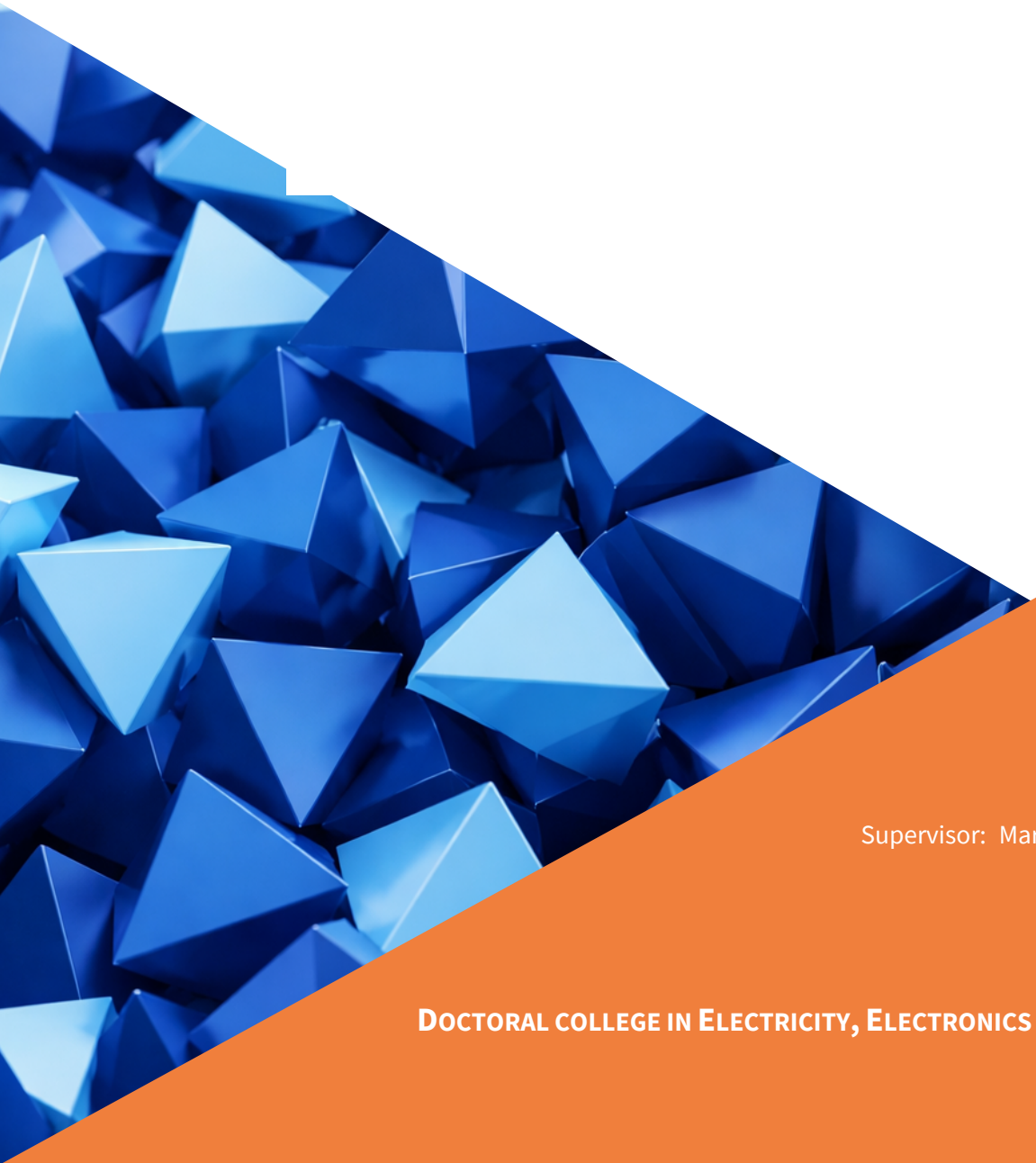
LIÈGE université
Sciences Appliquées

Probabilistic Theory of Performances, Rankings, and Domains, for the Evaluation, Comparison, and Adaptation of Methods

A thesis submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy (PhD) in Engineering Science

by

Sébastien PIÉRARD



Supervisor: Marc VAN DROOGENBROECK

DOCTORAL COLLEGE IN ELECTRICITY, ELECTRONICS AND COMPUTER SCIENCE

FEBRUARY 2026

About the front page illustration. The picture on the cover has been generated in a few seconds by a well-known "artificial intelligence" (AI) algorithm based on the prompt "I need an image of size 2,285 × 1,521 showing lots of tetrahedrons, in shades of blue, from light to dark."

- *What degree of satisfaction do we have with the result?* Clearly, the algorithm understood that it had to generate an image, and it did it fast. The requested colors have been respected, as has the request to draw a lot of polyhedra. However, if we look closely, we realize that the polyhedra drawn are not tetrahedra. We are not completely dissatisfied, but we are certainly not completely satisfied either. Specifying the degree of **satisfaction** that one has with each possible case that can arise comes to specify a task.

- *How important is this particular case for us?* As far as I am concerned, since I needed an image for the cover, I consider the importance as quite high, even though the errors made do not lead to any serious consequences. Another person might consider this case to be completely unimportant, or, on the contrary, extremely important. Specifying the degree of **importance** that one gives to each possible case that can arise comes to specify some application-related or user-related preferences.

Welcome to the fascinating world of performance and evaluation!

Jury Members

By alphabetical order:

Pr. Bernard Boigelot, president
Department of Electrical Engineering and Computer Science,
University of Liège (ULiège),
Liège, Belgium

Pr. Thierry Bouwmans
Laboratoire *Mathématiques Image Applications* (MIA),
La Rochelle Université,
La Rochelle, France

Dr. Adrien Deliège
Vision and Information Understanding Laboratory (VIULaB),
Department of Electrical Engineering and Computer Science,
University of Liège (ULiège),
Liège, Belgium

Pr. Pierre Geurts
Department of Electrical Engineering and Computer Science,
University of Liège (ULiège),
Liège, Belgium

Pr. Benoît Macq
Université catholique de Louvain (UCLouvain),
Louvain-la-Neuve, Belgium

Pr. Marc Van Droogenbroeck, advisor
Vision and Information Understanding Laboratory (VIULaB),
Department of Electrical Engineering and Computer Science,
University of Liège (ULiège),
Liège, Belgium

Abstract

This thesis addresses the reliability and trustworthiness of entities such as algorithms, devices, methods, models, procedures, systems, etc. Through nine articles, it presents a probabilistic approach (i) to evaluate entities and define their performances, (ii) to compare entities based on their performances and explore the diversity of meaningful rankings, and (iii) to adapt entities to applicative domains and characterize these domains from a physical point of view. A particular focus is placed on classification and segmentation tasks for computer vision applications. In this case, the entities are classifiers, change detection algorithms, or semantic segmentation methods, and the domains are represented by sets of images, videos, or categories of videos.

I begin by establishing the need to consider the performance as a mathematical object rather than as a list of numbers ("measures" or score values). I show it in the case of pixel-based two-class classification applied to videos. The expected relationships between scores (*e.g.*, F_1 is the harmonic mean of the precision and recall) are broken when one averages, score by score, the values obtained on different videos. In other words, averaging score values for different domains leads to inconsistent values. This problem is solved by viewing the performance as a mathematical object, defining a way to average these objects through a convex combination (the *summarization*), considering scores as real functions of performance, and calculating the values taken by the scores on the average performance rather than the averages of the values taken by the scores on all performances.

Then, I develop the first general theory of performance-based ranking. I introduce a rigorous framework built on top of both the probability and order theories. Our framework encompasses the elements necessary to (i) manipulate performances as mathematical objects, (ii) express which performances are worse than or equivalent to others, (iii) model tasks through a variable called *satisfaction*, (iv) consider properties of the evaluation, (v) define scores, and (vi) specify application-specific preferences through a variable called *importance*. On top of this framework, I propose an axiomatic definition of performance orderings and performance-based rankings. Then, I introduce a universal parametric family of scores, called *ranking scores*, that can be used to establish rankings satisfying our axioms while considering application-specific preferences. Finally, I show, in the case of two-class classification, that the family of ranking scores encompasses well-known performance scores, including the accuracy, the true positive rate (recall, sensitivity), the true negative rate (specificity), the positive predictive value (precision), and the F-score F_1. In addition, I also show that some other common scores are unsuitable to derive performance orderings satisfying the axioms.

Adopting the perspective of rankings rather than the one of values allows me to revisit a recurring need in the literature: that of finding a score that is a good trade-off between others. For example, F_1 is often chosen as a trade-off between the precision and the

recall since it is their harmonic mean. Interestingly, I discovered that while this is clearly the case from the perspective of values, from the perspective of rankings, it is not so. In fact, the F_1 score is rarely a good trade-off between precision and recall. In some cases, the ranking induced by F_1 perfectly mimics the ranking induced by one of the two and completely ignores the other. By studying the manifold of performance orderings induced by ranking scores, I am able to define what an optimal trade-off is and to provide the theoretical and algorithmic tools needed to find it for any reference distribution of performance or any finite set of entities to be ranked.

A recurring theme in this thesis is the exploration of the diversity of applicative preferences or the diversity of rankings induced by ranking scores. This exploration allows us to answer questions such as the following ones, in light of applicative preferences. How well is an entity ranked? What are the strengths and weaknesses of a given entity? For which applications do we have entities with sufficient performance? How different are the meaningful performance-based rankings? How easy or difficult is a domain to manage for a given entity? How well can we estimate or predict a ranking in a new domain? etc.

When particularized to two-class crisp classification, our theory leads to innovative practical tools: the Tile and its *flavors*. It turns out that, in this case, the performance orderings induced by the ranking scores form a 2D manifold that can be mapped on a 2D square (the Tile). Its axes can be interpreted as the relative importance between satisfying cases (true positive *vs* true negative) and between unsatisfying cases (false negative *vs* false positive). This space can be used to depict different types of information (the flavors), such as the best entity or the rank of a given entity *w.r.t.* the application- or user-related preferences. Rank correlations can also be depicted on the Tile.

It is just as useful to compare a set of entities based on their performances in a specific domain as it is to compare the performances of a specific entity across a set of domains. For this reason, I propose considering, in addition to the easiest and most difficult domains, the preponderant and bottleneck domains, the latter two being rigorously defined through the summarization.

Finally, three other works more focused on the domains are presented. In the first one, I present a methodology to evaluate strategies predicting rankings on previously unseen domains, based on Tiles. In the second one, I describe a method called *Mixture Domain Adaptation* (MDA) that can be used to dynamically adapt a classifier (*e.g.*, semantic segmentation method) to a target domain in which the distribution of cases that can arise is a mixture (convex combination) of the distribution in source domains. To end with, I describe techniques to characterize domains in a physically interpretable way, using deep learning methods applied on video streams. These techniques can be coupled with MDA.

Overall, this thesis makes it clear that a probabilistic approach is appropriate for defining and manipulating performances, rankings, and domains. These needs arise, respectively, when one wants to evaluate, compare, or adapt entities.

Acknowledgments

Acknowledgments will be added later.

Contents

Introduction 11

1. Averaging performances (1st article) 17

2. Performance-based rankings (2nd article) 25

3. Intermediate rankings (3rd article) 43

4. The Tile (4th article) 57

5. The flavors (5th article) 71

6. Multi-domain performance analysis (6th article) 85

7. Predicting rankings (7th article) 93

8. Mixture domain adaptation (8th article) 103

9. Domain characterization (9th article) 119

Conclusion 141

A. Supplementary material for the 2nd article 147

B. Supplementary material for the 3rd article 171

C. Supplementary material for the 4th article 247

D. Supplementary material for the 5th article 267

E. Supplementary material for the 7th article 293

F. Supplementary material for the 8th article 307

Bibliography 313

Introduction

In this thesis, I present in Chapters 1 to 9 a selection of 9 articles [1, 2, 3, 4, 5, 6, 7, 8, 9] in which I adopt a ***probabilistic view*** and develop the theory to ***evaluate***, ***compare***, and ***adapt*** entities. For engineers and computer scientists, the entities of interest can be algorithms, devices, methods, models, procedures, strategies, systems, solutions, etc. For the evaluation, the objective is to properly define and manipulate the ***performances*** of entities. For the comparison, the objective is to establish ***rankings*** of entities that are based on their performances and suitable for given application- or user-related preferences. And for the adaptation, the objective is to characterize the ***domains*** in which the entities have to operate and to optimize their behavior *w.r.t.* them. Each article covers one to three notions among those of ***performance*** (Chapters 1 to 8), ***ranking*** (Chapters 2 to 7), and ***domain*** (Chapters 1 and 6 to 9). Figure 1 presents the journey followed in this thesis. The a^{th} article [a] is presented in chapter a, $\forall a \in [1, 9]$.

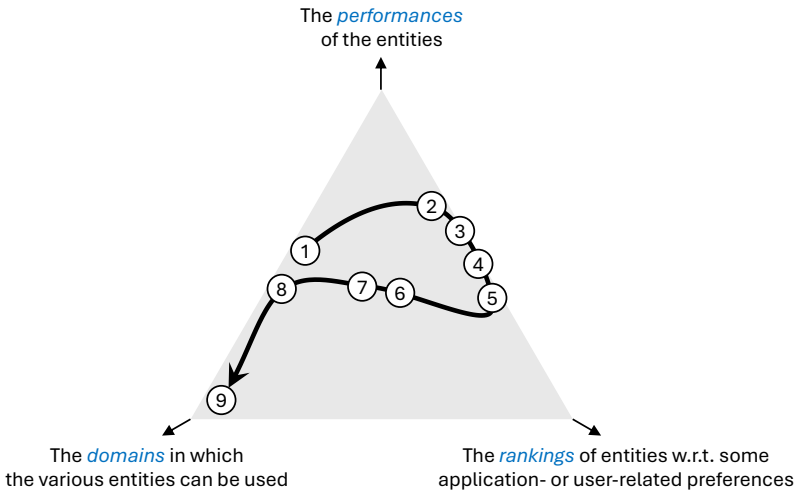


Figure 1.: The journey followed in this thesis. The numbers correspond to the nine presented articles. Their positions indicate the topics covered in each of them. The group [1, 8] in the middle of the left border is concerned with both the performances and the domains. The group [2, 3, 4, 5] in the middle of the right border is concerned with both the performances and the rankings. The central group [6, 7] is about the performances, the domains, and the rankings. Finally, the article in the bottom-left corner [9] is focused on the domains.

The nine articles are reproduced in this thesis exactly as they appear in their original version, except for the numbering of the pages and the bibliographical references. This

numbering has been standardized throughout the thesis. I have kept a bibliography at the end of each paper and added a comprehensive bibliography at the end of the thesis. As all the articles are based on probability theory, the notations are very similar from one article to another. Furthermore, the articles in Chapters 2 to 7 and those in Chapters 8 and 9 use the same notations.

Performance, a universal concept

Every day, millions of people either develop and optimize entities with respect to the expected performance, or establish rankings and select entities based on their *performance*. In engineering, entities of interest are called algorithms, devices, methods, models, procedures, strategies, systems, solutions, etc. In other disciplines, they may be people (athletes, students, workers, etc.) or groups of people (sports teams, etc.).

In computer science, there are two types of computation systems (entities):

- On the one hand, there are those guaranteed to be correct by design, obtained with classical algorithmic reasoning. For example, if one needs to sort an array of numbers, we know how to do it. It is generally agreed that the performance of this kind of systems is related to the amount of resources used (time, memory, etc.), these amounts varying from one input to another. The more resources we consume, the lower our satisfaction is.

- On the other hand, there are systems with no guarantee of correctness. As a first example, heuristics are used when correct algorithms are too costly in terms of resources. As a second example, there exist algorithms providing non-deterministic results (*e.g.*, the background subtraction algorithm ViBe [10]). As a third example, models obtained with machine learning are used when programmers are unable to convert their human expertise into an algorithm[1]. For over a dozen years, deep learning techniques have propelled artificial intelligence, revolutionizing computer vision and natural language processing. It is generally agreed that the performance of systems without guarantees is primarily a comparison between the computation results and the expected results, although the amount of resources used remains an important factor to consider. The closer the computation result and the expected result are, the higher our satisfaction is.

The examples that I develop in this thesis are all for systems that do not come with any guarantee of correctness. I choose to illustrate my developments on two common computer vision tasks: the *background subtraction* and the *semantic segmentation*. These tasks can be seen as some pixel-wise *classification*. In these illustrations, the entities are background subtraction *algorithms*, segmentation *models*, or *classifiers*. But, from my point of view, the name given to them is not relevant. I consider that developing specific

[1] Sometimes, humans can execute a task without being able to explain precisely how they do it. For example, it is easy for us to tell whether an image contains humans or not. But if we are asked to explain algorithmically how we do this, nobody can answer.

	Dominant approach	My probabilistic approach
task	given by a vague and verbose description or a loss function	modeled by a random variable (satisfaction)
evaluation	some tests in which a few quantities are "measured" or "counted"	a (thought) random experiment
performance	a finite list of values	a probability measure
comparison	based on values for arbitrarily chosen scores	based on probability measures
ranking	the applicative preferences are implicitly given by the score	the applicative preferences are explicitly given with a random variable (importance)
adaptation	without knowing the relationship between the sources and the target	based on the relationship (mixture) between the sources and the target
domain	implicitly represented by a finite dataset	either a probability measure or some non-deterministic source of data used in random experiments

Table 1.: Comparison between the dominant approach to evaluate, compare, and adapt entities, and my probabilistic approach.

theories for each type of task, or each name given to the entities, is like reinventing the wheel. In fact, I discovered that it is possible to develop a general theory for the evaluation, comparison, and optimization of entities.

A few words about my approach

Throughout this thesis, I adopt a probabilistic approach. While the notions of *task*, *evaluation*, *performance*, *comparison*, *ranking*, *adaptation*, and *domain* are common in the literature, adopting a probabilistic view for all of them is an original approach. This reformulation, detailed in Table 1, makes a big semantic difference and allows rigorous theory to be developed.

I am aiming for a mathematically rigorous theory. For its generic aspect, I base my work on the probability theory as well as on the order theory, which are themselves based on set theory. My approach is quantitative. Performances and domains are given by probability measures. Some application- or user-related preferences can be given in the form of a random variable called importance. In addition, the task is also modeled by a random variable that I call satisfaction. Once tasks, performances, applicative preferences, and domains are specified or modeled numerically, the level of abstraction becomes high, and one can develop a generic theory.

Why do I study the performances, the rankings, and the domains all together?

In a basic scenario, we start from a specific application. We begin by determining the domain we are working in (often by collecting a database of samples). Then, we determine the preferences related to the application or to its users in terms of performance to be achieved (for example, to what extent do we prefer false positives to false negatives in classification?), or we choose arbitrarily a score that we would like to maximize. Finally, we train a model or develop a tailor-made algorithm for this applicative domain and these applicative preferences.

Unfortunately, this basic scenario is sometimes very far from the real-world scenario. The people developing entities are certainly familiar with the task at hand but often know neither the applicative domain nor the applicative preferences. Preferences may depend on the users, who are unknown in advance, and the applicative domain may change over time even in a given application. In addition, it is very useful to be able to "recycle" entities previously developed for one application for other applications.

The challenge is therefore to perform performance analyses that take into account the uncertainty about both the applicative preferences and the applicative domains. In this thesis, I show how it can be done.

A little catalog of research questions

This field of research is very rich, and many interesting questions arise. Here are a few of them that I chose to explore, together with the pointers to the chapters in which I provide the answers.

About the performances and the evaluation of entities

Research question 1 *How can we mathematically define the notions of performance and score? The answer can be found in Chapters 1 and 2.*

Research question 2 *How can we interpret the scores that have been defined in the past in terms of importances and determine if they are suitable for ranking? The answer can be found in Chapters 2 to 4.*

Research question 3 *How can we handle any task, and in particular those apart from a "simple" classification? The answer can be found in Chapter 2.*

About the rankings and the comparison of entities

Research question 4 *How can we obtain a generic approach integrating the performances, the tasks, and the preferences? The answer can be found in Chapter 2.*

Research question 5 *How can we explore the diversity of meaningful rankings? The answer can be found in Chapters 2 to 5.*

Research question 6 *How can we tune the rankings with respect to the application- or user-related preferences? The answer can be found in Chapter 2.*

About the domains and the adaptation of entities

Research question 7 *How can we average the performances or scores that correspond to different domains? The answer can be found in Chapters 1 and 6.*

Research question 8 *How can we estimate the ranking of entities in a new domain? The answer can be found in Chapter 7.*

Research question 9 *How can we perceive the characteristics of the domain in which we are, and optimize the performance "on the fly"? The answer can be found in Chapters 8 and 9.*

1. Averaging performances (1st article)

This chapter reproduces the following article [1]:

> S. Piérard and M. Van Droogenbroeck, "Summarizing the performances of a background subtraction algorithm measured on several videos," in *IEEE International Conference on Image Processing (ICIP)*, (Abu Dhabi, United Arab Emirates), pp. 3234–3238, October 2020.

It covers both the notions of ***performance*** and ***domain***.

Context

For a given task, there is often a wide range of applications that differ by the domain. It is unrealistic to believe that, in practice, it would be possible to develop and optimize a customized entity for each applicative domain. There is therefore an enormous interest in developing versatile entities. In this scenario, it is reasonable to consider that there exist two possible objectives: loosely speaking, one may want to "maximize the worst-case performance" or one may want to "maximize the average performance". The article I present here relates to the latter (the former will be discussed in Chapter 6). While the underlying principles are generic, I decided to present them for a particular computer vision task known as background subtraction, foreground detection, or change detection, which can be seen as a two-class crisp classification task occurring in each pixel and in each frame of videos.

Motivation

The current issue is as follows. In order to get a global view of an entity's performance, the standard procedure in the literature consists in averaging performance scores or measures over some given set of domains, datasets, videos, etc. People choose these scores arbitrarily, and the choice differs from one study to another. Unfortunately, this results in inconsistent conclusions. The expected relationships between scores (*e.g.*, the F-score or F-measure F_1 is the harmonic mean of the precision and recall) are broken when one averages, score by score, the values (*a.k.a.* "measures") obtained on different domains. For example, in classification, users of the Receiver Operating Characteristic (ROC) space naturally tend to average the false positive and true positive rates, while users of the Precision/Recall (PR) space naturally tend to average the precision and the recall. The two results do not coincide in general. Because of these obvious biases, the current analyses cannot be compared and suffer from interoperability issues.

It turns out that the expected relationships between scores make it possible to deduce the value of any score once the values of a few are known. In the case of crisp classification with two classes, except in pathological cases, it is sufficient to know the values of three scores. This leads to the following idea. Consider any set of three scores that allows all the others to be found. If all possible triplets of values form a convex set, then an arithmetic mean can be used on these three scores. However, as a general rule, the probabilistic interpretation of probabilistic scores is broken by such averaging. This idea is therefore not recommended. We need something better.

Main contributions

I introduce a better idea for summarizing score values among diverse domains, which I call "summarization". After making an explicit choice for a first random evaluation experiment relative to any given domain and for a second random evaluation experiment specific to a set of domains, I give expressions for multi-domain score values based on domain-specific score values. It turns out that, for any unconditional probabilistic score, the multi-domain value corresponds to the arithmetic mean of the domain-specific values. For conditional probabilistic scores also, but the weights to be given to the values are not those that one would instinctively expect. Moreover, there exist scores for which the multi-domain value does not correspond to any average. It is a scientific fact, no matter if this is pleasant or not, it's like that.

Perspectives

The findings reported in this article pave the way to what I promote in Chapter 2: (1) precisely defining the thought random experiments used to evaluate the entities, (2) considering the performance as a mathematical object rather than just some list of numbers ("measures" or score values), (3) considering that the performance has a probabilistic meaning, (4) properly defining the summarization operation on these objects, and (5) defining scores as real functions of performances.

Remarks

1. Using the mathematical framework and notations that will be introduced in Chapter 2, considering a measurable space (Ω, Σ) and the performances (probability measures) $P_1, \ldots, P_n$ with respective weights $\lambda_1, \ldots, \lambda_n$, the summarized performance $\overline{P}$ is
$$\overline{P} : \Sigma \to [0,1] : E \mapsto \overline{P}(E) = \frac{\sum_{d=1}^{n} \lambda_d P_d(E)}{\sum_{d=1}^{n} \lambda_d}\,, \tag{1.1}$$
which can also be written, in short,
$$\overline{P} = \sum_{d=1}^{n} \frac{\lambda_d}{\sum_{d'=1}^{n} \lambda_{d'}} P_d\,. \tag{1.2}$$

2. The summarization is not a mean in the sense of Kolmogorov [11] as a performance is not a real number. However, it is a mean in the sense of Fréchet [12] and Karcher [13] for a particular distance between performances

$$d : \mathbb{P}^2_{(\Omega,\Sigma)} \to \mathbb{R}_{\geq 0} : (P_1; P_2) \mapsto \sqrt{\sum_{\omega \in \Omega} \left(P_1(\{\omega\}) - P_2(\{\omega\})\right)^2}. \tag{1.3}$$

   Note that the Fréchet variance could be used to indicate how widely the performances are dispersed.

3. Having a distance between performances and an average is particularly interesting. This allows us to structure the set $\mathbb{P}_{(\Omega,\Sigma)}$ of all possible performances. It is not just a unstructured set, but a well-structured space. With four samples, $\mathbb{P}_{(\Omega,\Sigma)}$ is a regular tetrahedron.

4. This article does not introduce the notion of average or intermediate domain. This is done in Chapters 8 and 9. When the behavior of the entity does not change from one domain to another (*i.e.*, there is no adaptation), the summarized performance is the performance of the entity in that intermediate domain.

5. Sometimes, several evaluation datasets are available. Suppose that the random evaluation experiment uses a non-deterministic data source once and only once, which consists of randomly selecting an object from a test set. In this case, the summarized performance is the performance obtained with a combined test set containing objects appropriately weighted.

SUMMARIZING THE PERFORMANCES OF A BACKGROUND SUBTRACTION ALGORITHM MEASURED ON SEVERAL VIDEOS

S. Piérard, and M. Van Droogenbroeck

Department of Electrical Engineering and Computer Science, University of Liège, Belgium
{S.Pierard, M.VanDroogenbroeck}@uliege.be

ABSTRACT

There exist many background subtraction algorithms to detect motion in videos. To help comparing them, datasets with ground-truth data such as CDNET or LASIESTA have been proposed. These datasets organize videos in categories that represent typical challenges for background subtraction. The evaluation procedure promoted by their authors consists in measuring performance indicators for each video separately and to average them hierarchically, within a category first, then between categories, a procedure which we name "summarization". While the summarization by averaging performance indicators is a valuable effort to standardize the evaluation procedure, it has no theoretical justification and it breaks the intrinsic relationships between summarized indicators. This leads to interpretation inconsistencies. In this paper, we present a theoretical approach to summarize the performances for multiple videos that preserves the relationships between performance indicators. In addition, we give formulas and an algorithm to calculate summarized performances. Finally, we showcase our observations on CDNET 2014.

***Index Terms*—** performance summarization, background subtraction, multiple evaluations, CDNET, classification performance

1. INTRODUCTION

A plethora of background subtraction algorithms have been proposed in the literature [14, 15, 16, 17, 18, 19, 20]. They aim at predicting, for each pixel of each frame of a video, whether the pixel belongs to the background, free of moving objects, or to the foreground. Background subtraction algorithms have to operate in various conditions (viewpoint, shadows, lighting conditions, camera jitter, etc). These conditions are covered by different videos in evaluation datasets, such as CDNET [21, 22] or LASIESTA [23].

Measuring the performance on a single video is done at the pixel level, each pixel being associated with a ground-truth class y (either negative c^- for the background, or positive c^+ for the foreground) and an estimated class $\hat{y}$ (the temporal and spatial dependences between pixels are ignored). Performance indicators adopted in the field of background subtraction are mainly those used for two-class crisp classifiers (precision, recall, sensitivity, F-score, error rate, etc).

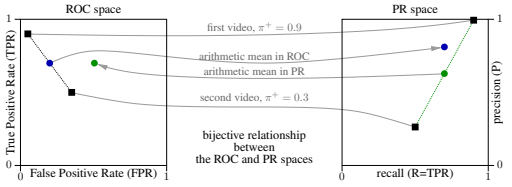


Fig. 1: How can we summarize the performances of a background subtraction algorithm, measured on several videos? This figure shows that arithmetically averaging the performances breaks the bijective relationship that exists between the ROC and Precision-Recall (PR) spaces [24]. This paper presents a new summarization technique that guarantees the coherence between any sets of performance indicators.

To compare the behavior of algorithms, it is helpful to "*summarize*" all the performance indicators that are originally measured on the individual videos. However, to the best of our knowledge, there is no theory for this summarization process, although an attempt to standardize it has been promoted with CDNET. In CDNET, videos are grouped into 11 categories, all videos in a given category having the same importance, and all categories having also the same importance. In other words, the videos are weighted regardless of their size. The standardized summarization process of CDNET is performed hierarchically by computing an arithmetic mean, first between the videos in a category, then between the categories. This averaging is done for seven performance indicators, one by one, independently. While the procedure of CDNET is valuable, it has two major drawbacks related to the interpretability of the summarized values.

First, the procedure breaks the intrinsic relationships between performance indicators. The case of the M4CD algorithm, as evaluated on CDNET, topically illustrates this inconsistency. Despite that the F-score is known to be the harmonic mean of precision and recall, the arithmetic mean of the F-scores across all videos is 0.69 while the harmonic mean between the arithmetically averaged precisions and re-

calls is 0.75. The difficulty to summarize F-scores has already been discussed in [25], when averaging between folds in cross-validation. Such a problem also occurs for other indicators. Moreover, Figure 1 shows that the summarization with the arithmetic mean breaks the bijective relationship between the ROC and PR spaces [24]. It is not equivalent to summarize the performances in one space or the other. In particular, one could obtain a meaningless arithmetically averaged performance point, located in the unachievable part of the PR space [26] (the achievable part is not convex). This leads to difficulties for the interpretation and, eventually, to contradictions between published summarized results.

Second, some indicators (such as the precision, recall, sensitivity, or error rate) have a probabilistic meaning [27]. Arithmetically averaging these indicators does not lead to a value that preserves the probabilistic meaning. This leads to interpretability issues. Strictly speaking, one cannot think in terms of probabilities unless a random experiment is defined very precisely, as shown by Bertrand's paradox [28].

This paper presents a better summarization procedure that avoids these interpretability issues. In Section 2, we present indicators to measure the performance of an algorithm applied on a single video and pose the random experiment that underpins the probabilistic meaning of these indicators. Then, in Section 3, we generalize this random experiment for several videos, and show that the resulting performance indicators can be computed based only on the indicators obtained separately for each video. In other words, we have established a theoretical model for summarization that guarantees a consistency between the indicators resulting from both random experiments. In Section 4, we showcase our new summarization procedure on CDNET and discuss how it affects the ranking between algorithms. Section 5 concludes the paper.

