 0.5) * 2$. Wu et al. [24] proposed a scored AUC (sAUC) that incorporates both the ranks and probability estimate of the classifiers. sAUC is calculated as

$$sAUC = \frac{1}{mk} \sum_{i=1}^{m} \sum_{j=1}^{k} 1_{p_i > n_j} * (p_i - n_j) \tag{12}$$

The only difference between sAUC and AUC is the addition of a factor $p_i - n_j$, which quantifies the probability deviation when the ranks are correct. pAUC (probabilistic AUC) is another AUC variant (Ferri et al.,2005) [23] which takes both rank performance and the magnitude of the probabilities into account. pAUC is defined as

$$pAUC = \frac{\frac{\sum_{i=1}^{m} p_i}{m} - \frac{\sum_{j=1}^{k} n_i}{k} + 1}{2} = \frac{\frac{\sum_{i=1}^{m} p_i}{m} + \frac{\sum_{j=1}^{k} (1 - n_i)}{k}}{2} \tag{13}$$

pAUC can be interpreted as the mean of the average of probabilities that positive instances are assigned to positive class, and negative instances are assigned to negative class.

ROC Convex Hull

For a set of classifiers with ROC curves in the same graph, the ROC convex hull (ROCCH) is constructed in a way that the line connects all corresponding points on the curves to ensure there are no points above the final curve. The ROC convex hull describes the potential optimal operating characteristics for the classifiers and provides a possible solution to form a better classifier by combining different classifiers. Classifiers below the ROCCH is always sub-optimal. As shown in Figure 3, classifiers B, D are sub-optimal, while classifiers A, C are potentially optimal classifiers.

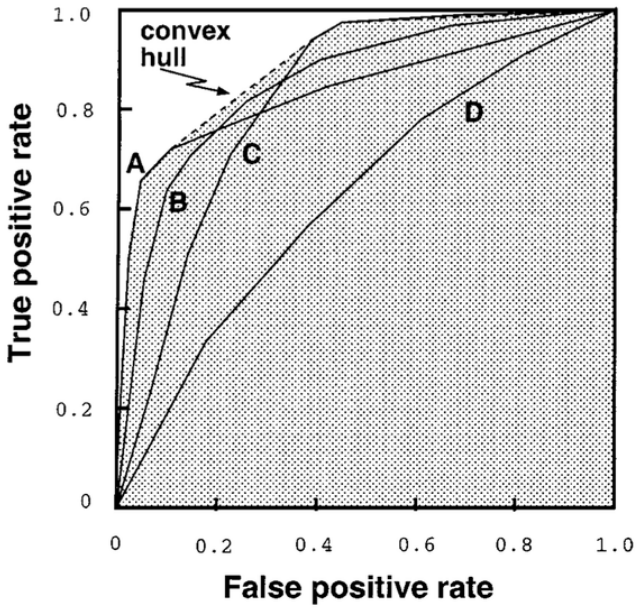

Figure 3: The ROC convex hull identifies potentially optimal classifier
Source: Provost and Fawcett, 2000

Since ROC curve does not rely on the cost matrix and class distribution, it is difficult to make decisions. By incorporating cost matrix and class distribution into ROC space, the decision goal can be achieved in the ROC space. Especially, the expected cost of applying the classifier represented by a point (FP, TP) in ROC space is: [25]

$$p(+) * (1 - TP) * C(-|+) + p(-) * FP * C(+|-) \tag{14}$$

Where $p(+)$ and $p(-)$ are the prior distribution of positives and negatives. $C(-|+)$ is the misclassification cost of false negative, while $C(+|-)$ is the misclassification cost of false positive. Therefore, two points, (FP_1, TP_1) and (FP_2, TP_2), have the same performance if

$$\frac{TP_2 - TP_1}{FP_2 - FP_1} = \frac{C(+|-)p(-)}{C(-|+)p(+)} \tag{15}$$

The above equation defines the slope of an iso-performance line. That is, all classifiers corresponding to points on the line have the same expected cost. Each set of class and cost distributions

defines a family of iso-performance lines. Lines "more northwest" (having a larger $TP - intercept$) are better because they correspond to classifiers with lower expected cost [25].

