WaAUROCC: measuring how steep the ROC curve is

Jože M. Rožanec
Jožef Stefan Institute
Jamova cesta 39, 1000 Ljubljana, Slovenia
joze.rozanec@ijs.si

Dunja Mladenić
Jožef Stefan Institute
Jamova cesta 39, 1000 Ljubljana, Slovenia
dunja.mladenic@ijs.si

Abstract. *The Area Under the Receiver Operating Characteristic Curve (ROC AUC) is a widely used performance measure summarizing classifiers' discriminative power. Among its strengths is that it does not depend on threshold settings. Nevertheless, machine learning models scoring the same ROC AUC value can display a different behavior along the curve - a characteristic that may be relevant to model selection. This paper introduces WaAUROCC, an alternative measure that builds upon the ROC AUC and mitigates some of its well-known weaknesses. In particular, it allows for assessing how steep the curve is without visually inspecting it. It thus enables the identification of models where the highest recall is achieved with the lowest False Positive Rate. Furthermore, the metric allows for contrasting the models' performance against performance acceptance criteria, providing insights on whether such criteria are met and how much is outperformed. The approach followed when creating the WaAUROCC metric can be followed with other AUC metrics, such as the Precision-Recall AUC. We validate the usefulness of the proposed metric on three real-world datasets. In addition, we illustrate its usefulness with four synthetic scenarios.*

Keywords. Metrics, Classification, Operational Performance, MlOps

1 Introduction

Background The ROC AUC (Bradley, 1997) is commonly used to assess the discriminative power of machine learning classifiers (Halligan et al., 2015) and select the best ones Muschelli III (2020); Carrington et al. (2021) by comparing how the performance of a model varies across the True Positive Rate (TPR) and the False Positive Rate (FPR). The goal of the metric is to achieve a perfect TPR with $FPR = 0$. Among the metric strengths is that it is scale-invariant (measures how predictions are ranked, regardless of their absolute values) and provides an aggregate measure of performance across all possible classification thresholds Hernández-Orallo et al. (2012). ROC AUC is unaffected by dataset skew but may mask poor performance (Jeni et al., 2013; Cook and Ramadas, 2020). This could be because ROC AUC considers the four quadrants of a confusion matrix (in contrast to PR AUC, which does not consider true negatives) Sofaer et al. (2019). Multiple authors have documented detailed analysis behind such behavior (Fawcett and Flach, 2005; Webb and Ting, 2005; Cook and Ramadas, 2020).

Motivation In use cases such as defect detection or medical diagnosis and treatment, machine learning models are used along with a manual evaluation to produce an outcome Luzio et al. (2024); Klawonn et al. (2011); Moons et al. (1997). For example, when dealing with defect detection, we may prefer models that allow for business rules, such as automatically rejecting (or accepting) products for which we are highly confident that they are defective (or not). On top of that, we expect to have a reasonable subset of cases that can be manually inspected where defect occurrence is highly probable so that upon human inspection, a decision is made whether to accept or reject the transaction. We expect that models achieving a high Recall at a low FPR enable greater automation, given (i) high Recall means a high proportion of True Positives has been identified while keeping the number of False Negatives low and (ii) the proportion of False Positives identified is low w.r.t. the True Negatives. While the ROC AUC is a single scalar value and doesn't directly tell about specific thresholds, the ROC curve (from which the AUC is derived) can be used to compare models at particular points. Furthermore, the ROC AUC provides no information on whether a particular model achieves higher Recall at a lower FPR when compared to others. Therefore, a different metric must be considered to account for such differences and enable model selection that takes such model behavior into account.

Limitations of the State of the Art To account for the different weights between sensitivity and specificity, Hand (2009) proposed an alternative metric that considers the relative misclassification cost distribution based on the assumption that cost should dominate specificity in the choice of measure. Such assumption contrasts with machine learning models bound to a manual revision setting, where decisions are made based on how many cases can undergo manual revision. Halligan et al. (2015) followed a different rationale and proposed a metric evaluating change in sen-

