# THE OPTIMAL RANKING TRADEOFFS BETWEEN TWO RANKING SCORES IN TWO-CLASS CLASSIFICATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

The accuracy, true negative rate (negative class recall), true positive rate (positive class recall), negative predictive value (negative class precision), positive predictive value (positive class precision), class-specific intersections over unions, class-specific F-scores, Cohen's kappa, and balanced accuracy are all scores that one intuitively wants to maximize. However, they lead to different rankings. So, how can we establish a ranking that is an optimal tradeoff between any pair of these scores? It turns out that all these scores are special cases of ranking scores: a continuum of scores parameterized by the relative importance values assigned to the true negative, false positive, false negative, and true positive outcomes. Since all these scores are known to induce meaningful rankings, we seek the optimal tradeoffs among them. We consider the tradeoff between two scores to be optimal when it minimizes the Fréchet variance with respect to the Kendall distance (also known as the bubble-sort distance) between rankings. For any given set of classifiers to rank, the optimal tradeoff lies at the center of the shortest path (geodesic) on the manifold of all rankings induced by the ranking scores. However, in the parameter space of the ranking scores, it is rarely achieved by averaging the importance values corresponding to the initial scores. In this paper, we present an exact analytical solution for determining a ranking score that is an optimal tradeoff, for any given pair of initial scores, in the common case where the class priors are the same for all compared performances. We also experimentally investigate what happens when performance follows some distributions with fixed priors.

## 1 INTRODUCTION

Benchmarks aim at sorting some compared methods from best (first rank) to worst (largest rank). A common way to do this is to choose a score and consider that a method is *worse than or equivalent to* another one if and only if the value taken by the score for the first method is *less than or equal to* the value taken for the second method. In other words, using a score is nothing more than a trick to induce a preorder on the methods based on the well-known natural order on real numbers.

Of course, not just any score or any preorder on performances can be used if the ranking is to be meaningful. When the performances have a probabilistic meaning, the conditions that must be satisfied have been formalized in Piérard et al. (2025c) in a generic way (*the axioms of performance-based ranking*), without assuming any particular task. Moreover, in the ubiquitous setting in which the evaluation behaves linearly with respect to the evaluated method (*i.e.*, a hybrid method obtained by blindly and non-deterministically combining some base methods has a performance that is a mixture of the performances of the base methods), it has been shown that a wide class of scores leads to meaningful rankings. These are called *ranking scores* and defined as the ratio of the expected value of the importance-satisfaction product to the expected value of the importance, the *satisfaction* being a random variable modeling the task and the *importance* being another random variable specifying the relative importance given to the various evaluation outcomes Piérard et al. (2025c).

In other words, the ranking scores $R_I$ form an infinite and continuous family of scores parameterized importance values. This offers great flexibility in freely choosing the relative weights that best align with the preferences associated with the target application or its users. However, we argue that



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

choosing the score (*i.e.*, the importance values in the case of ranking scores) remains a delicate matter for a challenge organizer or a benchmarker Halin et al. (2024), for example, who is unaware of the applications that the people consulting the proposed ranking are targeting. Consequently, further theoretical advances to guide this choice are still needed. This paper aims to contribute to that effort for two-class classification, with a special focus on the particular case in which all compared performances share the same class priors. This is a very common setting as good practice suggests to compare only performances determined by evaluating the methods on the same dataset, hence fixing the class priors.

In two-class classification, one usually considers a class as *negative* and the other one as *positive*. It should be stressed that this is just a convention. The choice of which class is positive is always arbitrary. Four evaluation outcomes are usually considered: true negative ($tn$), false positive ($fp$), false negative ($fn$), and true positive ($tp$). All of us will certainly agree that the outcomes $tn$ and $tp$ are fully satisfying (the satisfaction is set to one), and that the outcomes $fp$ and $fn$ are not satisfying at all (the satisfaction is set to zero). In this setting, the ranking scores include well-known scores such as the accuracy $A$, the true negative rate $TNR = Re^- = Sp$ (*a.k.a.*inverse recall and specificity), the true positive rate $TPR = Re^+ = Se$ (*a.k.a.*recall and sensitivity), the negative predictive value $NPV = Pr^-$ (*a.k.a.*inverse precision), the positive predictive value $PPV = Pr^+$ (*a.k.a.*precision), as well as the F-scores $F_\beta$ and the intersection over union $IoU$ (*a.k.a.*Jaccard coefficient) that can be defined for both classes. When the class priors are fixed, the balanced accuracy $BA$ also belongs to the family of ranking scores (the importance values are constants expressed in terms of the class priors), and Cohen's kappa $\kappa$ (*a.k.a.*Heidke Skill Score) is linearly increasing with a ranking score Piérard et al. (2024). All these classical scores are therefore associated with well-defined importance values. For any ranking score, its inverse (*i.e.*, the function composition of the score with the operation on performances that swaps the negative and positive classes Powers (2020)) is also a ranking score Piérard et al. (2024). For example, the true negative and positive rates are inverses of each other and the negative and positive predictive values are also inverses of each other.

The main objective of this paper is to provide all the theoretical and algorithmic tools to determine optimal tradeoffs between two scores. This indeed includes the determination of the tradeoff between a class-specific score (*e.g.*, the precision, recall, F-score, or intersection over union) and its inverse. But this also includes the determination of the tradeoff between two scores for the same class (*e.g.* between precision and recall).

There are mainly two reasons why one cannot just use any arbitrarily chosen mean (*e.g.*, arithmetic, geometric, harmonic, quadratic) to establish a tradeoff between scores. First, the performance ordering induced by a mean of two base scores does not necessarily satisfy the axioms of performance-based ranking even if the orderings induced by these two base scores do. Second, even if this was the case, the resulting ranking would not necessarily be intermediate between the rankings induced by the two base scores. For example, it has been shown in Piérard et al. (2026) that despite the fact that the F-score $F_1$ is the harmonic mean of the precision and recall and that $F_1$ is known to lead to meaningful rankings (it is a ranking score), the F-score sometimes mimics perfectly the precision or recall in terms of rankings, totally ignoring the other. It is very likely that this problem is not specific to the tradeoff between precision and recall. For this reason, it is far preferable to establish an optimal tradeoff from the perspective of ranking. We aim at providing all the theoretical and algorithmic tools to answer the following question.

**Question.** *In two-class classification, for any given set of performances of for any pair of scores perfectly rank-correlated with ranking scores (*e.g.*, $A$, $TNR$, $TPR$, $NPV$, $PPV$, $F_\beta^-$, $F_\beta^+$, $IoU^-$, $IoU^+$, $BA$, $\kappa^-$, $\kappa^+$), what ranking scores are optimal tradeoffs from a ranking point of view?*

To the best of our knowledge, only the very particular case of the optimal tradeoff between the precision and recall of the positive class has been extensively studied. This was done in Piérard et al. (2026). In this paper, we keep the same methodology and reasoning, generalizing it to any pair of ranking scores. Let us summarize the methodology in four points.

**(1) The search space.** How one interprets a ranking depends on how one interprets the score that induces it. For this reason, the use of rank aggregation methods, such as those developed by Kemeny Kemeny (1959), are inappropriate as they operate on the rankings only, ignoring their origins, and lead to a ranking that is not related to any score. Instead, we explicitly determine the tradeoff score that leads to the optimal tradeoff ranking. Moreover, as we want to be sure that the tradeoff

corresponds to one of the possible meaningful rankings, we seek for the optimal tradeoff within the family of ranking scores. More information about the search space can be found in section 3.1.