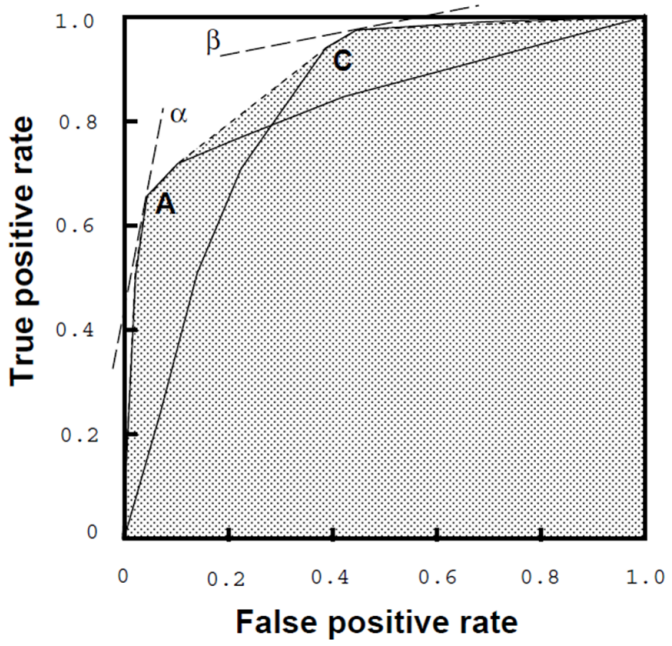

Figure 4: Line α and β show the optimal classifier under different sets of conditions
Source: Provost and Fawcett, 2000

Assume an imbalanced dataset with the negative observations 10 times of positive observations and equal misclassification cost. In this case, classifier A would be preferred. If the false negative costs 100 times of the false postive, classifier C would be a better choice. ROC convex hull provides a visual solution for the decision making under different set of cost and class distributions.

Concentrated ROC

The concentrated ROC (CROC) plot evaluates the early-retrieval performance of a classifier [26]. The findings are motivated by many practical problems ranging from marketing campaign to fraud detection, where only the top-ranked predictions are of any interest and ROC and AUCs, which are used to measure the overall ranking power, are not very useful.

Figure 5 shows that two ROC curves and the red rectangle area, the interested early retrieval area. The green curve outperforms the red curve in the early retrieval area. CROC adopts an approach whereby any portion of the ROC curve of interest is magnified smoothly using an appropriate continuous and monotone function f from [0,1] to [0,1], satisfying $f(0) = 0$ and $f(1) = 1$. Functions like exponential, power or logarithmic are popularly used to expand the early part of the $x - axis$ and contract the late part of the $x - axis$. The choice of magnification curve f and magnification parameters α are subject to different application scenarios. The early recognition performance of a

classifier can be measured by the AUC[CROC]. This area depends on the magnification parameter α and therefore both the area and α must be reported [26].