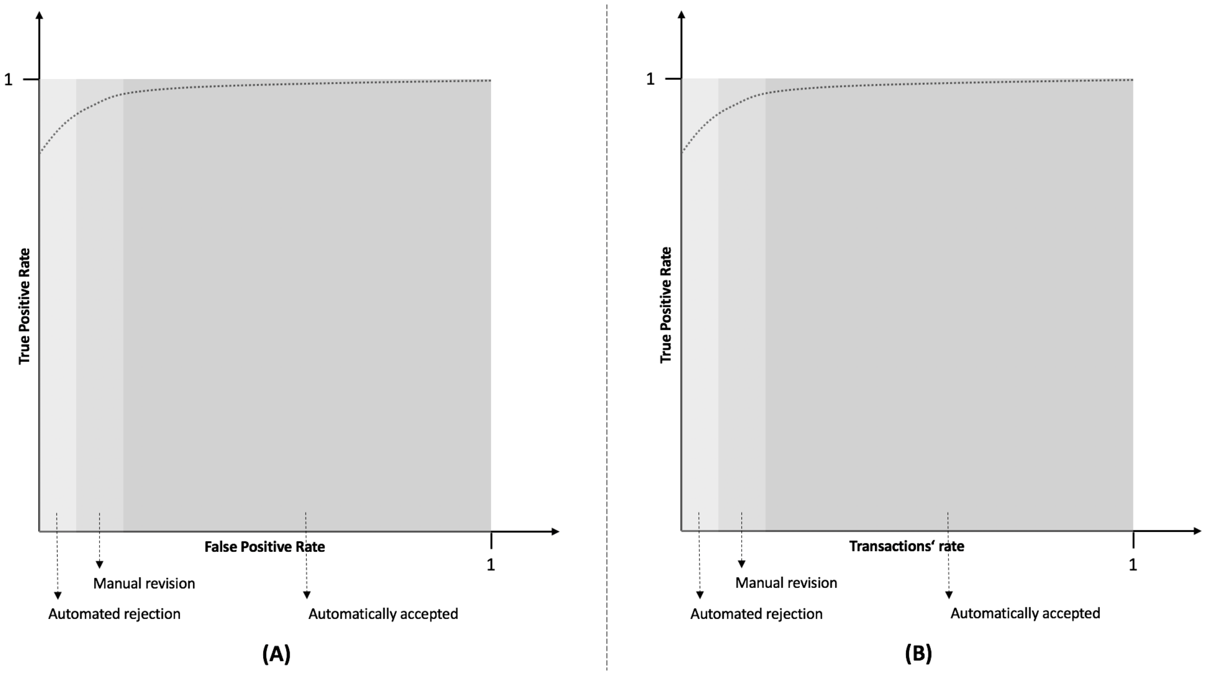

Figure 1: The images represent two plots that guide the evaluation of machine learning models for defect inspection: (A) ROC AUC, contrasting True Positive Rate against False Positive Rate, and (B) contrasting True Positive Rate against the ratio of items processed.

sitivity and specificity at clinically relevant thresholds. This approach is similar to the $Precision@Recall_K$ used by Luzio et al. (2024). Nevertheless, both fail to provide a global perspective of the models' performance (summarize in a score the behavior observed across all thresholds and observed in plots) - a key advantage of ROC AUC.

Key insights This paper introduces WaAUROCC [1] (/ˈwauɹɑk/) - *Wa*sserstein *A*rea *U*nder the *R*eceiver *O*perating *C*haracteristic *C*urve). It measures the Wasserstein distance between a given ROC AUC and the ideal case. By doing so, even when the ROC AUCs have the same value, WaAUROCC allows us to identify models with better operational performance (achieving high TPR and low FPR while considering the fewest samples possible).

2 WaAUROCC - how far is our model from ideal performance?

2.1 Comparing distributions

Two cases are known to us beforehand: ideal classifier performance (ROC AUC=1) and the random classifier case (ROC AUC=0.5). In addition, we can plot the ROC of the classifier we aim to evaluate. The ROC curve can be contrasted with the best and worst-case scenarios to obtain a score between zero and one, reflecting how close the performance of the classifier under consideration is to the best case. To do so, we follow a similar approach as presented in Rožanec et al. (2023): we can build bar plots that mirror the shape of the ROC curves for the three abovementioned cases and measure the effort required to turn one distribution into another. The bar plots are created considering a fixed number of equal-sized bins representing FPR values. When considering the effort required to turn one distribution into another, we look into the Optimal Transport mathematical problem, which aims to find the most efficient way to move mass between distributions. In this context, the Wasserstein distance measures the similarity between two distributions. For the WaAUROCC, we measure the distance between the actual ROC AUC curve and the optimal one.

2.2 Metric definition

We define the WaAUROCC metric for a multiclass scenario in Eq. 1. Classifiers close to an ideal performance will result in WaAUROCC values close to one, while models with little or no discriminative power will result in values close to zero. The cases that maximize TPR while minimizing FPR are rewarded.