**(2) The means.** The main challenge facing us is that the reasoning of Piérard et al. (2026) was heavily based on the observation that there exists a weighted mean (the weighted harmonic mean) such that, if applied to precision and recall, one obtains ranking scores. To be able to generalize it, for any pair of ranking scores we need to find a weighted mean such that, if applied to them, one obtains also a ranking score. This is done in section 3.2.

**(3) The shortest paths.** If there is a mean between the scores $R_{I_A}$ and $R_{I_B}$, then it is possible to determine scores $R_I$ such that $d_\tau(R_{I_A}; R_I) + d_\tau(R_I; R_{I_B}) = d_\tau(R_{I_A}; R_{I_B})$ where $d_\tau = \frac{1-\tau}{2}$ is Kendall distance Kendall (1938) (also known as the bubble-sort distance) between rankings. We say that these scores form a shortest path between $R_{I_A}$ and $R_{I_B}$. This is done in section 3.3.

**(4) The optimal tradeoffs.** The optimal tradeoff is the midpoint on the shortest path. It is given by the ranking score $R_I$ such that $d_\tau(R_{I_A}; R_I) = d_\tau(R_I; R_{I_B})$. It is a Karcher mean Karcher (1977) as it minimizes the Fréchet variance with respect to the Kendall distance between rankings:

$$R_{I^*} = \arg\min_{R_I} \sigma^2(I) \qquad \text{with} \qquad \sigma^2(I) = d_\tau^2(R_{I_A}; R_I) + d_\tau^2(R_I; R_{I_B})\,. \tag{1}$$

Establishing solutions for this equation is done in section 3.4.

Our theoretical contributions are threefold.

1. From the point of view of *values*, we introduce scores that are homographic transformations of ranking scores. We show that, when the class priors are fixed, for any pair of ranking scores $(R_{I_A}, R_{I_B})$ there exist ranking scores that are weighted $f$-means of some ranking scores $R'_{I_A}$ and $R'_{I_B}$ monotonically increasing with respect to $R_{I_A}$ and $R_{I_B}$, respectively, and that are generated by homographies $f$. This stands in contrast to the unjustifiable practice of considering only the arithmetic, geometric, and harmonic means.
2. From the point of view of *rankings*, we provide the equation for a shortest path between any pair of ranking scores $(R_{I_A}, R_{I_B})$, assuming the priors are fixed. Furthermore, we present a method for determining the optimal tradeoff when the set of performances to be ranked is known. This optimal tradeoff lies on the shortest path.
3. We show that, among all the ways to parameterize the two-dimensional manifold of performance orderings induced by ranking scores, there is one that has properties particularly interesting for finding optimal tradeoffs. We call this *the parameterization adapted to the priors*. Its usefulness should not be viewed as limited to simplifying equations, since a well-designed parameterization is key to obtaining ranking visualization tools that are well-suited to specific needs. Readers curious to discover such a tool, called *the Tile*, are invited to consult Piérard et al. (2024; 2025b) for the theory, Piérard et al. (2025d) for an open-source implementation, and Halin et al. (2024); Piérard et al. (2025a); Bouwmans et al. (2026); Zakharova et al. (2026) for some uses of it.

## 2 Related Work: The Risky Means

We dedicate this section to the tradeoffs between two scores that have been defined in the literature, in the case of two-class classification. As this list shows, these tradeoffs are all established using an arbitrarily chosen mean. The so-called "Pythagorean" means (*i.e.*, the arithmetic mean $\mu_A$, the geometric mean $\mu_G$, and the harmonic mean $\mu_H$) are systematically used, although there is no good reason to limit ourselves to the means that were known and studied by the ancient Greeks, some two and a half millennia ago. The base scores on which these tradeoffs are based are primarily the following: $TNR$, $TPR$, $NPV$, and $PPV$. Hereafter, we list not only the means but also their monotonous transformations as they induce the same ranking or its inverse. Scores related to $\mu_A(TNR; TPR)$. Many authors combine $TNR$ and $TPR$. Their weighted arithmetic mean is the weighted accuracy $WA$. When the weights are $0.5$, we obtain the balanced accuracy $BA$ Brodersen et al. (2010), and when the weights correspond to the priors, we get the accuracy $A$. Youden's index Youden (1950); Fluss et al. (2005) or Youden's $J_Y$ statistic is $J_Y = 2BA - 1$ . It is also called informedness Powers (2020) and Peirce Skill Score Peirce (1884). The determinant $|C|$ Foulle (2025) of

| score | classic formula | $I_{tn}$ | $I_{fp}$ | $I_{fn}$ | $I_{tp}$ | $a^{(\pi)}(I)$ | $b^{(\pi)}(I)$ |
|---|---|---|---|---|---|---|---|
| $TNR = Re^-$ | $P(Y = \hat{Y} \mid Y = c_-)$ | 1 | 1 | 0 | 0 | 0 | 0 |
| $TPR = Re^+$ | $P(Y = \hat{Y} \mid Y = c_+)$ | 0 | 0 | 1 | 1 | 1 | 1 |
| $NPV = Pr^-$ | $P(Y = \hat{Y} \mid \hat{Y} = c_-)$ | 1 | 0 | 1 | 0 | 0 | 1 |
| $PPV = Pr^+$ | $P(Y = \hat{Y} \mid \hat{Y} = c_+)$ | 0 | 1 | 0 | 1 | 1 | 0 |
| $F_\beta^-$ | $\frac{1}{F_\beta^-} = \frac{1}{1+\beta^2}\frac{1}{Pr^-} + \frac{\beta^2}{1+\beta^2}\frac{1}{Re^-}$ | 1 | $\frac{\beta^2}{1+\beta^2}$ | $\frac{1}{1+\beta^2}$ | 0 | 0 | $\frac{1}{1+\ell}$ with $\ell = \beta^2\frac{\pi_-}{\pi_+}$ |
| $F_\beta^+$ | $\frac{1}{F_\beta^+} = \frac{1}{1+\beta^2}\frac{1}{Pr^+} + \frac{\beta^2}{1+\beta^2}\frac{1}{Re^+}$ | 0 | $\frac{1}{1+\beta^2}$ | $\frac{\beta^2}{1+\beta^2}$ | 1 | 1 | $\frac{\ell}{1+\ell}$ with $\ell = \beta^2\frac{\pi_+}{\pi_-}$ |
| $IoU^-$ | $P(Y = \hat{Y} \mid Y = c_- \cup \hat{Y} = c_-)$ | 1 | 1 | 1 | 0 | 0 | $\pi_+$ |
| $IoU^+$ | $P(Y = \hat{Y} \mid Y = c_+ \cup \hat{Y} = c_+)$ | 0 | 1 | 1 | 1 | 1 | $\pi_+$ |
| $BA$ | $BA = 1/2TNR + 1/2TPR$ | $\pi_+$ | $\pi_+$ | $\pi_-$ | $\pi_-$ | 1/2 | 1/2 |
| $A$ | $P(Y = \hat{Y}) = \mathbf{E}_P[S]$ | 1 | 1 | 1 | 1 | $\pi_+$ | $\pi_+$ |
| Linear of $\kappa$ | $(\pi_-^2 + \pi_+^2)\kappa + 2\pi_-\pi_+$ | $\frac{\pi_+^2}{\pi_-^2+\pi_+^2}$ | $\frac{1}{2}$ | $\frac{1}{2}$ | $\frac{\pi_-^2}{\pi_-^2+\pi_+^2}$ | $\pi_-$ | $\pi_+$ |