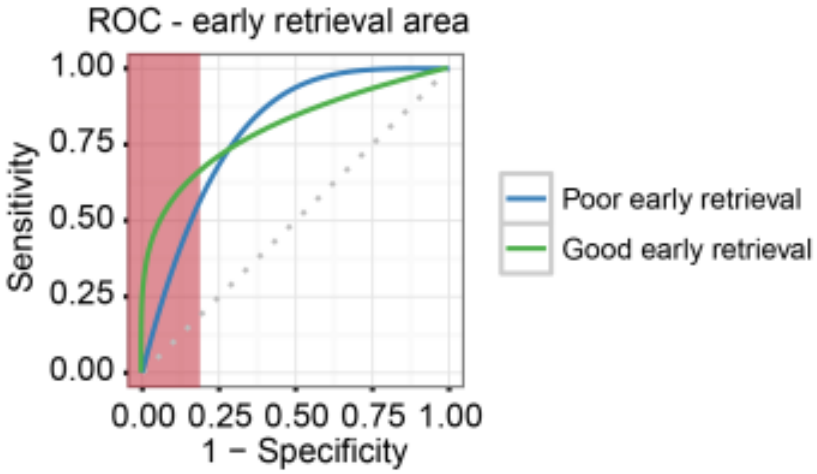

Figure 5: Early retrieval area in ROC curves. *Source: Swamidass et al., 2010*

Precision-Recall Curve

Like ROC curve, Precision-Recall curve (PRC) provides a model-wide evaluation of classifiers across different thresholds. The Precision-Recall curve shows a trade-off between precision and recall. The area under the curve (PRC) is called AUC (PRC). There are several aspects in which Precision-Recall is different from ROC curve. Firstly, the baseline for ROC curve is fixed, which is the diagonal connecting (0,0) and (1,1), while the baseline for PRC is correlated with the class distribution $\frac{P}{N+P}$. Secondly, interpolation methods for PRC and ROC curves are different-ROC analysis uses linear and PRC analysis uses non-linear interpolation. Interpolation between two points A and B in PRC space can be represented as a function $y = \frac{TP_A+x}{TP_A+x+FP_A+\frac{FP_B-FP_A}{TP_B-TP_A}*x}$ where x can be any value between TP_A and TP_B [27]. Thirdly, ROC curve is monotonic but PRC may not be monotonic, which leads to more crossover in PRC curves from different classifiers. As mentioned that the baseline for PRC is determined by the class distribution, for imbalance data PRC plots is believed to be more informative and powerful plot and can explicitly reveal differences in early-retrieval performance [28]. Figure 6 consists four panels for different visualization plots. Each panel contains two plots with balanced (left) and imbalanced (right) for (A) ROC, (B) CROC, (C) Cost Curve, and (D) PRC. Five curves represent five different performance levels: Random (Rand; red), Poor early retrieval (ER-; blue), Good early retrieval (ER+; green), Excellent (Excel; purple), and Perfect (Perf; orange) [28]. Figure 6 shows that PRC is changed, but other plots remain the same between balanced and imbalanced dataset, indicating that only precision-recall curve is sensitive to the class distribution changes.