2. PERFORMANCE INDICATORS FOR ONE VIDEO

The performance of a background subtraction algorithm on a video is measured by running it once (even for non-deterministic algorithms), and counting at the pixel level the amounts of true negatives TN ($y = \hat{y} = c^-$), false positives FP ($y = c^-$ and $\hat{y} = c^+$), false negatives FN ($y = c^+$ and $\hat{y} = c^-$), and true positives TP ($y = \hat{y} = c^+$).

The performance indicators are then derived from these amounts. For example, the prior of the positive class is $\pi^+ = \frac{\mathsf{FN}+\mathsf{TP}}{\mathsf{TN}+\mathsf{FP}+\mathsf{FN}+\mathsf{TP}}$, the rate of positive predictions is $\tau^+ = \frac{\mathsf{FP}+\mathsf{TP}}{\mathsf{TN}+\mathsf{FP}+\mathsf{FN}+\mathsf{TP}}$, the error rate is $\mathrm{ER} = \frac{\mathsf{FP}+\mathsf{FN}}{\mathsf{TN}+\mathsf{FP}+\mathsf{FN}+\mathsf{TP}}$, the true negative rate (specificity) is $\mathrm{TNR} = \frac{\mathsf{TN}}{\mathsf{TN}+\mathsf{FP}}$, the false positive rate is $\mathrm{FPR} = \frac{\mathsf{FP}}{\mathsf{TN}+\mathsf{FP}}$, the false negative rate is $\mathrm{FNR} = \frac{\mathsf{FN}}{\mathsf{FN}+\mathsf{TP}}$, the true positive rate (recall) is $\mathrm{TPR} = \mathrm{R} = \frac{\mathsf{TP}}{\mathsf{FN}+\mathsf{TP}}$, the positive predictive value (precision) is $\mathrm{PPV} = \mathrm{P} = \frac{\mathsf{TP}}{\mathsf{FP}+\mathsf{TP}}$, and the F-score is $\mathrm{F} = \frac{2\mathsf{TP}}{\mathsf{FP}+\mathsf{FN}+2\mathsf{TP}}$.

The $(\mathrm{FPR}, \mathrm{TPR})$ coordinates define the well-known Receiver Operating Characteristic (ROC) space [29, 30]. The precision P and recall R define the PR evaluation space [26], which is an alternative to the ROC space, as there exists a bijection between these spaces, for a given prior π^+ [24]. In fact, this bijection is just a particular case of the relationships that exist between the various indicators. There are other famous relationships. For example, the F-score is known to be the harmonic mean of the precision P and recall R.

Despite its importance for the interpretation, it is often overlooked that some indicators have a probabilistic meaning [27]. Let us consider the following random experiment.

Random Experiment 1 (for one video) *Draw one pixel at random (all pixels being equally likely) from the video and observe the ground-truth class Y as well as the predicted class $\hat{Y}$ for this pixel. The result of the random experiment is the pair $\Delta = (Y, \hat{Y})$.*

We use capital letters for Y, $\hat{Y}$, and Δ to emphasize the random nature of these variables. The outcome of this random experiment can be associated with a true negative $\mathrm{tn} = (c^-, c^-)$, a false positive $\mathrm{fp} = (c^-, c^+)$, a false negative $\mathrm{fn} = (c^+, c^-)$, or a true positive $\mathrm{tp} = (c^+, c^+)$. The family of *probabilistic indicators* can be defined based on this random experiment:

$$P\left(\Delta \in \mathcal{A} | \Delta \in \mathcal{B}\right) \text{ with } \emptyset \subsetneq \mathcal{A} \subsetneq \mathcal{B} \subseteq \{\mathrm{tn}, \mathrm{fp}, \mathrm{fn}, \mathrm{tp}\} \quad (1)$$

It includes $\pi^+ = P\left(\Delta \in \{\mathrm{fn}, \mathrm{tp}\}\right)$, $\tau^+ = P\left(\Delta \in \{\mathrm{fp}, \mathrm{tp}\}\right)$, $\mathrm{ER} = P\left(\Delta \in \{\mathrm{fp}, \mathrm{fn}\}\right)$, $\mathrm{TNR} = P\left(\Delta = \mathrm{tn} | \Delta \in \{\mathrm{tn}, \mathrm{fp}\}\right)$, $\mathrm{TPR} = P\left(\Delta = \mathrm{tp} | \Delta \in \{\mathrm{fn}, \mathrm{tp}\}\right)$, and $\mathrm{PPV} = P\left(\Delta = \mathrm{tp} | \Delta \in \{\mathrm{fp}, \mathrm{tp}\}\right)$. All the other performance indicators (the F-score, balanced accuracy, etc) can be derived from the probabilistic indicators.

3. SUMMARIZING THE PERFORMANCE INDICATORS FOR SEVERAL VIDEOS

Let us denote a set of videos by $\mathbb{V}$. We generalize the random experiment defined for one video to several videos as follows.

Random Experiment 2 (for several videos) *First, draw one video V at random in the set $\mathbb{V}$, following an arbitrarily chosen distribution $P\left(V\right)$. Then, draw one pixel at random (all pixels being equally likely) from V and observe the ground-truth class Y and the predicted class $\hat{Y}$ for this pixel. The result of the experiment is the pair $\Delta = (Y, \hat{Y})$.*

As the result of this second random experiment is again a pair of classes, as encountered for a single video (first random experiment), all performance indicators can be defined in the same way. This is important for the interpretability, since it guarantees that all indicators are coherent and that the probabilistic indicators preserve a probabilistic meaning.

The distribution of probabilities $P\left(V\right)$ parameterizes the random experiment. It can be arbitrarily chosen to reflect the relative weights given to the videos. For example, one could

use the weights considered in CDNET. An alternative consists in choosing weights proportional to the size of the videos (the number of pixels multiplied by the number of frames). This leads to identical distributions $P(\Delta)$ between the second random experiment and the first one with the fictitiously aggregated video $\bigcup_{v\in\mathbb{V}} v$.

Summarization formulas. We denote by $\mathrm{I}(v)$ the value of an indicator I obtained with our first random experiment on a video $v \in \mathbb{V}$, and by $\mathrm{I}(\mathbb{V})$ the value of I resulting from our second random experiment applied on the set $\mathbb{V}$ of videos.
We claim that the indicators $\mathrm{I}(\mathbb{V})$ summarize the performances corresponding to the various videos, because they can be computed based on some $\mathrm{I}(v)$, exclusively. To prove it, let us consider a probabilistic indicator $\mathrm{I}_{\mathcal{A}|\mathcal{B}}$ defined as $P(\Delta \in \mathcal{A}|\Delta \in \mathcal{B})$, and $\mathrm{I}_{\mathcal{B}}$ as $P(\Delta \in \mathcal{B})$. Thanks to our second random experiment, we have

$$\begin{aligned}
\mathrm{I}_{\mathcal{A}|\mathcal{B}}(\mathbb{V}) &= P(\Delta \in \mathcal{A}|\Delta \in \mathcal{B}) && (2)\\
&= \sum_{v\in\mathbb{V}} P(\Delta \in \mathcal{A}, V = v|\Delta \in \mathcal{B}) && (3)\\
&= \sum_{v\in\mathbb{V}} P(V = v|\Delta \in \mathcal{B})\ P(\Delta \in \mathcal{A}|\Delta \in \mathcal{B}, V = v)\\
&= \sum_{v\in\mathbb{V}} P(V = v|\Delta \in \mathcal{B})\ \mathrm{I}_{\mathcal{A}|\mathcal{B}}(v)\ , && (4)
\end{aligned}$$

with

$$\begin{aligned}
P(V = v|\Delta \in \mathcal{B}) &= \frac{P(V = v)\ P(\Delta \in \mathcal{B}|V = v)}{P(\Delta \in \mathcal{B})} && (5)\\
&= \frac{P(V = v)\ P(\Delta \in \mathcal{B}|V = v)}{\sum_{v'\in\mathbb{V}} P(V = v')\ P(\Delta \in \mathcal{B}|V = v')}\\
&= \frac{P(V = v)\ \mathrm{I}_{\mathcal{B}}(v)}{\sum_{v'\in\mathbb{V}} P(V = v')\ \mathrm{I}_{\mathcal{B}}(v')}\,. && (6)
\end{aligned}$$

Thus, the summarized probabilistic indicator $\mathrm{I}_{\mathcal{A}|\mathcal{B}}(\mathbb{V})$ is a weighted arithmetic mean of the indicators $\{\mathrm{I}_{\mathcal{A}|\mathcal{B}}(v)|v \in \mathbb{V}\}$, the weights being the (normalized) product between the relative importance $P(V = v)$ given to the video v and the corresponding $\mathrm{I}_{\mathcal{B}}(v)$:

$$\mathrm{I}_{\mathcal{A}|\mathcal{B}}(\mathbb{V}) = \sum_{v\in\mathbb{V}} \frac{P(V = v)\ \mathrm{I}_{\mathcal{B}}(v)}{\sum_{v'\in\mathbb{V}} P(V = v')\ \mathrm{I}_{\mathcal{B}}(v')}\,\mathrm{I}_{\mathcal{A}|\mathcal{B}}(v). \quad (7)$$

In the particular case of an unconditional probabilistic indicator $\mathrm{I}_{\mathcal{A}} = \mathrm{I}_{\mathcal{A}|\{\mathrm{tn,fp,fn,tp}\}}$, Equation (7) reduces to

$$\mathrm{I}_{\mathcal{A}}(\mathbb{V}) = \sum_{v\in\mathbb{V}} P(V = v)\ \mathrm{I}_{\mathcal{A}}(v)\,. \quad (8)$$

Note that we are able to summarize probabilistic indicators that are undefined for some, but not all, videos. The reason is that $\mathrm{I}_{\mathcal{A}|\mathcal{B}}(v)$ is undefined only when $\mathrm{I}_{\mathcal{B}}(v) = 0$, and we observe that only the product of these two quantities appears in Equation (7). This product is always well defined since it is an unconditional probability ($\mathrm{I}_{\mathcal{B}}(v)\,\mathrm{I}_{\mathcal{A}|\mathcal{B}}(v) = \mathrm{I}_{\mathcal{A}\cap\mathcal{B}}(v)$). To emphasize it, we rewrite Equation (7) as

$$\begin{aligned}
\mathrm{I}_{\mathcal{A}|\mathcal{B}}(\mathbb{V}) &= \sum_{v\in\mathbb{V}} \frac{P(V = v)\ \mathrm{I}_{\mathcal{B}}(v)\,\mathrm{I}_{\mathcal{A}|\mathcal{B}}(v)}{\mathrm{I}_{\mathcal{B}}(\mathbb{V})}\\
&= \frac{\sum_{v\in\mathbb{V}} P(V = v)\ \mathrm{I}_{\mathcal{A}\cap\mathcal{B}}(v)}{\mathrm{I}_{\mathcal{B}}(\mathbb{V})} = \frac{\mathrm{I}_{\mathcal{A}\cap\mathcal{B}}(\mathbb{V})}{\mathrm{I}_{\mathcal{B}}(\mathbb{V})}\,. \quad (9)
\end{aligned}$$

Applying our summarization in the ROC (FPR, TPR) and PR (R = TPR, P = PPV) spaces. Let us consider the case in which π^+, τ^+, FPR, TPR, and PPV are known for each video. The summarized indicators can be computed by Equations (7)-(8), since $\pi^+ = \mathrm{I}_{\{\mathrm{fn,tp}\}}$, $\tau^+ = \mathrm{I}_{\{\mathrm{fp,tp}\}}$, $\mathrm{FPR} = \mathrm{I}_{\{\mathrm{fp}\}|\{\mathrm{tn,fp}\}}$, $\mathrm{TPR} = \mathrm{I}_{\{\mathrm{tp}\}|\{\mathrm{fn,tp}\}}$, and $\mathrm{PPV} = \mathrm{I}_{\{\mathrm{tp}\}|\{\mathrm{fp,tp}\}}$. This yields:

$$\pi^+(\mathbb{V}) = \sum_{v\in\mathbb{V}} P(V = v)\ \pi^+(v)\,, \quad (10)$$

$$\tau^+(\mathbb{V}) = \sum_{v\in\mathbb{V}} P(V = v)\ \tau^+(v)\,, \quad (11)$$

$$\mathrm{FPR}(\mathbb{V}) = \frac{1}{\pi^-(\mathbb{V})} \sum_{v\in\mathbb{V}} P(V = v)\ \pi^-(v)\,\mathrm{FPR}(v)\,, \quad (12)$$

$$\mathrm{TPR}(\mathbb{V}) = \frac{1}{\pi^+(\mathbb{V})} \sum_{v\in\mathbb{V}} P(V = v)\ \pi^+(v)\,\mathrm{TPR}(v)\,, \quad (13)$$

$$\mathrm{PPV}(\mathbb{V}) = \frac{1}{\tau^+(\mathbb{V})} \sum_{v\in\mathbb{V}} P(V = v)\ \tau^+(v)\,\mathrm{PPV}(v)\,. \quad (14)$$

The passage from ROC to PR is given by:

$$\mathrm{PPV}(\mathbb{V}) = \frac{\pi^+(\mathbb{V})\,\mathrm{TPR}(\mathbb{V})}{\pi^-(\mathbb{V})\,\mathrm{FPR}(\mathbb{V}) + \pi^+(\mathbb{V})\,\mathrm{TPR}(\mathbb{V})}\,. \quad (15)$$

A generic algorithm for the computation of any summarized indicator. Let us assume that the elements of the normalized confusion matrix ($\mathrm{I}_{\{\mathrm{tn}\}}$, $\mathrm{I}_{\{\mathrm{fp}\}}$, $\mathrm{I}_{\{\mathrm{fn}\}}$, and $\mathrm{I}_{\{\mathrm{tp}\}}$) can be retrieved for each video. This could be done by computing them based on other indicators. In this case, an easy-to-code algorithm to summarize the performances is the following:

step 1: arbitrarily weight the videos with $P(V)$,

step 2: retrieve $\mathrm{I}_{\{\mathrm{tn}\}}(v)$, $\mathrm{I}_{\{\mathrm{fp}\}}(v)$, $\mathrm{I}_{\{\mathrm{fn}\}}(v)$, and $\mathrm{I}_{\{\mathrm{tp}\}}(v)$ for each video v,

step 3: then compute $\mathrm{I}_{\{\mathrm{tn}\}}(\mathbb{V})$, $\mathrm{I}_{\{\mathrm{fp}\}}(\mathbb{V})$, $\mathrm{I}_{\{\mathrm{fn}\}}(\mathbb{V})$, and $\mathrm{I}_{\{\mathrm{tp}\}}(\mathbb{V})$ with Equation (8),

step 4: and finally derive all the desired summarized indicators based on their relationships with the $\mathrm{I}_{\{\mathrm{tn}\}}, \mathrm{I}_{\{\mathrm{fp}\}}, \mathrm{I}_{\{\mathrm{fn}\}}, \mathrm{I}_{\{\mathrm{tp}\}}$ indicators. For example,

$$\mathrm{F}(\mathbb{V}) = \frac{2\mathrm{I}_{\{\mathrm{tp}\}}(\mathbb{V})}{\mathrm{I}_{\{\mathrm{fp}\}}(\mathbb{V}) + \mathrm{I}_{\{\mathrm{fn}\}}(\mathbb{V}) + 2\mathrm{I}_{\{\mathrm{tp}\}}(\mathbb{V})}\,.$$

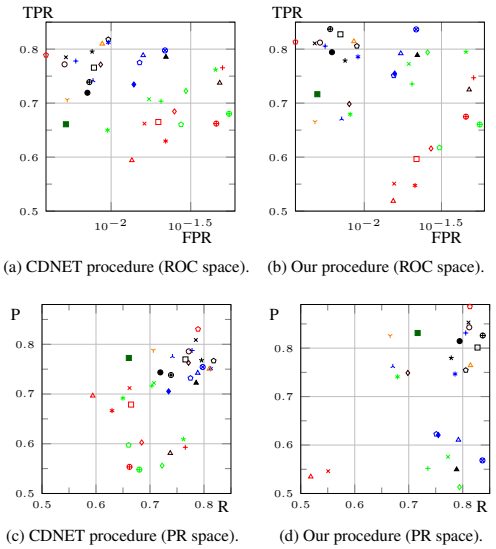
4. EXPERIMENTS WITH CDNET 2014

We illustrate the summarization with the CDNET 2014 dataset [22] (organized in 11 categories, containing each 4 to 6 videos). We recomputed all the performance indicators for the 36 unsupervised background subtraction algorithms (we could consider all the algorithms as well) whose segmentation maps are given on `http://changedetection.net`, compared to the ground-truth segmentation maps.

Our first experiment aims at determining if the summarized performance is significantly affected by the summarization procedure. Figure 2 compares the results of two summarization procedures, in both the ROC and PR spaces.

1. In the original CDNET procedure, performance indicators are obtained for each category by averaging the indicators measured on each video individually. A final summarized performance is then computed by averaging the performance indicators over the categories.

2. We applied our summarization, with the same weights for each category as in the original setup; thus, for each video, we have $P(V = v) = \frac{1}{11} \times \frac{1}{M}$, where M is the number of videos in the corresponding category.



(a) CDNET procedure (ROC space). (b) Our procedure (ROC space).

(c) CDNET procedure (PR space). (d) Our procedure (PR space).

Fig. 2: Summarized performances according to two different procedures, in the cropped ROC (upper row) or PR (lower row) spaces, for 36 classifiers evaluated on the CDNET 2014 dataset. The performances obtained by these procedures differ significantly. Only our summarization procedure preserves the bijection between ROC and PR (see text).

Algorithm	F of [22]	F [ours]
SemanticBGS	**0.8098 (1)**	**0.8479 (1)**
IUTIS-5	0.7821 (2)	0.8312 (3)
IUTIS-3	0.7694 (3)	0.8182 (5)
WisenetMD	0.7559 (4)	0.7791 (10)
SharedModel	0.7569 (5)	0.7885 (8)
WeSamBE	0.7491 (6)	0.7792 (9)
SuBSENSE	0.7453 (7)	0.7657 (12)
PAWCS	0.7478 (8)	0.8272 (4)

Table 1: Extract of F-scores (ranks) obtained with two summarization procedures on CDNET.

It can be seen that the results differ significantly (both in ROC and PR), which emphasizes the influence of the summarization procedure for the comparison between algorithms. However, with the original summarization procedure, the intrinsic relationships between indicators are not preserved at the summarized level, while ours preserves them.

Our second experiment aims at determining if the ranking between algorithms depends on the summarization procedure. Because the F-score is often considered as an appropriate ranking indicator for background subtraction, we performed our experiment with it. Table 1 shows the F-scores and ranks obtained according to the same two summarization procedures. We see that the summarization procedure affects the ranking.

5. CONCLUSION

In background subtraction, algorithms are evaluated by applying them on videos and comparing their results to ground-truth references. Performance indicators of individual videos are then combined to derive indicators representative for a set of videos, during a procedure called “summarization”.

We have shown that a summarization procedure based on the arithmetic mean leads to inconsistencies between summarized indicators and complicates the comparison between algorithms. Therefore, based on the definition of a random experiment, we presented a new summarization procedure and formulas that preserve the intrinsic relationships between indicators. These summarized indicators are specific for this random experiment and for the weights given to the various videos. An easy-to-code algorithm for our summarization procedure has been given. Our procedure is illustrated and commented on the CDNET dataset.

As a general conclusion, for background subtraction involving multiple videos, we recommend to always use our summarization procedure, instead of the arithmetic mean, to combine performance indicators calculated separately on each video. By doing so, the formulas that hold between them at the video level also hold at the summarized level.

6. REFERENCES

[14] Thierry Bouwmans, Fida El Baf, and Bertrand Vachon, "Statistical background modeling for foreground detection: A survey," in *Handbook of Pattern Recognition and Computer Vision (volume 4)*, chapter 3, pp. 181–199. World Scientific Publishing, January 2010.

[15] Thierry Bouwmans, "Traditional and recent approaches in background modeling for foreground detection: An overview," *Computer Science Review*, vol. 11-12, pp. 31–66, May 2014.

[16] 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*, vol. 117, pp. 8–66, September 2019.

[17] S. Elhabian, K. El-Sayed, and S. Ahmed, "Moving object detection in spatial domain using background removal techniques – State-of-art," *Recent Patents on Computer Science*, vol. 1, pp. 32–54, January 2008.

[18] Pierre-Marc Jodoin, Sébastien Piérard, Yi Wang, and Marc Van Droogenbroeck, "Overview and benchmarking of motion detection methods," in *Background Modeling and Foreground Detection for Video Surveillance*, chapter 24. Chapman and Hall/CRC, July 2014.

[19] Andrews Sobral and Antoine Vacavant, "A comprehensive review of background subtraction algorithms evaluated with synthetic and real videos," *Computer Vision and Image Understanding*, vol. 122, pp. 4–21, May 2014.

[20] N. Vaswani, T. Bouwmans, S. Javed, and P. Narayanamurthy, "Robust subspace learning: Robust PCA, robust subspace tracking, and robust subspace recovery," *IEEE Signal Processing Magazine*, vol. 35, no. 4, pp. 32–55, July 2018.

[21] 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)*, Providence, RI, USA, June 2012, pp. 1–8, Institute of Electrical and Electronics Engineers (IEEE).

[22] Nil Goyette, Pierre-Marc Jodoin, Fatih Porikli, Janusz Konrad, and Prakash Ishwar, "A novel video dataset for change detection benchmarking," *IEEE Transactions on Image Processing*, vol. 23, no. 11, pp. 4663–4679, November 2014.

[23] Carlos Cuevas, Eva María Yáñez, and Narciso García, "Labeled dataset for integral evaluation of moving object detection algorithms: LASIESTA," *Computer Vision and Image Understanding*, vol. 152, pp. 103–117, November 2016.

[24] Jesse Davis and Mark Goadrich, "The relationship between precision-recall and ROC curves," in *International Conference on Machine Learning (ICML)*, Pittsburgh, Pennsylvania, June 2006, pp. 233–240, ML Research Press.

[25] G. Forman and M. Scholz, "Apples-to-apples in cross-validation studies: Pitfalls in classifier performance measurement," *ACM SIGKDD Explorations Newsletter*, vol. 12, no. 1, pp. 49–57, November 2010.

[26] Kendrick Boyd, Victor Costa, Jesse Davis, and David Page, "Unachievable region in precision-recall space and its effect on empirical evaluation," in *International Conference on Machine Learning (ICML)*, Edinburgh, UK, June 2012, pp. 639–646.

[27] Cyril Goutte and Eric Gaussier, "A probabilistic interpretation of precision, recall and F-score, with implication for evaluation," in *Advances in Information Retrieval (Proceedings of ECIR)*. March 2005, vol. 3408 of *Lecture Notes in Computer Science*, pp. 345–359, Springer.

[28] Joseph Bertrand, *Calcul des probabilités*, Gauthier-Villars et fils, 1889.

[29] Peter A. Flach, "The geometry of ROC space: Understanding machine learning metrics through ROC isometrics," in *International Conference on Machine Learning (ICML)*, Washington, District of Columbia, USA, August 2003, pp. 194–201, ML Research Press.

[30] Tom Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, June 2006.

2. Performance-based rankings (2nd article)

This chapter reproduces the following article [2]:

> S. Piérard, A. Halin, A. Cioppa, A. Deliège, and M. Van Droogenbroeck, "Foundations of the theory of performance-based ranking," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (Nashville, Tennessee, USA), pp. 14293–14302, IEEE, June 2025.

It covers both the notions of ***performance*** and ***ranking***. The supplementary material of this article is reproduced in Chapter A.

Context

Every day, millions of people either develop and optimize entities with respect to the expected performance, or establish rankings and select entities based on their performance. This is indeed also the case in the scientific literature. Unfortunately, this is a colossus with feet of clay, as there does not seem to exist any rigorous mathematical theory for this.

Performance is a concept with no universally accepted definition. In management, a recent review [31] uncovered 239 (!) performance-related theories. These theories are qualitative, conceptual, and informal. They are certainly useful in some fields, but in others, like engineering, computer science, etc, we need a quantitative and mathematical approach. In this case —which is the one in which I place myself in this thesis—, the traditional approach is to "measure" certain quantities related to the proper functioning of the entities being evaluated and to imagine functions of these quantities, known as scores. These can be anything.

Motivation

When we attend a scientific conference such as CVPR[1] and look at the thousands of posters presented, you realize that most authors present a table of values, with rows corresponding to entities (their method, some "baseline" methods, and some competing methods) and columns corresponding to arbitrarily chosen scores that suggest that the

[1] The IEEE/CVF Conference on Computer Vision and Pattern Recognition is one of the top-tier international conferences in computer vision and machine learning. Its current CVPR h5-index is 450, just behind the well-known journal Nature, whose h5-index is 490.

entity developed by the authors is the best. As a general rule, the reasons for choosing these scores are unclear. It may be a matter of personal taste or imitation of what is done in a particular scientific community. The interpretation of the values is often very difficult for non-probabilistic scores. Moreover, there is no reason to believe that the chosen scores are sufficiently diverse to have a broad picture of performance and no reason to believe that they are suitable for establishing a meaningful ranking.

This is also the case in scientific challenges where online platforms report benchmarks. When the ranking of entities changes from one column to another, some compromise must be devised to obtain a single ranking of entities. This is traditionally done in two ways: (1) either by averaging the scores between columns, which amounts to ranking according to a single score (often difficult to interpret), or (2) by averaging the ranks induced by the scores between columns, which does not lead to a stable ranking. The insertion of a new entity can cause swaps of entities previously present in the ranking. This is a real disaster for the scientific community, since conclusions such as "method A is better than method B" are not perennial (see Fig. 2.1).

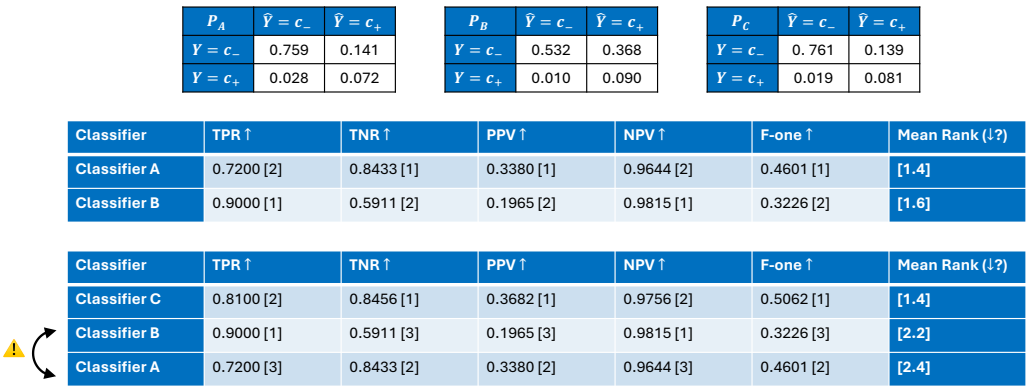
Main contributions

My ambition is to introduce the first generic mathematical theory for the performance-based ranking of entities.

I start by taking a step back and introducing a generic mathematical framework that integrates the notions of performance, task, and applicative preferences. I do this in a rigorous probabilistic framework [32, 33]. ***Performances*** are modeled as probability measures P, tasks are described through a random variable S called ***satisfaction***, and some preferences related to the applications or to the users are quantified in a positive random variable I called ***importance***. To the best of my knowledge, my work is the first[2] to propose a generic and non-trial (*i.e.*, not degenerated to a list of real numbers) mathematical nature for the performance.

Then, I take another step back and propose three axioms to specify what a performance-based ranking of entities is. My 1st axiom states that any performance-based ranking should be derived from a preorder on performances, which ensures the stability of the rankings *w.r.t.* insertions and deletions of ranked entities. This prevents the issue of ephemeral rankings shown in Fig. 2.1. My 2nd axiom gives compatibility conditions between the preorders and the considered task, modeled by the random variable satisfaction. My 3rd axiom gives compatibility conditions between the preorders and known properties about the evaluation (*i.e.*, the mapping of the entities to their performances). In particular, it prevents the issue of blind load balancing shown in Fig. 2.2. It is in this last axiom that the summarization (Chapter 1) is used. Interestingly, no score appears in the proposed axioms.

[2]The proposed probabilistic nature is in line with the reformulation proposed in [27] for two peculiar classification scores: instead of viewing the precision and recall as values computed from the observed data, they view them as probabilities. The authors underline that it makes a big semantic difference but do not formally ground their developments in the probability theory. In particular, they do not



P_A	$\hat{Y} = c_-$	$\hat{Y} = c_+$
$Y = c_-$	0.759	0.141
$Y = c_+$	0.028	0.072

P_B	$\hat{Y} = c_-$	$\hat{Y} = c_+$
$Y = c_-$	0.532	0.368
$Y = c_+$	0.010	0.090

P_C	$\hat{Y} = c_-$	$\hat{Y} = c_+$
$Y = c_-$	0. 761	0.139
$Y = c_+$	0.019	0.081

Classifier	TPR ↑	TNR ↑	PPV ↑	NPV ↑	F-one ↑	Mean Rank (↓?)
Classifier A	0.7200 [2]	0.8433 [1]	0.3380 [1]	0.9644 [2]	0.4601 [1]	[1.4]
Classifier B	0.9000 [1]	0.5911 [2]	0.1965 [2]	0.9815 [1]	0.3226 [2]	[1.6]

Classifier	TPR ↑	TNR ↑	PPV ↑	NPV ↑	F-one ↑	Mean Rank (↓?)
Classifier C	0.8100 [2]	0.8456 [1]	0.3682 [1]	0.9756 [2]	0.5062 [1]	[1.4]
Classifier B	0.9000 [1]	0.5911 [3]	0.1965 [3]	0.9815 [1]	0.3226 [3]	[2.2]
Classifier A	0.7200 [3]	0.8433 [2]	0.3380 [2]	0.9644 [3]	0.4601 [2]	[2.4]

Figure 2.1.: **The issue of ephemeral rankings:** instability of rankings based on averaged ranks. **Top.** This example is based on three classifiers for a two-class crisp classification task. Their confusion matrices are given above. **Middle.** First, let us consider a leaderboard that includes only the classifiers A and B. We arbitrarily take six scores that we want to maximize: the true positive rate (TPR), the true negative rate (TNR), the positive predictive value (PPV), the negative predictive value (NPV), and the F_1 score. The middle table gives the value of each of these scores for each ranked entity and gives in brackets the rank obtained for each score. The last column shows the arithmetic mean value of these ranks. Intuitively, we would expect this quantity to be maximized, but this is not the case. **Bottom.** To see the issue, let us invert the classifier C to the leaderboard. The result of this insertion is shown in the bottom table. We can see that the average rank of A, which was lower than the average rank of B, is now higher. The insertion of C to the leaderboard has caused A and B to swap places. The average rank, for an arbitrary set of scores, therefore does not lead to perennial rankings.