Table 1: Some ranking scores, when the class priors $(\pi_-, \pi_+)$ are fixed, with corresponding importance values, and the coordinates of the induced performance ordering in the $(a^{(\pi)}, b^{(\pi)})$ space.

the (normalized) confusion matrix is $|C| = \pi_-\pi_+ J_Y$. Scores related to $\mu_G(TNR;TPR)$. Instead of taking an arithmetic mean, it has been proposed to take the geometric mean Kubat (1997); Barandela et al. (2003); Guo et al. (2008), which leads to the G-measure Canbek et al. (2017). Scores related to $\mu_H(PPV;TPR)$. This is the F-one score van Rijsbergen (1979); Christen et al. (2023), also called Dice-Sørensen coefficient Dice (1945). It is well known that that Jaccard coefficient Jaccard (1908) leads to the same performance ordering as $F_1$, due to a strict increasing monotonic relationship between the two scores. $\mu_G(PPV;TPR)$ is Fowlkes-Mallows index Fowlkes & Mallows (1983). $\mu_A(PPV;TPR)$ is Kulczynski index Kulczyński (1927). The Individual Classification Success Index (ICSI) is a monotonous transformation of Kulczynski index. Scores related to $\mu_A(NPV;PPV)$. The markedness Powers (2020) is defined as $NPV + PPV - 1$ and is also known as the Clayton Skill Score Canbek et al. (2017); Wilks (2020).

## 3 Theory

**Probabilistic grounding and notations (following Piérard et al. (2025c)).** Let us consider a measurable space $(\Omega, \Sigma)$ with a sample space (*a.k.a.* universe) made of four evaluation outcomes $\Omega = \{tn, fp, fn, tp\}$ and the event space $\Sigma = 2^\Omega$. By definition, performances are probability measures on $\Omega$, and scores are real functions whose domain is the set of all possible performances. We denote by $\mathbf{E}_P[\cdot]$ the expected value of random variables for the performance $P$. In order to establish a connection with two-class classification, we introduce a set $\mathbb{C} = \{c_-, c_+\}$ that we interpret as the set of classes, $c_-$ being called the *negative class* and $c_+$ being called the *positive class*. Moreover, since it is common to refer to ground-truth and predicted classes, we introduce symbolic variables for them. The ground-truth class is $Y : \Omega \to \mathbb{C}$ with $Y(tn) = Y(fp) = c_-$ and $\hat{Y}(fn) = Y(tp) = c_+$. The predicted class is $\hat{Y} : \Omega \to \mathbb{C}$ with $\hat{Y}(tn) = \hat{Y}(fn) = c_-$ and $\hat{Y}(fp) = \hat{Y}(tp) = c_+$. The symbols $tn$, $fp$, $fn$, and $tp$ can thus be interpreted, respectively, as a *true negative*, *false positive*, *false negative*, and *true positive*. For a performance $P$, the class priors are given by $\pi_- = P(Y = c_-)$ and $\pi_+ = P(Y = c_+)$. The classification task is modeled by a binary random variable called *satisfaction* $S : \Omega \to \mathbb{R} : \omega \mapsto \mathbf{1}_{\{tn,tp\}}(\omega)$. In other words $S = 1$ if and only if $Y = \hat{Y}$ and $S = 0$ if and only if $Y \neq \hat{Y}$. We also introduce another random variable, positive but non-zero, called *importance* $I : \Omega \to [0, \infty[$. To shorten the notations, we pose $I_\omega = I(\omega)$ and $P_\omega = P(\{\omega\})$ for all $\omega \in \Omega$.

### 3.1 The Performance Orderings Induced by Ranking Scores

**A space of scores.** To ensure meaningful ranking tradeoffs, we seek for optimal tradeoffs within the family of ranking scores. These scores are generic, but we are only interested in their particularization to two-class classification. They have been defined in Piérard et al. (2025c) as

$$R_I(P) = \frac{\mathbf{E}_P[SI]}{\mathbf{E}_P[I]} = \frac{I_{tn}\,P_{tn} + I_{tp}\,P_{tp}}{I_{tn}\,P_{tn} + I_{fp}\,P_{fp} + I_{fn}\,P_{fn} + I_{tp}\,P_{tp}}\,. \tag{2}$$

A few examples of ranking scores or of scores strictly increasing with them (*i.e.*, perfectly rank-correlated with them) are given in table 1. The ranking scores are insensitive to the scaling of

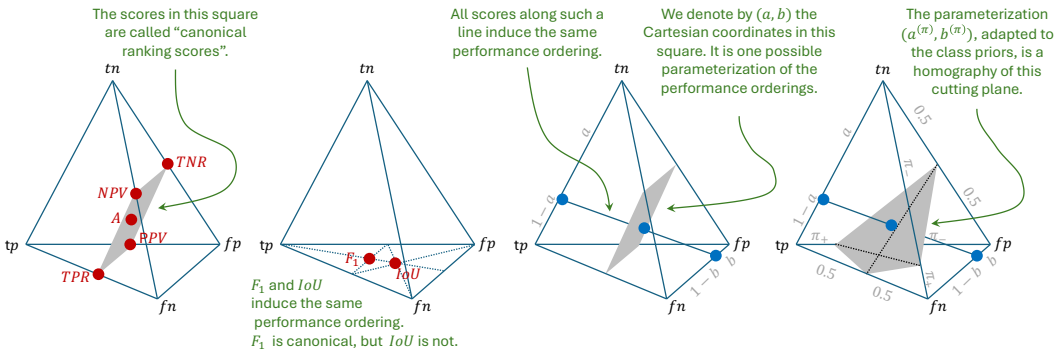



Figure 1: The tetrahedron of ranking scores, or the simplex of importance values in two-class classification. Be careful not to confuse this with the tetrahedron of performances or confusion simplex Chicco et al. (2021); Chicco & Jurman (2023); Lovell et al. (2023); Brzezinski et al. (2017; 2019; 2018). In this one, each point corresponds to a score, not a performance.

the importance $I$, so we can limit ourselves to considering that $I_{tn} + I_{fp} + I_{fn} + I_{tp} = 1$. As the importance values must be positive, the ranking scores form a three-dimensional simplex (see fig. 1). In the tetrahedron $(I_{tn}, I_{fp}, I_{fn}, I_{tp})$, all scores on the line between $(1-a, 0, 0, a)$ and $(0, 1-b, b, 0)$ induce the same performance ordering. We call them *iso-ranking lines*.

**A space of performance orderings.** The performance orderings can be parameterized in different ways. A quite intuitive choice is with the pair $(a, b) \in [0,1]^2$. These are Cartesian coordinates in the square $I_{tn} + I_{tp} = I_{fp} + I_{fn}$ cutting the tetrahedron in the middle and intersecting each iso-ranking line in one point. But we discovered that there is a much more interesting cutting plane to consider for our purpose: $I_{tn}\,\pi_- + I_{tp}\,\pi_+ = I_{fp}\,\pi_- + I_{fn}\,\pi_+$. Its advantages will become clear in the next subsections. This plane form a convex kite that can be mapped to the square $[0,1]^2$ with a homography (thus a collineation; this will be important for us a little bit further) $(a^{(\pi)}, b^{(\pi)})$:

$$a^{(\pi)}(I) = \frac{I_{tp}\,\pi_+}{I_{tn}\,\pi_- + I_{tp}\,\pi_+} \in [0,1] \qquad b^{(\pi)}(I) = \frac{I_{fn}\,\pi_+}{I_{fp}\,\pi_- + I_{fn}\,\pi_+} \in [0,1]\,. \tag{3}$$

Note that $a^{(\pi)}$ and $b^{(\pi)}$ are strictly increasing with $a$ and $b$, respectively.