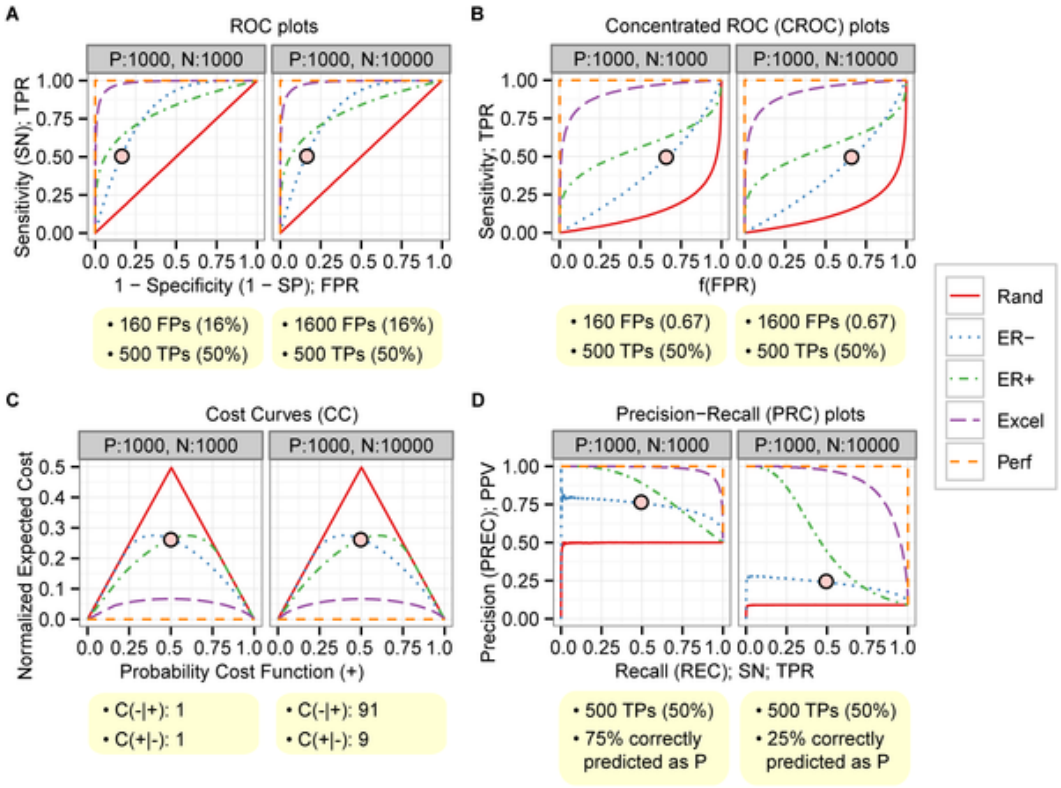

Figure 6: PRC is changed but other plots remain the same between balanced and imbalanced dataset
Source: Saito & Rehmsmeier, 2015

The study in paper [29] shows a deep connection between ROC space and PRC space, stating that a curve dominates in ROC space if and only if it dominates in PRC space. And the same paper demonstrates that PRC space analog to the convex hull in ROC space is achievable.

Alternative Graphical Performance Measures

There are many other graphical performance measures, such as cost curve, lift chart, PN-curve, K-S curve, calibration curves, etc. These visualization tools are discussed as below:

- **Cost curve**. The paper [30] proposed an alternative graphical measure to ROC analysis, which represents the cost explicitly. The ROC curve can be transformed into a cost curve. The x-axis in a cost curve is the probability-cost function for cost examples, $PCF(+) = p(+)C(-|+)/(p(+)C(-|+) + p(-)C(+|-))$. The y-axis is expected cost normalized with respect to the cost incurred when every example is incorrectly classified, which is defined as below:

$$NE[C] = \frac{(1 - TP)p(+)C(-|+) + FPp(-)C(+|-)}{p(+)C(-|+) + p(-)C(+|-)} \tag{16}$$

where $p(+)$ is the probability of positive instance (prior class probability) and $C(-|+)$ is the false negative misclassifying cost. And $p(-)$ is the probability of negative instance (prior class probability) and $C(+|-)$ is the false positive misclassifying cost.

- **Lift Chart**. Lift chart is a graphical depiction of how the model improves proportion of positive instances when a subset of data is selected. The x-axis is the % of the dataset and y-axis is the number of true positives. Not like ROC curve, the diagonal line for lift chart is not necessarily at 45^0. The diagonal line estimates what would be the number of true positives if a sample size of x is randomly selected. The far above the diagonal, the greater the model improves the lift. The lift chart is widely used in response models to make decisions on marketing campaigns. The same principle of ROC convex hull with iso-performance can be applied to lift chart to decide the cut-off point for maximizing profit [31].
- **PN-Curve**. PN-curve is similar to ROC. Instead of using normalized TPR and FPR, PN-curve directly has TP on the y-axis and FP on the x-axis. PN-curve can be transformed to a ROC after normalized using prior class distribution.
- **Kolmogorov–Smirnov alike curve**. Kolmogorov–Smirnov (K-S) graph is widely used in credit risk scoring to measure the discrimination. The K-S stat measures the maximum difference between the cumulative distribution function of positives and negatives. A similar graph is proposed in [32], where true positive rate and false positive rate are drawn separately against the probability threshold ([0,1]) in the same graph 7.