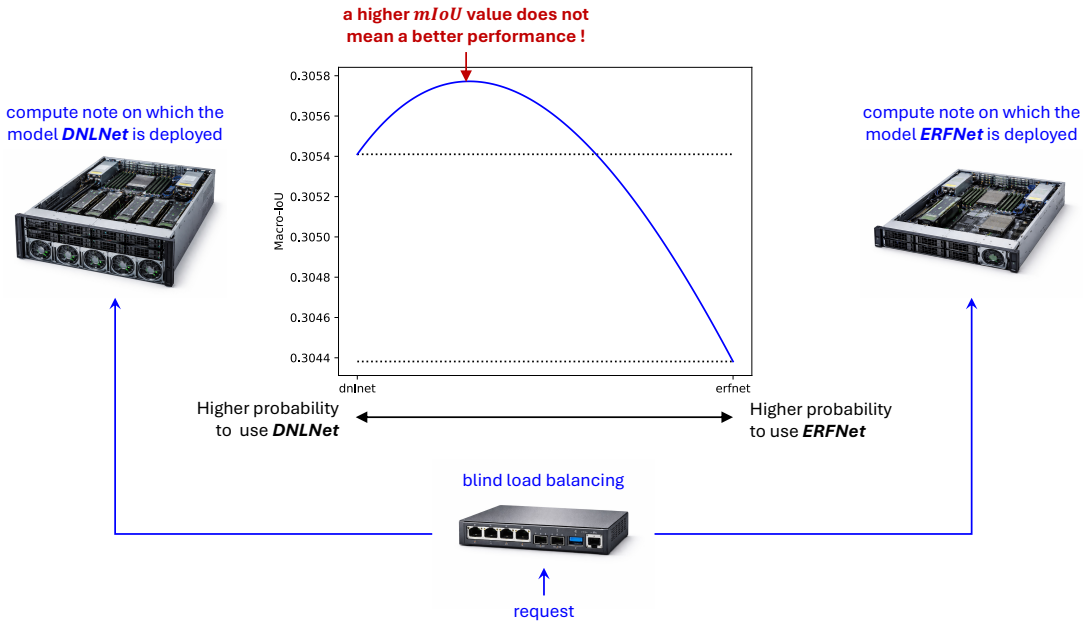


Figure 2.2.: **The issue of blind load balancing.** The macro-averaged intersection over union ($mIoU$) is an example of a problematic score for ranking. To see this, consider a system composed of several computing nodes on which semantic segmentation models are deployed. DNLNet [34] runs on some nodes and ERFNet [35] on others. According to $mIoU$, DNLNet (left) appears to be better than ERFNet (on the right). Knowing the performance of both models, it is possible to calculate the performance of a system implementing a load balancing that randomly and blindly (*i.e.*, without analyzing the requests) redirects the requests to one or the other type of computing nodes, *w.r.t.* the probabilities of using each model. Given this blind balancing, it is clear that the system implementing it cannot be better than a system using only DNLNet, nor can it be worse than a system using only ERFNet. However, as this curve shows, $mIoU$ suggests that the system's performance is maximized when DNLNet and ERFNet are used for approximately 70% and 30% of requests, respectively, and that it outperforms the best of both nodes. From a theoretical perspective, this is not acceptable.

Then, I introduce a universal parametric family of scores, called *ranking scores*, that can be used to establish rankings satisfying our axioms while considering application-specific preferences. All in all, what is proposed in this article is an integrated and generic approach, as shown in Fig. 2.3.

Finally, I show, in the case of two-class classification, that the family of ranking scores encompasses well-known performance scores, including the accuracy, the true positive rate (recall, sensitivity), the true negative rate (specificity), the positive predictive value (precision), and F_1. However, I also show that some other common scores are unsuitable to derive performance orderings satisfying the axioms.

Perspectives

The generic theoretical foundations provided in this paper are at the heart of all chapters concerned with the performance-based ranking (Chapters 2 to 7).

The particularization of my theoretical foundations to the two-class crisp (*a.k.a.* hard) classification task is rich in insights and leads to innovative practical tools (the Tile presented in Chapter 4 and its flavors presented in Chapters 5 to 7) in comparison to what exists [36]. However, the two-class crisp classification is a very particular case. Thus, I expect that the rigorous study of other tasks will be challenging but impactful for a large community. However, this is left for future work (see Fig. 2.4).

Remarks

1. If we have a ranking that does not respect the axioms, it means that if an entity ϵ_A appears at a lower (resp. higher) rank than an entity ϵ_B, then we cannot claim that ϵ_A is better (resp. worse) than ϵ_B. On the other hand, if the axioms are respected, then we can claim that ϵ_A is better (resp. worse) than ϵ_B, but we must be careful to specify that this is for a certain preorder. With rankings induced by ranking scores, we specify the preorder through the random variable I.

2. Ranking cannot be done rigorously without the order (lattice) theory. Traditionally, performance comparison is done with scores assigning a numerical value to each performance. This allows the natural order on real numbers to be used on performances. The scientific literature is full of arbitrarily defined scores. For example, several dozen have been defined for the two-class crisp classification task [37, 38]. Their properties and their suitability to derive meaningful rankings remain poorly understood. In a recent paper [39], Gösgens *et al.* listed nine intuitively desirable properties for classification scores but proved that they cannot be simultaneously satisfied. Taking the point of view of rankings simplifies the reasoning [2] as only three axioms about performance orderings, which are consistent, seem sufficient. The exact relationships between the properties of [39] and the axioms of [2] have yet to be studied.

specify any measurable space or any probability measure. This is properly done here.

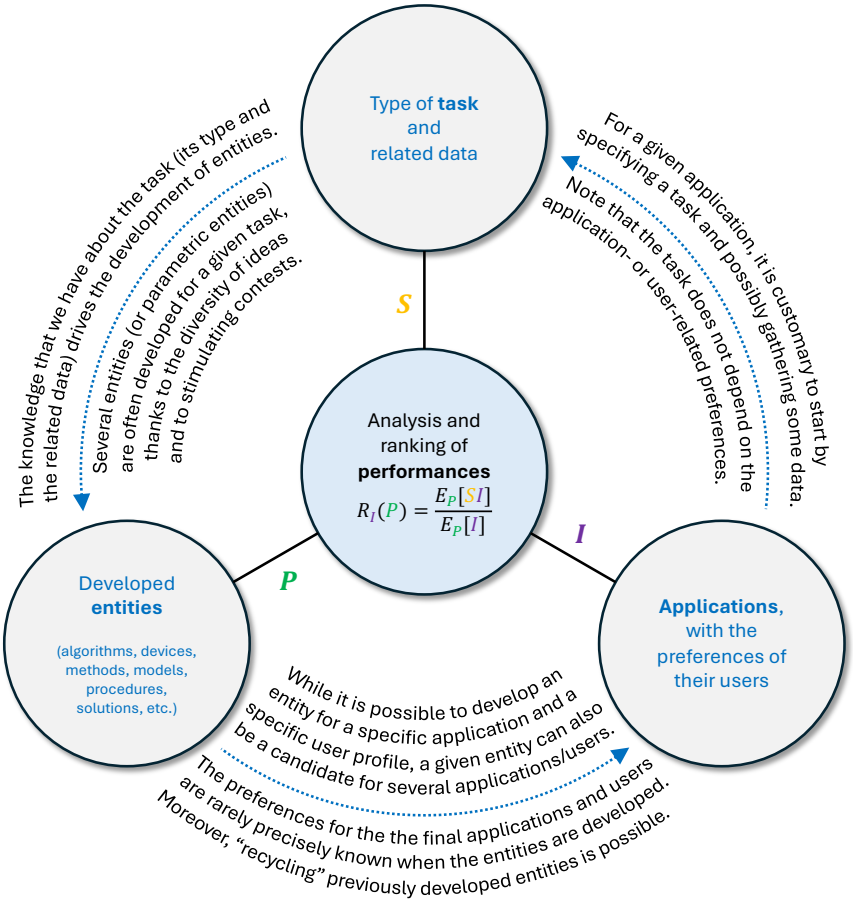


Figure 2.3.: My generic and integrated approach for the evaluation of entities, the analysis of their performances, and their comparisons with rankings. It is aligned with the three pillars depicted by the outer disks: (1) the applications and their users, (2) the tasks and data, and (3) the developed entities. The preferences associated with the applications and their users are characterized by the importance I, the task by the satisfaction S, and the entities by their performances P. The integration of I, S, and P is represented at the center, where we see the formula for the ranking scores R_I, a continuum of scores parameterized by I that induce meaningful rankings of entities. The symbol $\mathbf{E}_P$ stands for the expected value with the probability measure P.

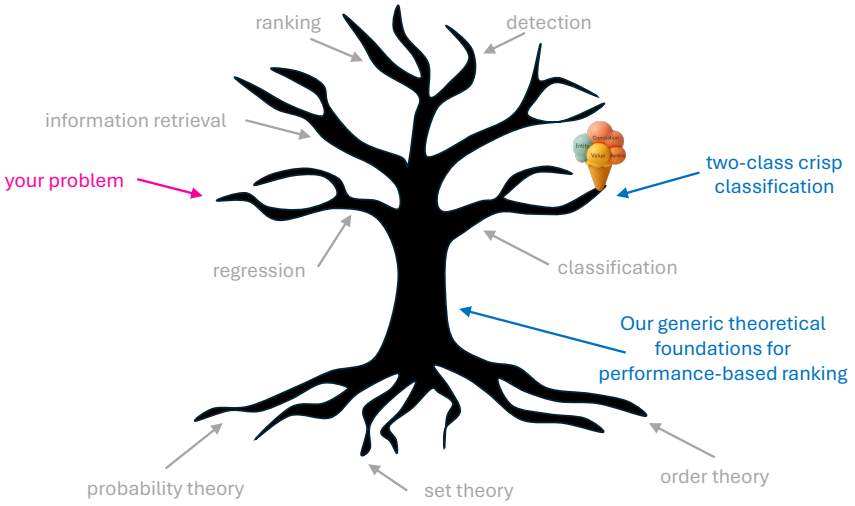


Figure 2.4.: My global vision for the research topics related to both the rankings and the performances depicted as a tree. The trunk represents the generic theoretical foundations for performance-based ranking introduced in Chapter 2. The roots correspond to the mathematical theories on top of which it is built: the set, probability, and order theories. The branches correspond to various problems or tasks. Chapter 4 particularizes my generic theory to the particular case of tasks similar to two-class crisp classification. The ice cream cone with four scoops of different flavors, that stands at a branch extremity, represents *Sorbetto*, an open-source library implementing the results of this particularization: the Tile (*cf* Chapter 4) and its flavors (*cf* Chapter 5). For the other tasks listed on the branches, see the section entitled "a Little Catalog of Problems" in supplementary material of [2].

3. Tasks are generally only implicitly defined. Pan and Yang [40] proposed to define it as a function that should be mimicked, a definition in line with a learning objective. Unfortunately, perfectly mimicking a function is often unreachable, so it is important to know how satisfying a prediction that differs from the ground truth is. To the contrary, in [2], the task is only characterized by this satisfaction, in line with a performance analysis objective.

Foundations of the Theory of Performance-Based Ranking

Sébastien Piérard, Anaïs Halin, Anthony Cioppa, Adrien Deliège, and Marc Van Droogenbroeck
Montefiore Institute, University of Liège, Liège, Belgium
{S.Pierard,Anais.Halin,Anthony.Cioppa,Adrien.Deliege,M.VanDroogenbroeck}@uliege.be

Abstract

Ranking entities such as algorithms, devices, methods, or models based on their performances, while accounting for application-specific preferences, is a challenge. To address this challenge, we establish the foundations of a universal theory for performance-based ranking. First, we introduce a rigorous framework built on top of both the probability and order theories. Our new framework encompasses the elements necessary to (1) manipulate performances as mathematical objects, (2) express which performances are worse than or equivalent to others, (3) model tasks through a variable called satisfaction, (4) consider properties of the evaluation, (5) define scores, and (6) specify application-specific preferences through a variable called importance. On top of this framework, we propose the first axiomatic definition of performance orderings and performance-based rankings. Then, we introduce a universal parametric family of scores, called ranking scores*, that can be used to establish rankings satisfying our axioms, while considering application-specific preferences. Finally, we show, in the case of two-class classification, that the family of ranking scores encompasses well-known performance scores, including the accuracy, the true positive rate (recall, sensitivity), the true negative rate (specificity), the positive predictive value (precision), and F_1. However, we also show that some other scores commonly used to compare classifiers are unsuitable to derive performance orderings satisfying the axioms.*

1. Introduction

Every day, millions of people are faced with choices to make. Often, these choices are between entities (*e.g.*, algorithms, devices, methods, models, options, procedures, solutions, strategies, *etc.*) considered to be interchangeable, although not necessarily equivalent in terms of performance. One of the main difficulties arises from the uncertainty that people have regarding the use that will be made of the entity to choose. A widespread approach to objectifying these choices is to (1) perform an evaluation to determine (*i.e.*, assume, calculate, estimate, predict, *etc.*) a *performance*, encompassing the necessary uncertainty, for each of these entities; (2) choose a way of comparing these performances with each other; and (3) assume that an entity is preferable to others if it has the best performance. A more general problem is to establish an order of preference between the entities: this is the *performance-based ranking*.

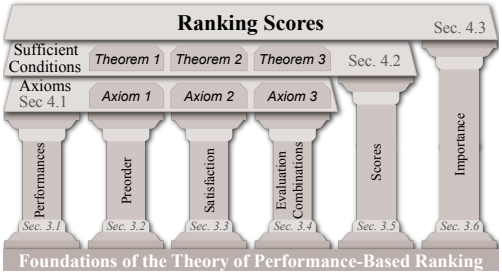


Figure 1. This work establishes the foundations of the theory of performance-based ranking. We do this in two steps. First, we introduce a new mathematical framework with 6 main elements, as depicted here by the pillars. Second, we build on top of it: (1) a set of three axioms for the ordering of performances and for the performance-based ranking of entities, (2) sufficient conditions for them when the performance ordering is induced by a score, and (3) a family of scores, named *ranking scores* that consider the application-specific preferences. This theory is universal in the sense that it is applicable to any task.

The approach of performance-based ranking is common in many fields and has proved its usefulness, especially in scientific communities that organize themselves around competitions [22, 41–43] for the development of algorithms for specific tasks. Nevertheless, several studies [44, 45] have alerted the scientific community about the ranking methodology used in these competitions.

A critical analysis [44] of common practices for 150 biomedical image analysis challenges reveals that the scores used are justified in only 23% of the cases, and that the rank computation method is reported in only 36% of the cases. Moreover, there are at least 10 different methods for deter-

mining the rank of an algorithm based on multiple scores. The properties of these methods are largely unknown.

There remains an obvious lack of theoretical foundations for the performance-based ranking. From our perspective, there is a common and detrimental confusion between the concept of performance and the numerical scores (also called metrics, measures, indicators, criteria, factors, and indices [46, 47]). Moreover, the way of comparing performances is often chosen by intuition [45] or by imitation of what has been chosen in the previously published related works. This inevitably leads to a drastic loss of diversity in the rankings that can be presented. From our perspective, considering the diversity of possible rankings is healthy, and also desirable for those who want to subsequently choose an entity based on their own application-specific preferences.

The lack of theoretical foundations is profound. Loosely speaking, the *performance* of a given entity can be defined as the information necessary to determine the degree of *satisfaction* one has with it, relatively to application-specific preferences tuned by the relative *importance* given to the various cases that can occur when the entity is used. Although such a definition sheds interesting light on the subject, it does not specify the mathematical nature of the performance, the space in which it is, or the operations that are permitted, particularly those that enable the various performances to be compared with one another.

The aim of this paper is to look at the performance-based ranking from a broader angle. As shown in Fig. 1, we establish mathematically rigorous theoretical foundations that can be applied to a wide range of problems. Throughout the paper, we exemplify our theory with the task of two-class classification. More tasks are detailed in Sec. A.2 about: multi-class classification with links to micro- and macro-averaging; regression with links to the mean squared error and the mean absolute error; information retrieval; detection with links to the intersection over-union and the F-score; clustering with a link to the Fowlkes-Mallows index; ranking with a link to Kendall's τ.

Contributions. Our contributions are threefold. (1) First, in Sec. 3, we present a new mathematical framework that encompasses the evaluation of entities, the performances, the space in which they are, the notions of *satisfaction* and *importance*, the scores that characterize numerically the performances, and some operations permitted on performances such as those for combining several performances and those for comparing performances with one another (*is equivalent to*, *is worse than*, *is better than*, *is incomparable with*, *etc*.). (2) Second, we introduce the foundations of the theory of performance-based ranking in Sec. 4. We innovate with an *axiomatic definition for both the ordering of performances and the ranking of entities based on their performances*. We also present a new family of scores, named *ranking scores*, that can be used to induce performance orderings that satisfy our axioms. These scores are parameterized by application-specific preferences, and are universal, *i.e*., applicable to any task. (3) Third, in Sec. 5, we study the particularization of our theory of performance-based ranking to the popular case of the two-class crisp classification.

2. Related Work

Mathematical foundations for our work can be found in the probability theory, in the order theory, and in statistics. The probability theory [32, 33] provides the tools needed to consider the uncertainty that one has about how an entity will be used. However, to be rigorous, probability measures cannot be used without defining a measurable space. Surprisingly, after reviewing hundreds of papers that use the notion of performance, mainly in the fields of computer vision, medicine and physics, we found none that explicitly gives such a space and explains how to express, based on it, what the performances are and what operations are allowed on them. The order theory [48, 49] provides the basis for defining and manipulating homogeneous binary relations such as *is equivalent to*, *is worse than*, *is better than*, and *is incomparable with* that underpin rankings. Statistics provide tools to compare rankings through rank correlations, in particular Kendall's τ [50] and Spearman's ρ [51].

In a recent attempt to formalize the notion of ranking, for basic vision tasks, Nguyen *et al*. [45] proposed to impose three properties for ranking: (1) reliability ("a small change in parameter values, should not result in a drastic change in rankings"), (2) meaningfulness (evaluated by humans), and (3) mathematical consistency ("use scores that satisfy certain properties"). The mathematical framework introduced in this paper helps to clarify these three requirements. (1) Regarding the reliability, which is also a matter of concern in [44], we argue that one should distinguish between two types of parameters: those involved in the evaluation (*i.e*. the step in which the performance of an entity is determined) and those involved in the ranking of the entities based on their previously determined performances. The case actually discussed in [45] is of the first type. The second type of parameters can be useful to adapt the ranking to application-specific preferences: the *importance*. (2) Regarding the meaningfulness, in the absence of analytical means to ensure the meaningfulness of scores, Nguyen *et al*. [45] suggest testing the ranking procedure on sanity tests with pre-determined desired rankings. In contrast, we propose to start by modeling the task through a variable called *satisfaction* and then to derive meaningful ranking scores, by construction. (3) Concerning the mathematical consistency, we show that it is possible to impose it based on axioms, at a level just below the scores (see Fig. 1).

3. Mathematical Framework

We now present the mathematical framework in which the theory of performance-based rankings will be established in Sec. 4.

To the best of our knowledge, it is the first time that a rigorous mathematical framework is conceived for the universal comparison of performances. All mathematical symbols used in this paper are defined where they first appear. For convenience, we also provide a list of them in Sec. A.1. Our framework involves six components that correspond to the pillars depicted in Fig. 1.

Performance to address uncertainty. First, we introduce the notion of *performance*, which is our main object of interest. To anchor it on a solid mathematical ground, we leverage probability theory to benefit from its established expressivity and rigor, and thus define performances as probability measures. Probability theory is indeed the ideal framework for studying uncertainty and randomness, which naturally pertain to performances, hence our choice.

The essence of performance. Then, we define the three components that really differentiate performances from ordinary probability measures: performances should be comparable/rankable (*e.g.*, we need notions of *better* and *worse* performances) through a *preorder* $\lesssim$; performances should be related to a task, that we model by a random variable called *satisfaction* S; and performances should be related to entities (algorithms, devices, *etc.*) through an *evaluation*, that we model by a function Φ. We show that there exist compatibility conditions to be met between $\lesssim$ and S as well as between $\lesssim$ and Φ, which lead us to formalize the *axioms* at the basis of our theory in Sec. 4.

Performance in practical applications. Finally, we define two components that connect our theory to practical applications: the *scores* X, which are functions mapping performances to numerical values (*e.g.* accuracy, error rate,*etc.*); and the *importance* I, a random variable that encodes applicative preferences about the possible outcomes of the process. Later in Sec. 4, we show how these two components can be defined to satisfy sufficient conditions (through three theorems) to fulfill our three axioms, thus further serving as basis for our performance-based ranking theory and the definition of new universal ranking scores.

3.1. The Performance P as a Probability Measure

For us, a performance is not a real number or a collection of them, as sometimes assumed in the literature. Instead, we choose to design our mathematical framework specifically to compare performances having a probabilistic meaning. Thus, we ground our framework in probability theory [32, 33] and consider that *performances* P are probability measures. For two performances to be comparable, they should be defined on a common measurable space (Ω, Σ), where Ω is the sample space, or universe, and Σ is a σ-algebra on Ω called event space. Without loss of generality, when Ω is finite, one can choose $\Sigma = 2^{\Omega}$. We note $\mathbb{P}_{(\Omega,\Sigma)}$ the set of probability measures on (Ω, Σ).

Example 1. *For the popular case of two-class crisp classification, several choices can be made for* Ω*. In the simplest setting, this set contains two elements interpreted as "correct result" and "incorrect result". In this case, the performance analysis can only be based on the proportion of correct results,* i.e.*, the accuracy. In another setting, one can choose to have three elements in* Ω*, making the distinction between the two types of incorrect results, namely* false positive *(*a.k.a. *type I error) and* false negative *(*a.k.a. *type II error). Finally, one could prefer having four elements in* Ω*, one for each pair of ground-truth and predicted classes (in the frequency-based approach, this corresponds to the normalized contingency table or confusion matrix). This enlarges the flexibility in the analysis of performances.*

3.2. Ordering Performances with a Preorder $\lesssim$

Our framework is not only grounded in probability theory, as explained above, but also in order theory [48, 49]. We aim at being able to decide if a performance is, *e.g.*, worse than, equivalent, or better than another one. Mathematically, these comparisons are done thanks to binary homogeneous relations on $\mathbb{P}_{(\Omega,\Sigma)}$. Indeed, all binary relations used to compare performances should be coherent, *i.e.*, they should all correspond to a common *performance ordering*. Following this path, in this paper, the binary relations $\sim$, $>$, $<$, and $\gtrless$ on $\mathbb{P}_{(\Omega,\Sigma)}$ will be implicitly considered as derived from a common binary relation $\lesssim$ on $\mathbb{P}_{(\Omega,\Sigma)}$ as follows:

- $P_1 \sim P_2$ if and only if (*iif*) $P_1 \lesssim P_2 \wedge P_2 \lesssim P_1$;
- $P_1 > P_2$ *iif* $P_1 \not\lesssim P_2 \wedge P_2 \lesssim P_1$;
- $P_1 < P_2$ *iif* $P_1 \lesssim P_2 \wedge P_2 \not\lesssim P_1$;
- $P_1 \gtrless P_2$ *iif* $P_1 \not\lesssim P_2 \wedge P_2 \not\lesssim P_1$.

With such a construction, if $\lesssim$ is a preorder (*i.e.*, reflexive and transitive), then $\sim$ is an equivalence (*i.e.* reflexive, transitive, and symmetric), $>$ and $<$ are converse strict partial orders (*i.e.*, irreflexive, asymmetric, and transitive), and $\gtrless$ is irreflexive and symmetric. For the proof, see Sec. A.3.4. These are the intuitively expected properties that justify to interpret $\lesssim$ as *worse or equivalent*, $>$ as *better*, $<$ as *worse*, and $\gtrless$ as *incomparable*.

Example 1 (continued)**.** *In the case of two-class classification, two spaces are commonly used to depict the performances as points, at fixed priors: the* Receiver Operating Characteristic *(ROC) space and the* Precision-Recall *(PR) space. On the one hand, ROC users will all intuitively agree that a performance* P_1 *is better than, equivalent to, or worse than a performance* P_2 *when both the values of* TNR *(true negative rate) and the* TPR *(true positive rate) of* P_1 *are greater, equal, or smaller than those of* P_2*, respectively.*

Few of these users will risk deciding when one of the two scores is higher while the other is lower, as this depends on application-specific preferences. On the other hand, PR users will perform a similar intuitive reasoning based on the TPR (recall) and the PPV (precision). A careful comparison reveals that ROC and PR users do not intuitively make the same decisions. This is surprising since, at fixed priors, both spaces show the same thing [24]. Our theory clarifies what are the suitable performance orderings.

3.3. Modeling the Task as a Random Variable S

The random variable *satisfaction*, $S : \Omega \to \mathbb{R}$, is task-specific. The user is responsible for assigning satisfaction values that are meaningful for the task at hand. We argue that a task is ill-defined when the satisfaction is not specified, either explicitly or implicitly. In this paper, we pose $s_{min,\Omega} = \min_{\omega\in\Omega} S(\omega)$ and $s_{max,\Omega} = \max_{\omega\in\Omega} S(\omega)$.

Example 1 (continued)**.** *In the case of two-class classification, we expect that most people will agree that (1) the satisfaction takes the same value for all samples corresponding to incorrect results (*e.g. $S = 0$*), (2) the satisfaction takes the same value for all samples corresponding to correct results (*e.g. $S = 1$*), and that (3) the satisfaction is strictly greater for correct results than for incorrect results.*

3.4. Modeling the Evaluation as a Function Φ

We now have all the elements necessary to introduce the notion of *evaluation*. Let us denote, by $\mathbb{E}$, the set of entities of interest. In our framework, the evaluation is modeled by function $\mathrm{eval} : \mathbb{E} \to \mathbb{P}_{(\Omega,\Sigma)} : \epsilon \mapsto \mathrm{eval}(\epsilon)$. We find it convenient to think in terms of a (thought) random experiment involving an entity ϵ and having outcomes that allow to determine how satisfying the various realizations are, *e.g.*, if the result is correct, how accurate it is, or how many resources are used. We define the performance $P = \mathrm{eval}(\epsilon)$ of an entity ϵ as the distribution of these outcomes.

Example 1 (continued)**.** *In the particular case of classification, a typical random experiment implicitly considered in the literature is in five steps. (1) Draw a sample s at random from a source. (2) Apply the oracle on s to obtain ground-truth class $y(s)$. (3) Apply a descriptor on s to obtain the features (*a.k.a. *attributes) $x(s)$. (4) Feed the classifier with $x(s)$ to obtain the predicted class $\hat{y}(x(s))$. (5) Set the outcome of the experiment to the pair $(y, \hat{y})$.*

Our framework can be further enriched by integrating some knowledge about the function eval. By definition, a performance P is achievable when there exists an entity ϵ whose evaluation leads to it: $\exists\epsilon : \mathrm{eval}(\epsilon) = P$. It often happens that, based solely on the performances of the evaluated entities, we can be sure that some other performances are also achievable, by combining and/or disrupting the entities we have at our disposal. To take this knowledge into account, we define the function $\Phi : 2^{\mathbb{P}_{(\Omega,\Sigma)}} \to 2^{\mathbb{P}_{(\Omega,\Sigma)}}$ that gives the set of performances that are achievable for sure, for some set of achievable performances given in input. Note that Φ is idempotent, *i.e.* $\Phi \circ \Phi = \Phi$.

Consider a random experiment using a black box entity ϵ only once (this is the knowledge we have about eval). Let $\lambda_1, \lambda_2, \ldots \lambda_n$ be positive values summing up to one. All other things remaining identical, if one can achieve the performances $P_1, P_2, \ldots P_n$ with, respectively, the entities $\epsilon_1, \epsilon_2, \ldots \epsilon_n$, then the performance $\sum_i \lambda_i P_i$ is achievable with a hybrid entity that randomly selects an entity among $\epsilon_1, \epsilon_2, \ldots \epsilon_n$ with the series of respective selection probabilities $\lambda_1, \lambda_2, \ldots \lambda_n$ before running the corresponding entity. In this case, denoting the set of all possible convex combinations by conv, we can take $\Phi = \mathrm{conv}$.

Example 1 (continued)**.** *In the particular case of two-class crisp classification, when the source of samples, the oracle, and the descriptor are kept unchanged,* Fawcett's interpolation *[30] allows interpolating linearly between performances with a hybrid classifier. And, when the oracle, the descriptor, and the classifier are kept unchanged,* Piérard's summarization *[1] allows interpolating linearly between performances with a hybrid source of samples.*

3.5. The Scores as Functions X of Performances

In our framework, *scores*[1] (also called metrics, measures, indicators, criteria, factors, and indices in the literature [46, 47]) are functions associating a real value to performances, that is, $X : \mathrm{dom}(X) \to \mathbb{R} : P \mapsto X(P)$ with $\mathrm{dom}(X) \subseteq \mathbb{P}_{(\Omega,\Sigma)}$. One can define different parametric families of scores, such as:

- *Expected value scores* are parameterized by a random variable V. We define them as $X_V^E : \mathbb{P}_{(\Omega,\Sigma)} \to [\min_{\omega\in\Omega} V(\omega), \max_{\omega\in\Omega} V(\omega)] : P \mapsto X_V^E(P) = \mathbf{E}_P[V]$, where $\mathbf{E}$ denotes the mathematical expectation. The score X_S^E, that we call the *expected satisfaction*, is a universal score in the sense that it exists for all sample spaces Ω. It will be further studied in Sec. 4.
- *Probabilistic scores* are parameterized by two events $E_1, E_2 \in \Sigma$ such that $\emptyset \subsetneq E_1 \subsetneq E_2 \subseteq \Omega$. We define them as $X_{E_1|E_2}^P : \{P \in \mathbb{P}_{(\Omega,\Sigma)} : P(E_2) \neq 0\} \to [0, 1] : P \mapsto X_{E_1|E_2}^P(P) = P(E_1|E_2)$. All probabilistic scores can be expressed as a ratio of two expected value scores.

Example 1 (continued)**.** *In the case of two-class classification, the expected satisfaction is both the expected value score X_S^E and the probabilistic score $X_{S=1|\Omega}^P$. This is because S is a $\{0, 1\}$-binary random variable. In this case, the expected satisfaction is called* accuracy *and is noted A.*

[1] We choose the term *score* to avoid any possible confusion with the mathematical meaning of the terms *metric*, *measure*, and *indicator*.