### 3.2 Means of Ranking Scores

Let us remind that for real numbers (the co-domain of scores), a large class of means is known as the generalized $f$-means, quasi-arithmetic means, or Kolmogorov-Nagumo-de Finetti means Kolmogorov (1930). The generalized $f$-mean $\overline{X}$, on an interval $[l, u]$, between $X_1$ and $X_2$ is defined as $f(\overline{X}) = \frac{1}{2}f(X_1) + \frac{1}{2}f(X_2)$, where $f$ is monotone over $[l, u]$. These means include the arithmetic, geometric, harmonic, and quadratic means, to cite only a few examples. In the case of a binary satisfaction, in which we are, it has been shown in Piérard et al. (2025c) that, for the generators $f : x \mapsto x^{-1}$ and $f : x \mapsto (1-x)^{-1}$, some ranking scores are means of others. When the priors are fixed, we discovered that the means to consider are more generaly the $f$-means generated by some homographies $f$.[1] For this reason, we introduce the homographic transformations of ranking scores:

$$f \circ R_I = \frac{A\,R_I + B}{C\,R_I + D} = \frac{(A+B)\,I_{tn}\,P_{tn} + B\,I_{fp}\,P_{fp} + B\,I_{fn}\,P_{fn} + (A+B)\,I_{tp}\,P_{tp}}{(C+D)\,I_{tn}\,P_{tn} + D\,I_{fp}\,P_{fp} + D\,I_{fn}\,P_{fn} + (C+D)\,I_{tp}\,P_{tp}}\,, \tag{4}$$

with $AD > BC$. This constraint on the transformation parameters ensures that the transformation is strictly increasing and therefore that the score $f \circ R_I$ induces the same performance ordering as $R_I$.

From now, we leave aside the unnecessary constraint $\sum_\omega I(\omega) = 1$. To ensure $I_{tn}\,\pi_- + I_{tp}\,\pi_+ = I_{fp}\,\pi_- + I_{fn}\,\pi_+$, for the performance ordering $(a^{(\pi)}, b^{(\pi)})$, we consider only

$$I_{tn} = \pi_+\,(1 - a^{(\pi)}) \qquad I_{fp} = \pi_+\,(1 - b^{(\pi)}) \qquad I_{fn} = \pi_-\,b^{(\pi)} \qquad I_{tp} = \pi_-\,a^{(\pi)}\,. \tag{5}$$

[1] If two weighted means are equal, their generators are related by homographies Grünwald & Páles (2019). Also, the composition of two homographies is also a homography. That is why we consider all homographies.

We now take a look at the conditions on $(A, B, C, D)$ to obtain a score $R_I$ with the $f$-mean of some given scores $R_{I_A}$ and $R_{I_B}$. The equation $f \circ R_I = {}^1\!/\!_2 f \circ R_{I_A} + {}^1\!/\!_2 f \circ R_{I_B}$ has solutions when $I$ is the arithmetic mean of $I_A$ and $I_B$ and the denominator of eq. (4) is the same for $f \circ R_{I_A}$ and $f \circ R_{I_B}$, for any given performance. Thus, $(C + D)\,\Delta_{tn}\,P_{tn} + D\,\Delta_{fp}\,P_{fp} + D\,\Delta_{fn}\,P_{fn} + (C + D)\,\Delta_{tp}\,P_{tp} = 0$, with $\Delta_\omega = I_A(\omega) - I_B(\omega)\ \forall \omega \in \Omega$. As $P_{tn} + P_{fp}$ and $P_{fn} + P_{tp}$ are constant when the priors are fixed, the coefficients in front of $P_{tn}$ and $P_{fp}$ should be equal, as well as those of $P_{fn}$ and $P_{tp}$:

$$(C + D)\,\Delta_{tn} = D\,\Delta_{fp} \qquad\qquad D\,\Delta_{fn} = (C + D)\,\Delta_{tp} \tag{6}$$

Let us pose $\Delta_{a^{(\pi)}} = a^{(\pi)}(I_B) - a^{(\pi)}(I_A)$ and $\Delta_{b^{(\pi)}} = b^{(\pi)}(I_B) - b^{(\pi)}(I_A)$. We have $\Delta_{tn} = -\pi_+\,\Delta_{a^{(\pi)}}$, $\Delta_{fp} = -\pi_+\,\Delta_{b^{(\pi)}}$, $\Delta_{fn} = \pi_-\,\Delta_{b^{(\pi)}}$, and $\Delta_{tp} = \pi_-\,\Delta_{a^{(\pi)}}$. So, the two constraints given in eq. (6) are equivalent to $(C + D)\,\Delta_{a^{(\pi)}} = D\,\Delta_{b^{(\pi)}}$. We choose the solution

$$A = \Delta_{a^{(\pi)}} \qquad B = \Delta_{a^{(\pi)}} - \Delta_{b^{(\pi)}} \qquad C = \Delta_{b^{(\pi)}} - \Delta_{a^{(\pi)}} \qquad D = \Delta_{a^{(\pi)}}\,. \tag{7}$$

This homography $f$ is an involution (*i.e.*, $f^{-1} = f$). Its pole $z = -D/C$ can be in the $[0, 1]$ range (*i.e.*, the image of ranking scores). So, $f$ is monotonously increasing on $]-\infty, z[$ and on $]z, \infty[$, but not on $[0, 1]$. However, this is not a problem for us because it can be shown that

$$R_{I_A}(P) < z \Leftrightarrow R_{I_B}(P) < z \quad R_{I_A}(P) = z \Leftrightarrow R_{I_B}(P) = z \quad R_{I_A}(P) > z \Leftrightarrow R_{I_B}(P) > z\,. \tag{8}$$

### 3.3 Universal Shortest Paths Between Rankings

We have just seen that, when the class priors are fixed, for any pair of ranking scores $(R_{I_A}, R_{I_B})$ there exist ranking scores $R_I$ that are weighted means of some ranking scores $R'_{I_A}$ and $R'_{I_B}$ monotonically increasing with respect to $R_{I_A}$ and $R_{I_B}$, respectively. The reasoning of Piérard et al. (2026) can be applied as is: it follows immediately that $d_\tau(R_{I_A}; R_I) + d_\tau(R_I; R_{I_B}) = d_\tau(R_{I_A}; R_{I_B})$ for all these means $R_I$. This shortest path is obtained with a weighted linear interpolation of importance values when they satisfy eq. (5). However, taking $I = \frac{1}{2}I_A + \frac{1}{2}I_B$ does not necessarily lead to optimal tradeoff (this will be the topic of the next subsection). It is interesting to note that the passage from the plane of eq. (5) to the cutting plane shown on the right hand side tetrahedron of fig. 1 is a homography. As the mapping from this cutting plane to the space $(a^{(\pi)}, b^{(\pi)})$ is also a homography, it follows that the shortest paths are line segments in $(a^{(\pi)}, b^{(\pi)})$. Aside from the ease of graphical representation this provides (see section 4), it also sheds light on the manifolds of performance orderings induced by the ranking scores when the class priors are fixed. According to Beltrami's theorem Beltrami (1868), this implies that these manifolds have a constant curvature.