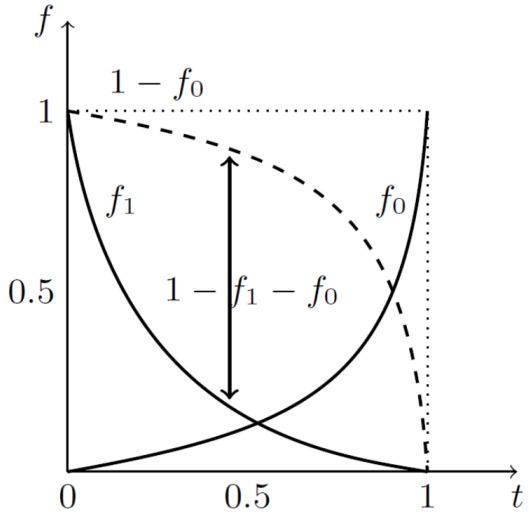

Figure 7: False positive/True positive Curve
Source: Desen & Bintong, 2017

In the graph, f_1 is true positive rate, monotonically decreasing when the threshold is set higher and higher. f_0 stands for true negative rate and $1 - f_0$ is the false positive rate. The distance between the two curves is $1 - f_1 - f_0$, which equals to $fpr - tpr$. Recall the distance between

ROC curve and the diagonal is $tpr - fpr$ and $AUC = (tpr - fpr + 1)/2$ [20]. Minimizing the distance in figure 7 is the same as maximizing the distance between ROC curve and the diagonal as well as the AUC. The k-s alike is another visual representation of ROC curve.

- **Calibration Curves**. Calibration curves are a method that will show whether the predicted probability is well calibrated that there are actually p proportional events happening when they are predicted probability is p. The x-axis is true probability, which counts the subset of examples with the same score. And y-axis is the predicted probability. Calibration curves focus on the classification bias, not the classification quality. A perfect calibration does not mean a perfect classifier [31, 33].

2.4 Metrics based on distribution stability

In order to ensure that a model continues to perform as intended, it is important to monitor the stability of the model output between the development data and more recent data. The model stability can be assessed either at the overall system level by examining the population distribution changes across model outcomes or at the individual model variable or characteristic level. The Population Stability Index (PSI) is a common metric used to measure overall model population stability while the Characteristic Stability Index (CSI) can be used to assess the stability at the individual variable level.

PSI and CSI are widely accepted metrics in banking industry. The formula is displayed as below:

$$PSI = \sum_{i}^{n}((Actual\% - Expected\%)ln(\frac{Actual\%}{Expected\%}))$$

Where the samples are divided into n bins based on the model output distribution (PSI) or variable distribution (CSI). The percentage of actual and expected events in each bin is calculated. The PSI and CSI values are sensitive to the number of bins. In general, 8-12 bins are mostly used, which should take various operating points into account.

3 Measure of measures

We can see no metric can absolutely dominate another. The metrics all have advantages and disadvantages. As Ferri et al. (2009) [1] point out, several works have shown that given a dataset, the learning method that obtains the best model according to a given measure, may not be the best method if another different measure is employed. A handful of examples have been discussed to demonstrate how challenging it is to compare the performance of the measures. For example, Huang et al. (2003) [34] reveal that Naive Bayes and decision tree perform similarly in predicting accuracy. However, Naive Bayes is significantly better than decision trees in AUC. Even surprisingly, Rosset (2004) [35] shows that if we use AUC for selecting models based on a validation dataset, we obtain betters results in accuracy (in a different test dataset) than using accuracy for selecting the models.