2.3 WaAUROCC vs. Partial ROC plots

ROC AUC interpretation The ROC AUC has at least three interpretations: (i) the average Recall over all possible FPR values between zero and one, (ii) the

[1] A scikit-learn compatible implementation will be made available upon the paper's acceptance.

$$WaAUROCC = \sum_{i=1}^{n} \frac{1 - \frac{W_i(b_i, b_{best})}{W_i(b_{worst}, b_{best})}}{n} \qquad (1)$$

Equation 1: $W_i(b_i, b_{ref})$ is the Wasserstein distance between the histogram h_i (representing normalized ROC AUC values for class i) and the reference bar plots (b_{best} and b_{worst} for best and worst cases, respectively), and n is the number of classes.

probability of correctly ranking a particular class based on the observed cases in the test set, and (iii) given a set of classes, how different the distributions of the predicted values are for them (relates to the Mann-Whitney statistic) Bamber (1975).

ROC AUC drawbacks Nevertheless, given Recall and FPR evaluate complementary aspects of the model performance, a classifier is frequently expected to have a high Recall and a low FPR. Furthermore, there are many cases (e.g., medical diagnosis) where only a limited range of FPR must be considered (either due to the values observed or some restrictions), making it unreasonable to use a metric that reflects the performance of higher FPR values Fahey et al. (1995); Scheidler et al. (1997); Carrington et al. (2022). Therefore, attention must be paid to the upper-left area of a ROC AUC diagram Yang et al. (2021).

Overcoming ROC AUC limitations with pROCAUC To address these ROC AUC flaws, the partial ROC AUC has been proposed, restricting the ROC AUC metric to a particular FPR range Walter (2005). Nevertheless, given the pAUC requires defining FPR ranges, comparisons between tests may lack homogeneity. Furthermore, the pROCAUC lacks the symmetry property of the ROCAUC, and the effective use of less information has been the source of concern about whether it results in a loss of statistical precision compared to the ROC AUC Obuchowski and McClish (1997); Ma et al. (2013). To mitigate such issues, several variations have been proposed. E.g., Ma et al. (2013) suggested a standardized pAUCROC by dividing the partial area of the curve by the partial area of the random case. While such a solution can estimate the quality of the predictor between 0.5 to 1 regardless of the FPR interval, the actual value of the AUC could increase or decrease when changing the interval size. A different approach was suggested by Carrington et al. (2020), who suggests not only performing the integration over the x-axis but a horizontal integration over the y-axis, thereby capturing the area at the top-left corner. Such an area provides insight into the improvement opportunity for a particular model compared to the best case. Based on the horizontal and vertical AUC, they define the concordant partial AUC as half the sum of the vertical partial area under the ROC curve pAUC and the horizontal partial area under the ROC curve.

Toward a better metric: WaAUROCC As described above, pROCAUC and the derivative metrics suffer from several shortcomings. Nevertheless, most of them could be overcome if a metric could (i) measure not only the ROC AUC but also how steep the curve is through FPR on a continuous range without defining particular segments and (ii) if a custom baseline curve could be specified so that a comparison could be drawn between the baseline and the performance of the model under evaluation. WaAUROCC achieves both by computing the Wasserstein distance between (i) the distributions that emulate the ROC AUC curve and the ideal case and (ii) the distributions that emulate the ROC AUC curve and the custom baseline curve. Through (i), the metric takes into account that lower FPRs are preferred and penalizes curves that achieve higher Recall at higher FPRs. Through (ii), domain knowledge and business requirements can be introduced, allowing us to measure how far the models' performance is satisfying a particular Recall and FPR levels.

2.4 WaAUROCC: measuring compliance with acceptance criteria

Intuition In Eq. 1, we show how a single metric can summarize the classifiers' ROC behavior, considering the AUC and how steep such a curve is over all of the False Positive Rate values. Similarly, acceptance criteria could be established considering AUC and the minimal expected curve steepness across False Positive Rate sections. Models whose WaAUROCC would be equal or greater than the WaAUROCC of such curve would be guaranteed to satisfy or supersede the acceptance criteria behavior.

How is this useful? In many real-world cases, knowing the models' overall performance is insufficient. More fine-grained insights are required to decide how such a model should be deployed and used in production environment settings. E.g., (i) what threshold interval should be considered for automated decision-making?, (ii) what threshold interval should be considered to seek complementary evaluation (e.g., with a model whose inference has a higher cost)?, (iii) what threshold interval should be considered for decision-making that requires human intervention? WaAUROCC enables drafting ROC AUC curves that reflect the expected performance across the whole FPR continuum. The WaAUROCC score of such curves establishes a baseline against which the models can be evaluated.