### 3.4 Optimal Tradeoffs Between Two Ranking Scores

Again, the reasoning that was proposed in Piérard et al. (2026) to determine the optimal tradeoff between $Pr$ and $Re$ can be easily generalized to the determine the optimal tradeoff between any pair of ranking scores thanks to the fact that we found means. If there is a score $R_I$ such that $d_\tau(R_{I_A}; R_I) = d_\tau(R_I; R_{I_B})$, then it is a solution to eq. (1).

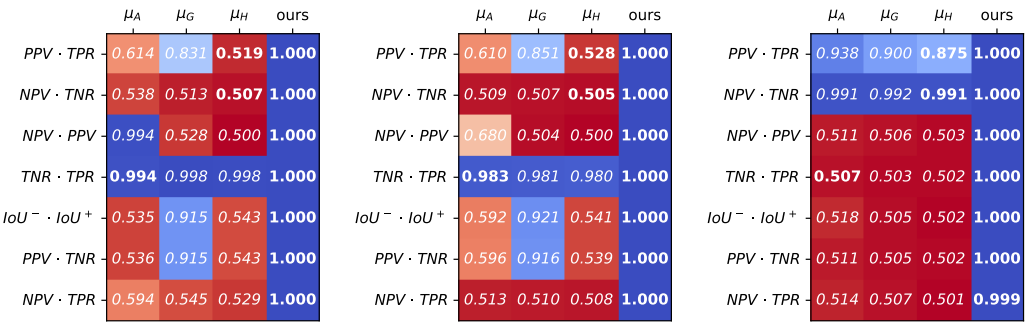
Let us consider any score $X$ that is a weighted $f$-mean of some scores $X_1$ and $X_2$, that is $f \circ X = (1 - w)\,f \circ X_1 + w\,f \circ X_2$ with $w \in [0, 1]$. The performances $P_i$ and $P_j$ are put on an equal footing by $X$ when

$$X(P_i) = X(P_j) \Leftrightarrow \frac{w}{w - 1} = \frac{(f \circ X_1)(P_i) - (f \circ X_1)(P_j)}{(f \circ X_2)(P_i) - (f \circ X_2)(P_j)} = \psi(P_i; P_j)\,. \tag{9}$$

The weight $w$ depends on the pair of performances $(P_i; P_j)$. We have $w(P_i; P_j) = \frac{\psi(P_i;P_j)}{\psi(P_i;P_j)-1}$. If both $X_1$ and $X_2$ agree on the order between $P_i$ and $P_j$, then $w(P_i; P_j) \notin [0, 1]$.

The optimal tradeoffs between $X_1$ and $X_2$, from a ranking perspective, are the scores on the shortest path such that there is the same amount of swaps on both sides. Denoting multisets by $\{\cdot\}$, a weight leading to an optimal tradeoff is given by

$$w^* = \text{median}\left(\{w(P_i, P_j)\,|\,i \neq j \wedge w(P_i, P_j) \in [0, 1]\} \cup \{0, 1\}\right)\,. \tag{10}$$




(a) With $\Pi_A$. (b) With $\Pi_B$. (c) With $\Pi_C$.

Figure 2: Optimality $\mathcal{O}$ obtained with each of the three Pythagorean means ($\mu_A$: arithmetic, $\mu_G$: geometric, $\mu_H$: harmonic) and with our method, for seven pairs of scores and three sets of 1000 performances corresponding to $\pi_+ = 0.01$. Our method always yields an optimal result ($\mathcal{O} = 1$, in blue), but arbitrarily chosen means are only approximately optimal in a few specific cases. In fact, the optimality value is often barely above 0.5 (the worst, in red). The meaningful rankings (*i.e.*, the tradeoffs satisfying the axioms of performance-based ranking Piérard et al. (2025c)) are in bold.

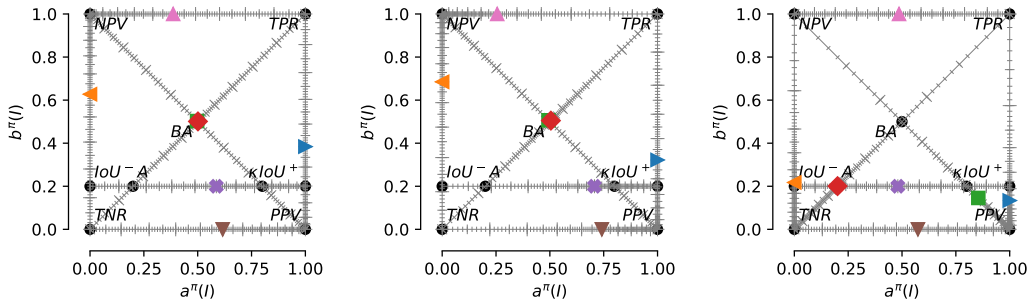
# 4 Experiments

## 4.1 Arbitrarily chosen means: a recipe for trouble

The purpose of our 1st experiment is to highlight the problems that arise when a tradeoff is made between two scores $(R_{I_A}, R_{I_B})$ using an arbitrarily chosen mean $\mu$. We illustrate this using three common means (the arithmetic, the geometric, and the harmonic) for seven tradeoffs (three of them are tradeoffs between a score and its inverse), based on three large sets of performances. These sets consist of performances drawn at random, *i.i.d.*, following three performance distributions that have been studied theoretically in Piérard et al. (2026) in the specific case of the $Pr$-$Re$ tradeoff. In the case of the set $\Pi_A$, the distribution is uniform for fixed priors; this corresponds to a uniform distribution in $ROC = (FPR, TPR) \in [0,1]^2$. For $\Pi_B$, the distribution remains uniform with fixed priors but is restricted to performances above the no-skills (*i.e.*, $TPR \geq FPR$). Finally, for $\Pi_C$, we restrict even further and consider a uniform distribution over $(FPR, TPR) \in [0, \pi_+] \times [\pi_+, 1]$.

For each pair of scores for which a compromise must be determined, we calculate the values taken by these two scores as well as their means, across all performances. Then, we calculate three Kendall correlations: between the two base scores, between the first score and the mean, and between the mean and the second score. From these values, we derive the optimality. It is defined in Piérard et al. (2026) and can be generalized as: $\mathcal{O}(X) = 1 - |d_\tau(R_{I_A}; X) - d_\tau(X; R_{I_B})| / 2d_\tau(R_{I_A}; R_{I_B})$, where $X$ is the proposed tradeoff. For a mean $\mu$, $X = \mu(R_{I_A}; R_{I_B})$. On any shortest path, the optimality increases linearly with Kendall's distance from both base scores (where it takes the minimal value $\mathcal{O} = 1/2$) towards the optimal tradeoff (where it takes the maximal value $\mathcal{O} = 1$). The optimality can be interpreted as the proportion of pairs of performances for which $X$ and the optimal tradeoff do not agree, among those for which the ordering is not trivial (*i.e.*, $R_{I_A}$ and $R_{I_B}$ do not agree). Indeed, the optimality can also be computed for any ranking score that our theory recommends.