3.6. Modeling Application-Specific Preferences as a Random Variable I

It is well known that the ranking of entities does not only have to be specific for the task, but that it should also be sensitive to application-specific preferences. To encode these preferences, we propose to rely on a second random variable that we call *importance*: $I : \Omega \to \mathbb{R}_{\geq 0}$. We require that $I \neq 0$, *i.e.* $\exists \omega : I(\omega) \neq 0$.

Further in this paper, we will describe a new family of scores, called *ranking scores*, which are parameterized by this random variable. We would, however, like to draw the reader's attention to the fact that the importance is something that cannot, in general, be deduced from a score. In particular, the visual inspection of a formula for a given score could be misleading. Therefore, in this paper, we present a technique to analyze the behavior of any score by computing the rank correlations with the ranking scores for which the importances are well-defined.

Example 1 (continued)**.** *In the case of two-class classification, we provide examples of misleading formulas in Sec. A.4, in particular two equivalent formulas for the accuracy and two for the true positive rate. In both cases, by visually inspecting them, one would intuitively draw different conclusions about the importance given to the true negatives, false positives, false negatives, and true positives.*

4. Performance-Based Ranking Theory

In this section, we build our theory in three steps, represented by the three lintels of Fig. 1, on top of the six pillars depicting our mathematical framework. First, we present a universal axiomatic definition of performance orderings and performance-based rankings of entities (Sec. 4.1). Then, we link these axioms with the scores, and give a sufficient condition per axiom (Sec. 4.2). Finally, we account for the application-specific preferences and provide an infinite, diversified, and universal family of scores that induce performance orderings satisfying our axioms (Sec. 4.3).

4.1. Axiomatic Definition

We propose an axiomatic definition (the first as far as we know) of both performance orderings and performance-based rankings. The axioms do not involve any X or I.

4.1.1 Leveraging the Preorder $\lesssim$

We argue that if several entities from a set $\mathbb{E}$ have been ranked, then removing or adding an entity should not affect the relative order of those ranked entities. To guarantee it, our first axiom imposes that the ranking function is based on a preorder $\lesssim$ on $\mathbb{P}_{(\Omega,\Sigma)}$. There is no consensus in the literature on the rank value to consider when several entities have equivalent performances. Thus, instead of setting an arbitrarily chosen value, our axiom specifies bounds for it.

Axiom 1. *The ranking function* $\mathrm{rank}_{\mathbb{E}} : \mathbb{E} \to [1, |\mathbb{E}|] : \epsilon \mapsto \mathrm{rank}_{\mathbb{E}}(\epsilon)$ *satisfies* $|\{\epsilon' \in \mathbb{E} : \mathrm{eval}(\epsilon) < \mathrm{eval}(\epsilon')\}| + 1 \leq \mathrm{rank}_{\mathbb{E}}(\epsilon) \leq |\{\epsilon' \in \mathbb{E} : \mathrm{eval}(\epsilon) \lesssim \mathrm{eval}(\epsilon')\}|$*, where* $\lesssim$ *is a preorder on* $\mathbb{P}_{(\Omega,\Sigma)}$*.*

4.1.2 Leveraging the Satisfaction S

We argue that if the satisfaction that can be obtained with an entity ϵ_1 is for sure less or equal than the satisfaction that can be obtained with an entity ϵ_2, then the performance P_1 of ϵ_1 cannot be better than the performance P_2 of ϵ_2.

Axiom 2. *For* $P_1, P_2 \in \mathbb{P}_{(\Omega,\Sigma)}$ *such that* $P_1(S \leq s) = 1$ *and* $P_2(S \geq s) = 1$ *for some* s*, then* $P_1 \lesssim P_2$ *or* $P_1 \lneqq P_2$*.*

This implies that, thanks to S, we can use the natural order $\leq$ that exists on $\mathbb{R}$ to obtain a preorder on Ω, and the axiom states that the preorder on $\mathbb{P}_{(\Omega,\Sigma)}$ is coherent with it.

Let us now take a look at three implications that are clearly intuitively expected.

Corollary 1. *For any* s*, all performances* P *such that* $P(S = s) = 1$ *are either equivalent or incomparable.*

Corollary 2. *We have* $P(S = s_{min,\Omega}) = 1 \Rightarrow \nexists P' : P' < P$*. In other words,* $P(S = s_{min,\Omega}) = 1$ *means that* P *belongs to the set of the worst performances, among* $\mathbb{P}_{(\Omega,\Sigma)}$*.*

Corollary 3. *We have* $P(S = s_{max,\Omega}) = 1 \Rightarrow \nexists P' : P' > P$*. In other words,* $P(S = s_{max,\Omega}) = 1$ *means that* P *belongs to the set of the best performances, among* $\mathbb{P}_{(\Omega,\Sigma)}$*.*

4.1.3 Leveraging the Combinations Φ

We argue that, for any set of achievable performances, it must be impossible to obtain with certainty a performance better than the best of them, or worse than the worst of them, by sequentially combining operations from an arbitrary set of possible operations to perturb (*e.g.*, add noise to their output) the entities corresponding to that initial set. Expressing this requirement in terms of $\lesssim$ leads to our third axiom.

Axiom 3. *Let* P *be a performance, and* Π *be a set of performances on* $\mathbb{P}_{(\Omega,\Sigma)}$ *such that* $P' \lesssim P \vee P \lesssim P' \, \forall P' \in \Pi$*.*

- $P' \lesssim P \, \forall P' \in \Pi \Rightarrow \overline{P} \lesssim P \, \forall \overline{P} \in \Phi(\Pi)$*;*
- $P' \not\lesssim P \, \forall P' \in \Pi \Rightarrow \overline{P} \not\lesssim P \, \forall \overline{P} \in \Phi(\Pi)$*;*
- $P \lesssim P' \, \forall P' \in \Pi \Rightarrow P \lesssim \overline{P} \, \forall \overline{P} \in \Phi(\Pi)$*;*
- *and* $P \not\lesssim P' \, \forall P' \in \Pi \Rightarrow P \not\lesssim \overline{P} \, \forall \overline{P} \in \Phi(\Pi)$*.*

With the definitions for $\sim$, $>$, $<$, and $\lneqq$ given in Sec. 3.2, the following corollary can be derived.

Corollary 4. *Let* $P \in \mathbb{P}_{(\Omega,\Sigma)}$ *and* $\Pi \subseteq \mathbb{P}_{(\Omega,\Sigma)}$ *such that* $P' \lesssim P \vee P \lesssim P' \, \forall P' \in \Pi$*. We have:*

- $P' \sim P \,\forall P' \in \Pi \Rightarrow \overline{P} \sim P \,\forall \overline{P} \in \Phi(\Pi)$;
- $P' > P \,\forall P' \in \Pi \Rightarrow \overline{P} > P \,\forall \overline{P} \in \Phi(\Pi)$;
- $P' < P \,\forall P' \in \Pi \Rightarrow \overline{P} < P \,\forall \overline{P} \in \Phi(\Pi)$;
- $P' \lesseqgtr P \,\forall P' \in \Pi \Rightarrow \overline{P} \lesseqgtr P \,\forall \overline{P} \in \Phi(\Pi)$.

4.1.4 Consistency

The three axioms are consistent in the sense that they do not contradict each others. A trivial preorder $\lesssim$ for which all axioms are satisfied is the one such that all performances are equivalent, regardless of what S and Φ are.

4.2. Sufficient Conditions for Score-Based Rankings

We can now connect the axioms and scores and give sufficient conditions to satisfy our axioms. The proofs for the three following theorems are given in Sec. A.5.

The 1st theorem explains how performance orderings $\lesssim$ can be induced from scores X, which allows capitalizing on the natural order $\leq$ on $\mathbb{R}$ to obtain a preorder $\lesssim$ on $\mathbb{P}_{(\Omega,\Sigma)}$.

Theorem 1 (Sufficient condition for Axiom 1)**.** *A binary relation $\lesssim_X$ on $\mathbb{P}_{(\Omega,\Sigma)}$ induced by a score X as $P_1 \lesssim_X P_2$* iif *either $P_1 = P_2$ or $P_1 \in \mathrm{dom}(X)$ and $P_2 \in \mathrm{dom}(X)$ and $X(P_1) \leq X(P_2)$, is a preorder satisfying Axiom 1.*

The 2nd theorem makes the connection between the properties of the scores X and the satisfaction S.

Theorem 2 (Sufficient condition for Axiom 2)**.** *If a score X satisfies $\min_{\omega\in E} S(\omega) \leq X(P) \leq \max_{\omega\in E} S(\omega)$ for all events $E \in \Sigma$ and all performances $P \in \mathrm{dom}(X)$ such that $P(E) = 1$, then the ordering $\lesssim_X$ satisfies Axiom 2.*

The 3rd theorem makes the connection between the properties of the scores X and the function Φ.

Theorem 3 (Sufficient condition for Axiom 3)**.** *If a score X is such that $\Pi \subseteq \mathrm{dom}(X) \Rightarrow \Phi(\Pi) \subseteq \mathrm{dom}(X)$ and $\min_{P\in\Pi} X(P) \leq X(\overline{P}) \leq \max_{P\in\Pi} X(P)$ for all $\Pi \subseteq \mathrm{dom}(X)$ and all $\overline{P} \in \Phi(\Pi)$, then the ordering $\lesssim_X$ satisfies Axiom 3.*

4.3. The Ranking Scores: Solutions for $\Phi = \mathrm{conv}$

We now aim to define scores that satisfy our theorems and thus our axioms. We provide such scores in the case of $\Phi = \mathrm{conv}$, shown particularly relevant in our catalog of practical examples (see Sec. A.2). Beyond that, we include the application-specific preferences modeled by the random variable I. We introduce a new family of scores, the *ranking scores* R_I, that are parameterized by the *importance*, a non-negative random variable $I \neq 0$:

$$R_I : \mathrm{dom}(R_I) \to [s_{min,\Omega}, s_{max,\Omega}] :$$
$$P \mapsto R_I(P) = \frac{\mathbf{E}_P[IS]}{\mathbf{E}_P[I]} = \frac{\sum_{\omega\in\Omega} I(\omega)S(\omega)P(\{\omega\})}{\sum_{\omega\in\Omega} I(\omega)P(\{\omega\})}$$

where $\mathrm{dom}(R_I) = \{P \in \mathbb{P}_{(\Omega,\Sigma)} : \mathbf{E}_P[I] \neq 0\}$ and the second equality holds in the common case where Ω is finite and $\Sigma = 2^\Omega$. The expected satisfaction X_S^E corresponds to R_I when all samples are equally important.

The scores R_I and the performance orderings $\lesssim_{R_I}$ satisfy the conditions of Theorem 1, 2, and 3 for $\Phi = \mathrm{conv}$ (proofs in Sec. A.6.1). Thus, the performance orderings $\lesssim_{R_I}$ induced by the ranking scores satisfy all our axioms. Hence, the chosen name. Remarkably, these scores are universal: they can be used for performance-based ranking, regardless of the sample space Ω.

Properties. Let us now give some selected properties (proofs are in Sec. A.6.1). The first one clarifies how the importance can be interpreted. Consider any random experiment for which the set of possible outcomes is Ω, their distribution being given by P. One can choose to *filter* the outcomes as follows: run the experiment, and if the outcome is ω, end the experiment with probability $I(\omega)/\sum_{\omega'\in\Omega} I(\omega')$, otherwise restart everything. The new distribution of outcomes is given by the operation $\mathrm{filter}_I(P)$.

Property 1. *The ranking scores can be decomposed into an operation on performances considering the importance I and a score (the expected satisfaction) considering the satisfaction S. We have $R_I = X_S^E \circ \mathrm{filter}_I$, with*

$$\mathrm{filter}_I : \mathrm{dom}(R_I) \to \mathbb{P}_{(\Omega,\Sigma)} :$$
$$P \mapsto \left(\Sigma \to [0,1] : E \mapsto \frac{\sum_{\omega\in E} P(\{\omega\})I(\omega)}{\sum_{\omega\in\Omega} P(\{\omega\})I(\omega)}\right).$$

So far, the satisfaction was fixed. But what if one hesitates about the objective of the task, *i.e.* with its modeling through S? The next property clarifies what really matters.

Property 2. *Linearly transforming the satisfaction results in the same linear transformation of the ranking score. It is something that does not affect the ordering.*

The next property is about the scale invariance of I.

Property 3. $R_{kI} = R_I, \forall k \neq 0$. *We can thus restrict the study of ranking scores to the case in which the total importance $\sum_{\omega\in\Sigma} I(\omega)$ is constant.*

We can go further in the case of a binary satisfaction.

Property 4. *For a binary satisfaction, the performance ordering induced by a ranking score is insensitive to the uniform scaling of the importance given to the unsatisfying ($S^{-1}(0)$) or to the satisfying ($S^{-1}(1)$) samples.*

Most scores obtained by averaging or integrating ranking scores are not themselves ranking scores, and are unsuited for ranking. Thanks to Theorem 2, we know that the performance orderings induced by them satisfy Axiom 2. However, most often, they do not satisfy Axiom 3. Yet, two exceptions occur in the case of a binary satisfaction.

Property 5. *If I is the arithmetic mean of I_1 and I_2, then R_I is the f-mean of R_{I_1} and R_{I_2} with $f : x \mapsto x^{-1}$,* i.e., *the harmonic mean, when $S(\omega) = 1 \Rightarrow I_1(\omega) = I_2(\omega)$.*

Property 6. *If I is the arithmetic mean of I_1 and I_2, then R_I is the f-mean of R_{I_1} and R_{I_2} with $f : x \mapsto (1-x)^{-1}$, when $S(\omega) = 0 \Rightarrow I_1(\omega) = I_2(\omega)$.*

For any homogeneous binary relation $\mathcal{R}$ (*e.g.* $\leq, <, =, >, \geq, \ldots$) on $\mathbb{R}$, let us define the set $\phi_{\mathcal{R}}(P) = \{P' \in \mathbb{P}_{(\Omega,\Sigma)} : R_I(P')\mathcal{R}R_I(P)\}$. Based on these sets, the following property helps to understand why the ranking scores are suitable when $\Phi = \mathrm{conv}$. This is indeed related to the fact that the scores R_I are pseudolinear functions.

Property 7. *With the performance orderings $\lesssim_{R_I}$ induced by the ranking scores R_I and for any given performance $P \in \mathbb{P}_{(\Omega,\Sigma)}$, the worse set $\phi_<(P)$* (a.k.a. *the strictly lower contour set), the worse or equivalent set $\phi_\leq(P)$ (the lower contour set), the equivalent set $\phi_=(P)$, the better or equivalent set $\phi_\geq(P)$ (the upper contour set) and the better set $\phi_>(P)$ (the strict upper contour set) are all convex.*

5. Observations for Two-Class Classification

We now examine the particular case of two-class crisp classification. First, we compare the classical formulation of this task with ours. Then, we examine more closely what our ranking theory teaches us for this task.

5.1. Classical vs. our Formulation

Limitation of the classical formulation. Usual two-class classification settings first define a set $\mathbb{C} = \{c_-, c_+\}$ in which c_- and c_+ are named *negative class* and *positive class*. Then, pairs $(y, \hat{y}) \in \mathbb{C}^2$ composed of a "ground truth" y and a "prediction" $\hat{y}$ are interpreted as: a *true negative* $tn = (c_-, c_-)$, a *false positive* $fp = (c_-, c_+)$ (type I error), a *false negative* $fn = (c_+, c_-)$ (type II error), and a *true positive* $tp = (c_+, c_+)$. Finally, scores are defined as formulas involving these elements, upon which rankings of entities are based. However, there is no mathematical guarantee of the meaningfulness of such rankings.

Advantage of our formulation. In our framework, the two-class classification settings are as follows. We consider the sample space $\Omega = \{tn, fp, fn, tp\}$ and the event space $\Sigma = 2^\Omega$. The samples tn, fp, fn, and tp are interpreted as in the classical case. The natural choice for the satisfaction is such that $S(fp) = S(fn) = 0$ and $S(tn) = S(tp) = 1$. After modeling application-specific preferences through a choice of importances I, we can derive rankings scores R_I to rank entities based on their performances, with the certainty that the ranking is mathematically valid.

Link between the two formulations. Fortunately, the two formulations can be easily connected, which allows translating concepts from one to the other whenever necessary. Indeed, to go from our formulation to the classical one, we only need to define a "ground-truth" random variable $Y : \Omega \to \mathbb{C}$ such that $Y(tn) = Y(fp) = c_-$ and $Y(fn) = Y(tp) = c_+$, as well as the "prediction" random variable $\hat{Y} : \Omega \to \mathbb{C}$ such that $\hat{Y}(tn) = \hat{Y}(fn) = c_-$ and $\hat{Y}(fp) = \hat{Y}(tp) = c_+$. Conversely, moving from the classical formulation to ours just requires considering $\Omega = \mathbb{C}^2$ and $S = \mathbf{1}_{Y=\hat{Y}}$. This link is represented in Fig. A.7.1.

5.2. What Does our Ranking Theory Teach us?

We now examine some common scores used in the literature for two-class classification, from the ranking standpoint. Methodologically, we choose the set of numerical quantities that were listed in a recent review [47], and focus on those that are scores (this excludes the base measures and the 1st level measures as they are called in that paper).

We consider three sets of performances: (1) all performances $\mathbb{P}_{(\Omega,\Sigma)}$; (2) all performances for fixed and unbalanced priors (we set arbitrarily a positive prior of 0.2); (3) all performances for fixed and balanced priors. For each score X and each set of performances Π, we report in Tab. 1 the result of three tests on X as well as the minimum and maximum rank correlations X has with the ranking scores.

The tests are the following. The 1st test determines if $\lesssim_X$ satisfies Axiom 2 when the set $\mathbb{P}_{(\Omega,\Sigma)}$ is restricted to Π. The 2nd and 3rd tests relate to Theorem 3 (and thus to Axiom 3). The 2nd test determines if, for any subset Π' of Π, $\max_{P' \in \Pi'} X(P')$ is greater or equal to $X(P)$ for all performances $P \in \mathrm{conv}(\Pi')$. The 3rd test determines if, for any subset Π' of Π, $\min_{P' \in \Pi'} X(P')$ is less or equal to $X(P)$ for all performances $P \in \mathrm{conv}(\Pi')$.

For the rank correlations, we first try to determine analytically if the score is monotonically increasing or decreasing with one of the ranking scores. If it is the case, we report a maximum correlation of 1 or a minimum correlation of -1, respectively. The proofs can be found in Secs. A.7.3 and A.7.4. Otherwise, we report empirical values obtained by optimizing Kendall's τ. We consider a uniform distribution of performances within Π. Kendall's τ is computed using a function provided in SciPy [52], with its default parameters, and fed with the values of R_I and X for about 6,550 performances regularly placed in Π. Note that Kendall's τ is not a continuous function of I when estimated on a finite set of performances. We designed a custom optimizer to estimate τ, detailed in Sec. A.7.2.

The result of our review is given in Tab. 1, where the scores have been grouped in three categories: in green are the scores satisfying the three tests in all cases. In orange are those that satisfy the three tests only for fixed priors. In black are the others. We draw 4 conclusions. (1) Performance orderings satisfying our axioms can be induced by several classical scores for two-class classification (in green). (2) There exist scores that are commonly used in the

Table 1. Properties of some common scores defined in the literature for two-class crisp classification. The symbol †indicates a value that has been obtained theoretically, the others have been obtained empirically. Our conclusion is that, for the purpose of ranking, the scores in green can always be used, those in orange should only be used when the priors are fixed, and those in black cannot be used even when the priors are fixed. See Sec. 5 for the detailed description.

	without any constraint: all performances					with constraint: positive prior = 0.2					with constraint: positive prior = 0.5				
score	1st test	2nd test	3rd test	τ_{min}	τ_{max}	1st test	2nd test	3rd test	τ_{min}	τ_{max}	1st test	2nd test	3rd test	τ_{min}	τ_{max}
Accuracy	**V**	**V**	**V**	**0.469**	**1**†	**V**	**V**	**V**	**0.157**	**1**†	**V**	**V**	**V**	**0.505**	**1**†
F-score for $\beta = 0.5$	**V**	**V**	**V**	**0.079**	**1**†	**V**	**V**	**V**	**0.451**	**1**†	**V**	**V**	**V**	**0.352**	**1**†
F-score for $\beta = 1.0$	**V**	**V**	**V**	**0.161**	**1**†	**V**	**V**	**V**	**0.352**	**1**†	**V**	**V**	**V**	**0.194**	**1**†
F-score for $\beta = 2.0$	**V**	**V**	**V**	**0.079**	**1**†	**V**	**V**	**V**	**0.194**	**1**†	**V**	**V**	**V**	**0.072**	**1**†
Negative Predictive Value (NPV)	**V**	**V**	**V**	**0.000**	**1**†	**V**	**V**	**V**	**0.503**	**1**†	**V**	**V**	**V**	**0.503**	**1**†
Positive Predictive Value (PPV)	**V**	**V**	**V**	**0.000**	**1**†	**V**	**V**	**V**	**0.503**	**1**†	**V**	**V**	**V**	**0.503**	**1**†
True Negative Rate (TNR)	**V**	**V**	**V**	**0.000**	**1**†	**V**	**V**	**V**	**0.000**	**1**†	**V**	**V**	**V**	**0.000**	**1**†
True Positive Rate (TPR)	**V**	**V**	**V**	**0.000**	**1**†	**V**	**V**	**V**	**0.000**	**1**†	**V**	**V**	**V**	**0.000**	**1**†
Balanced Accuracy	V	X	X	0.486	0.713	V	V	V	0.504	1†	V	V	V	0.505	1†
Cohen's κ	X	X	X	0.476	0.697	V	V	V	0.503	1†	V	V	V	0.505	1†
Informedness	V	X	X	0.486	0.713	V	V	V	0.504	1†	V	V	V	0.505	1†
Positive Likelihood Ratio (PLR)	V	X	X	0.420	0.677	V	V	V	0.491	1†	V	V	V	0.491	1†
Probability of True Negative (PTN)	X	V	V	-0.007	0.818	V	V	V	0.000	1†	V	V	V	0.000	1†
Probability of True Positive (PTP)	X	V	V	-0.006	0.818	V	V	V	0.000	1†	V	V	V	0.000	1†
Chance in Cohen's κ	X	X	X	0.194	0.498	X	V	V	-0.157	0.849	V	V	V	0†	0†
Error Rate	X	V	V	-1†	-0.469	X	V	V	-1†	-0.157	X	V	V	-1†	-0.505
False Discovery Rate (FDR)	X	V	V	-1†	0.000	X	V	V	-1†	-0.503	X	V	V	-1†	-0.503
False Negative Rate (FNR)	X	V	V	-1†	0.000	X	V	V	-1†	0.000	X	V	V	-1†	0.000
False Omission Rate (FOR)	X	V	V	-1†	0.000	X	V	V	-1†	-0.503	X	V	V	-1†	-0.503
False Positive Rate (FPR)	X	V	V	-1†	0.000	X	V	V	-1†	0.000	X	V	V	-1†	0.000
Geometric mean of TNR and TPR	V	X	X	0.461	0.653	V	X	V	0.503	0.831	V	X	V	0.503	0.830
Markedness	V	X	X	0.486	0.713	V	X	X	0.418	0.887	V	X	X	0.503	0.913
Matthews Correlation Coefficient (MCC)	V	X	X	0.503	0.746	V	X	X	0.458	0.944	V	X	X	0.503	0.963
Negative Likelihood Ratio (NLR)	X	X	X	-0.677	-0.418	X	V	V	-1†	-0.491	X	V	V	-1†	-0.491
Odds Ratio (OR)	V	X	X	0.499	0.671	V	X	X	0.503	0.894	V	X	X	0.503	0.892
Rate of positive predictions	X	V	V	-0.469	0.469	X	V	V	-0.849	0.157	X	V	V	-0.504	0.505
Sensitivity Index Estimate (d')	V	X	X	0.502	0.786	V	X	X	0.503	0.926	V	X	X	0.503	0.924

literature (in orange) that cannot be used for ranking classifiers, unless the priors are fixed. (3) There exist scores that are often used to compare classifiers (*e.g.*, the geometric mean of TNR and TPR, the markedness, Matthews Correlation Coefficient, the Odds Ratio) but that cannot be used to rank, even when the priors are fixed. (4) In all studied cases where our axioms are satisfied, there is a perfect correlation with a ranking score. This shows that our family of scores covers at least a broad part of the needs. Only one exception occurred (in gray): the accuracy achievable by chance, as considered by Cohen in the definition of his κ [53], when the classes are balanced. In this case, the score takes a constant value, that is why it satisfies our axioms.

6. Conclusion

We present a mathematical framework for a theory of performance-based ranking, that comes with several practical benefits. Notably, it provides a universal language to properly define tasks and characterize applications. Also, separating these concepts allows evaluating entities (algorithms, devices, *etc.*) independently of application-specific preferences, as an entity can be used in various use cases, and various entities can be candidates for a same use case.

Importantly, our axioms are crucial for organizers of challenges to ensure sound rankings, to avoid, *e.g.*, that the relative order between two methods changes when a new method appears and thus prevents drawing perennial conclusions on which is better. Our axioms are minimal requirements to satisfy, and we provide practical theorems to help check if that is the case.

Besides, practitioners often face many scores in the literature for a task, but none comes with an analysis of their usability for ranking. We prove that our theory can help find appropriate scores (we propose an infinite family of them) or legitimate existing ones. As the chosen performance score for ranking may have a major impact on the growth of a research field, our framework clarifies best practices.

We can further deepen our axiomatic framework and infinite family of scores for ranking classifiers. In [4], we particularize extensively this framework to binary classification and present the *Tile*, a visualization tool that organizes these scores (among which the precision PPV, the true positive rate TPR, the true negative rate TNR, the scores F_β, and the accuracy A) in a single plot. Finally, in [5], we provide a comprehensive guide to using the Tile according to four practical scenarios. For that purpose, we present different Tile flavors on a real example, analyzing and ranking 74 segmentation classifiers. We now wish to build upon this trilogy of papers to reach various research communities and impact significantly their way of establishing performance-based rankings.

Acknowledgments. S. Piérard and A. Halin are funded by the SPW EER, Wallonia, Belgium (grant 2010235, ARIAC by DIGITALWALLONIA4.AI); A. Cioppa and A. Deliège (grant T.0065.22) by the F.R.S.-FNRS.

References

[55] Davide Ballabio, Francesca Grisoni, and Roberto Todeschini. Multivariate comparison of classification performance measures. *Chemom. Intell. Lab. Syst.*, 174:33–44, 2018.

[47] Gürol Canbek, Seref Sagiroglu, Tugba Taskaya Temizel, and Nazife Baykal. Binary classification performance measures/metrics: A comprehensive visualized roadmap to gain new insights. In *Int. Conf. Comput. Sci. Eng. (UBMK)*, pages 821–826, Antalya, Turkey, 2017.

[41] Anthony Cioppa, Silvio Giancola, Vladimir Somers, Floriane Magera, Xin Zhou, Hassan Mkhallati, Adrien Deliège, Jan Held, Carlos Hinojosa, Amir M. Mansourian, Pierre Miralles, Olivier Barnich, Christophe De Vleeschouwer, Alexandre Alahi, Bernard Ghanem, Marc Van Droogenbroeck, Abdullah Kamal, Adrien Maglo, Albert Clapés, Amr Abdelaziz, Artur Xarles, Astrid Orcesi, Atom Scott, Bin Liu, Byoungkwon Lim, Chen Chen, Fabian Deuser, Feng Yan, Fufu Yu, Gal Shitrit, Guanshuo Wang, Gyusik Choi, Hankyul Kim, Hao Guo, Hasby Fahrudin, Hidenari Koguchi, Håkan Ardö, Ibrahim Salah, Ido Yerushalmy, Iftikar Muhammad, Ikuma Uchida, Ishay Be'ery, Jaonary Rabarisoa, Jeongae Lee, Jiajun Fu, Jianqin Yin, Jinghang Xu, Jongho Nang, Julien Denize, Junjie Li, Junpei Zhang, Juntae Kim, Kamil Synowiec, Kenji Kobayashi, Kexin Zhang, Konrad Habel, Kota Nakajima, Licheng Jiao, Lin Ma, Lizhi Wang, Luping Wang, Menglong Li, Mengying Zhou, Mohamed Nasr, Mohamed Abdelwahed, Mykola Liashuha, Nikolay Falaleev, Norbert Oswald, Qiong Jia, Quoc-Cuong Pham, Ran Song, Romain Hérault, Rui Peng, Ruilong Chen, Ruixuan Liu, Ruslan Baikulov, Ryuto Fukushima, Sergio Escalera, Seungcheon Lee, Shimin Chen, Shouhong Ding, Taiga Someya, Thomas B. Moeslund, Tianjiao Li, Wei Shen, Wei Zhang, Wei Li, Wei Dai, Weixin Luo, Wending Zhao, Wenjie Zhang, Xinquan Yang, Yanbiao Ma, Yeeun Joo, Yingsen Zeng, Yiyang Gan, Yongqiang Zhu, Yujie Zhong, Zheng Ruan, Zhiheng Li, Zhijian Huang, and Ziyu Meng. SoccerNet 2023 challenges results. *Sports Eng.*, 27(2):1–18, 2024.

[53] Jacob Cohen. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.*, 20(1):37–46, 1960.

[56] Andrew R. Conn, Katya Scheinberg, and Luis N. Vicente. *Introduction to Derivative-Free Optimization*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2009.

[48] Brooke A. Davey and Hilary A. Priestley. *Introduction to Lattices and Order*. Camb. Univ. Press, second edition, 2002.

[24] Jesse Davis and Mark Goadrich. The relationship between precision-recall and ROC curves. In *Int. Conf. Mach. Learn. (ICML)*, pages 233–240, Pittsburgh, Pennsylvania, 2006.

[30] Tom Fawcett. An introduction to ROC analysis. *Pattern Recognit. Lett.*, 27(8):861–874, 2006.

[54] Edward B. Fowlkes and Colin L. Mallows. A method for comparing two hierarchical clusterings. *J. Am. Stat. Assoc.*, 78(383):553–569, 1983.

[22] Nil Goyette, Pierre-Marc Jodoin, Fatih Porikli, Janusz Konrad, and Prakash Ishwar. A novel video dataset for change detection benchmarking. *IEEE Trans. Image Process.*, 23 (11):4663–4679, 2014.

[49] George Grätzer. *General Lattice Theory*. Birkhäuser Basel, second edition, 2003.

[5] 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.

[50] Maurice George Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81, 1938.

[32] Andrey Nikolaevich Kolmogorov. *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Springer Berl. Heidelb., Berlin, Ger., 1933.

[33] Andrey Nikolaevich Kolmogorov. *Foundations of the theory of probability*. Chelsea Publishing Company, New York City, NY, USA, 1950.