Synthetic case	Predicted values	Ground truth
1	[0.1, 0.1, 0.1, 0.5, 0.7, 0.2, 0.2, 0.5, 0.55, 0.75]	[0, 0, 0, 0, 0, 1, 1, 1, 1, 1]
2	[0.25, 0.25, 0.45, 0.55, 0.50, 0.5, 0.55, 0.35, 0.55, 0.75]	[0, 0, 0, 0, 0, 1, 1, 1, 1, 1]
3	[0.1, 0.1, 0.1, 0.3, 0.1, 0.3, 0.1, 0.1, 0.7, 0.5]	[0, 0, 0, 0, 0, 1, 1, 1, 1, 1]
4	[0.1, 0.4, 0.2, 0.25, 0.55, 0.35, 0.25, 0.75, 0.8, 0.35]	[0, 0, 0, 0, 0, 1, 1, 1, 1, 1]

Table 1: Synthetically crafted samples of predictions and ground truths. The examples result in the same ROC AUC but display different curve shapes.

3 Operational meaning of the proposed metric

Let's consider the ROC AUC curve, which can be discretized into bins (like a bar chart), and compare two different ROC AUC bar charts. The Wasserstein distance between both distributions measures the effort required to transform one distribution into another (e.g., a given ROC AUC into a perfect ROC AUC (the distribution for the latest one is always the same)). Such effort conveys information on how early the model can achieve a perfect True Positive Rate: the lower the distance, the smaller the difference (between our ROC AUC and a perfect model), and the closer we are to the perfect model. We considered an arbitrary cost (e.g., a cost of "one" per bin) to compute the effort required to move cases between bins. It must be noted that the ROC AUC is a monotonically non-decreasing function: the True Positive Rate values can remain the same (e.g., a perfect ROC curve) or increase when the False Positive Rate increases.

Computing the proposed metric is not equivalent to computing the ROC AUC: the synthetic examples presented in Section 6 show that curves with the same ROC AUC result in different values of the proposed metric, which can be associated with a particular behavior (lower metric values correspond to models achieving higher True Positive Rate at a lower False Positive Rate) as confirmed by visual analysis.

To compute the abovementioned metric, we propose normalizing the ROC AUC histograms to ensure the area under the curve is always the same. By doing this, we lose information regarding the original area under the curve. Nevertheless, we keep information regarding the curve shape - a key aspect of our metric. We can estimate the effort required to reshape the distribution to the ideal case by keeping the shape. The lower the required effort, the better the model: the model will achieve a higher True Positive Rate at a low False Positive Rate.

3.1 Putting it all together: WaAUROCC explained through the defect detection use case

When evaluating machine learning models for defect detection, we are interested in identifying the models that help us capture the highest percentage of True Positives while keeping False Positives low to avoid a bad user experience. This can be done by comparing the ROC AUC across models: models with a higher ROC AUC score display such characteristics. Nevertheless, there may be cases where models with an identical ROC AUC score display a different behavior along the curve. For such cases, we propose the WaAUROCC metric, where at an equal ROC AUC score, a higher WaAUROCC score will be assigned to models with a higher True Positive Rate at a lower False Positive Rate. Therefore, the WaAUROCC metric score could be used to select a subset of models. Once such selection is performed, we consider there are at least two plots of interest (see Fig. 1):

(A) ROC AUC helps us understand how the model performs along different thresholds when comparing the tradeoff between the True Positive Rate (Recall) and the False Positive Rate. This plot is of primary importance in defining the threshold for automated rejections: for a set of items, we may reject the ones we are most confident are defective, tolerating a small number of false positives. The amount of false positives to be tolerated is a business criterion.

(B) Recall vs. inspection rate, which helps us understand the number of items that fall within a particular Recall range. This plot is particularly relevant to defining the thresholds for cases undergoing manual revision. Usually, the manual revision capacity is fixed; therefore, we seek models that allow the widest possible Recall range for the same capacity.

Considering the same rationale as the one followed when creating the partial ROC plots and the fact that some business criteria may fix a False Positive Rate threshold, we may evaluate the models considering a partial ROC AUC and the corresponding WaAUROCC score. Furthermore, an analogous metric could be computed to identify the most promising models for the threshold range for manual revision.