The results of this first experiment are given in fig. 2. One can see that the ranking score recommended by our theory has an optimality of 1, as expected, no matter the set of performances and the pair of scores that are considered. This is never the case with arbitrarily chosen means. With a prior $\pi_+ = 0.01$, the optimality is often close to 0.5 (the worst scenario) with arbitrarily chosen means. In this figure, we have written in bold the values for the cases in which the tradeoff score satisfies the axioms of performance-based ranking. It is always the case with the tradeoffs we recommend as they are ranking scores. However, it is rarely the case with the classic means (for the readers interested to dive deeper into this topic, this is because the third axiom is not satisfied).




(a) $\Pi_A$: uniform (b) $\Pi_B$: uniform, above no-skills (c) $\Pi_C$: uniform , close to oracle

Figure 3: Case studies with large performances sets randomly drawn from reference distributions with fixed priors ($\pi_+ = 0.2$). The colored markers indicate, in the $(a^{(\pi)}, b^{(\pi)})$, the performance orderings that are the optimal tradeoffs for the pairs $PPV \cdot TPR$ (▸), $NPV \cdot TNR$ (◂), $NPV \cdot PPV$ (■), $TNR \cdot TPR$ (◆), $IoU^- \cdot IoU^+$ (✖), $PPV \cdot TNR$ (▼), and $NPV \cdot TPR$ (▲).

### 4.2 Case studies: with reference distributions

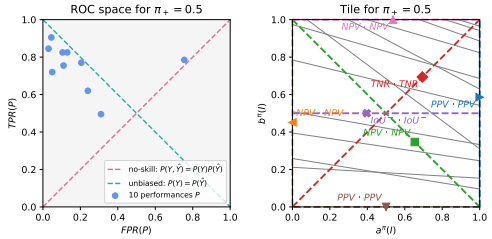
The goal of our 2nd experiment is to illustrate, within the $(a^{(\pi)}, b^{(\pi)})$ space, where the optimal tradeoffs lie for the seven pairs of base scores and for the three distributions of performances of the previous experiment. For each, We draw 1000 performances to have large sets. Considering reference distributions is interesting because the organizer of a challenge is often required to announce the score that will be used to rank participants before receiving their actual performance results. For the organizer, one option is to choose a reference distribution and use it to select a score.

In fig. 3, the rankings induced by the base scores, $A$, $BA$, and $\kappa$ are shown with labeled black dots. Their locations can depend on the priors, but never on the set of performances. In the $(a^{(\pi)}, b^{(\pi)})$ space, the shortest path is the line segment connecting the base scores. The position of the optimal tradeoff on it depends on the set of performances. Along the seven studied segments, we draw graduated rulers such that the (not uniform) spacing between the marks is proportional to $d_\tau$. The optimal tradeoff (colored point) is thus at the median mark. The marks on the rulers allow to visualize how the optimal tradeoff would move if one weighted non-uniformly the two base scores in $\sigma^2(I)$ (see eq. (1)). In this case, one needs to replace the median by a quantile in eq. (10).

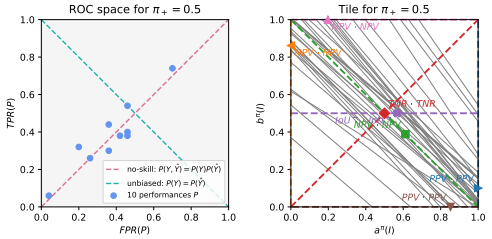
We can see that the optimal tradeoff between precision ($PPV$) and recall ($TPR$) is not $F_1$ (same ranking as $IoU^+$), but always corresponds to one of the $F_\beta$ scores ($a^{(\pi)} = 1$ line). This confirms the results of Piérard et al. (2026). We also see that the behavior is, broadly speaking, quite similar between $\Pi_A$ and $\Pi_B$, but quite different with $\Pi_C$, particularly for the optimal tradeoff between $TNR$ and $TPR$ (close to the balanced accuracy $BA$ for the first two sets; close to the accuracy $A$ for the third) and for the optimal tradeoff between $NPV$ and $PPV$ (close to the balanced accuracy $BA$ for the first two datasets; in the lower right quadrant of Cohen's $\kappa$ for the third). There is generally no universal optimal tradeoff for a given pair of scores.

### 4.3 Case studies: with real-world rankings

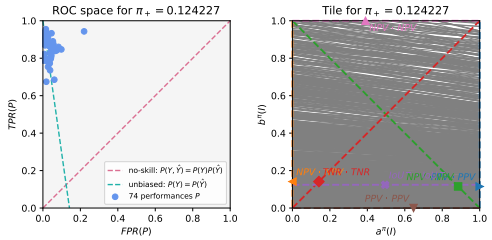
The goal of our 3rd experiment is to show what happens with real-world performance sets whose size is significantly smaller than those we studied in the previous experiment. We considered four published or publicly available performance sets related to computer vision problems. The results are shown in fig. 4. There, we have plotted in the $(a^{(\pi)}, b^{(\pi)})$ space the frontiers between rankings. They correspond to the scores for which any two performances are considered equivalent. It turns out that, in the case of our parameterization adapted to the priors, $R_I(P_i) = R_I(P_j)$ holds if and only if the performance orderings lie on the line $\zeta_a\, a^{(\pi)}(I) + \zeta_b\, b^{(\pi)}(I) + \zeta = 0$ with $\zeta_a = FPR(P_i)\,FNR(P_j) - FPR(P_j)\,FNR(P_i)$, $\zeta_b = TPR(P_i)\,TNR(P_j) - TPR(P_j)\,TNR(P_i)$, and $\zeta = TNR(P_i)\,FPR(P_j) - TNR(P_j)\,FPR(P_i)$. On any given shorter path, the number of times the path crosses one of these lines on either side of the optimal tradeoffs is equal.



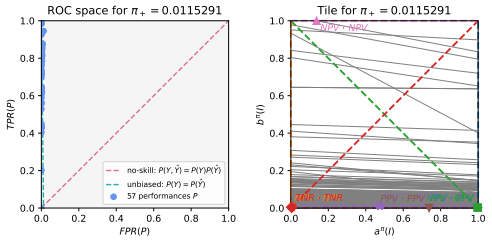

(a) With the 10 performances reported in Bouwmans et al. (2026) for the *Illegal Waste Dumping Detection* challenge, when the evaluation is done on the LS.




(b) With the 10 performances reported in Bouwmans et al. (2026) for the *Illegal Waste Dumping Detection* challenge, when the evaluation is done on the TS.




(c) With the performances given in Halin et al. (2024) for 74 semantic segmentation models from *MMSegmentation* MMSegmentation Contributors (2020).




(d) With the performances of 57 background subtraction methods reported on changedetection.net (CDnet 2014 Wang et al. (2014)) for the *blizzard* video.

Figure 4: Case studies with 4 real-world performance sets. For each of them, the performances are shown in the ROC space, and the optimal tradeoffs for 7 pairs of scores are shown in $(a^{(\pi)}, b^{(\pi)})$.

The two first cases are based on the 10 performances reported in Bouwmans et al. (2026) regarding the results of the Illegal Waste Dumping Detection (IWDD) challenge 2026. The authors reported the performances measured on the learning set (LS) and on the test set (TS) to highlight the discrepancy between the two (overfitting). Both LS and TS are balanced ($\pi_- = \pi_+$).We examine the LS case in fig. 4a and the TS case in fig. 4b. In the latter case, all performances are close to the no-skills (rising diagonal of ROC, *i.e.* $Y \perp\!\!\!\perp_P \hat{Y}$), and the optimal tradeoffs approach the descending diagonal of $(a^{(\pi)}, b^{(\pi)})$ where the no-skill performances are put on an equal footing. In the extreme case where all compared performances were no-skill, all the boundaries between the rankings would coincide with the descending diagonal, and consequently, all optimal tradeoffs would lie on it.