[42] Matej Kristan, Jiri Matas, Ale Leonardis, Michael Felsberg, Luka Cehovin, Gustavo Fernandez, Toma Vojir, Gustav Hager, Georg Nebehay, Roman Pflugfelder, Abhinav Gupta, Adel Bibi, Alan Lukezic, Alvaro Garcia-Martin, Amir Saffari, Alfredo Petrosino, and Andrés Solis Montero. The visual object tracking VOT2015 challenge results. In *IEEE Int. Conf. Comput. Vis. Work. (ICCV Work.)*, pages 564–586, Santiago, Chile, 2015.

[43] Matej Kristan, Jiri Matas, Pavel Tokmakov, Michael Felsberg, Luka Čehovin Zajc, Alan Lukežič, Khanh-Tung Tran, Xuan-Son Vu, Johanna Bjorklund, Hyung Jin Chang, and Gustavo Fernandez. The second visual object tracking segmentation VOTS2024 challenge results. In *Eur. Conf. Comput. Vis. Work. (ECCV Work.)*, Milan, Italy, 2024.

[44] Lena Maier-Hein, Matthias Eisenmann, Annika Reinke, Sinan Onogur, Marko Stankovic, Patrick Scholz, Tal Arbel, Hrvoje Bogunovic, Andrew P. Bradley, Aaron Carass, Carolin Feldmann, Alejandro F. Frangi, Peter M. Full, Bram van Ginneken, Allan Hanbury, Katrin Honauer, Michal Kozubek, Bennett A. Landman, Keno März, Oskar Maier, Klaus Maier-Hein, Bjoern H. Menze, Henning Müller, Peter F. Neher, Wiro Niessen, Nasir Rajpoot, Gregory C. Sharp, Korsuk Sirinukunwattana, Stefanie Speidel, Christian Stock, Danail Stoyanov, Abdel Aziz Taha, Fons van der Sommen, Ching-Wei Wang, Marc-André Weber, Guoyan Zheng, Pierre Jannin, and Annette Kopp-Schneider. Why rankings of biomedical image analysis competitions should be interpreted with care. *Nat. Commun.*, 9(1):1–13, 2018.

[45] Tran Thien Dat Nguyen, Hamid Rezatofighi, Ba-Ngu Vo, Ba-Tuong Vo, Silvio Savarese, and Ian Reid. How trustworthy are performance evaluations for basic vision tasks? *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(7):8538–8552, 2023.

[1] Sébastien Piérard and Marc Van Droogenbroeck. Summarizing the performances of a background subtraction algorithm measured on several videos. In *IEEE Int. Conf. Image Process. (ICIP)*, pages 3234–3238, Abu Dhabi, United Arab. Emir., 2020.

[4] 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.

[38] Marina Sokolova and Guy Lapalme. A systematic analysis of performance measures for classification tasks. *Inf. Process. & Manag.*, 45(4):427–437, 2009.

[51] Charles Spearman. The proof and measurement of association between two things. *Am. J. Psychol.*, 15(1):72–101, 1904.

[46] Putnam P. Texel. Measure, metric, and indicator: An object-oriented approach for consistent terminology. *Proc. IEEE Southeastcon*, pages 1–5, 2013.

[52] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C. J. Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, Aditya Vijaykumar, Alessandro Pietro Bardelli, Alex Rothberg, Andreas Hilboll, Andreas Kloeckner, Anthony Scopatz, Antony Lee, Ariel Rokem, C. Nathan Woods, Chad Fulton, Charles Masson, Christian Häggström, Clark Fitzgerald, David A. Nicholson, David R. Hagen, Dmitrii V. Pasechnik, Emanuele Olivetti, Eric Martin, Eric Wieser, Fabrice Silva, Felix Lenders, Florian Wilhelm, G. Young, Gavin A. Price, Gert-Ludwig Ingold, Gregory E. Allen, Gregory R. Lee, Hervé Audren, Irvin Probst, Jörg P. Dietrich, Jacob Silterra, James T. Webber, Janko Slavič, Joel Nothman, Johannes Buchner, Johannes Kulick, Johannes L. Schönberger, José Vinícius de Miranda Cardoso, Joscha Reimer, Joseph Harrington, Juan Luis Cano Rodríguez, Juan Nunez-Iglesias, Justin Kuczynski, Kevin Tritz, Martin Thoma, Matthew Newville, Matthias Kümmerer, Maximilian Bolingbroke, Michael Tartre, Mikhail Pak, Nathaniel J. Smith, Nikolai Nowaczyk, Nikolay Shebanov, Oleksandr Pavlyk, Per A. Brodtkorb, Perry Lee, Robert T. McGibbon, Roman Feldbauer, Sam Lewis, Sam Tygier, Scott Sievert, Sebastiano Vigna, Stefan Peterson, Surhud More, Tadeusz Pudlik, Takuya Oshima, Thomas J. Pingel, Thomas P. Robitaille, Thomas Spura, Thouis R. Jones, Tim Cera, Tim Leslie, Tiziano Zito, Tom Krauss, Utkarsh Upadhyay, Yaroslav O. Halchenko, and Yoshiki Vázquez-Baeza. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods*, 17(3):261–272, 2020.

3. Intermediate rankings (3rd article)

This chapter reproduces the following article [3]:

> S. Piérard, A. Deliège, and M. 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), IEEE, June 2026.

It covers both the notions of ***performance*** and ***ranking***. The supplementary material of this article is reproduced in Chapter B.

Context

In the previous chapter, we saw that the rankings of entities do not necessarily make sense in terms of performances, but that if we know the values for the importance I, then the preorder on performances induced by the ranking scores R_I allows one to rank in a meaningful way. However, there remains the question of how to rank when the importance is uncertain. Picking a score arbitrarily can be misleading in terms of the expected ranking. In particular, the ranking induced by the F-score F_1, which is the harmonic mean between the precision Pr and the recall Re, is rarely a good trade-off between the rankings induced by Pr and Re.

Motivation

The F_1 score (*a.k.a.* Dice) is definitely standard in computer vision, serving as ranking criteria in benchmarks and challenges on tasks such as image/video classification in medical imaging (COVID-CT, PAIP, ECDP, TN-SCUI), and behavior analysis (ABAW), segmentation in video surveillance (IWDD), medical imaging (3D-TEETH-SEG, DRAC, UNICORN), and visual quality control (VAND), detection in medical imaging (MIDOG, UNICORN). Clarifying that F_1 is rarely a good trade-off between the rankings induced by Pr and Re, and explaining how one can obtain the optimal trade-off is therefore of first importance.

Main contributions

The idea developed in this article is to start by specifying a set of ranking scores of interest and then determine the optimal trade-off between them. We do this in the specific case of two-class crisp classification, and seek the optimal trade-off between precision (*i.e.*,

the probability that a positive decision is correct, $Pr(P) = P(S = 1|\hat{Y} = c_+)$ and recall (*i.e.*, the probability that a decision is correct for a positive, $Re(P) = P(S = 1|Y = c_+)$.

The issue is that a trade-off obtained by averaging values is generally not appropriate because an average of scores inducing meaningful performance orderings does not necessarily produce a score that itself induces a meaningful ordering (see Section 4.3 in [2]). In particular, means of ranking scores are generally not ranking scores. Instead, I propose to consider a trade-off from the ranking point of view, looking for it only among ranking scores. I consider the manifold of performance orderings induced by the ranking scores R_I, for which the distances between orderings are closely related to rank correlations (Kendall's τ and Spearman's ρ), and I motivate the choice of defining the optimal trade-off as the Karcher-Fréchet mean (*i.e.*, the ranking score that minimizes the Fréchet variance) relative to the Kendall distance (*a.k.a.* bubble-sort distance).

As shown in the article, the optimal tradeoff between the rankings induced by Pr and Re is a ranking induced by some F_β, but in general $\beta \neq 1$.

Perspectives

While only the trade-off between precision and recall is studied in this article, the principles appear to be generic. The same approach could be used to define the trade-off between any set of ranking scores, no matter if the task is a two-class classification task or any other task. However, this generalization is left for future work.

Remarks

1. The manifold of the performance orderings induced by the ranking scores deforms when the reference distribution of performances changes: some rankings are getting closer, others are drifting away. This sensitivity of the manifold to the distribution of performances explains why the optimal trade-off depends on this distribution.

2. The theory presented in this article can be extended to different downstream objectives. For example, the organizers of the CADA-RRE challenge (see Fig. 3.1) chose F_2 to rank the participants. We can assume that, through this choice, the organizers wanted to weight 4 times more the recall than the precision as $F_2^{-1} = {}^1\!/_5\, Pr^{-1} + {}^4\!/_5\, Re^{-1}$. We can generalize Eq. 11 in the article by weighting d_τ^2 and Eq. 15 by replacing the median by a quantile. For instance, to weight 4 times more Re than Pr, from the *point of view of values* one uses $b = 0.8$ in Eq. 1 (*i.e.*, $\beta = 2$, as in CADA-RRE). But, from the *point of view of ranks*, we can take the 80^{th} percentile of ϑ values; this leads to $\beta = 0.914$.

3. For given priors, our heuristic (Eq. 30) can be depicted in ROC: in Fig. 3.2, the suggested β is read at the centroid of the point cloud, which corresponds to the average performance (summarization).

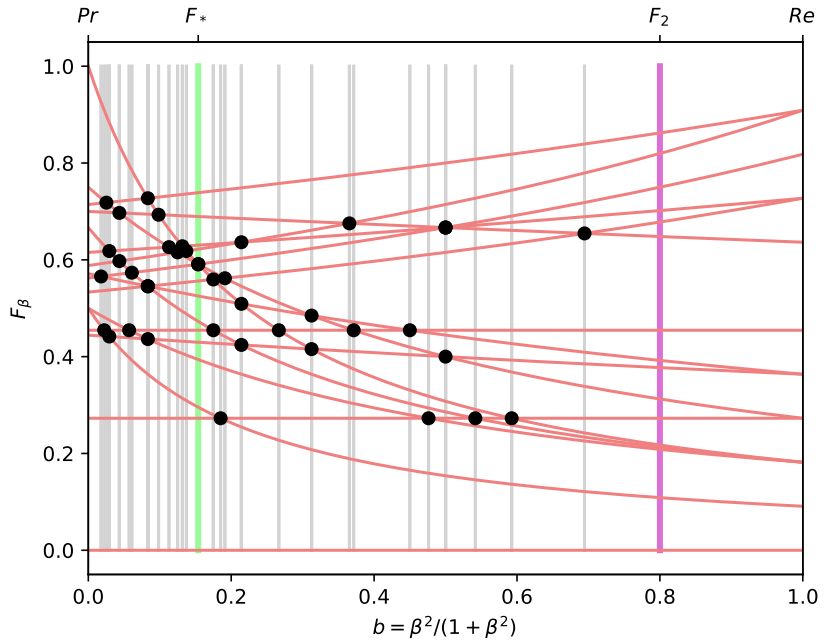


Figure 3.1.: Leaderboard for the challenge CADA-RRE (medical imaging). The organizers arbitrarily chose F_2 to rank the participants. The red curves shown on this graph correspond to $F_\beta(P)$ *w.r.t.* β for the 16 performances P to rank. We see that the ranking with F_2 perfectly mimics Re, totally ignoring Pr! The values ϑ (Eq. 14 in the paper) for the swaps in the ranking are shown by the black dots and the gray vertical lines, which delimit 44 different rankings. There is the same amount of black dots on both sides of the optimal compromise (green vertical bar, $\beta = 0.426$ by Eq. 15).

4. The computation time is not a problem in practice. Once the classifiers to rank have been evaluated, all the quantities that appear in Eq. 14 (*i.e.*, the probability of a false positive $PFP(P)$, the probability of a false negative $PFN(P)$, and the probability of a true positive $PTP(P)$) are known for all of them. The straightforward way of computing the optimal value for β requires calculating Eq. 14 for each swap (at most $n(n-1)/2$ times). This has the same time complexity as the computation of Kendall's τ value, *i.e.* the counting of the swaps. This is not a problem at all in practice. It is reasonable to assume that there will never be a benchmark with more than $n = 1{,}000$ classifiers. In this case, on my laptop, the calculation of Eq. 14 and 15 takes about 0.3 seconds (without parallelization).

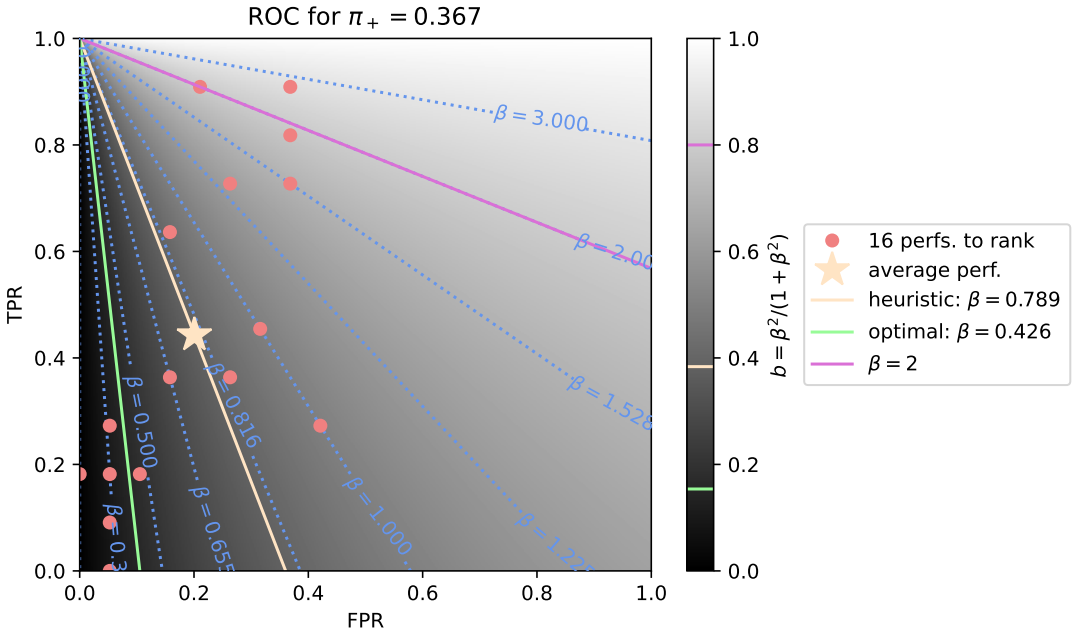


Figure 3.2.: Visualization of the heuristic in the ROC space. This example corresponds to the challenge CADA-RRE. The centroid of the point cloud (average performance) suggests $\beta = 0.789$.

What Is the Optimal Ranking Score Between Precision and Recall? We Can Always Find It and It Is Rarely F_1

Sébastien Piérard, Adrien Deliège, and Marc Van Droogenbroeck
Montefiore Institute, University of Liège, Liège, Belgium
{S.Pierard,Adrien.Deliege,M.VanDroogenbroeck}@uliege.be

Abstract

Ranking methods or models based on their performance is of prime importance but is tricky because performance is fundamentally multidimensional. In the case of classification, precision and recall are scores with probabilistic interpretations that are both important to consider and complementary. The rankings induced by these two scores are often in partial contradiction. In practice, therefore, it is extremely useful to establish a compromise between the two views to obtain a single, global ranking. Over the last fifty years or so, it has been proposed to take a weighted harmonic mean, known as the F-score, F-measure, or F_β. Generally speaking, by averaging basic scores, we obtain a score that is intermediate in terms of values*. However, there is no guarantee that these scores lead to meaningful* rankings *and no guarantee that the rankings are good* tradeoffs *between these base scores. Given the ubiquity of F_β scores in the literature, some clarification is in order. Concretely: (1) We establish that F_β-induced rankings are meaningful and define a shortest path between precision- and recall-induced rankings. (2) We frame the problem of finding a tradeoff* between *two scores as an* optimization *problem expressed with Kendall rank correlations. We show that F_1 and its skew-insensitive version are far from being optimal in that regard. (3) We provide theoretical tools and a closed-form expression to* find *the optimal value for β for any distribution or set of performances, and we illustrate their use on six case studies.*

1. Introduction

The *precision* Pr (also called *positive predictive value*) and *recall* Re (also called *true positive rate*) are of first importance in classification and other related problems. These scores have a straightforward interpretation, as they give the probabilities of taking the correct decision when the predicted or ground-truth class is positive, respectively.

Unfortunately, working with several scores is often impractical. For this reason, it is common to average the values of Pr and Re. Most often, this is done with a harmonic mean, leading to the family of F-scores F_β, with $\beta \geq 0$:

$$F_\beta^{-1} = (1-b)\, Pr^{-1} + b\, Re^{-1} \quad \text{with } b = \frac{\beta^2}{1+\beta^2}\,. \tag{1}$$

These last decades, this family of scores has become very popular in the literature[1]. It forms a continuum between pre-

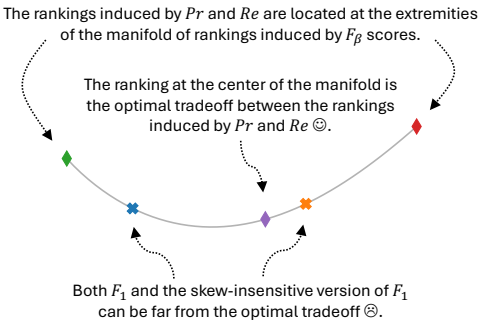


Figure 1. The manifold of rankings inducible with the F_β scores is a curve (drawn as ⌣) that depends on the set or distribution of performances that is considered. The distance along the curve is proportional to the number of swaps of neighboring classifiers needed to transform a ranking into another. The rankings induced by precision ♦ and recall ♦ are at the extremities of the manifold. The rankings induced by the traditional (balanced) F_1 ✖ and by the skew-insensitive version of F_1 ✖ [29] can be anywhere, depending on the set or distribution of performances that are compared, and cannot be considered as the optimal tradeoff between precision and recall. We consider that the ranking located at equal distance, along the curve, from precision and recall is the optimal tradeoff ♦ to rank classifiers.

[1] With the requests "F-measure + classification", "F-score + classification", and "F-beta + classification", Google Scholar reported on July 23th 2025 that there was about 315,000, 205,000, and 9,080 matching documents, respectively. The score F_1 is also popular in the CVPR community:

cision $Pr = F_0$ and recall $Re = F_\infty$. The traditional (balanced) F-score, F_1, gives equal weights to Pr and Re. But is F_1 really the optimal tradeoff between precision and recall? The answer to this question depends on the point of view.

The point of view of values. The F-scores are obviously between Pr and Re in the sense that, for any given performance P (*i.e.*, a confusion matrix), the value $F_\beta(P)$ varies monotonically with β, between $Pr(P)$ and $Re(P)$.

The point of view of ranks. This is the point of view that we develop in this paper. Although many authors use the F_1 or F_β scores to rank and compare methods, the claim that they are an appropriate compromise for ranking deserves to be studied. Indeed, several questions arise. How are the rankings induced by the F_β scores spatially organized? Are the F_β scores suitable to induce performance-based rankings? Do the F_β scores induce rankings that form a shortest path between those induced by Pr and Re? For which β(s) do we obtain the optimal tradeoff between the rankings induced by Pr and Re?

The objective of this paper is to provide answers, teased in Fig. 1, to all these questions. For that purpose, we develop the theory, and we provide experimental results for various distributions and sets of two-class classification performances. We summarize our contributions as follows:

1. We establish that the F_β-family is the right set of scores to consider when looking for an optimal tradeoff between Pr and Re for ranking, as, on the one hand, all F_β lead to meaningful rankings according to the theory of performance-based ranking [2] and, on the other hand, the rankings induced by F_β always form a shortest path between the rankings induced by Pr and Re.
2. We show that neither F_1, traditionally considered as a balanced compromise between Pr and Re, nor its skew-insensitive version (denoted by $SIVF$ in this paper) introduced in [29] are acceptable tradeoffs for ranking, in general. These scores can lead to rankings far from an optimal one.
3. We provide the theory and methods for defining and finding (with a closed-form expression) the optimal tradeoffs between the rankings induced by Pr and Re. These methods are then successfully applied to numerous sets and distributions of performances.

2. Related Work

2.1. Useful References

The meaning of F-scores is discussed in countless papers; see, for example, the work by Christen *et al.* [57] for a recent review about the F-scores. A plethora of other scores have been defined for classification. They can be found in several reviews [37, 38, 47, 55, 58–62], to cite only a few.

Regarding the ranking of methods or models based on their performances, theoretical foundations have been provided in [2] and the theory has been particularized to the case of two-class crisp classification in [4, 5]. When comparing performances with the same class priors, the isometrics depicted in the ROC (Receiver Operating Characteristic) space is a graphical representation of the performance ordering induced by a given score. The isometrics of Pr and F_1 have been depicted in [29] for different class priors.

There are only a few studies of the correlations between the rankings induced by scores in classification. Both Ferri *et al.* [59] and Liu *et al.* [63] used Pearson linear correlations and Spearman rank correlations. To the best of our knowledge, there is no study comparing classification scores based on Kendall rank correlations.

2.2. Building Upon the Related Work

2.2.1. Sometimes, F_1 Mimics the Rankings Induced by Pr or Re and Ignores the Other One

Let us consider the particular case in which the class priors are fixed and denote them by π_- and π_+ for the negative and positive classes, respectively. Extending the work of [29], we depict, in Fig. 2, the isometrics of Pr, F_β for different values of β, Re, and $SIVF$ in the ROC space. For all these scores, we found that the isometrics form pencils of lines with vertices located at $(FPR, TPR) = (-\ell, 0)$ with some $\ell \geq 0$. The performance orderings, and the resulting rankings, depend only on the location of the vertex, and thus on ℓ. For Pr, $\ell = 0$. For Re, $\ell = \infty$. For $SIVF$, $\ell = 1$. And for F_β,

$$\ell = \beta^2 \frac{\pi_+}{\pi_-} = \beta^2 \frac{\pi_+}{1 - \pi_+} . \tag{2}$$

When $\pi_+ \to 0$, we see that $\ell \to 0$ and that all F_β mimic the ranking of $Pr = F_0$ and ignore Re. Similarly, when $\pi_+ \to 1$, we see that $\ell \to \infty$ and that all F_β mimic the ranking of $Re = F_\infty$, ignoring Pr. We conclude that *a single value of β cannot provide the optimal ranking for all priors*.

2.2.2. All F_β Scores Lead to Meaningful Rankings, Without Any Constraint on the Performances

There are some conditions that a performance ordering must fulfill to be meaningful. They have been given in the form of axioms in [2]. We found that all these axioms are satisfied by the performance orderings induced by F_β, $\forall \beta \geq 0$. In other words, *all F_β lead to meaningful performance orderings*. It was not a foregone conclusion, because the scores that are, from the point of view of values, between two scores suitable for ranking are not guaranteed to be themselves suitable for ranking. For example, the performance orderings induced by the arithmetic and geometric means of Pr and

we manually estimated that about 10% of the papers accepted at CVPR 2025 refer to or use it.

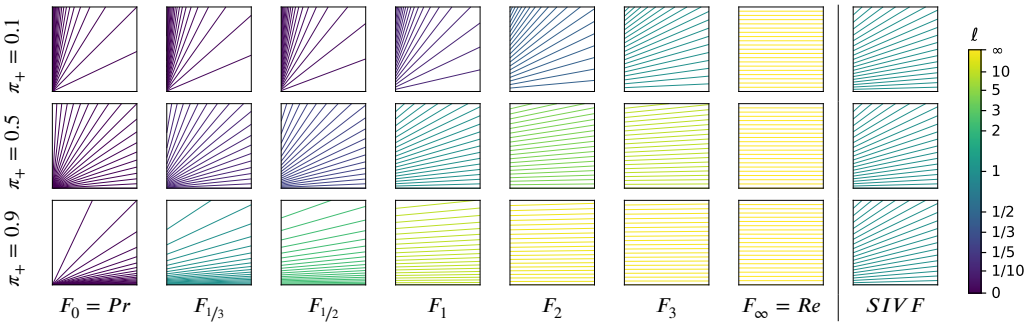


Figure 2. Isometrics of F_β and $SIVF$ in the ROC space $(FPR, TPR) \in [0,1]^2$, for three class priors. An isometric is a locus of equivalent performances, according to the score value. We plotted here those corresponding to the values $0.05, 0.10, 0.15, \ldots, 0.95$. The isometrics of F_β form a pencil of lines intersecting at $(FPR, TPR) = (-\ell, 0)$ with $\ell = \beta^2 \pi_+ / \pi_-$. The isometrics of $SIVF$ intersect at $(FPR, TPR) = (-1, 0)$, no matter what the class priors are. The score $SIVF$ induces the same performance ordering as F_β with $\beta^2 = \pi_- / \pi_+$.

Re do not satisfy the axioms of performance-based ranking. As the ranking induced by $SIVF$ is the same as the ranking induced by F_β when β^2 is equal to the skew ratio [29] π_- / π_+, $SIVF$ leads to a meaningful performance ordering when the priors are fixed. However, we found that it does not satisfy the axioms of [2] without this constraint.

3. Theory

3.1. Preliminaries

We use the notations of [2]. $\mathbb{P}_{(\Omega,\Sigma)}$ is the set of all possible performances P in two-class crisp classification, *i.e.* the set of all probability measures with a sample space Ω containing the four cases that can arise: tn (true negative), fp (false positive), fn (false negative), and tp (true positive). The negative and positive class priors are $\pi_- = P(\{tn, fp\})$ and $\pi_+ = P(\{fn, tp\})$, respectively.

We are interested in the rankings induced by scores. Scores are functions $X : \mathbb{P}_{(\Omega,\Sigma)} \to \mathbb{R} : P \mapsto X(P)$. The unconditional probabilities of a false positive, false negative, and true positive are given by the scores PFN, PFP, and PTP, respectively. The precision (Pr), also called positive predictive value (PPV), is defined as

$$Pr = PPV = F_0 : P \mapsto P(\{tp\} | \{fp, tp\}) . \tag{3}$$

The recall (Re), also called true positive rate (TPR), is

$$Re = TPR = F_\infty : P \mapsto P(\{tp\} | \{fn, tp\}) . \tag{4}$$

The skew-insensitive version of F_1 [29] is

$$SIVF = \frac{2TPR}{TPR + FPR + 1} , \tag{5}$$

where FPR is the false positive rate.

3.2. The Rankings and Their Manifold

Without loss of generality, let us first consider a finite set $\Pi = \{P_1, \ldots, P_n\} \subset \mathbb{P}_{(\Omega,\Sigma)}$ of performances (for infinitely many performances, we just need to consider a distribution[2] $\mathcal{P}$ instead of a set Π). The rankings of Π are its permutations, in finite number.

Let us consider a score X that induces meaningful rankings. Assuming that the value of X is defined for all $P \in \Pi$, and that X does not assign the same value to two performances of Π (absence of ties), X induces the ranking:

$$\mathbf{x} = (\mathrm{rank}_X(P_1), \ldots, \mathrm{rank}_X(P_n)) , \tag{6}$$

where $\mathrm{rank}_X(P)$ denotes the number of performances in Π for which X takes a value greater or equal to $X(P)$.

A set of scores (*e.g.* all F_β) corresponds to a set of points as written in Eq. (6). The Euclidean distance (*i.e.* along a straight line) between two rankings is their Spearman distance d_ρ, related to Spearman's ρ [51]. *The distance measured along the manifold of rankings is Kendall's distance d_τ (a.k.a. bubble-sort distance), related to Kendall's τ [50].* We can explain it as follows. Let us create the path graph that connects the rankings that are neighbors. Two rankings $\mathbf{x}_a$ and $\mathbf{x}_b$ are neighbors if and only if they differ only by one swap of two classifiers at consecutive ranks. The geodesic distance, *i.e.* the length of the shortest path, between $\mathbf{x}_a$ and $\mathbf{x}_b$, is the minimum number of swaps needed to transform $\mathbf{x}_a$ into $\mathbf{x}_b$, and it is proportional to the Kendall distance d_τ. For this reason, we consider this distance to be adequate for specifying the optimal tradeoff between two rankings.

[2] Technically, we need to consider a measurable space whose sample space is $\mathbb{P}_{(\Omega,\Sigma)}$. A performance distribution $\mathcal{P}$ then corresponds to a probability measure. Scores, defined as real functions on $\mathbb{P}_{(\Omega,\Sigma)}$, become random variables, and we can rigorously speak of correlation between scores.

As we observed in all our case studies that the manifold is curved, as depicted in Fig. 1, we will define the optimal tradeoff in terms of Kendall's distance d_τ or in terms of Kendall's rank correlation τ.

3.3. The Shortest Paths and the Means of Scores

The score F_β is the weighted harmonic mean between precision Pr and recall Re. The harmonic mean is the generalized f-mean, also called the regular mean [11], with $f : x \mapsto x^{-1}$. This has an important implication for the performance orderings induced by Pr, Re, and F_β.

Consider an interval $\xi \subseteq \mathbb{R}$, m scores $X_1, \ldots, X_m$ whose images are included in ξ, and a continuous strictly monotonic real function f defined on ξ. If a performance P_A is better than, equivalent to, or worse than a performance P_B according to all the scores $X_1, \ldots, X_m$, then so is it according to all their generalized f-means $\overline{X}$. This extends to weighted f-means. So, *if a performance P_A is better than, equivalent to, or worse than a performance P_B according to Pr and to Re, then so is it according to all the F_β scores*.