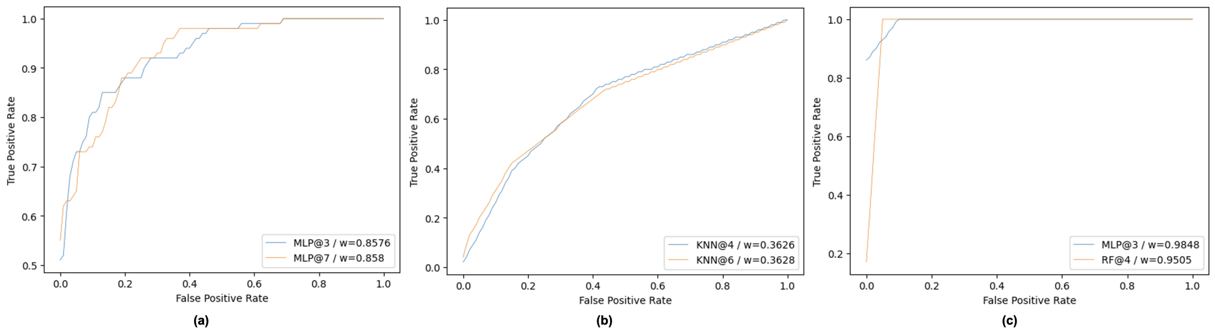

Figure 2: The plots correspond to the three cases reported in Table 2 where the measured ROC AUC values for the models are the same, but the resulting curves are different. In particular, the cases correspond to the following datasets: (a) DEFECTS, (b) SUPPORT2, and (c) CANCER.

USE CASE	**MODEL**	**FOLDS**									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
DEFECTS	**KNN**	0,8794	0,9073	0,8592	0,9054	0,8249	0,8943	0,8554	0,8833	0,9096	0,8496
	MLP	0,9076	0,9304	**0,9306**	0,9568	0,8984	0,9118	**0,9306**	0,9364	0,9480	0,9039
	RF	0,8653	0,8836	0,8980	0,9395	0,8364	0,8949	0,8856	0,8406	0,9186	0,8626
SUPPORT2	**KNN**	0,6732	0,6753	0,6445	**0,6830**	0,6291	**0,6830**	0,6697	0,6157	0,6523	0,6856
	MLP	0,7605	0,7718	0,7912	0,7704	0,7506	0,7581	0,7971	0,7755	0,7824	0,7908
	RF	0,7416	0,7538	0,7643	0,7631	0,7456	0,7418	0,7799	0,7696	0,7721	0,7598
CANCER	**KNN**	0,9414	0,9750	0,9930	0,9792	0,9554	0,9784	0,9992	0,9836	0,9864	0,9465
	MLP	0,9801	0,9850	**0,9937**	0,9973	0,9622	0,9972	1,0000	0,9943	0,9946	0,9924
	RF	0,9587	0,9866	0,9854	**0,9937**	0,9527	0,9914	0,9973	0,9972	0,9924	0,9960

Table 2: ROC AUC scores obtained for the use cases presented by machine learning model and fold. Bolded values correspond to cases where the models resulted in the same ROC AUC score for that particular dataset.

USE CASE	**MODEL**	**FOLDS**									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
DEFECTS	**KNN**	0,7535	0,8111	0,7135	0,8071	0,6451	0,7826	0,7059	0,7602	0,8133	0,6945
	MLP	0,8103	0,8564	**0,8576**	0,9101	0,7925	0,8188	**0,8580**	0,8683	0,8931	0,8038
	RF	0,7265	0,7640	0,7923	0,8756	0,6693	0,7855	0,7671	0,6770	0,8345	0,7212
SUPPORT2	**KNN**	0,3432	0,3473	0,2865	**0,3626**	0,2558	**0,3628**	0,3362	0,2295	0,3018	0,3677
	MLP	0,5162	0,5390	0,5770	0,5358	0,4966	0,5113	0,5883	0,5459	0,5594	0,5766
	RF	0,4788	0,5032	0,5236	0,5214	0,4867	0,4790	0,5547	0,5341	0,5392	0,5150
CANCER	**KNN**	0,8749	0,9420	0,9848	0,9505	0,9032	0,9489	0,9982	0,9671	0,9711	0,8850
	MLP	0,9572	0,9665	**0,9867**	0,9941	0,9202	0,9941	1,0000	0,9881	0,9887	0,9832
	RF	0,9083	0,9713	0,9699	**0,9861**	0,9036	0,9828	0,9941	0,9941	0,9832	0,9911

Table 3: WaAUROCC scores obtained for the use cases, presented by machine learning model and fold. Bolded values correspond to cases where the models resulted in the same ROC AUC score for that particular dataset.