The last two cases are for the performances reported by Halin et al. (2024) for 74 semantic segmentation models from *MMSegmentation* MMSegmentation Contributors (2020), evaluated on BDD100K Yu et al. (2020) after grouping the semantic labels into two classes, and the performances of 57 background subtraction methods (classifying pixels as background $c_-$ or foreground $c_+$; for more information, see Bouwmans et al. (2019); Bouwmans & Garcia-Garcia (2019a;b)) measured on the CDnet video "blizzard" Goyette et al. (2012); Wang et al. (2014). They correspond, respectively, to $\pi_+ \simeq 0.124$ and $\pi_+ \simeq 0.012$. The comparison with fig. 3c shows that the optimal tradeoffs are close to those predicted by the uniform distribution of performance, close to the oracle ($\Pi_C$).

## 5 Conclusion

Two-class classification is, conceptually, a simple task. However, the evaluation of the classifiers, as well as the analysis and comparison of their performances, remains relatively poorly understood. We have presented theoretical contributions that address the question: "Which scores correspond to the optimal tradeoff between the rankings induced by two arbitrarily chosen scores?" We did so in the specific but very common case of two-class classification with fixed priors, deliberately restricting ourselves to the family of ranking scores known to yield meaningful rankings (*i.e.*, rankings interpretable from a performance perspective). We first presented the theory, then presented the results of three experiments demonstrating the practical applicability of the proposed technique.

## AI USE STATEMENT

In this work, we used generative AI tools to assist with translation. We have not used generative AI tools to help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, and interpret results. Generate synthetic data sets, clean and reformat dataset, support qualitative and thematic data analysis are not applicable to this work. Additionally, we used generative AI tools to identify relevant literature. We have reviewed all AI-assisted work and read the suggested papers. We take responsibility for the final content of this work.

## REFERENCES


Ricardo Barandela, Josep Salvador Sánchez, Vicente García, and Erick Rangel. Strategies for learning in class imbalance problems. *Pattern Recognition*, 36(3):849–851, March 2003. doi: 10.1016/S0031-3203(02)00257-1. URL https://doi.org/10.1016/S0031-3203(02)00257-1.

Eugenio Beltrami. Teoria fondamentale degli spazii di curvatura costante. *Annali di Matematica Pura ed Applicata*, 2(1):232–255, August 1868. doi: 10.1007/bf02419615. URL https://doi.org/10.1007/BF02419615.

Thierry Bouwmans and Belmar Garcia-Garcia. Visual surveillance of human activities: Background subtraction challenges and methods. In *From Visual Surveillance to Internet of Things: Technology and Applications*, pp. 1–24. Taylor & Francis Group, October 2019a. doi: 10.1201/9780429297922-4. URL https://doi.org/10.1201/9780429297922-4.

Thierry Bouwmans and Belmar Garcia-Garcia. Visual surveillance of natural environments: Background subtraction challenges and methods. In *From Visual Surveillance to Internet of Things: Technology and Applications*, pp. 1–20. Taylor & Francis, October 2019b. doi: 10.1201/9780429297922-5. URL https://doi.org/10.1201/9780429297922-5.

Thierry Bouwmans, Sajid Javed, Maryam Sultana, and Soon Ki Jung. Deep neural network concepts for background subtraction:a systematic review and comparative evaluation. *Neural Networks*, 117:8–66, September 2019. doi: 10.1016/j.neunet.2019.04.024. URL https://doi.org/10.1016/j.neunet.2019.04.024.

Thierry Bouwmans, Antonio Greco, Sébastien Piérard, Andrea Vincenzo Ricciardi, Carlo Sansone, Marc Van Droogenbroeck, and Bruno Vento. Illegal waste dumping detection. In *IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, pp. 539–548, Tucson, Arizona, USA, March 2026.

Kay Henning Brodersen, Cheng Soon Ong, Klaas Enno Stephan, and Joachim M. Buhmann. The balanced accuracy and its posterior distribution. In *IEEE International Conference on Pattern Recognition (ICPR)*, pp. 3121–3124, Istanbul, Turkey, August 2010. Institute of Electrical and Electronics Engineers (IEEE). doi: 10.1109/icpr.2010.764. URL https://doi.org/10.1109/ICPR.2010.764.

Dariusz Brzezinski, Jerzy Stefanowski, Robert Susmaga, and Izabela Szczch. Tetrahedron: Barycentric measure visualizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 419–422. Springer, 2017.

Dariusz Brzezinski, Jerzy Stefanowski, Robert Susmaga, and Izabela Szczch. Visual-based analysis of classification measures and their properties for class imbalanced problems. *Information Sciences*, 462:242–261, 2018.

Dariusz Brzezinski, Jerzy Stefanowski, Robert Susmaga, and Izabela Szczech. On the dynamics of classification measures for imbalanced and streaming data. *IEEE transactions on neural networks and learning systems*, 31(8):2868–2878, 2019.

Gürol Canbek, Seref Sagiroglu, Tugba Taskaya Temizel, and Nazife Baykal. Binary classification performance measures/metrics: A comprehensive visualized roadmap to gain new insights. In *International Conference on Computer Science and Engineering (UBMK)*, pp. 821–826, Antalya, Turkey, October 2017. IEEE. doi: 10.1109/UBMK.2017.8093539. URL https://doi.org/10.1109/UBMK.2017.8093539.

Davide Chicco and Giuseppe Jurman. The matthews correlation coefficient (mcc) should replace the roc auc as the standard metric for assessing binary classification. *BioData mining*, 16(1):4, 2023.

Davide Chicco, Valery Starovoitov, and Giuseppe Jurman. The benefits of the matthews correlation coefficient (mcc) over the diagnostic odds ratio (dor) in binary classification assessment. *Ieee Access*, 9:47112–47124, 2021.

Peter Christen, David J. Hand, and Nishadi Kirielle. A review of the F-measure: Its history, properties, criticism, and alternatives. *ACM Computing Surveys*, 56(3):1–24, October 2023. doi: 10.1145/3606367. URL https://doi.org/10.1145/3606367.

Lee R. Dice. Measures of the amount of ecologic association between species. *Ecology*, 26:297–302, 1945. doi: 10.2307/1932409. URL https://doi.org/10.2307/1932409.

Ronen Fluss, David Faraggi, and Benjamin Reiser. Estimation of the Youden index and its associated cutoff point. *Biometrical Journal*, 47(4):458–472, August 2005. doi: 10.1002/bimj.200410135. URL https://doi.org/10.1002/bimj.200410135.

Sébastien Foulle. Mathematical characterization of better-than-random multiclass models. *Transactions on Machine Learning Research*, pp. 1–17, June 2025. URL https://openreview.net/forum?id=VdW9SkALSd.

Edward B. Fowlkes and Colin L. Mallows. A method for comparing two hierarchical clusterings. *Journal of the American Statistical Association*, 78(383):553–569, September 1983. doi: 10.1080/01621459.1983.10478008. URL https://doi.org/10.1080/01621459.1983.10478008.