Let us introduce the indicator $\Delta^{P_A,P_B}_{X_1,X_2} \in \{0,1\}$ that specifies if the scores X_1 and X_2 disagree on the relative order between the performances P_A and P_B. Given the implication that a mean of scores has on the rankings, we have

$$\Delta^{P_A,P_B}_{Pr,Re} = \Delta^{P_A,P_B}_{Pr,F_\beta} + \Delta^{P_A,P_B}_{F_\beta,Re} \qquad \forall \beta \geq 0\,. \tag{7}$$

And, as [50]

$$d_\tau(X_1;X_2) = \frac{2}{n(n-1)} \sum_{i<j} \Delta^{P_i,P_j}_{X_1,X_2} \qquad \in [0,1]\,, \tag{8}$$

$$d_\tau(Pr;Re) = d_\tau(Pr;F_\beta) + d_\tau(F_\beta;Re) \qquad \forall \beta \geq 0\,. \tag{9}$$

This means that *the rankings induced by the F_β scores form one shortest path (i.e. a geodesic) between the rankings induced by Pr and Re*. This is a key result for this paper, as it proves meaningful the search for the optimal tradeoff between Pr and Re in the family of F_β scores. As $\tau = 1-2d_\tau$, Eq. (9) can be rewritten in terms of correlations as

$$1 + \tau(Pr;Re) = \tau(Pr;F_\beta) + \tau(F_\beta;Re) \qquad \forall \beta \geq 0\,. \tag{10}$$

3.4. The Optimal Tradeoff Between Pr and Re

3.4.1. Optimal Tradeoffs as Karcher Means

The optimal tradeoffs are the F_β scores that minimize the Fréchet variance [64]:

$$\sigma^2(\beta) = d^2_\tau(Pr;F_\beta) + d^2_\tau(F_\beta;Re)\,. \tag{11}$$

The solutions $F_\beta = F_*$ of Eq. (11), known as the Karcher means [13] are those that are equidistant of Pr and Re (this is detailed in supplementary material), *i.e.* such that

$$d_\tau(Pr;F_*) = d_\tau(F_*;Re) = \frac{d_\tau(Pr;Re)}{2} \tag{12}$$

$$\Leftrightarrow\ \tau(Pr;F_*) = \tau(F_*;Re) = \frac{1+\tau(Pr;Re)}{2}\,. \tag{13}$$

3.4.2. A Closed-Form Expression for the Optimal β

Let us, by thought, place ourselves at the ranking induced by $Pr = F_0$ and then continuously increase β, following the path of rankings induced by the F_β scores, towards the one induced by $Re = F_\infty$. With a finite number of performances, we stay at a given ranking during a range of values for β and then, suddenly, move to the next ranking. The value of β at which such a transition occurs is a value for which two performances are put on an equal footing by F_β. Let us consider any two performances P_1 and P_2. We have $F_\beta(P_1) = F_\beta(P_2)$ if and only if $\beta^2 = \vartheta(P_1,P_2)$ with

$$\vartheta(P_1,P_2) = -\frac{PTP(P_1)PFP(P_2) - PTP(P_2)PFP(P_1)}{PTP(P_1)PFN(P_2) - PTP(P_2)PFN(P_1)}\,, \tag{14}$$

when this value is positive. Otherwise, there is no score F_β that puts the performances P_1 and P_2 in equivalence. Kendall's distance between Pr and F_β increases linearly with the number of swaps. This leads us to the conclusion that *the optimal tradeoff is $F_* = F_\beta$ with*

$$\beta^2 = \text{median}\left(\{\vartheta(P_i,P_j) \mid i \neq j \wedge \vartheta(P_i,P_j) \geq 0\}\right)\,. \tag{15}$$

3.4.3. Assessing the Degree of Optimality of Some F_β

The choice of β comes down to choosing a permutation of the classifiers. Looking at the level of pairwise comparisons, we can identify three mutually exclusive cases that we denote by ✠, ✗, and ✓.

✠: the pairs of classifiers that are ordered in the same way by Pr and Re. As it is also the case for all β, there is no choice to make. The proportion of such pairs is given by

$$\begin{aligned} P(✠) &= 1 - d_\tau(Pr;Re) \\ &= 1/2\,[1+\tau(Pr;Re)] = 1/2\left[\tau(Pr;F_\beta) + \tau(F_\beta;Re)\right]\,. \end{aligned} \tag{16}$$

✗: the pairs of classifiers for which a choice has to be made because Pr and Re contradict each other, and for which the choice is not optimal (F_β disagrees with F_*). The proportion of such pairs is given by

$$P(✗) = d_\tau(F_\beta;F_*) = 1/4\left|\tau(Pr;F_\beta) - \tau(F_\beta;Re)\right|\,. \tag{17}$$

✓: the pairs of classifiers for which a choice has to be made and for which the choice is optimal. The proportion of such pairs is given by

$$P(✓) = d_\tau(Pr;Re) - d_\tau(F_\beta;F_*) = 1 - P(✠) - P(✗)\,. \tag{18}$$

It appears from Eqs. (16) and (17) that *if one uses some F_β to rank and reports the values for both $\tau(Pr; F_\beta)$ and $\tau(F_\beta; Re)$, then we can determine the degree of optimality for the chosen β*. We define it as follows:

$$\mathcal{O} = P(✓|\overline{✠}) = \frac{P(✓)}{P(✓) + P(✗)} = 1 - \frac{P(✗)}{1 - P(✠)} . \quad (19)$$

Note that $\mathcal{O} = 1$ if and only if $\tau(Pr; F_*) = \tau(F_*; Re)$, which is what is targeted by minimizing the Fréchet variance defined in Eq. (11).

4. Case Studies

We now perform case studies for several distributions (see Fig. 3) and many sets of performances. The results have been obtained either analytically (with the help of Wolfram 14.2 software; all details in supplementary material) or numerically based on Monte Carlo simulations. A summary of the results, emphasizing the degree of optimality and providing the links to results that can be used to determine the optimal tradeoff, is provided in Tab. 1.

4.1. Uniform Distribution Over All Performances

Let us start by considering the uniform distribution over the set Π_1 of all performances:

$$\Pi_1 = \mathbb{P}_{(\Omega,\Sigma)} . \quad (20)$$

It is the Dirichlet distribution with all concentration parameters set to 1, and corresponds to a uniform distribution in a tetrahedron (see Fig. 3a). With this distribution, precision and recall have uniformly distributed values and are positively correlated: $\tau(Pr; Re) = 1/3$.

The F_1 score has not uniformly distributed values, but is the optimal tradeoff between Pr and Re:

$$\tau(Pr; F_1) = \tau(F_1; Re) = 2/3 . \quad (21)$$

The skew-insensitive version of F_1, $SIVF$, does not lead to meaningful rankings as it does not satisfy the 3rd axiom of the theory of performance-based ranking [2] on Π_1. Moreover, we computed $\tau(Pr; SIVF) \simeq 0.43$ and $\tau(SIVF; Re) \simeq 0.81$, which implies that $SIVF$ it is not located on a shortest path between Pr and Re and is not at equidistance from them.

4.2. Uniform Distributions With Fixed Probability of True Negatives

Motivated by the fact that Pr, Re, and F_β give no importance to true negatives, we now consider a second type of distributions, the uniform distribution over the sets

$$\Pi_2(ptn) = \{ P \in \mathbb{P}_{(\Omega,\Sigma)} : P(\{tn\}) = ptn \} \subset \Pi_1 . \quad (22)$$

Note that $\Pi_2(0)$ is the set of performances considered in detection problems. As shown in Fig. 3b, these distributions correspond to uniform distributions over equilateral triangles. Pr and Re have uniformly distributed values and are positively correlated: $\tau(Pr; Re) = 1/3$.

For F_1, our observations remain the same as what we had without any constraint, *i.e.* it is the optimal tradeoff:

$$\tau(Pr; F_1) = \tau(F_1; Re) = 2/3 . \quad (23)$$

For $SIVF$, we found that it does not satisfy the 3rd axiom of the theory of performance-based ranking [2] on $\Pi_2(ptn)$, unless $ptn = 0$. But, to the contrary of what we had without any constraint, $SIVF$ is now located on a shortest path between Pr and Re. However, it is still far from being at equidistance from them. For $ptn = 0$, we found $\tau(Pr; SIVF) = 1/3$ and $\tau(SIVF; Re) = 1$.

4.3. Uniform Distributions With Fixed Class Priors

We now move to more realistic distributions for classification, and consider that one needs to rank only performances corresponding to some given class priors (π_-, π_+). For this reason, we consider the uniform distributions over the sets

$$\Pi_3(\pi_+) = \{ P \in \mathbb{P}_{(\Omega,\Sigma)} : P(\{fn, tp\}) = \pi_+ \} \subset \Pi_1 . \quad (24)$$

As shown in Fig. 3c, these distributions correspond to uniform distributions over rectangles. It turns out that their axes correspond to $FPR \in [0, 1]$ and to $TPR \in [0, 1]$. The uniform distributions over $\Pi_3(\pi_+)$ correspond thus to uniform distributions of points in ROC. We found, no matter what the class priors are, that $\tau(Pr; Re) = 1/2$.

The manifold can be easily visualized by PCA. For example, if we apply a two-components PCA to the rankings obtained with $\Pi_3(\pi_+)$ for $\pi_+ = 0.1$, we obtain the curve depicted in Fig. 1 (98.77% of total variance explained).

We stress that, to the contrary of what we concluded for the two previous distributions, when the class priors are fixed, F_1 is far from the optimal tradeoff, as shown in Fig. 4a: unless $\pi_+ \simeq 0.381$, $\tau(Pr; F_1) \neq \tau(F_1; Re)$. In the extreme cases for which the prior of the negative (positive) class tends towards 1.0, the ranking induced by F_1 perfectly mimics the ranking induced by Pr (Re), thus totally ignoring the ranking induced by Re (Pr).

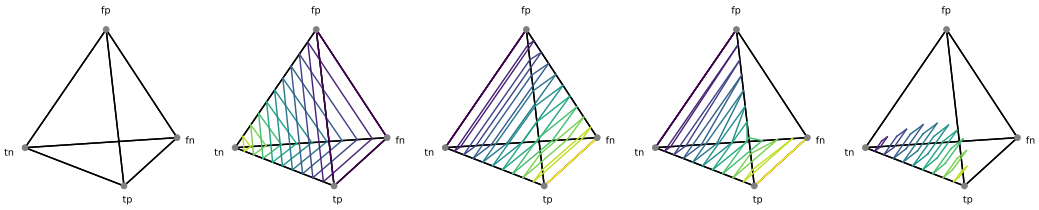
To find the optimal threshold, we need first to compute either $\tau(Pr; F_\beta)$ or $\tau(F_\beta; Re)$. The other can be found using Eq. (10). Analytically, we found

$$\tau(F_\beta; Re) = \frac{1}{2} + \ell - \ell^2 \log \frac{1 + \ell}{\ell} , \quad (25)$$

with ℓ defined as in Eq. (2). Then, we minimized the Fréchet variance, given in Eq. (11). If one considers some given class priors, then it is a function of β (see Fig. 4b) and the minimization leads to the optimal value of β for $F_* = F_\beta$. However, it is also possible to see the Fréchet variance as a function of ℓ. Its minimization leads to the solution $\ell \simeq$

Case study	$\tau(Pr; Re)$	F_1 ($\beta^2 = 1$)	$SIVF$ ($\equiv \beta^2 = \frac{\pi_-}{\pi_+}$)	Heuristic $\beta^2 = \frac{\mathbb{E}[PFP]}{\mathbb{E}[PFN]}$	Optimal tradeoff, $\mathcal{O} = 100\%$
uniform distribution over all performances	$1/3$	$\mathcal{O} = 100\%$ (optimal)	NA (meaningless ranking)	$\mathcal{O} = 100\%$ (optimal) (it selects F_1)	our analytical result *cf*. Eq. (21)
uniform distributions with fixed probability of true negatives	$1/3$	$\mathcal{O} = 100\%$ (optimal)	NA (meaningless ranking)	$\mathcal{O} = 100\%$ (optimal) (it selects F_1)	our analytical result *cf*. Eq. (23)
uniform distributions with fixed class priors	$1/2$	$\mathcal{O} \in [50\%, 100\%]$ optimal only for $\pi_+ \simeq 0.381$	$\mathcal{O} = \log(4) - 1/2 \simeq 88.63\%$	$\mathcal{O} = \log(4) - 1/2 \simeq 88.63\%$ (it selects $SIVF$)	our analytical result *cf*. Eq. (26) and Fig. 4c
uniform distributions with fixed class priors, above no-skill	0	$\mathcal{O} \in [50\%, 100\%]$ optimal only for $\pi_+ \simeq 0.325$	$\mathcal{O} = 5/6 \simeq 83.33\%$	$\mathcal{O} = 5/6 \simeq 83.33\%$ (it selects $SIVF$)	our analytical result *cf*. Eq. (28)
uniform distributions with fixed class priors, close to oracle	$\in (0, 1/2)$	$\mathcal{O} \in [87.85\%, 100\%]$ optimal only for $\pi_+ \to 1$	$\mathcal{O} \in [50\%, 100\%]$ optimal only for $\pi_+ \simeq 0.561$	$\mathcal{O} \in [87.85\%, 100\%]$ (it selects F_1)	our numerical result *cf*. Fig. 5c
53 real sets of about 60 performances	$\in [-0.31, 0.65]$	$\mathcal{O} \in [52.23\%, 100.00\%]$ (mean: 78.44%)	$\mathcal{O} \in [50.00\%, 99.13\%]$ (mean: 56.71%)	$\mathcal{O} \in [69.69\%, 100.00\%]$ (mean: 89.53%)	our closed-form expression for the optimal β, Eq. (15)

Table 1. Summary of our results emphasizing the degree of optimality $\mathcal{O}$ introduced in Eq. (19). As shown by the last column, we can always find the optimal ranking score (in green) between Pr and Re. Relying on a fixed and arbitrarily chosen score like F_1 or $SIVF$ can lead to catastrophic cases (in red). In our case studies, the simple heuristic introduced in Sec. 4.6 never led to such catastrophic cases, but nevertheless appeared to be suboptimal in some cases (in orange).



(a) The set Π_1. (b) The sets $\Pi_2(ptn)$. (c) The sets $\Pi_3(\pi_+)$. (d) The sets $\Pi_4(\pi_+)$. (e) The sets $\Pi_5(\pi_+)$.

Figure 3. Graphical representation of some sets of two-class crisp classification performances considered in this paper. (a) All possible performances (see Eq. (20)) correspond to points in a regular tetrahedron. (b) The performances corresponding to fixed probabilities of true negatives (see Eq. (22)) are points in parallel equilateral triangles. (c) The performances corresponding to fixed class priors (see Eq. (24)) are points in parallel rectangles. (d) The performances corresponding to fixed class priors and "above" no-skills (see Eq. (27)) are points in parallel right triangles. (e) The performances for given class priors, and "close to the oracle" (see Eq. (29)) are points in parallel squares. The color represents the value of the fixed parameter considered (either ptn or π_+).

0.61585. Injecting this value into Eq. (2), we obtained the link between the optimal β and the class priors:

$$\beta^2 = \ell \frac{\pi_-}{\pi_+} \Rightarrow b \simeq \frac{0.61585\,(1-\pi_+)}{\pi_+ + 0.61585\,(1-\pi_+)} \tag{26}$$

It is depicted in Fig. 4c. We see that the optimal relationship is quite close to that of $SIVF$ (*i.e.*, $b = 1 - \pi_+$), but very different from that of F_1 (*i.e.*, $b = 1/2$).

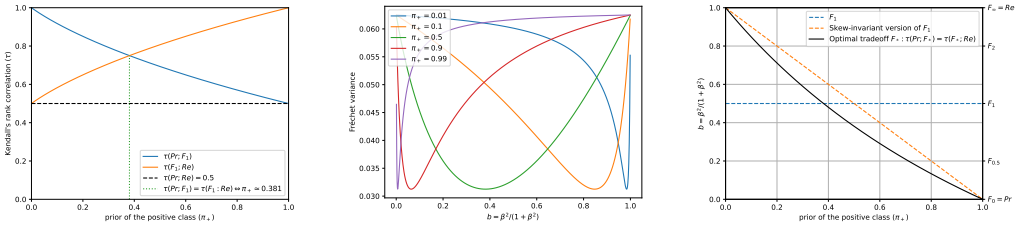
4.4. Uniform Distributions With Fixed Class Priors, Above No-Skill

Performances below the rising diagonal of ROC, where the no-skill performances lie, are usually of little interest. Systematically inverting the output of a classifier is equivalent to applying a central symmetry in ROC. Thus, from any performance below the no-skills, it is easy to obtain a performance above the no-skills. But the rankings induced by Pr, Re, and F_β are sensitive to this operation. For this reason, we study the case of the uniform distributions over the sets

$$\Pi_4(\pi_+) = \{P \in \mathbb{P}_{(\Omega,\Sigma)} : P(\{fn, tp\}) = \pi_+ \\ \wedge\, TPR(P) \geq FPR(P)\} \subset \Pi_3(\pi_+)\,. \tag{27}$$

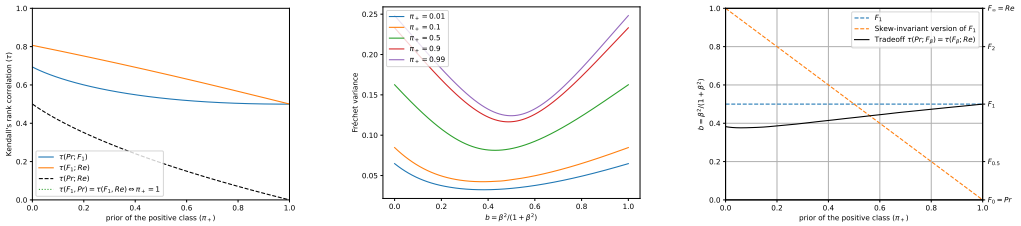
These sets are depicted in Fig. 3d. No matter what the class priors are, we obtain $\tau(Pr; Re) = 0$. All detailed results are in supplementary material. In a nutshell, the observations are very close to those reported for the uniform distributions with fixed class priors. The overall shape of the manifold, as observed by PCA, is very similar. $\tau(Pr; F_1)$ monotonically increases from 0 to 1 (instead of 0.5 to 1) and $\tau(F_1; Re)$ monotonically decreases from 1 to 0 (instead of 1 to 0.5) when π_+ sweeps the $[0, 1]$ interval. The F_1 score is the optimal tradeoff only for $\pi_+ \simeq 0.325$. Regarding the adaptation, the curve is slightly more curved and lower than the one we depicted in Fig. 4c, with the optimal value $\ell \simeq 0.48$:

$$\beta^2 = \ell \frac{\pi_-}{\pi_+} \Rightarrow b \simeq \frac{0.48\,(1-\pi_+)}{\pi_+ + 0.48\,(1-\pi_+)}\,. \tag{28}$$



(a) F_1 is not the optimal tradeoff. (b) Fréchet variance for various priors. (c) Adaptation of β w.r.t. class priors.

Figure 4. Results for uniform distributions over the performances with fixed class priors, *i.e.* $\Pi_3(\pi_+)$.



(a) F_1 is not the optimal tradeoff. (b) Fréchet variance for various priors. (c) Adaptation of β w.r.t. class priors.

Figure 5. Results for uniform distributions over the performances with fixed class priors and close to the oracle, *i.e.* $\Pi_5(\pi_+)$.

4.5. Uniform Distributions With Fixed Class Priors, Close to Oracle

In competitions, a setting in which ranking is omnipresent, contenders usually refrain from choosing decision thresholds that would lead either to a high TPR at the cost of a high FPR or to a low FPR at the cost of a low TPR. In our opinion, this means that $FPR < \pi_+$ and $TPR > \pi_+$. In ROC, these performances are above and on the left-hand side of the no-skill performance that results from randomly choosing the negative or positive class with respective probabilities equal to π_- and π_+. For this reason, we now discuss the case of uniform distributions over the sets

$$\Pi_5(\pi_+) = \{P \in \mathbb{P}_{(\Omega,\Sigma)} : P(\{fn, tp\}) = \pi_+ \\ \wedge\, FPR(P) < \pi_+ \wedge TPR(P) > \pi_+\} \subset \Pi_4(\pi_+)\,. \quad (29)$$

As shown in Fig. 3e, these distributions correspond to uniform distributions over squares. Drawing a performance at random can be achieved by drawing independent values uniformly in $[0, \pi_+]$ for FPR and in $[\pi_+, 1]$ for TPR.

The results, shown in Fig. 5, are remarkably different from the previous ones. In comparison to Fig. 4, the optimal β now increases from about 0.8 to 1, when π_+ increases from 0 to 1, while it was previously decreasing from ∞ to 0.

4.6. Some Real Sets of Performances

We now report results for performances encountered in existing rankings. They are either performances of some baseline methods or performances that people considered good enough for a competition. More specifically, we considered a computer vision task known as *background subtraction*, similar to a pixelwise classification between background and foreground [18, 65]. The results of about 60 methods (classifiers), on 53 videos —a dataset known as *CDnet 2014* [66]— are publicly available on `changedetection.net`. According to [21], the multi-criteria ranking proposed by this platform is well correlated with F_1. On any given video, all performances belong to $\Pi_3(\pi_+)$. Moreover, 99.9% and 99.4% of them belong to $\Pi_4(\pi_+)$ and $\Pi_5(\pi_+)$, respectively. Depending on the video, $\tau(Pr; Re) \in [-0.31, 0.65]$.

Figure 6 shows the results for the set of performances reported for one video; more results are to be found in the supplementary material. The optimal β is between 0.29 and 22.79 depending on the video. It is in the $[0.5, 2]$ range for only 35 videos among the 53.

An interesting observation, shown in Fig. 7, is that the class priors (π_-, π_+) are poor predictors for the optimal value of β. However, b is well correlated with

$$\frac{\mathbf{E}[PFP]}{\mathbf{E}[PFN] + \mathbf{E}[PFP]} = \frac{\sum_{P\in\Pi} P(\{fp\})}{\sum_{P\in\Pi} P(\{fn, fp\})}\,, \quad (30)$$

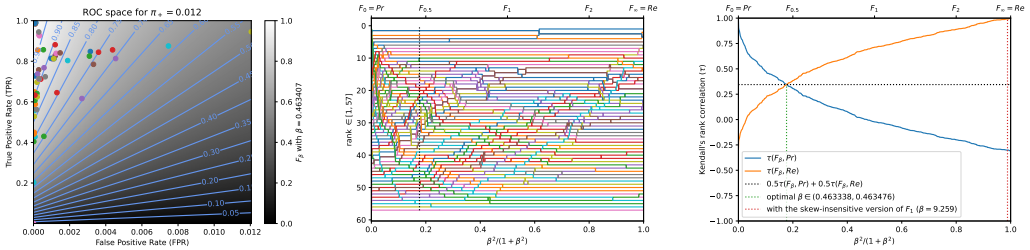


Figure 6. What is the optimal tradeoff for the performances of 57 background subtraction methods (pixelwise classifiers) on the *blizzard* video of *CDnet 2014*? Left: the 57 performances to be ranked, depicted in the ROC space, with the isometrics of the optimal tradeoff. Center: the ranks of each classifier, w.r.t. β. Right: $\tau(Pr; F_\beta)$ and $\tau(F_\beta; Re)$ w.r.t. β. The ranking with F_1, at the center, is too far from Pr and too close to Re. The optimal tradeoff, for which $\tau(Pr; F_\beta) = \tau(F_\beta; Re)$ is at $\beta \simeq 0.463$.

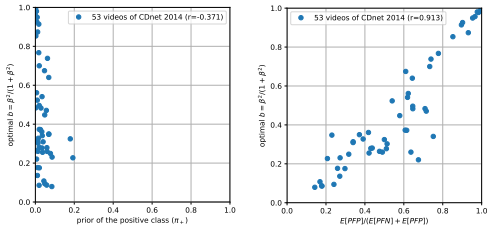


Figure 7. Which statistic about the sets of performances can be used to predict the optimal value of β? These results are based on 53 sets of performances (those of about 60 methods on each video of *CDnet 2014*). Left: the class priors are not suitable. Right: the ratio of the average probability of false positives ($\mathbf{E}[PFP]$) to the average probability of errors ($\mathbf{E}[PFN] + \mathbf{E}[PFP]$) seems to be a good predictor for β, as Pearson's linear correlation between $\frac{\mathbf{E}[PFP]}{\mathbf{E}[PFN]+\mathbf{E}[PFP]}$ and $\frac{\beta^2}{1+\beta^2}$ is $r \simeq 0.913$.

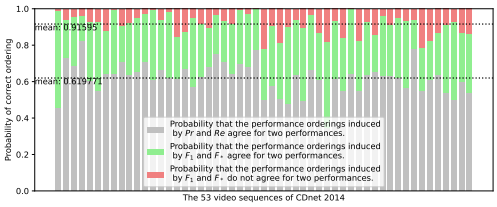


Figure 8. How far is F_1 from the optimal tradeoff F_*? Each vertical bar is for a specific set of performances. ▒: P(✠), ▒: P(✓), ▒: P(✗).

which suggests that $\beta^2 = \frac{\mathbf{E}[PFP]}{\mathbf{E}[PFN]}$ is a simple heuristic that is worth trying as an alternative to the closed-form and optimal solution provided in Eq. (15). For the sake of completeness, the degree of optimality $\mathcal{O}$ achieved by this heuristic is provided in Tab. 1 for all our case studies.

Figure 8 shows the consequences of using F_1 for ranking instead of the optimal tradeoff F_*. On average, there is 61.98% (▒) of chance that a randomly chosen pair of methods is ranked the same by Pr and Re. In this case, which tradeoff is chosen among all F_β is not relevant. But when Pr and Re disagree, the value of β matters. The degree of optimality for F_1 is only about $\mathcal{O} \simeq 78\%$ ▒/(▒+▒) on average.

5. Conclusion

Nobody would ever choose a classifier solely based on the precision Pr or the recall Re. The traditional approach of calculating the harmonic mean F_1 of these two scores and then ranking the classifiers based on it is, unfortunately, nothing but a red herring. The colossal problem, ignored in the literature until now, is that the ranking built upon the balanced average of two scores is generally not a good tradeoff between the rankings obtained with the two base scores. In extreme cases, the ranking obtained with F_1 copies the ranking based on Pr or Re and totally ignores the other.

To solve these issues, we defined and described the manifold of rankings induced by scores (see Fig. 1) and found that the distance along it is linearly related to Kendall's rank correlation τ. Besides, we showed that all F_β scores induce meaningful rankings, even without any constraint on the performances. We further proved that the F_β scores induce rankings that form a shortest path between the rankings of Pr and Re. Also, we provided the theory and methods (including a closed-form expression) for defining and identifying the optimal tradeoffs between the rankings induced by Pr and Re. Lastly, we exemplified six case studies, covering numerous distributions and sets of performances. A summary of our results is provided in Tab. 1.

Acknowledgments. S. Piérard is funded by grants 8573 (ReconnAIssance project) and 2010235 (ARIAC by DIGITALWALLONIA4.AI) of the SPW EER, Wallonia, Belgium; A. Deliège is a F.R.S.-FNRS postdoc researcher.

References

[55] Davide Ballabio, Francesca Grisoni, and Roberto Todeschini. Multivariate comparison of classification performance measures. *Chemometrics and Intelligent Laboratory Systems*, 174:33–44, 2018.

[47] 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)*, pages 821–826, Antalya, Turkey, 2017. IEEE.

[37] Seung-Seok Choi, Sung-Hyuk Cha, and Charles C. and Tappert. A survey of binary similarity and distance measures. *Journal of Systemics, Cybernetics and Informatics*, 8(1): 43–48, 2010.

[57] 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, 2023.

[58] Luciana Ferrer. Analysis and comparison of classification metrics. *arXiv*, abs/2209.05355, 2022.

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

[29] Peter A. Flach. The geometry of ROC space: Understanding machine learning metrics through ROC isometrics. In *International Conference on Machine Learning (ICML)*, pages 194–201, Washington, District of Columbia, USA, 2003. ML Research Press.

[64] Maurice Fréchet. Les éléments aléatoires de nature quelconque dans un espace distancié. *Annales de l'institut Henri Poincaré*, 10(4):215–310, 1948.

[65] Belmar Garcia-Garcia, Thierry Bouwmans, and Alberto Jorge Rosales Silva. Background subtraction in real applications: Challenges, current models and future directions. *Computer Science Review*, 35:1–42, 2020.

[21] 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)*, pages 1–8, Providence, RI, USA, 2012. Institute of Electrical and Electronics Engineers (IEEE).

[5] 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.

[60] Mohammad Hossin and M. N. Sulaiman. A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining & Knowledge Management Process*, 5(2):1–11, 2015.

[61] Nathalie Japkowicz and Mohak Shah. *Evaluating Learning Algorithms: A Classification Perspective*. Cambridge University Press, 2011.

[18] Pierre-Marc Jodoin, Sébastien Piérard, Yi Wang, and Marc Van Droogenbroeck. Overview and benchmarking of motion detection methods. In *Background Modeling and Foreground Detection for Video Surveillance*, chapter 24. Chapman and Hall/CRC, 2014.

[13] Hermann Karcher. Riemannian center of mass and mollifier smoothing. *Communications on Pure and Applied Mathematics*, 30(5):509–541, 1977.

[50] Maurice George Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81, 1938.

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

[63] Yangguang Liu, Yangming Zhou, Shiting Wen, and Chaogang Tang. A strategy on selecting performance metrics for classifier evaluation. *International Journal of Mobile Computing and Multimedia Communications*, 6(4):20–35, 2014.

[4] 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.

[2] 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)*, pages 14293–14302, Nashville, Tennessee, USA, 2025. IEEE.

[62] David M. W. Powers. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness & correlation. *Journal of Machine Learning Technologies*, 2(1): 37–63, 2011.

[38] Marina Sokolova and Guy Lapalme. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4):427–437, 2009.

[51] Charles Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904.

[66] 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)*, pages 393–400, Columbus, Ohio, USA, 2014. Institute of Electrical and Electronics Engineers (IEEE).

4. The Tile (4th article)

This chapter reproduces the following article [4]:

> S. Piérard, A. Halin, A. Cioppa, A. Deliège, and M. Van Droogenbroeck, "The Tile: A 2D map of ranking scores for two-class classification," *arXiv*, vol. abs/2412.04309, 2024.

It covers both the notions of ***performance*** and ***ranking***. The supplementary material of this article is reproduced in Chapter C.

Context

In the previous chapter, I used the manifold of performance orderings induced by the ranking scores R_I to choose a ranking when the application preferences I are not specified. In this chapter, I will make another use of this manifold: I develop an innovative and versatile graphical tool that I call Tile and that is suitable for tasks similar to two-class classification.

Motivation

For tasks similar to two-class crisp classification, it can be shown that the manifold is two-dimensional and bounded by the constraints $I(tn) \geq 0$, $I(fp) \geq 0$, $I(fn) \geq 0$, and $I(tp) \geq 0$. The parameterization $a(I) = \frac{I(tp)}{I(tn)+I(tp)}$, $b(I) = \frac{I(fn)}{I(fp)+I(fn)}$ establishes a bijection between the manifold of performance orderings and the square $(a, b) \in [0, 1]^2$. All ranking scores R_I leading to the same performance ordering are mapped to the same point. All points correspond to different orderings (if we consider all possible performances), but zones of the square correspond to the same ranking (if we consider a finite set of entities, the amount of possible rankings is finite).

Main contributions

I introduce the Tile. It is essentially a map of applicative preferences on which various functions of I (from the next chapter onwards, we will refer to them as flavors) can be represented (in false colors) and on top of which various annotations can be placed. This makes it a versatile tool.