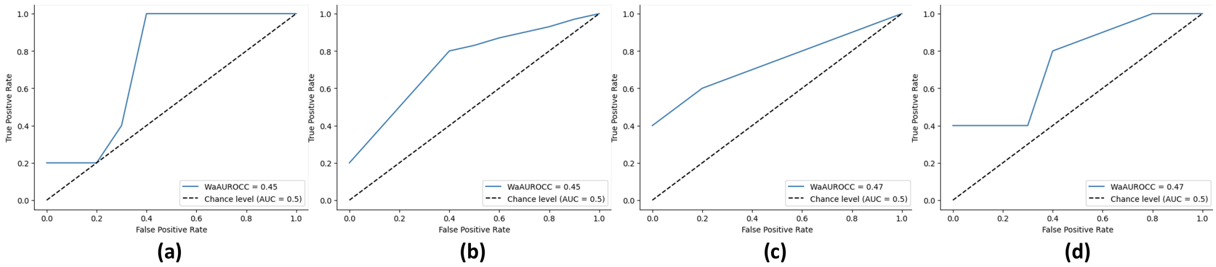

Figure 3: The images correspond to four synthetic scenarios with the same ROC AUC (0.74) but different WaAUROCC. The plots are sorted in ascending order based on WaAUROCC values. The dotted line indicates the chance level.

4 Methodology and experimental setup

4.1 Real-world datasets

We considered ten-fold stratified cross-validation, devoting nine folds to train K-Nearest Neighbors (KNN), Multi-Layer Perceptron (MLP), and Random Forest (RF) classifiers, and one fold to test them. We experimented on three datasets: (i) a dataset introduced by Connors et al. (1995), predicting death in hospital (SUPPORT2), (ii) the breast cancer dataset Street et al. (1993) (CANCER), and (iii) a real-world dataset of printed logos on shaver images provided by an industrial partner within project cooperation to research automated defect inspection (DEFECTS). Images were converted into feature vectors by leveraging ResNet-18 embeddings. Non-numerical features were removed, and features selection considering top K ranking features based on mutual information and a maximum of $\sqrt{N}$, with N equal to the number of instances in the training set. We executed the abovementioned setup multiple times considering different seed numbers, until getting cases where the models' performance would result in the same ROC AUC values. We then focused on those cases and analyzed whether did the WaAUROCC provide a better intuition behind them.

4.2 Synthetic samples

In addition to the real-world datasets, we crafted four synthetic samples of predictions and expected ground truth to showcase different scenarios in which the ROC AUC would result in the same value despite a different curve shape, but the WaAUROCC would detect such discrepancies. We showcase the examples in Table 1.

5 Results

5.1 Real-world datasets

We detail the results obtained in Table 2 and Table 3. The tables present the results obtained by measuring ROC AUC and WaAUROCC across datasets and machine learning models over ten folds of a cross-validation setting. When performing the experiments, we could find one case per dataset where the machine learning models resulted in the same ROC AUC score, but the curves were noticeably different. For ease of analysis, those cases are bolded in the abovementioned tables and we plot their ROC curves in Fig. 2.

Same ROC AUC, different operational performance. For the DEFECTS dataset, we found that the MLP model reported the same ROC AUC value at third and seventh fold, but the WaAUROCC scores showed the model tested at the seventh fold would achieve a higher Recall at a lower FPR. This is validated with the plot at Fig. 2 (a). A similar case is observed for the SUPPORT2 dataset but for the KNN model: while the ROC AUC score for the fourth and sixth folds is the same, the WaAUROCC score indicates the model tested at the sixth fold achieves a higher TPR for the same FPR at least until a FPR of 0.2 (see Fig. 2 (b)). Finally, for the CANCER dataset, the same ROC AUC is reported by two different kinds of models at different test folds: MLP at the third fold shows the same performance as the RF model at the fourth fold. Nevertheless, the WaAUROCC score shows a preference for the MLP model tested at the third fold - which is consistent with the ROC curves in 2 (c).