Nil Goyette, Pierre-Marc Jodoin, Fatih Porikli, Janusz Konrad, and Prakash Ishwar. Changedetection.net: A new change detection benchmark dataset. In *IEEE International Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1–8, Providence, Rhode Island, USA, June 2012. IEEE. doi: 10.1109/CVPRW.2012.6238919. URL https://doi.org/10.1109/CVPRW.2012.6238919.

Richárd Grünwald and Zsolt Páles. On the equality problem of generalized Bajraktarević means. *Aequationes Mathematicae*, 94(4):651–677, August 2019. doi: 10.1007/s00010-019-00670-9. URL https://doi.org/10.1007/s00010-019-00670-9.

Xinjian Guo, Yilong Yin, Cailing Dong, Gongping Yang, and Guangtong Zhou. On the class imbalance problem. In *International Conference on Natural Computation*, pp. 192–201, Jinan, China, October 2008. Institute of Electrical and Electronics Engineers (IEEE). doi: 10.1109/icnc.2008.871. URL https://doi.org/10.1109/ICNC.2008.871.

Anaïs Halin, Sébastien Piérard, Anthony Cioppa, and Marc Van Droogenbroeck. A hitchhiker's guide to understanding performances of two-class classifiers. *arXiv*, abs/2412.04377, 2024. doi: 10.48550/arXiv.2412.04377. URL https://doi.org/10.48550/arXiv.2412.04377.

Paul Jaccard. Nouvelles recherches sur la distribution florale. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 44(163):223–270, 1908.

Hermann Karcher. Riemannian center of mass and mollifier smoothing. *Communications on Pure and Applied Mathematics*, 30(5):509–541, September 1977. doi: 10.1002/cpa.3160300502. URL https://doi.org/10.1002/cpa.3160300502.

John G. Kemeny. Mathematics without numbers. *Daedalus*, 88(4):577–591, 1959. URL http://www.jstor.org/stable/20026529.

Maurice George Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81, June 1938. doi: 10.2307/2332226. URL https://doi.org/10.2307/2332226.

Andrey Nikolaevich Kolmogorov. Sur la notion de la moyenne. *Atti della Academia Nazionale dei Lincei*, 12(9):388–391, 1930.

Miroslav Kubat. Addressing the curse of imbalanced training sets: one-sided selection. In *International Conference on Machine Learning (ICML)*, pp. 179–186. Morgan Kaufmann, 1997.

Stanisław Kulczyński. Die pflanzenassoziationen der pieninen. *Bulletin International de l'Academie Polonaise des Sciences et des Lettres, Classe des Sciences Mathematiques et Naturelles, B (Sciences Naturelles)*, II:57–203, 1927.

David Lovell, Dimity Miller, Jaiden Capra, and Andrew P. Bradley. Never mind the metrics – what about the uncertainty? visualising binary confusion matrix metric distributions to put performance in perspective. In *International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 22702–22757, 2023. URL https://proceedings.mlr.press/v202/lovell23a.html.

MMSegmentation Contributors. MMSegmentation: OpenMMLab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation, 2020.

Charles S. Peirce. The numerical measure of the success of predictions. *Science*, 4(93):453–454, November 1884. doi: 10.1126/science.ns-4.93.453.b. URL http://www.jstor.org/stable/1760565.

Sébastien Piérard, Anaïs Halin, Anthony Cioppa, Adrien Deliège, and Marc Van Droogenbroeck. The Tile: A 2D map of ranking scores for two-class classification. *arXiv*, abs/2412.04309, 2024. doi: 10.48550/arXiv.2412.04309. URL https://doi.org/10.48550/arXiv.2412.04309.

Sébastien Piérard, Adrien Deliège, Anaïs Halin, and Marc Van Droogenbroeck. A methodology to evaluate strategies predicting rankings on unseen domains. In *IEEE International Conference on Multimedia and Expo Workshops (ICMEW), Workshop on Big Surveillance Data Analysis and Processing (BIG-Surv)*, pp. 1–6, Nantes, France, June 2025a. IEEE. doi: 10.1109/icmew68306.2025.11152102. URL https://doi.org/10.1109/ICMEW68306.2025.11152102.

Sébastien Piérard, Adrien Deliège, and Marc Van Droogenbroeck. Multi-domain performance analysis with scores tailored to user preferences. *arXiv*, abs/2512.08715, 2025b. doi: 10.48550/arXiv.2512.08715. URL https://doi.org/10.48550/arXiv.2512.08715.

Sébastien Piérard, Anaïs Halin, Anthony Cioppa, Adrien Deliège, and Marc Van Droogenbroeck. Foundations of the theory of performance-based ranking. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14293–14302, Nashville, Tennessee, USA, June 2025c. IEEE. doi: 10.1109/cvpr52734.2025.01333. URL https://doi.org/10.1109/CVPR52734.2025.01333.

Sébastien Piérard, Anaïs Halin, François Marelli, Simon Pernas, and Jérôme Pierre. Sorbetto: a Python library for producing classification tiles with different flavors to visualize, analyze, and compare performances in two-class classification problems. *Zenodo*, September 2025d. doi: 10.5281/zenodo.17591788. URL https://doi.org/10.5281/zenodo.17591788.

Sébastien Piérard, Adrien Deliège, and Marc Van Droogenbroeck. What is the optimal ranking score between precision and recall? we can always find it and it is rarely $F_1$. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Denver, Colorado, USA, June 2026. IEEE.

David M. W. Powers. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv*, abs/2010.16061, 2020. doi: 10.48550/arXiv.2010.16061. URL https://doi.org/10.48550/arXiv.2010.16061.

Cornelis Joost van Rijsbergen. *Information Retrieval*. Butterworths, London, England, second edition, 1979.

Yi Wang, Pierre-Marc Jodoin, Fatih Porikli, Janusz Konrad, Yannick Benezeth, and Prakash Ishwar. CDnet 2014: An expanded change detection benchmark dataset. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 393–400, Columbus, Ohio, USA, June 2014. Institute of Electrical and Electronics Engineers (IEEE). doi: 10.1109/cvprw.2014.126. URL https://doi.org/10.1109/CVPRW.2014.126.

Daniel S. Wilks. *Statistical methods in the atmospheric sciences*. Elsevier, fourth edition, 2020. doi: 10.1016/C2017-0-03921-6. URL http://dx.doi.org/10.1016/C2017-0-03921-6.

William John Youden. Index for rating diagnostic tests. *Cancer*, 3(1):32–35, January 1950. doi: 10.1002/1097-0142(1950)3:1⟨32::aid-cncr2820030106⟩3.0.co;2-3. URL https://doi.org/10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3.

Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. BDD100K: A diverse driving dataset for heterogeneous multitask learning. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2633–2642, Seattle, Washington, USA, June 2020. IEEE. doi: 10.1109/cvpr42600.2020.00271. URL https://doi.org/10.1109/CVPR42600.2020.00271.

Anastasia Zakharova, Thierry Bouwmans, Anthony Cioppa, Adrien Deliège, Antonio Greco, Anaïs Halin, Kamil Jeziorek, Meghna Kapoor, Tomasz Kryjak, Islam Osman, Sébastien Piérard, Carlo Sansone, Mohamed S. Shehata, Renaud Vandeghen, Marc Van Droogenbroeck, and Bruno Vento. Event detection in videos: A framework for the development of new methods. *arXiv*, abs/2607.04372, 2026. doi: 10.48550/arXiv.2607.04372. URL https://doi.org/10.48550/arXiv.2607.04372.