From my point of view, the Tile is a game changer. On the one hand, the ROC space that people have been using for 85 years is a square in which classification performances

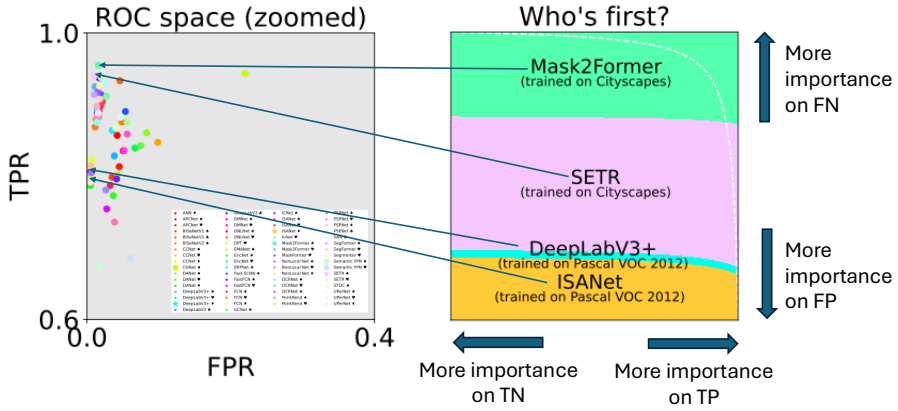


Figure 4.1.: Comparison between the ROC space (left) and the Tile (right). It is a real-world example (its detailed description is in the next chapter) involving 74 segmentation models provided by MMSegmentation, with their semantic labels arbitrarily regrouped in two classes. In ROC (left), the results are cluttered, and it is hard to rank the models, while the Tile (right) makes it easy ("Who's first?" shows usually a small subset of models—here $^4/_{74}$).

are projected as points and on which it is difficult to visualize rankings and applicative preferences. On the other hand, the Tile is a square in which the points correspond to applicative preferences, to ranking scores, or even to the performance orderings induced by them. For example, Fig. 4.1 provides a comparison between the ROC space and the Tile when one depicts on it the entity that is ranked number one, depending on the applicative preferences. While the information depicted on the Tile is present on the ROC plot, it is clear that reading the Tile is much easier than analyzing ROC.

A significant review was conducted to identify the scores that induce a meaningful performance ordering (both in general and under the constraint of fixed class priors) among the many scores that have been previously defined in the literature (see Fig. 4.2). It turns out that these scores form clusters, as many of them are strict monotonous functions of others (*e.g.*, Jaccard J_+ and the F-score F_1).

Perspectives

- When we look at the figures in the article, we see that Figs. 4 and 7 show correlation values, Figs. 1, 6, and 9 show score values (ranking scores, class priors, and prediction rates), and Fig. 5 shows entities. Clearly, the Tile is a versatile tool. Practical uses will be discussed in Chapters 5 to 7.

- At the TRAIL summer workshop 2025, I gathered a team of 5 people interested in developing a new open-source python library providing implementation for the Tile and some flavors. It is called *Sorbetto* [67]. It is currently under development, but a first version is already available on *GitHub*, *PyPI*, *Read the Docs*, and *Zenodo*.

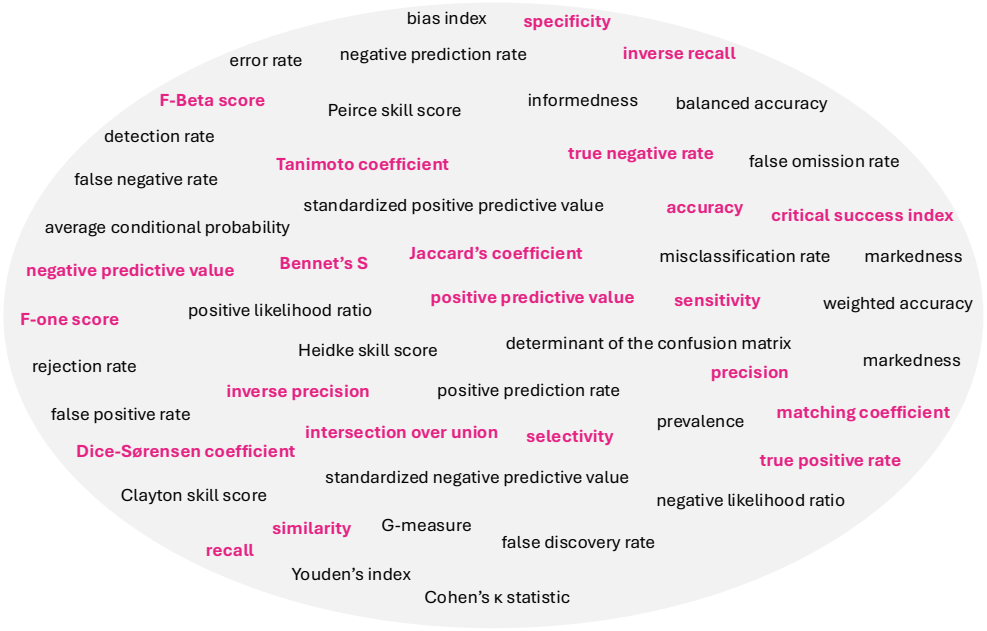


Figure 4.2.: Lost in an ocean of scores? This is a small part of the scores produced by the human imagination during a few decades and just for the two-class classification tasks. The scores that are written in color are those inducing a meaningful performance ordering; they can be placed onto the Tile.

Remarks

1. As a map of scores (Fig. 3 of the article), the Tile is a real innovation that was made possible by my willingness, on the one hand, to take the perspective of performance orderings and rankings, and on the other hand, to focus on the orderings induced by scores for which it is possible to know what the associated applicative preferences are. This map provides a better understanding of the meaning and structure of scores (thanks to the correlation Tiles, this is not limited to ranking scores). In comparison, the literature contains a few attempts to organize scores, which I, personally, find rather disappointing, partly because the scores considered are not part of a well-defined family and partly because the authors force the representation to be two-dimensional when there is no justification for this (without a well-defined family, it seems hopeless). This is the case in [59] that uses dendrograms, [47] that proposes a "chart like the periodic table of elements", [55] that proposes 2D PCA of the values of 12 scores on 20 datasets, and resources for the public available online such as Wikipedia that shows 2D tables. These representations are arbitrary and/or empirical, and without any theoretical foundations.

2. As explained in the previous chapter, the ranking scores are insensitive to any strictly positive scaling of I, which means that we can limit ourselves to studying scores for which $\sum_{\omega\in\Omega} I(\omega) = 1$. Furthermore, we know that I must be positive.

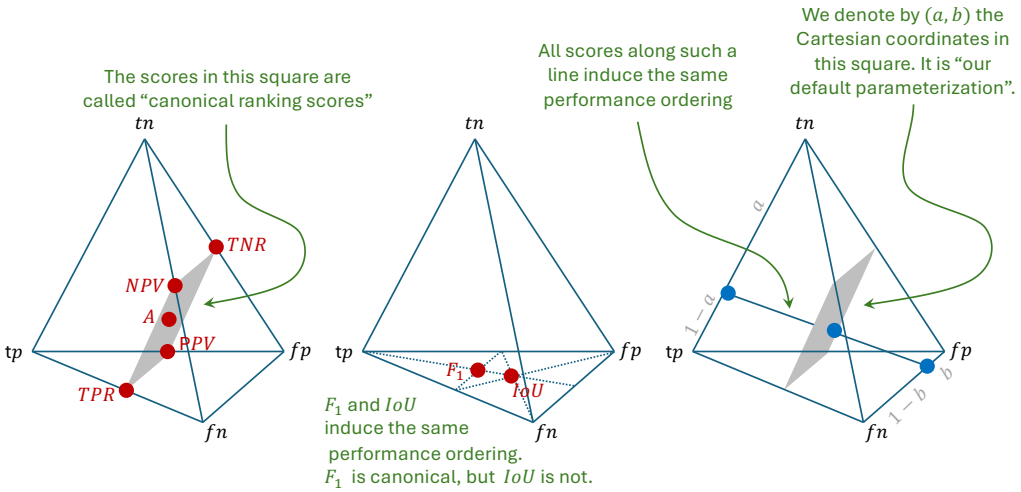


Figure 4.3.: The tetrahedron of ranking scores, or simplex of importances.

We are therefore dealing with a simplex. For crisp classification with two classes, the ranking scores are points in a tetrahedron. Figure 4.3 shows this tetrahedron, shows the set of ranking scores that are said to be canonical, shows the sets of scores that lead to the same ranking (and are therefore superimposed on the tile), and shows also the parameterization (a, b). The map of ranking scores appears to be a projection of this tetrahedron. It turns out that, for any ranking score inside the tetrahedron, there exists a single line passing through it and intersecting both the $tn - tp$ and $fp - fn$ edges. And, all the ranking scores along this line induce the same performance ordering. So, they are projected onto the same point of the map. In the tetrahedron, the locus of canonical ranking scores is a square that intersects such a line in only one point. One can consider this square as the surface on which the projection is performed: we obtain the Tile.

The *Tile*: A 2D Map of Ranking Scores for Two-Class Classification

Sébastien Piérard, Anaïs Halin, Anthony Cioppa, Adrien Deliège, and Marc Van Droogenbroeck
Montefiore Institute, University of Liège, Liège, Belgium
{S.Pierard,Anais.Halin,Anthony.Cioppa,Adrien.Deliege,M.VanDroogenbroeck}@uliege.be

Abstract

In the computer vision and machine learning communities, as well as in many other research domains, rigorous evaluation of any new method, including classifiers, is essential. One key component of the evaluation process is the ability to compare and rank methods. However, ranking classifiers and accurately comparing their performances, especially when taking application-specific preferences into account, remains challenging. For instance, commonly used evaluation tools like Receiver Operating Characteristic (ROC) and Precision/Recall (PR) spaces display performances based on two scores. Hence, they are inherently limited in their ability to compare classifiers across a broader range of scores and lack the capability to establish a clear ranking among classifiers. In this paper, we present a novel versatile tool, named the Tile*, that organizes an infinity of ranking scores in a single 2D map for two-class classifiers, including common evaluation scores such as the accuracy, the true positive rate, the positive predictive value, Jaccard's coefficient, and all F_β scores. Furthermore, we study the properties of the underlying ranking scores, such as the influence of the priors or the correspondences with the ROC space, and depict how to characterize any other score by comparing them to the Tile. Overall, we demonstrate that the Tile is a powerful tool that effectively captures all the rankings in a single visualization and allows interpreting them.*

1. Introduction

Two-class classification is a fundamental task, encountered in numerous real-world scenarios. For instance, it plays a vital role in medical diagnostics, such as blood tests, MRI scans, and other imaging techniques to determine whether a patient has a disease or is healthy. In security systems, alarms must activate only when intrusions are detected. Similarly, in quality control, identifying defects in manufactured items is crucial to ensure faulty products do not reach the market. To address these challenges, selecting the right classifier is essential. However, this requires ranking classifiers in the context of application-specific preferences. For example, in medical testing, minimizing false negatives is critical since failing to diagnose a patient could have life-threatening consequences. In security systems, the focus is often on maximizing true negatives, accepting occasional false alarms as a trade-off for ensuring safety. Meanwhile, in quality control, false positives can be costly, as they may trigger unnecessary halts in production. Each application thus has unique requirements regarding the types of errors a classifier can tolerate.

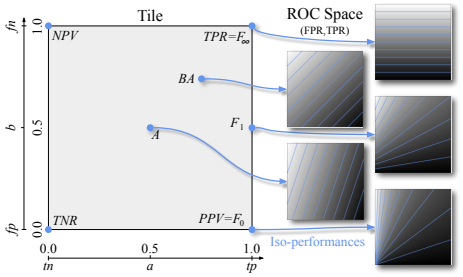


Figure 1. **Introducing the Tile.** We introduce a new visual tool, called the Tile, representing an infinite family of ranking scores to evaluate the performances of two-class classifiers at a glance. In this figure, we highlight the correspondences between specific ranking scores on the Tile and their corresponding set of iso-performance lines in the ROC space. Notably, the variation of iso-performance lines along the right border of the Tile demonstrates the limitations of the ROC space for ranking performance. This visualization illustrates how the Tile simplifies the task of ranking classifiers and enhances the interpretation of performance scores across various evaluation spaces, such as the ROC space.

A wide range of scores penalizing different types of errors are available in the literature. However, selecting the appropriate score to rank classifiers taking application-specific preferences into account can be challenging. Additionally, the common practice of ranking based on a single score, as done in most benchmarks, can lead to a suboptimal classifier choice. Moreover, so-called evaluation spaces combining two scores, such as Receiver Operating Char-

acteristic (ROC) and Precision/Recall (PR) spaces, do not allow ranking classifiers. This raises a recurring question: how can classifiers be effectively ranked to best align with the specific needs of each application?

In this paper, we introduce a novel visual tool for two-class classification, called the *Tile*, which organizes an infinity of performance orderings derived from ranking scores into a two-dimensional map. Our Tile is parametrized by two parameters reflecting application-specific preferences: the first controls the trade-off between true positives and true negatives, while the second balances false positives and false negatives. This parametrization allows mapping common performance scores from the literature, such as the accuracy or F_1, as illustrated on the left side of Fig. 1. Next, we analyze the correspondences between the Tile and standard evaluation spaces such as ROC and PR, with a particular emphasis on iso-performance lines. As shown in Fig. 1, iso-performance lines can be drawn on the ROC space for each score of the Tile. Even though rankings of classifiers on the ROC space is dependent on the choice of a score, our Tile enables a straightforward, unified interpretation of the ranking of two-class classifiers at a glance. Finally, we demonstrate that our Tile's organization allows to easily rank classifiers and study some ranking properties such as the orderings induced by all ranking scores, the characterization of any score, and the robustness of ranking scores.

Contribution. We summarize our contributions as follows. **(i)** We introduce a novel visual tool, called the Tile, that organizes an infinity of ranking scores on a two-dimensional map. **(ii)** We analyze the correspondences between the Tile and common evaluation spaces such as ROC and PR, with a particular focus on iso-performance lines. **(iii)** We show that our Tile's organization allows to easily compare ranking scores, rank classifiers, and study ranking properties.

2. Preliminaries and Related Work

We first present the necessary preliminaries, including the mathematical framework, definitions of the relevant scores, their underlying structure, and a review of related work to set the context. By coherence, we adopt the mathematical framework, terminology, and notations introduced in [2]. Note that all acronyms and mathematical symbols used in this paper are defined where they appear[1].

2.1. Mathematical Framework

The framework is based on the probability theory. A score X is a real-valued function defined on a subset $\mathrm{dom}(X)$ of the performance space $\mathbb{P}_{(\Omega,\Sigma)}$. The latter is the set of all possible probability measures on the measurable space (Ω, Σ), where the sample space (*a.k.a.* universe) is Ω, and the event space (*a.k.a.* σ-algebra) is Σ. A performance P is an element of $\mathbb{P}_{(\Omega,\Sigma)}$, thus, a probability measure.

Moreover, the framework is also based on the order theory. The symbols $\lesssim_X$ and $\gtrsim_X$ are used to denote, respectively, the ordering and dual ordering induced by the score X. When both performances $P_1, P_2 \in \mathrm{dom}(X)$, P_1 is *worse than*, *equivalent to*, or *better than* P_2 when $X(P_1) < X(P_2)$, $X(P_1) = X(P_2)$, or $X(P_1) > X(P_2)$, respectively. When $P_1 \notin \mathrm{dom}(X)$ or $P_2 \notin \mathrm{dom}(X)$, they are *equivalent* if $P_1 = P_2$, and *incomparable* otherwise.

Two-class crisp classification. We particularize this probabilistic framework to the case in which the sample space contains two satisfying and two unsatisfying elements, thus $|\Omega| = 4$ and $\Sigma = 2^\Omega$. Using the binary random variable $S : \Omega \to \{0, 1\}$ to denote the satisfaction, we have $|\{\omega \in \Omega : S(\omega) = 0\}| = |S^{-1}(\{0\})| = 2$, and $|\{\omega \in \Omega : S(\omega) = 1\}| = |S^{-1}(\{1\})| = 2$. Variable S determines how the elements of Ω are interpreted. The most popular interpretation is undoubtedly the popular *two-class crisp classification*. For this choice, we take $\Omega = \{tn, fp, fn, tp\}$. The samples tn, fp, fn and tp are interpreted as, respectively, a *true negative*, a *false positive* (type I error, false alarm), a *false negative* (type II error, miss), and a *true positive* (hit). In that case, $S(tn) = S(tp) = 1$ and $S(fp) = S(fn) = 0$.

Classes and predictions. We can introduce the set of classes $\mathbb{C} = \{c_-, c_+\}$, as well as the random variables $Y : \Omega \to \mathbb{C}$ and $\hat{Y} : \Omega \to \mathbb{C}$ such that $Y(tn) = Y(fp) = c_-$, $Y(fn) = Y(tp) = c_+$, $\hat{Y}(tn) = \hat{Y}(fn) = c_-$, and $\hat{Y}(fp) = \hat{Y}(tp) = c_+$. Indeed, Y and $\hat{Y}$ can be interpreted as the ground-truth and predicted classes. The *no-skill performances* are those such that $P(Y, \hat{Y}) = P(Y)P(\hat{Y})$. The satisfaction is the indicator $S = \mathbf{1}_{Y=\hat{Y}}$.

2.2. Scores

In the literature, numerous performance scores[2] (also called metrics, measures, indicators, criteria, factors, and indices [46, 47]) have been introduced for two-class crisp classification. In fact, the list of scores is almost endless, as can be seen from numerous reviews [37, 38, 47, 55, 59–62]. Choosing one score over another depends on the application field (medical, machine learning, statistics, *etc*.).

Unconditional probabilistic scores. Unconditional probabilistic scores, denoted by X_E^U, are parameterized by an event $E \in \Sigma$ such that $\emptyset \subsetneq E \subsetneq \Omega$:

$$X_E^U : \mathbb{P}_{(\Omega,\Sigma)} \to [0, 1] : P \mapsto P(E)\,. \tag{1}$$

There exist only 14 unconditional probabilistic scores. For singleton events, there is $PTN = X^U_{\{tn\}}$ (called *rejection rate*), $PFP = X^U_{\{fp\}}$, $PFN = X^U_{\{fn\}}$, and $PTP = X^U_{\{tp\}}$ (called *detection rate*). There are also the *priors* for

[1] A list of them is provided in the supplementary material.

[2] In this paper we choose the term *score* to avoid any possible confusion with the mathematical meaning of the terms *metric*, *measure* and *indicator*.

the negative and positive classes, given by $\pi_- = X^U_{\{tn,fp\}}$ and $\pi_+ = X^U_{\{fn,tp\}}$ respectively, as well as the *negative* and *positive prediction rates*, given by $\tau_- = X^U_{\{tn,fn\}}$ and $\tau_+ = X^U_{\{fp,tp\}}$ respectively. The *prevalence* is a synonym used for π_+. The *accuracy* (*a.k.a. matching coefficient*) corresponds to the expected value of S and is given by $A = X^U_{\{tn,tp\}}$. Its complement [47] is the *error rate* or *misclassification rate* .

Conditional probabilistic scores. The conditional probabilistic scores $X^C_{E_1|E_2}$ are parameterized by two events, $E_1, E_2 \in \Sigma$ such that $\emptyset \subsetneq E_1 \subsetneq E_2 \subseteq \Omega$:

$$X^C_{E_1|E_2} : \mathrm{dom}(X^C_{E_1|E_2}) \to [0,1] : P \mapsto P(E_1|E_2)\,, \quad (2)$$

with $\mathrm{dom}(X^C_{E_1|E_2}) = \{P \in \mathbb{P}_{(\Omega,\Sigma)} : P(E_2) \neq 0\}$. There exist only 50 conditional probabilistic scores (including the 14 unconditional ones, as $X^U_E = X^C_{E|\Omega}$). The probabilities of making a correct decision for negative and positive inputs are given by the *True Negative Rate* $TNR = X^C_{\{tn\}|\{tn,fp\}}$ (*a.k.a.* specificity, selectivity, inverse recall [62]) and the *True Positive Rate* $TPR = X^C_{\{tp\}|\{fn,tp\}}$ (*a.k.a. sensitivity*, *recall*). Their complements are the *False Positive Rate* FPR and the *False Negative Rate* FNR respectively. The probabilities of negative and positive predictions being correct are given by the *Negative Predictive Value* $NPV = X^C_{\{tn\}|\{tn,fn\}}$ (*a.k.a.* inverse precision [62]) and the *Positive Predictive Value* $PPV = X^C_{\{tp\}|\{fp,tp\}}$ (*a.k.a.* precision). Their complements are the *False Omission Rate* FOR , and the *False Discovery Rate* FDR respectively. *Jaccard's coefficient* is $J_- = X^C_{\{tn\}|\{tn,fp,fn\}}$ for the negative class and $J_+ = X^C_{\{tp\}|\{fp,fn,tp\}}$ for the positive class. The latter is also called *Tanimoto coefficient*, *similarity*, *intersection over union*, *critical success index* [68], as well as *G-measure* in [29].

Non-probabilistic scores. There is also an infinite number of scores that have no probabilistic significance.

We start with transformations of probabilistic scores. *Bennett's* S_B [69] is related to the accuracy by $S_B = 2A - 1$. The *F-one* score , also called *Dice-Sørensen coefficient*, is related to Jaccard by $F_1 = {}^{2J_+}/_{J_+ + 1}$. The *standardized negative and predictive values* are transformations of the negative and predictive values given by $SNPV = \frac{TNR}{TNR+FNR} = \frac{NPV\pi_+}{NPV(\pi_+ - \pi_-)+\pi_-}$ and $SPPV = \frac{TPR}{FPR+TPR} = \frac{PPV\pi_-}{PPV(\pi_- - \pi_+)+\pi_+}$ [70]. The *likelihood ratios* [62, 71–73] are also transformations of these scores. The *Negative Likelihood Ratio* is $NLR = \frac{FNR}{TNR} = \frac{1-SNPV}{SNPV}$. The *Positive Likelihood Ratio* [74] is $PLR = \frac{TPR}{FPR} = \frac{SPPV}{1-SPPV}$. *Cohen's* κ *statistic* [53, 75] is a transformation of the accuracy: $\kappa = \frac{A - A \circ \text{no–skill}}{1 - A \circ \text{no–skill}}$ where $\circ$ is the function composition operator and no–skill is the operation that transforms a performance P into P' such that $P'(Y, \hat{Y}) = P(Y)P(\hat{Y})$. It is also known as *Heidke Skill Score* [47, 76]. It is also common to combine probabilistic scores, either linearly or by averaging them. For example, the *Bias Index* BI has been defined as $\tau_+ - \pi_+$ in [77].

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 [78], and when the weights correspond to the priors, we get the *accuracy*. Instead of taking an arithmetic mean, it has been proposed to take the geometric mean [79–81], which leads to the *G-measure* [47] (not to be confused with the G-measure of [29]). *Youden's index* [82, 83] or *Youden's* J_Y *statistic* is $J_Y = TNR + TPR - 1 = 2BA - 1$. It is also called *informedness* and *Peirce Skill Score* [47, 76]. The *determinant* $|C|$ [84] of the (normalized) confusion matrix is $|C| = \pi_- \pi_+ J_Y$.

Some authors prefer to combine PPV with TPR. Their weighted harmonic mean is the *F-score* F_β . Others prefer to combine NPV with PPV. The *markedness* [85] is defined as $NPV + PPV - 1$ and is also known as the *Clayton Skill Score* [47, 76]. It has also been proposed to combine the 4 probabilistic scores TNR, TPR, NPV, and PPV. Their arithmetic mean is called *Average Conditional Probability* ACP [86], and their harmonic mean is P_4 [87]. Finally, a plethora of other scores can be found in the literature concerning similarity scores [37, 88–92]. Many of them are peculiar cases of the ranking scores that are discussed in detail in this paper.

2.3. Structuring the Scores

In [59], 18 scores have been experimentally structured in the form of histograms of linear and rank correlations, based on 30 datasets. Similarly, Choi *et al.* [37] made a hierarchical clustering of 76 binary similarity and distance scores based on a random binary dataset. In [91], 7 properties have been arbitrarily defined, and 11 scores have been classified into 5 classes based on whether the properties are verified. In [46], taking the object-oriented software development standpoint, the structure takes the form of a *Unified Modeling Language* (UML) diagram to represent the concepts of measure, metric, and indicator, as well as the relationships between the three concepts. [47] proposes to distinguish between only two types of scores: the measures and the metrics. Based on a sample of 44 scores and related quantities (22 measures and 22 metrics), they propose to divide measures into 4 levels (base, 1st, 2nd and 3rd), and metrics into 3 levels (base, 1st and 2nd). The 4 basic measures are the elements of the confusion matrix (also known as the contingency matrix). Based on this, the authors draw what they describe as "the periodic table of elements in binary classification performance". It is a map of the 44 scores, organized around the confusion matrix, the vertical dimension being related to the ground-truth class and the horizontal dimension to the predicted class. In this work, we avoid mixing scores

of different types and only consider those suitable for performance ordering and performance-based ranking of classifiers.

2.4. Reminders on the Axioms of Ranking

We also leverage the axiomatic definitions around the notion of performance introduced in [2]. In a nutshell, the 1st axiom states that if several performances have been ranked, then removing or adding a performance should not affect the relative order of the previously present performances. We will reuse the second and third axioms later in this paper, rephrased as follows.

Axiom 2 (reminder)**.** *If the degree of satisfaction that can be obtained with a 1st classifier is for sure less or equal than the degree of satisfaction that can be obtained with a 2nd one, then the former classifier is not better than the latter.*

Axiom 3 (reminder)**.** *Given a set of classifiers, and a arbitrary set of possible operations to perturb (*e.g.*, add noise to their output) and/or combine them, then, on the basis of their performances only, it must be impossible to determine a sequence of these operations that would lead with certainty to a classifier either better than the best of the initial ones or worse than the worst of the initial ones.*

Moreover, we show for the first time that it is possible to establish a continuous, two-dimensional map of an infinity of scores, and that the map includes many scores that are widespread in the literature.

3. Ranking Scores for Two-Class Classification

3.1. Particularization of Ranking Scores

To build the Tile, we consider the family of ranking scores in the special case of the two-class crisp classification task:

$$R_I(P) = \frac{\sum_{\omega\in\{tn,tp\}} I(\omega)P(\{\omega\})}{\sum_{\omega\in\{tn,fp,fn,tp\}} I(\omega)P(\{\omega\})} \tag{3}$$

where $I : \Omega \to \mathbb{R}_{\geq 0}$ is a non-negative random variable, called *importance*, such that $I(tn) + I(tp) \neq 0$ and $I(fp) + I(fn) \neq 0$. These scores allow for the comparison of all two-class classification performances in $\mathrm{dom}(R_I)$, even when the classes are imbalanced ($\pi_- \neq \pi_+$).

Example 1 (Probabilistic ranking scores)**.** *The family of ranking scores includes the following* 9 *probabilistic scores:* NPV, $X^C_{\{tn,tp\}|\{tn,fn,tp\}}$, TPR, J_-, A, J_+, TNR, $X^C_{\{tn,tp\}|\{tn,fp,tp\}}$, *and* PPV.

Example 2 (F-scores)**.** *For all $\beta \geq 0$, F_β is a ranking score.*

Example 3 (PABDC)**.** *The ranking scores for which the importance values $I(\omega)$ are rational numbers correspond to the class of* Presence/Absence Based Dissimilarity Coefficients *(PABDC) satisfying the first* 9 *properties listed in [93] (see Prop. 1 in that paper).*

3.2. Properties

We start by examining the effect of the target/prior shift operation [94], denoted by $\mathrm{shift}_{\pi\to\pi'}$. It transforms a performance P into P' such that $P'(E) = P(E)\frac{\pi'_-}{\pi_-}$ for all $E \in 2^{\{tn,fp\}}$ and $P'(E) = P(E)\frac{\pi'_+}{\pi_+}$ for all $E \in 2^{\{fn,tp\}}$.

Property 1. *The composition of a ranking score with a target/prior shift operation is a ranking score. We have $R_I \circ \mathrm{shift}_{\pi\to\pi'} = R_{I'}$ with $I'(\omega) = I(\omega)\frac{\pi'_-}{\pi_-}$ for all $\omega \in \{tn, fp\}$ and $I'(\omega) = I(\omega)\frac{\pi'_+}{\pi_+}$ for all $\omega \in \{fn, tp\}$.*

As we have particularized the axiomatic framework to the particular case of two-class crisp classification, the following property holds.

Property 2. *Multiplying both $I(tn)$ and $I(tp)$, or both $I(fp)$ and $I(fn)$, by the same strictly positive constant leads to another ranking score that is monotonously increasing with the original one. The induced performance ordering is thus unaffected by such a transformation.*

Thanks to the last property, we can get rid of the redundancy between the performance orderings induced by the ranking scores by focusing on the *canonical ranking scores*.

Definition 1. *Canonical ranking scores are given by*

$$R_{I_{a,b}} = \frac{(1-a)PTN + aPTP}{(1-a)PTN + (1-b)PFP + bPFN + aPTP},$$

where $a, b \in [0, 1]$ and $I_{a,b}$ is the importance given by $I_{a,b}(tn) = 1 - a$, $I_{a,b}(fp) = 1 - b$, $I_{a,b}(fn) = b$, and $I_{a,b}(tp) = a$.

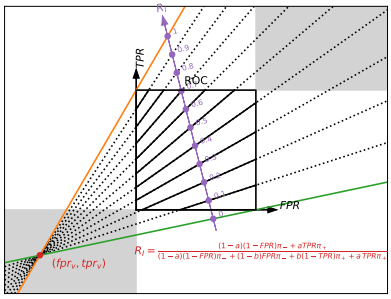
Consequently, the NPV, PPV, TNR, TPR, A, and F_β scores belong to the canonical ranking scores.

To avoid potential issues, it should be noted that averaging ranking scores cannot be done without precautions. In fact, there are only a few special cases in which a score obtained in this way can be used for ranking. The same precaution also applies for canonical ranking scores. For example, because of this issue, the orderings induced by ACP, P_4, and $VUT = \int_0^1 \int_0^1 R_{I_{a,b}}\, da\, db$ are incompatible with the axioms of ranking. If a compromise has to be found between several canonical ranking scores, we recommend averaging the importances rather than the scores.

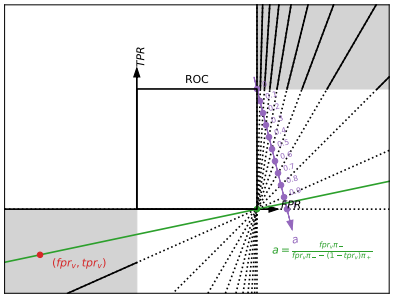
3.3. Correspondences with the ROC Space

We now study the correspondences between the canonical ranking scores and the *Receiver Operating Characteristic* (ROC) space (*i.e.*, $FPR \times TPR$) for fixed priors.