5.2 Synthetic samples

We detail the results obtained for the synthetic samples in Fig. 3. By analyzing the plots, we observe that (a) has a low initial TPR but gradually improves and reaches 0.8 TPR at 0.4 FPR. In contrast, (b) and (c) have a higher initial TPR. While both reach a 0.8 TPR at 0.4 FPR, (b) has a flat 0.4 TPR between 0 and 0.4 FPR, while (c) shows a mild increasing slope in that same range. Among the three models, (c) is preferable, given it achieves the highest TPR at the lowest FPR.

Same ROC AUC, different operational performance. Four synthetic scenarios were considered, ensuring the same ROC AUC values but different operational performances (see Fig. 3). Considering overall performance, (c) and (d) are preferable, given that they achieve the highest TPR at the lowest FPR when considering the whole FPR range.

Partial ROC plots. Nevertheless, (c) and (d) display a different operational performance if considering a specific region of the ROC plot. If we constrain the metric computation to the FPR $[0 - 0.2]$ range, the WaAUROCC values change ((a)=0.11, (b)=0.28, (c)=0.44, and (d)=0.33), showing that model (c) is preferred to the rest and that model (b) is preferred to (a). Such insights could be particularly relevant, e.g., if interested in automating decisions for ranges with a high TPR and FPR below a certain threshold.

6 Discussion and conclusion

Contribution The ROC AUC can sometimes mask critical performance differences between models, especially regarding Recall and False Positive Rates. To address this issue, we propose the WaAUROCC metric. By framing the ideal and actual ROC AUC curves as two distributions, we can compute the Wasserstein distance between the ideal scenario and the actual one. Doing so provides insights into the ROC curve steepness (higher Recall at a lower FPR), enabling a better model selection based on operational performance criteria. We showcase the usefulness of the proposed metric by evaluating machine learning models on three real-world datasets. Furthermore, we also crafted four synthetic scenarios where each ROC curve has a distinct shape but results in the same ROC AUC scores. In all cases, the WaAUROCC metric correctly identifies which curves are steeper at the beginning of the curve.

Finally, we show the WaAUROCC metric could also be applied for partial ROC curves, leading to scores that are analogous to pROCAUC while correctly preferring curves with higher initial steepness.

Limitations The proposed metric has certain limitations, among which we should mention the fact that (i) the ROC plot is discretized into bins and (ii) an arbitrary cost is established to compute the effort required to move cases between bins, under the assumption that the same cost is incurred to move an instance to a better score as to a lower score. Nevertheless, it must be noticed that strongly misclassified instances will finally incur a high cost, given they must be moved a longer way to guarantee a perfect Recall without False Positives. Another limitation of this work is the assumption that the classifiers operate within the convex hull of their ROC curve. While it is possible to find nonlinear combinations of classifiers producing ROC curves that exceed their particular convex hulls Scott et al. (1998), we consider such ensembles to be a different model on their own.

Future work We plan to contribute the implementation of this metric to the scikit-learn repository. Our future work will explore means to mitigate the limitations of the proposed metric, such as the costs of the misclassified cases being moved through bins.

Acknowledgments

This work was supported by the Slovenian Research Agency, the European Union's Horizon 2020 project STAR under grant agreement H2020-956573, and the Horizon Europe project HumAIne, Grant Agreement No. 101120218. We want to express our gratitude to all of the project partners for their contributions and collaboration.

References

Bamber, D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of mathematical psychology*, 12(4):387–415.

Bradley, A. P. (1997). The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145 – 1159.

Carrington, A. M., Fieguth, P. W., Qazi, H., Holzinger, A., Chen, H. H., Mayr, F., and Manuel, D. G. (2020). A new concordant partial auc and partial c statistic for imbalanced data in the evaluation of machine learning algorithms. *BMC medical informatics and decision making*, 20:1–12.

Carrington, A. M., Manuel, D. G., Fieguth, P. W., Ramsay, T., Osmani, V., Wernly, B., Bennett, C., Hawken, S., Magwood, O., Sheikh, Y., et al. (2022). Deep roc analysis and auc as balanced average accuracy, for improved classifier selection, audit and explanation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):329–341.

Carrington, A. M., Manuel, D. G., Fieguth, P. W., Ramsay, T., Osmani, V., Wernly, B., Bennett, C., Hawken, S., McInnes, M., Magwood, O., et al. (2021). Deep roc analysis and auc as balanced average accuracy to improve model selection, understanding and interpretation. *arXiv preprint arXiv:2103.11357*.

Connors, A. F., Dawson, N. V., Desbiens, N. A., Fulkerson, W. J., Goldman, L., Knaus, W. A., Lynn, J., Oye, R. K., Bergner, M., Damiano, A., et al. (1995). A controlled trial to improve care for seriously iii hospitalized patients: The study to understand prognoses and preferences for outcomes and risks of treatments (support). *Jama*, 274(20):1591–1598.