Huang et al. (2005) [23] propose a theoretical framework for comparing two different measures for learning algorithms. The authors define the criteria, degree of consistency and degree of discriminancy, as follows:

- **Degree of Consistency**. For two measures f and g on domain Ψ, let $R = \{(a,b)|a,b \in \Psi, f(a) > f(b), g(a) > g(b)\}$ and $S = \{(a,b)|a,b \in \Psi, f(a) > f(b), g(a) < g(b)\}$. The degree of consistency of f and g is $C(0 \leq \mathbf{C} \leq 1)$, where $\mathbf{C} = \frac{|R|}{|R|+|S|}$

- **Degree of Discriminancy**.For two measures f and g on domain Ψ, let $P = \{(a,b)|a,b \in \Psi, f(a) > f(b), g(a) = g(b)\}$ and $Q = \{(a,b)|a,b \in \Psi, g(a) > g(b), f(a) = f(b)\}$. The degree of discriminancy for f over g is $\mathbf{D} = \frac{P}{Q}$

Intuitively, for two classifiers (a & b), consistence requires that when one measure indicates one classifier is strictly better than the other ($f(a) > f(b)$), the other measure would not tell the opposite way. Discriminancy requires that when one measure can't tell the difference between classifiers, the other measure can differentiate. In order to conclude a measure f is better than measure g, we require $\mathbf{C} > 0.5$ and $\mathbf{D} > 1$.

4 A novel performance measure

For selecting a better classifier, the decision made based on one measure may be different if employing a different measure. The metrics mentioned in Section 2 have different focus. Confusion-matrix-based metrics (e.g., accuracy, precision, recall, F-measures, etc.) focus on errors, probability-based matrix (e.g., Brier score, cross-entropy, etc.) focus on probability, and ranking tools (e.g., AUC/VUS, lift chart, PRC, K-s, etc.) focus on ranking and visualization. Some measures combine those focuses. For example, sAUC combines the probability and ranking and AUC:acc proposed by Huang et al.(2007) [36] uses AUC as a dominant measure and acc as a tie-breaker if AUC ties.

Here we propose a new measure, which combines accuracy, MCC, and AUC. The new measure will be based on the voting of three metrics. Given two classifiers (a,b) on the same dataset, the new measure will be defined as:

$$NewMeasure = 1(AUC(a), AUC(b)) + 1(MCC(a), MCC(b)) + 1(acc(a), acc(b)) \quad (17)$$

We can empirically prove that the voting methods for three metrics would be better than accuracy in terms of discriminancy and consistency as defined in [23].

5 Experimental results and discussion

Following a similar method in [36], we compared the new measures with accuracy on artificial test datasets. The datasets are balanced datasets with equal numbers of positive and negative instances. We tested on datasets with 8, 10, 12, 14, 16 instances, respectively. We exhausted all the ranking orders of the instances and calculated accuracy/MCC/AUC for each ranking order. The two criteria (consistency and discriminancy) as defined in Section 3 are computed and the results are summarized in Table 2 and 3.

For consistency, we count the number of pairs that satisfy "$new(a) > new(b) \& Acc(a) > Acc(b)$ and "$new(a) > new(b) \& Acc(a) < Acc(b)$. For discriminancy, we count the number of pairs that satisfy "$new(a) > new(b) \& Acc(a) = Acc(b)$ and "$new(a) = new(b) \& Acc(a) > Acc(b)$.

Two measures are consistent when comparing two models a and b, if one measure claims that model a is better than model b, the other measure indicates that model a is better than model b. As Table 2 shows, our proposed new measure demonstrates excellent consistency with accuracy with C value significantly greater than 0.5. Take the 8 instances case for example. we exhausted all ranking orders of the instances and calculated accuracy/MCC/AUC for each ranking order. The results show that there are 184 times where our new metric stipulates that a is better than b and accuracy indicates the same, while there are only 35 times that the decision from our new metric is

not in line with the accuracy. The results demonstrate that in most cases, the decision based on our new metric is well aligned with the decision from accuracy. Therefore, we claim our new metric is consistent with accuracy.

Stastical Consistency between Accuracy and the new measure

#	New(a)>New(b) & Acc(a)>Acc(b)	New(a)>New(b) & Acc(a)<Acc(b)	C
8	184	35	0.84
10	1122	137	0.89
12	3355	688	0.83
14	9546	1200	0.89
16	11213	2035	0.85

Table 2: Experimental results for verifying statistical consistency between new metric and accuracy for the balanced dataset

The degree of discriminancy is defined as the ratio of cases where our new metric can tell the difference but accuracy cannot, over the cases where accuracy can tell the difference but our new metric cannot. Table 3 indicates that our new measure is much better in terms of discriminancy given D is significantly higher than 1. We use the 8 instances case for example. The results show that there are 78 cases where our new metric can tell the difference between a and b, but accuracy cannot, while there are only 3 cases that accuracy can tell the differences, but our new metric cannot. Therefore, we claim our new metric is much more discriminant than accuracy.

Stastical Discriminancy between Accuracy and new measure

#	New(a)>New(b) & Acc(a)=Acc(b)	Acc(a)>Acc(b) & New(a)=New(b)	D
8	78	3	26.0
10	386	8	48.3
12	1063	46	23.1
14	2238	57	39.3
16	2268	60	37.8

Table 3: Experimental results for verifying statistical discriminancy between new metric and accuracy for the balanced dataset

6 Conclusion

Performance measures are essential in model development, selection and evaluation. There are no gold standard rules for the best performance measure. Most of classifiers are trained to optimize

accuracy, but the accuracy is more biased towards majority class. Balanced performance measures such as F_1 and MCC are preferred in imbalanced data. However, all confusion matrix based metrics are subject to threshold selection, which is very difficult to determine especially when misclassification cost and prior probability are unknown. In contrast, the probability metrics do not rely on how the probability will fall relative to a threshold. However, the probability metrics are not intuitive and it is more commonly used in objective function instead of measuring the outcome of machine learning models. Visualization tools such as ROC, precision-recall complement the confusion-matrix, which are threshold free and independent of prior class distribution. We can see no metric in one category can completely dominate another. They all have its advantages and disadvantages.

In this paper, we provide a detailed review of commonly used performance measures and highlighted their advantages and disadvantages. To address the deficiency of one single measure, we propose a voting method by combining multiple measures. Our proposed new measure that incorporates several measures proves to be more effective in terms of consistency and discriminancy.

7 Compliance with Ethical Standards

Funding: Not Applicable.
Conflict of Interest: Author declares that he/she has no conflict of interest.
Ethical approval: This article does not contain any studies with human participants or animals performed by any of the authors.

References

[1] César Ferri, José Hernández-Orallo, and R Modroiu. An experimental comparison of performance measures for classification. *Pattern Recognition Letters*, 30(1):27–38, 2009.

[2] Chang Cao, Davide Chicco, and Michael M Hoffman. The mcc-f1 curve: a performance evaluation technique for binary classification. *arXiv preprint arXiv:2006.11278*, 2020.

[3] Davide Chicco and Giuseppe Jurman. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, 21(1):6, 2020.

[4] Yan Wang and Xuelei Sherry Ni. Predicting class-imbalanced business risk using resampling, regularization, and model emsembling algorithms. *International Journal of Managing Information Technology (IJMIT) Vol*, 11, 2019.

[5] A Cecile JW Janssens and Forike K Martens. Reflection on modern methods: revisiting the area under the roc curve. *International Journal of Epidemiology*, 2020.

[6] Sarang Narkhede. Understanding auc-roc curve. *Towards Data Science*, 26, 2018.

[7] Matjaž Majnik and Zoran Bosnić. Roc analysis of classifiers in machine learning: A survey. *Intelligent data analysis*, 17(3):531–558, 2013.

[8] Jonathan E Fieldsend and Richard M Everson. Formulation and comparison of multi-class roc surfaces. 2005.

[9] Thomas Landgrebe and R Duin. A simplified extension of the area under the roc to the multiclass domain. In *Seventeenth annual symposium of the pattern recognition association of South Africa*, pages 241–245, 2006.

[10] Kun Deng, Chris Bourke, Stephen Scott, and NV Vinodchandran. New algorithms for optimizing multi-class classifiers via roc surfaces. In *Proceedings of the ICML workshop on ROC analysis in machine learning*, pages 17–24, 2006.