Cook, J. and Ramadas, V. (2020). When to consult precision-recall curves. *The Stata Journal*, 20(1):131–148.

Fahey, M. T., Irwig, L., and Macaskill, P. (1995). Meta-analysis of pap test accuracy. *American journal of epidemiology*, 141(7):680–689.

Fawcett, T. and Flach, P. A. (2005). A response to webb and ting's on the application of roc analysis to predict classification performance under varying class distributions. *Machine Learning*, 58:33–38.

Halligan, S., Altman, D. G., and Mallett, S. (2015). Disadvantages of using the area under the receiver operating characteristic curve to assess imaging tests: a discussion and proposal for an alternative approach. *European radiology*, 25:932–939.

Hand, D. J. (2009). Measuring classifier performance: a coherent alternative to the area under the roc curve. *Machine learning*, 77(1):103–123.

Hernández-Orallo, J., Flach, P., and Ferri Ramírez, C. (2012). A unified view of performance metrics: Translating threshold choice into expected classification loss. *Journal of Machine Learning Research*, 13:2813–2869.

Jeni, L. A., Cohn, J. F., and De La Torre, F. (2013). Facing imbalanced data–recommendations for the use of performance metrics. In *2013 Humaine association conference on affective computing and intelligent interaction*, pages 245–251. IEEE.

Klawonn, F., Höppner, F., and May, S. (2011). An alternative to roc and auc analysis of classifiers. In

International Symposium on Intelligent Data Analysis, pages 210–221. Springer.

Luzio, E., Ponti, M. A., Arevalo, C. R., and Argerich, L. (2024). Decoupling decision-making in fraud prevention through classifier calibration for business logic action. *arXiv preprint arXiv:2401.05240*.

Ma, H., Bandos, A. I., Rockette, H. E., and Gur, D. (2013). On use of partial area under the roc curve for evaluation of diagnostic performance. *Statistics in medicine*, 32(20):3449–3458.

Moons, K. G., Stijnen, T., Michel, B. C., Büller, H. R., Van Es, G.-A., Grobbee, D. E., and Habbema, J. D. F. (1997). Application of treatment thresholds to diagnostic-test evaluation: an alternative to the comparison of areas under receiver operating characteristic curves. *Medical Decision Making*, 17(4):447–454.

Muschelli III, J. (2020). Roc and auc with a binary predictor: a potentially misleading metric. *Journal of classification*, 37(3):696–708.

Obuchowski, N. A. and McClish, D. K. (1997). Sample size determination for diagnostic accuracy studies involving binormal roc curve indices. *Statistics in medicine*, 16(13):1529–1542.

Rožanec, J. M., Bizjak, L., Trajkova, E., Zajec, P., Keizer, J., Fortuna, B., and Mladenić, D. (2023). Active learning and novel model calibration measurements for automated visual inspection in manufacturing. *Journal of Intelligent Manufacturing*, pages 1–22.

Scheidler, J., Hricak, H., Kyle, K. Y., Subak, L., and Segal, M. R. (1997). Radiological evaluation of lymph node metastases in patients with cervical cancer: a meta-analysis. *Jama*, 278(13):1096–1101.

Scott, M. J., Niranjan, M., and Prager, R. W. (1998). Realisable classifiers: Improving operating performance on variable cost problems. In *BMVC*, pages 1–10. Citeseer.

Sofaer, H. R., Hoeting, J. A., and Jarnevich, C. S. (2019). The area under the precision-recall curve as a performance metric for rare binary events. *Methods in Ecology and Evolution*, 10(4):565–577.

Street, W. N., Wolberg, W. H., and Mangasarian, O. L. (1993). Nuclear feature extraction for breast tumor diagnosis. In *Biomedical image processing and biomedical visualization*, volume 1905, pages 861–870. SPIE.

Walter, S. D. (2005). The partial area under the summary roc curve. *Statistics in medicine*, 24(13):2025–2040.

Webb, G. I. and Ting, K. M. (2005). On the application of roc analysis to predict classification performance under varying class distributions. *Machine learning*, 58:25–32.

Yang, Z., Xu, Q., Bao, S., He, Y., Cao, X., and Huang, Q. (2021). When all we need is a piece of the pie: A generic framework for optimizing two-way partial auc. In *International Conference on Machine Learning*, pages 11820–11829. PMLR.