[11] Jia Hua and Lili Tian. A comprehensive and comparative review of optimal cut-points selection methods for diseases with multiple ordinal stages. *Journal of Biopharmaceutical Statistics*, 30(1):46–68, 2020.

[12] Bilal Yurdakul. Statistical properties of population stability index. 2018.

[13] Luzia Gonçalves, Ana Subtil, M Rosário Oliveira, and P d Bermudez. Roc curve estimation: An overview. *REVSTAT–Statistical Journal*, 12(1):1–20, 2014.

[14] Amalia Luque, Alejandro Carrasco, Alejandro Martín, and Ana de las Heras. The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognition*, 91:216–231, 2019.

[15] M Hossin and MN Sulaiman. A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining & Knowledge Management Process*, 5(2):1, 2015.

[16] Guzman Santafe, Iñaki Inza, and Jose A Lozano. Dealing with the evaluation of supervised classification algorithms. *Artificial Intelligence Review*, 44(4):467–508, 2015.

[17] Henrik Strøm, Steven Albury, and Lene Tolstrup Sørensen. Machine learning performance metrics and diagnostic context in radiology. In *2018 11th CMI International Conference: Prospects and Challenges Towards Developing a Digital Economy within the EU*, pages 56–61. IEEE, 2018.

[18] Ashok Suragala, P Venkateswarlu, and M China Raju. A comparative study of performance metrics of data mining algorithms on medical data. In *ICCCE 2020*, pages 1549–1556. Springer, 2021.

[19] Shadi Diab and Badie Sartawi. Classification of questions and learning outcome statements (los) into blooms taxonomy (bt) by similarity measurements towards extracting of learning outcome from learning material. *arXiv preprint arXiv:1706.03191*, 2017.

[20] David Martin Powers. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. 2011.

[21] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

[22] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[23] Jin Huang and Charles X Ling. Using auc and accuracy in evaluating learning algorithms. *IEEE Transactions on knowledge and Data Engineering*, 17(3):299–310, 2005.

[24] Cèsar Ferri, Peter Flach, José Hernández-Orallo, and Athmane Senad. Modifying roc curves to incorporate predicted probabilities. In *Proceedings of the second workshop on ROC analysis in machine learning*, pages 33–40, 2005.

[25] Foster Provost and Tom Fawcett. Robust classification for imprecise environments. *Machine learning*, 42(3):203–231, 2001.

[26] S Joshua Swamidass, Chloé-Agathe Azencott, Kenny Daily, and Pierre Baldi. A croc stronger than roc: measuring, visualizing and optimizing early retrieval. *Bioinformatics*, 26(10):1348–1356, 2010.

[27] Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240. ACM, 2006.

[28] Takaya Saito and Marc Rehmsmeier. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, 10(3):e0118432, 2015.

[29] Tom Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

[30] Chris Drummond and Robert C Holte. Explicitly representing expected cost: An alternative to roc representation. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 198–207. ACM, 2000.

[31] Miha Vuk and Tomaz Curk. Roc curve, lift chart and calibration plot. *Metodoloski zvezki*, 3(1):89, 2006.

[32] Desen Wang and Bintong Chen. Cost-sensitive learning algorithm. In *Working paper*. University of Delaware, 2017.

[33] Ira Cohen and Moises Goldszmidt. Properties and benefits of calibrated classifiers. In *PKDD*, volume 3202, pages 125–136. Springer, 2004.

[34] Charles X Ling, Jin Huang, and Harry Zhang. Auc: a statistically consistent and more discriminating measure than accuracy. In *IJCAI*, volume 3, pages 519–524, 2003.

[35] Saharon Rosset. Model selection via the auc. In *Proceedings of the twenty-first international conference on Machine learning*, page 89. ACM, 2004.

[36] Jin Huang and Charles X Ling. Constructing new and better evaluation measures for machine learning. In *IJCAI*, pages 859–864, 2007.