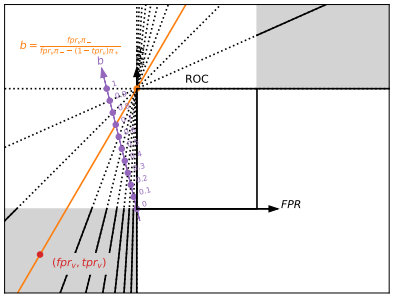
Three pencils of lines are depicted in Fig. 2. The first pencil (see Fig. 2a) can be used to read the value of $R_{I_{a,b}}$ in any point of ROC (on any line of slope $-\pi_-/\pi_+$, in purple



(a) The locus of equivalent performances (*i.e.*, those for which the ranking score takes a given value) are lines (restricted to ROC). These lines form a pencil whose vertex (fpr_v, tpr_v) (in red) is located outside the ROC space, either in the bottom-left or upper-right areas (in gray). The value taken by the score varies linearly along any line of slope $-\pi_-/\pi_+$ (in purple).



(b) The locus of the vertices (fpr_v, tpr_v) (red point) for all ranking scores with a given value of a are lines (restricted to the gray areas). These lines form a pencil whose vertex is located at the bottom-right corner of ROC. The value of a varies linearly along any line of slope $-\pi_-/\pi_+$ (in purple).



(c) The locus of vertices (fpr_v, tpr_v) (red point) for all ranking scores with a given value of b are lines (restricted to the gray areas). These lines form a pencil whose vertex is located at the top-left corner of ROC. The value of b varies linearly along any line of slope $-\pi_-/\pi_+$ (in purple).

Figure 2. The geometry of the ranking scores $R_{I_{a,b}}$ in the ROC plane (FPR, TPR). Example given for the class priors $\pi_+ = 1 - \pi_- = 0.2$ and the importance given by $(a, b) = (0.95, 0.7)$.

in the figure), when the location of the red point is known. The red point corresponds to the locus of iso-performance lines for a given ranking score. It corresponds to the intersection of the green and orange lines, that is, the iso-performance lines for which the score is equal to 0 and 1, respectively. The equations of these lines can be obtained by looking at other pencils, as following. The second pencil (see Fig. 2b) can be used to find the green line based on the value of the parameter a (the green line is vertical when $a = 0$ and horizontal when $a = 1$). Finally, the third pencil (see Fig. 2c) can be used to find the orange line based on the value of the parameter b (the orange line is vertical when $b = 0$ and horizontal when $b = 1$). This procedure generalizes the geometric constructions provided by Flach in [29] for A, F_1, and PPV.

Contour plots. Figure 2a can be considered as a *contour plot* depicting the preorder induced by a score. The depicted curves (lines in our case) are called *iso-performance lines* in [95] and *isometrics* in [29]. Applying any monotonic function to the score leaves these curves unchanged. Hence, a line corresponds to a set of *equivalent* performances. All performances that are on the top-left side of the line are *better than* them, and all performances that are on the bottom-right side of the line are *worse than* them. Note that this geometric analysis is not peculiar to the canonical ranking scores. It is valid for all ranking scores, as for any ranking score there exists a canonical ranking score such that the orderings induced by them are equal.

Interpretation with respect to Axiom 2. It is also interesting to note that, for any I, the red point is in one of the gray areas (extending to infinity). When $a < b$, the red point is on the right and above ROC, and when $a > b$, it is on the left and below ROC. When $a = b$, all lines in Fig. 2a (including the green and orange ones) are parallel (*e.g.*, for TNR, A, and TPR) and the red point is a point at infinity. This is related to Axiom 2. Indeed, in the case of fixed priors, the performance orderings induced by any score that has a pencil of iso-performance lines is a performance ordering that can be induced by a ranking score. This is because the family of ranking scores covers all cases where the vertex of the pencil is in one of the gray areas. Putting the vertex outside these areas would be illogical as there would either be better performances than the one located in the upper-left corner of ROC, or worse than the one located in the lower-right corner of ROC, which is prohibited by Axiom 2.

Interpretation with respect to Axiom 3. As the ROC space is, for fixed priors, a linear projection of $\mathbb{P}_{(\Omega,\Sigma)}$, any convex in the ROC space corresponds to a convex in $\mathbb{P}_{(\Omega,\Sigma)}$. Therefore, a convex combination of performances cannot be better than the best of the combined performances and cannot be worse than the worst of the combined performances.

On the diversity. The performance orderings induced by

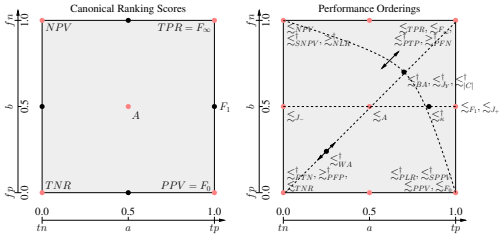


Figure 3. Placement of the canonical ranking scores (left) and of some performance orderings (right) on the Tile. The symbol † indicates the orderings that are specific for given priors. For the orderings whose locations are prior-dependent, we arbitrarily chose a negative prior of 0.7. Double arrows ↔ indicate the direction in which $\lesssim_{WA}$ moves when the weights are tuned and how the curve on which $\lesssim_{BA}$ and $\lesssim_{\kappa}$ moves when the priors are tuned. The colored points correspond to probabilistic scores.

the scores $R_{I_{a,b}}$ are all different. For two different ranking scores, one has different pencil vertices (red points), leading to different iso-performance lines, sets of equivalent performances, and performance orderings. Hence, there is no redundancy in the canonical ranking scores.

4. The Tile for Two-Class Classification

The Tile, which is depicted in Fig. 3, is defined as follows.

Definition 2. *The Tile is the mapping* $[0,1]^2 \to \mathbb{X}_{(\Omega,\Sigma)} : (a,b) \mapsto R_{I_{a,b}}$, *where* $\mathbb{X}_{(\Omega,\Sigma)}$ *denotes the set of all scores.*

Although attempts to organize a selection of two-class classification scores in the 2D plane are common, to our knowledge, this is the first time that it is done mathematically, quantitatively, and automatically, without the intervention of a human expert. The Tile is not limited to the spatial organization of scores, however.

4.1. Canonical Ranking Scores on the Tile

The left-hand side of Fig. 3 shows the layout of the Tile with some canonical ranking scores on it. Two opposite corners correspond to NPV and PPV, and the other corners to TNR and TPR. The accuracy A is in the center. These are the only 5 canonical ranking scores that are also probabilistic scores. F_β scores are between TPR and PPV ($\beta = \sqrt{b/1-b}$), F_1 being in the middle of the right side.

Interpolations. By construction, the canonical scores can be interpolated, in the Tile, as follows:

- horizontally: with the f-mean such that $f : x \mapsto x^{-1}$;
- vertically: with the f-mean such that $f : x \mapsto (1-x)^{-1}$.

The fact that F_1 (the harmonic mean between TPR and PPV) appears at the mid-position between TPR and PPV is a consequence of this property.

Indistinguishable samples. The scores that can be calculated when the two unsatisfying outcomes (fp and fn) are grouped together (*i.e.*, $\Omega = \{tn, tp, incorrect\}$) are those for which $I(fp) = I(fn)$. They are located on the median horizontal, passing through A. Likewise, the scores that can be calculated when two satisfying outcomes (tn and tp) are grouped together (*i.e.*, $\Omega = \{fp, fn, correct\}$) are those for which $I(tn) = I(tp)$. They are located on the median vertical, passing through A.

Operations on performances. Some remarkable geometric properties of the Tile are related to the operations that can be applied on performances. They are given hereafter.

- Changing either the predicted or ground-truth class amounts to, respectively, flipping the Tile around the raising diagonal (TNR-TPR axis) or falling diagonal (NPV-PPV axis), and complementing the scores to 1.
- Swapping the predicted and ground-truth classes is equivalent to vertical mirroring. Canonical ranking scores on the horizontal midline are *symmetric*, in the sense of Gosgens *et al*. [39].
- Swapping the positive and negative classes is equivalent to applying a central symmetry. The sole *class-symmetric* canonical ranking score, in the sense of Gosgens *et al*. [39], is thus the accuracy.
- The target/prior shift operation [94] moves the performance orderings on the Tile. We found it very useful, when priors are fixed, to represent the displacement that would have occurred if we had started with uniform priors and applied the target/prior shift.

4.2. Performance Orderings on the Tile

We have put some performance orderings on the right-hand side of Fig. 3 on the Tile. For a position (a,b), we have mentioned the orderings induced by the canonical ranking score $R_{I_{a,b}}$, and all the other orderings that are equal to it, either on $\mathbb{P}_{(\Omega,\Sigma)}$ or on a subset of it (fixed priors).

With all performances. The ordering $\lesssim_{R_I}$ corresponds to location $(a,b) = (\frac{I(tp)}{I(tn)+I(tp)}, \frac{I(fp)}{I(fn)+I(fp)})$. Those induced by the 9 probabilistic ranking scores given in Example 1 are depicted by colored points on Fig. 3. While it was not possible to place J_- and J_+ on the Tile of canonical ranking scores, the corresponding orderings $\lesssim_{J_-}$ and $\lesssim_{J_+}$ are on the Tile of performance orderings. The well-known fact that F_1 and J_+ lead to the same ordering [29] can be easily visualized on the Tile since $\lesssim_{F_1}$ and $\lesssim_{J_+}$ are at the same place. All orderings induced by the similarity coefficients of the T_θ and S_θ families, defined in [90], are equal to $\lesssim_{F_1}$ and $\lesssim_A$, respectively. The orderings induced by the family of similarity coefficients defined in [88] are the ones on the line segment between $\lesssim_A$ and $\lesssim_{F_1}$.

With fixed priors. It is also possible to place, on the Tile, the orderings that are equal to $\lesssim_{R_{I_{a,b}}}$ for a given

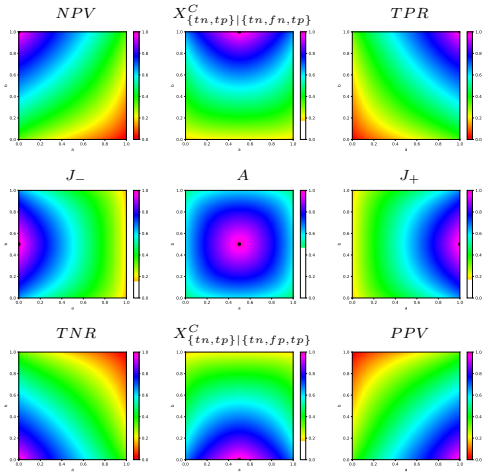
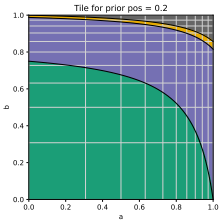


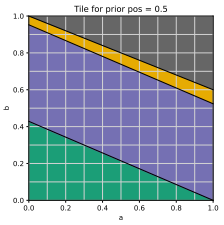
Figure 4. Tiles showing the rank correlations (Kendall τ) between 9 probabilistic scores (those that belong to the ranking scores, as given in Example 1), and all ranking scores, for a uniform distribution of performances. The correlation values have been estimated based on 10,000 performances drawn at random.

subset of performances only. An important practical case is when the priors are fixed. In this case, the orderings induced by the unconditional probabilistic scores PTN, PFP (dual order), PFN (dual order), and PTP can be placed on the Tile. Also, the standardized negative predictive value $SNPV$, the negative likelihood ratio NLR (dual order), the standardized positive predictive value $SPPV$, the positive likelihood ratio PLR, the weighted accuracy $WA = \lambda_- TNR + \lambda_+ TPR$, the balanced accuracy BA, the score $|C|$, Youden's index J_Y, and Cohen's κ can be placed on the Tile. Some of them have a fixed position, while for others, the position depends on the priors: $\lesssim_{WA}$, $\lesssim_{BA}$ and $\lesssim_{J_Y}$ sweep the ascending diagonal, while $\lesssim_\kappa$ sweeps the median horizontal. The performance ordering $\lesssim_{WA}$ for WA is at $(a,b) = (\frac{\lambda_+\pi_-}{\lambda_+\pi_- + \lambda_-\pi_+}, \frac{\lambda_+\pi_-}{\lambda_+\pi_- + \lambda_-\pi_+})$, $\lesssim_{BA}$ is at $(a,b) = (\pi_-, \pi_-)$, and $\lesssim_\kappa$ is at $(a,b) = (\frac{\pi_-^2}{\pi_-^2+\pi_+^2}, \frac{1}{2})$.

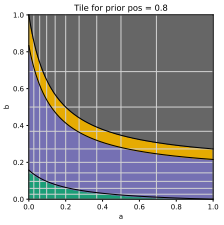
Characterizing scores. The Tile can be used to characterize any score, showing the rank correlations between that score and all canonical ranking scores, for a given performance distribution. For example, Fig. 4 shows the results obtained with the 9 probabilistic scores that belong to the ranking scores (see Example 1), for a uniform distribution of performances (*i.e.*, a symmetric Dirichlet distribution with all concentration parameters set to one). For this distribution, these rank correlations are either null (NPV with PPV, TNR with TPR) or positive. Such an analysis



	performances			
	P_- ●	P_1 ●	P_2 ●	P_+ ●
PTN	0.80	0.56	0.40	0.00
PFP	0.00	0.24	0.40	0.80
PFN	0.20	0.06	0.04	0.00
PTP	0.00	0.14	0.16	0.20



	performances			
	P_- ●	P_1 ●	P_2 ●	P_+ ●
PTN	0.50	0.35	0.25	0.00
PFP	0.00	0.15	0.25	0.50
PFN	0.50	0.15	0.10	0.00
PTP	0.00	0.35	0.40	0.50



	performances			
	P_- ●	P_1 ●	P_2 ●	P_+ ●
PTN	0.20	0.14	0.10	0.00
PFP	0.00	0.06	0.10	0.20
PFN	0.80	0.24	0.16	0.00
PTP	0.00	0.56	0.64	0.80

Figure 5. Toy examples showing on the Tile which of the 4 performances P_-, P_1, P_2, and P_+ is the best. The 3 examples differ in the class priors (either 0.2, 0.5, or 0.8 for the positive class). In all examples, P_- (●) is the performance of classifiers predicting always the negative class, P_1 (●) is such that $TNR(P_1) = 0.7$ and $TPR(P_1) = 0.7$, P_2 (●) is such that $TNR(P_2) = 0.5$ and $TPR(P_2) = 0.8$, and P_+ (●) is the performance of classifiers predicting always the positive class.

can be easily performed with any score; the Tile turns out to be a very practical visualization tool to gain intuition about the behavior of the plethora of scores that exist.

Visualizing the robustness. The importance values $I(\omega)$ are design choices for competitions. Several recent papers [44, 45] have alerted the scientific community about the necessary robustness: the performance-based rankings should not vary much when parameters are slightly perturbed. Figure 4 makes clear that, for a uniform distribution of performances, the performance orderings do not vary much when the parameters a and b are slightly perturbed.

4.3. Rankings on the Tile

For a given set of classifiers to rank, one can use the Tile to show which classifier is ranked first, second, third, *etc.*, according to $\lesssim_{R_{I_{a,b}}}$, in (a,b). This is shown in Fig. 5 with a

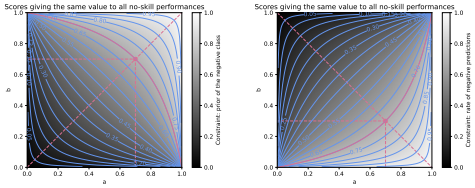


Figure 6. In the Tile, the ranking scores that put all no-skill performances on an equal footing are along a curve γ_π between NPV (upper-left corner) and PPV (lower-right corner) when the class priors $P(Y)$ are fixed, and along a curve γ_τ between TNR (lower-left corner) and TPR (upper-right corner) when the rates of predictions $P(\hat{Y})$ are fixed. The pink curves correspond to the constraints $\pi_- = 0.7$ (on the left) and $\tau_- = 0.7$ (on the right).

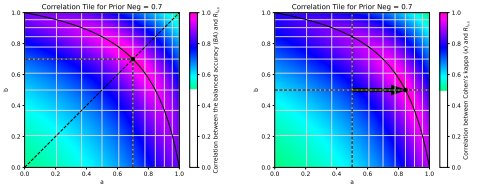


Figure 7. Tiles showing the Kendall rank correlation coefficient τ for BA (left) and κ (right), for a uniform distribution of performances P such that $\pi_- = 0.7$. The correlation values have been estimated based on 10,000 performances drawn at random.

toy example. When $\pi_- = \pi_+$, the regions where the different classifiers are ranked first are convex polygons. When $\pi_- \neq \pi_+$, the borders between these regions are curved.

4.4. More About No-Skill Performances

How can we rank no-skill performances ex aequo? The ranking scores allow ranking all no-skill performances (*i.e.*, those for which the groundtruth and predicted classes are independent) ex aequo when some constraints on the compared performances are added. The more common constraint is undoubtedly that the priors are fixed. In this case, the canonical ranking scores that put the no-skill performances on the same footing are located on the curve $\gamma_\pi : \pi_+^2\, a\, b = \pi_-^2\, (1-a)\, (1-b)$. Another interesting constraint is that the rates of predictions are fixed. By symmetry, the canonical ranking scores that put the no-skill performances on the same footing are located on the curve $\gamma_\tau : \tau_+^2\, a\, (1-b) = \tau_-^2\, (1-a)\, b$. The γ_π and γ_τ curves are depicted in Fig. 6.

A new look at the balanced accuracy BA and Cohen's kappa κ. Figure 7 shows the rank correlations for both BA (on the left) and κ (on the right). We can see that BA is perfectly correlated with the ranking scores at the intersection between the curve γ_π and the rising diagonal, which is at (π_-, π_-), and that κ is perfectly correlated with the ranking scores at the intersection between the curve γ_π and the median horizontal, which is at $(\frac{\pi_-^2}{\pi_-^2+\pi_+^2}, \frac{1}{2})$.

Correction for chance. The idea of correcting a score for what can be achieved by chance is common in the literature. Scott [96] and Fleiss [97] proposed a correction for accuracy. Cohen [53] proposed another correction for it, that differs in what is considered to be achievable by chance. We noticed that Scott's π and Fleiss's κ do not satisfy the axioms of ranking, even in the case of fixed priors. Cohen's correction, on the other hand, allows nothing more than what we can do with ranking scores, with fixed priors: correcting $R_{I_{a,b}}$ in the same way as Cohen [53] did with A, that is $\frac{R_{I_{a,b}} - R_{I_{a,b}}\circ\text{no–skill}}{1 - R_{I_{a,b}}\circ\text{no–skill}}$ leads to a score that is perfectly rank-correlated with $R_{I_{a',b'}}$, where $a' = \frac{\pi_-^2(1-b)}{\pi_-^2(1-b)+\pi_+^2 b}$ and $b' = b$. Geometrically, an entire horizontal line of the Tile is crushed into a single point (the intersection it has with the γ_π curve). There is thus an enormous loss of diversity when applying Cohen's correction to our scores.

5. Conclusion

In this paper, we presented the *Tile*, which organizes ranking scores for two-class classification task, according to the random variable called *importance* capable of considering application-specific preferences. The scores organized on this Tile are called the *canonical ranking scores* and include well-known scores, such as the accuracy or F_β scores. These canonical ranking scores lead to performance orderings that are different in each point. The Tile is a visual tool that can be used in different ways. It can be used to establish correspondences of the Tile with the ROC space and to show how to read the values taken by the canonical ranking scores in any point of the ROC space. We also showed how to use the Tile to (1) study the behavior of any score, by depicting rank correlations between a score and all ranking scores, (2) rank classifiers, using a toy example, (3) study the influence of priors on the ranking scores, (4) study properties of no-skill performances, and (5) clarify what the balanced accuracy and Cohen's kappa are. In summary, the Tile offers a comprehensive visual framework for ranking two-class classifiers, empowering researchers and practitioners to make informed, application-specific decisions, and ultimately driving advancements in the field of machine learning.

References

[74] Douglas G. Altman and Martin Bland. Diagnostic tests 2: Predictive values. *Br. Medical J.*, 309(6947):102–102, 1994.

[55] Davide Ballabio, Francesca Grisoni, and Roberto Todeschini. Multivariate comparison of classification performance measures. *Chemom. Intell. Lab. Syst.*, 174:33–44, 2018.

[80] Ricardo Barandela, Josep Salvador Sánchez, Vicente García, and Erick Rangel. Strategies for learning in class imbalance problems. *Pattern Recognit.*, 36(3):849–851, 2003.

[88] Ildar Z. Batyrshin, Nailya Kubysheva, Valery Solovyev, and Luis A. Villa-Vargas. Visualization of similarity measures for binary data and 2x2 tables. *Comput. Y Sist.*, 20(3):345–353, 2016.

[93] Forrest B. Baulieu. A classification of presence/absence based dissimilarity coefficients. *J. Classif.*, 6(1):233–246, 1989.

[78] Kay Henning Brodersen, Cheng Soon Ong, Klaas Enno Stephan, and Joachim M. Buhmann. The balanced accuracy and its posterior distribution. In *IEEE Int. Conf. Pattern Recognit. (ICPR)*, pages 3121–3124, Istanbul, Turkey, 2010.

[73] Christopher D. Brown and Herbert T. Davis. Receiver operating characteristics curves and related decision measures: A tutorial. *Chemom. Intell. Lab. Syst.*, 80(1):24–38, 2006.

[89] Michael Brusco, J. Dennis Cradit, and Douglas Steinley. A comparison of 71 binary similarity coefficients: The effect of base rates. *PLOS ONE*, 16(4):e0247751, 2021.

[86] Moisès Burset and Roderic Guigó. Evaluation of gene structure prediction programs. *Genomics*, 34(3):353–367, 1996.

[77] Ted Byrt, Janet Bishop, and John B. Carlin. Bias, prevalence and kappa. *J. Clin. Epidemiology*, 46(5):423–429, 1993.

[47] Gürol Canbek, Seref Sagiroglu, Tugba Taskaya Temizel, and Nazife Baykal. Binary classification performance measures/metrics: A comprehensive visualized roadmap to gain new insights. In *Int. Conf. Comput. Sci. Eng. (UBMK)*, pages 821–826, Antalya, Turkey, 2017.

[37] Seung-Seok Choi, Sung-Hyuk Cha, and Charles C. and Tappert. A survey of binary similarity and distance measures. *J. Syst. Cybern. Informatics*, 8(1):43–48, 2010.

[53] Jacob Cohen. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.*, 20(1):37–46, 1960.

[59] Cèsar Ferri, José Hernández-Orallo, and Ramona Modroiu. An experimental comparison of performance measures for classification. *Pattern Recognit. Lett.*, 30(1):27–38, 2009.

[29] Peter A. Flach. The geometry of ROC space: Understanding machine learning metrics through ROC isometrics. In *Int. Conf. Mach. Learn. (ICML)*, pages 194–201, Washington, DC, USA, 2003.

[97] Joseph L. Fleiss. Measuring nominal scale agreement among many raters. *Psychol. Bull.*, 76(5):378–382, 1971.

[83] Ronen Fluss, David Faraggi, and Benjamin Reiser. Estimation of the Youden index and its associated cutoff point. *Biom. J.*, 47(4):458–472, 2005.

[71] Ian A. Gardner and Matthias Greiner. Receiver-operating characteristic curves and likelihood ratios: improvements over traditional methods for the evaluation and application of veterinary clinical pathology tests. *Veterinary Clin. Pathol.*, 35(1):8–17, 2006.

[72] Afina S. Glas, Jeroen Lijmer, Martin H. Prins, Gouke Bonsel, and Patrick M. M. Bossuyt. The diagnostic odds ratio: a single indicator of test performance. *J. Clin. Epidemiology*, 56(11):1129–1135, 2003.

[39] Martijn Gösgens, Anton Zhiyanov, Alexey Tikhonov, and Liudmila Prokhorenkova. Good classification measures and how to find them. In *Adv. Neural Inf. Process. Syst. (NeurIPS)*, pages 1–12, Vancouver, Can., 2021.

[90] John C. Gower and Pierre Legendre. Metric and euclidean properties of dissimilarity coefficients. *J. Classif.*, 3(1):5–48, 1986.

[81] Xinjian Guo, Yilong Yin, Cailing Dong, Gongping Yang, and Guangtong Zhou. On the class imbalance problem. In *Int. Conf. Nat. Comput.*, pages 192–201, Jinan, China, 2008.

[70] Thomas F. Heston. Standardizing predictive values in diagnostic imaging research. *J. Magn. Reson. Imaging*, 33(2): 505–505, 2011.

[68] Robin J. Hogan, Christopher A. T. Ferro, Ian T. Jolliffe, and David B. Stephenson. Equitability revisited: Why the "equitable threat score" is not equitable. *Weather. Forecast.*, 25(2):710–726, 2010.

[60] Mohammad Hossin and M. N. Sulaiman. A review on evaluation metrics for data classification evaluations. *Int. J. Data Min. & Knowl. Manag. Process.*, 5(2):1–11, 2015.

[61] Nathalie Japkowicz and Mohak Shah. *Evaluating Learning Algorithms: A Classification Perspective*. Camb. Univ. Press, 2011.

[75] Uzay Kaymak, Arie Ben-David, and Rob Potharst. The AUK: A simple alternative to the AUC. *Eng. Appl. Artif. Intell.*, 25(5):1082–1089, 2012.

[79] Miroslav Kubat. Addressing the curse of imbalanced training sets: one-sided selection. In *Int. Conf. Mach. Learn. (ICML)*, pages 179–186. Morgan Kaufmann, 1997.

[44] Lena Maier-Hein, Matthias Eisenmann, Annika Reinke, Sinan Onogur, Marko Stankovic, Patrick Scholz, Tal Arbel, Hrvoje Bogunovic, Andrew P. Bradley, Aaron Carass, Carolin Feldmann, Alejandro F. Frangi, Peter M. Full, Bram van Ginneken, Allan Hanbury, Katrin Honauer, Michal Kozubek, Bennett A. Landman, Keno März, Oskar Maier, Klaus Maier-Hein, Bjoern H. Menze, Henning Müller, Peter F. Neher, Wiro Niessen, Nasir Rajpoot, Gregory C. Sharp, Korsuk Sirinukunwattana, Stefanie Speidel, Christian Stock, Danail Stoyanov, Abdel Aziz Taha, Fons van der Sommen, Ching-Wei Wang, Marc-André Weber, Guoyan Zheng, Pierre Jannin, and Annette Kopp-Schneider. Why rankings of biomedical image analysis competitions should be interpreted with care. *Nat. Commun.*, 9(1):1–13, 2018.

[91] Ivan Ramirez Mejia and Ildar Batyrshin. Towards a classification of binary similarity measures. In *Adv. Soft Comput.*, pages 325–335. Springer Int. Publ., 2018.

[45] Tran Thien Dat Nguyen, Hamid Rezatofighi, Ba-Ngu Vo, Ba-Tuong Vo, Silvio Savarese, and Ian Reid. How trustworthy are performance evaluations for basic vision tasks? *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(7):8538–8552, 2023.

[2] Sébastien Piérard, Anaïs Halin, Anthony Cioppa, Adrien Deliège, and Marc Van Droogenbroeck. Foundations of the

theory of performance-based ranking. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 14293–14302, Nashville, TN, USA, 2025.

[62] David M. W. Powers. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness & correlation. *J. Mach. Learn. Technol.*, 2(1):37–63, 2011.

[85] David M. W. Powers. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv*, abs/2010.16061, 2020.

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

[96] William A. Scott. Reliability of content analysis: The case of nominal scale coding. *Public Opin. Q.*, 19(3):321–325, 1955.

[94] Tomas Sipka, Milan Sulc, and Jiri Matas. The hitchhiker's guide to prior-shift adaptation. In *IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, pages 2031–2039, Waikoloa, HI, USA, 2022.

[87] Mikołaj Sitarz. Extending F1 metric, probabilistic approach. *Adv. Artif. Intell. Mach. Learn.*, 3(2):1025–1038, 2023.

[38] Marina Sokolova and Guy Lapalme. A systematic analysis of performance measures for classification tasks. *Inf. Process. & Manag.*, 45(4):427–437, 2009.

[46] Putnam P. Texel. Measure, metric, and indicator: An object-oriented approach for consistent terminology. *Proc. IEEE Southeastcon*, pages 1–5, 2013.

[69] Matthijs J. Warrens. The effect of combining categories on Bennett, Alpert and Goldstein's S. *Stat. Methodol.*, 9(3): 341–352, 2012.

[92] Matthijs J. Warrens. A comparison of multi-way similarity coefficients for binary sequences. *Int. J. Res. Rev. Appl. Sci.*, 16(1):64–75, 2013.

[76] Daniel S. Wilks. *Statistical methods in the atmospheric sciences*. Elsevier, fourth edition, 2020.

[84] Matthias Wimmer, Bernd Radig, and Michael Beetz. A person and context specific approach for skin color classification. In *Int. Conf. Pattern Recognit. (ICPR)*, pages 39–42, Hong Kong, China, 2006.

[82] William John Youden. Index for rating diagnostic tests. *Cancer*, 3(1):32–35, 1950.

5. The flavors (5th article)

This chapter reproduces the following article [5]:

> A. Halin*, S. Piérard*, A. Cioppa, and M. Van Droogenbroeck, "A hitchhiker's guide to understanding performances of two-class classifiers," *arXiv*, vol. abs/2412.04377, 2024.

It covers both the notions of ***performance*** and ***ranking***. The supplementary material of this article is reproduced in Chapter D.

Context

In Chapter 4, we saw that, for tasks similar to two-class crisp classification, we can map the manifold of meaningful rankings (at least those induced by ranking scores) onto a square. The position of a ranking is determined solely by the relative importance given to a true positive *vs* a true negative and the relative importance given to a false negative *vs* a false positive. This map offers various flavors by covering the square with textures showing functions of the importance in false colors.

Motivation

The Tile is an extremely versatile and practical tool, due to the ease with which truly useful flavors can be defined. When we look at the figures in the article [4] presented in the previous chapter, we see textures showing correlation values, values taken by scores (ranking scores, class priors, and prediction rates), and entities. It remains to explain how these flavors can be used in practice, what we do in this chapter.

Main contributions

While the article [4] presented in the previous chapter was theory-oriented, this one is intended to be practical. We present several flavors of Tiles: *correlation Tiles*, *(canonical ranking score) value Tiles*, *baseline value Tiles*, *state-of-the-art value Tiles*, *ranking Tiles*, *entity Tiles*, *no-skill Tiles*, and *relative-skill Tiles*. We have deliberately simplified things by presenting four of these flavors as the solution to a need corresponding to a user profile: a theoretical analyst, a method designer, a benchmarker, and an app developer.

Note that the stacking of all entity Tiles represents graphically all the rankings induced by the ranking scores. This stacking is called *mille-feuille* in Chapter 7. One is shown in Fig. 5.1. The ranking Tiles can de deduced from it.

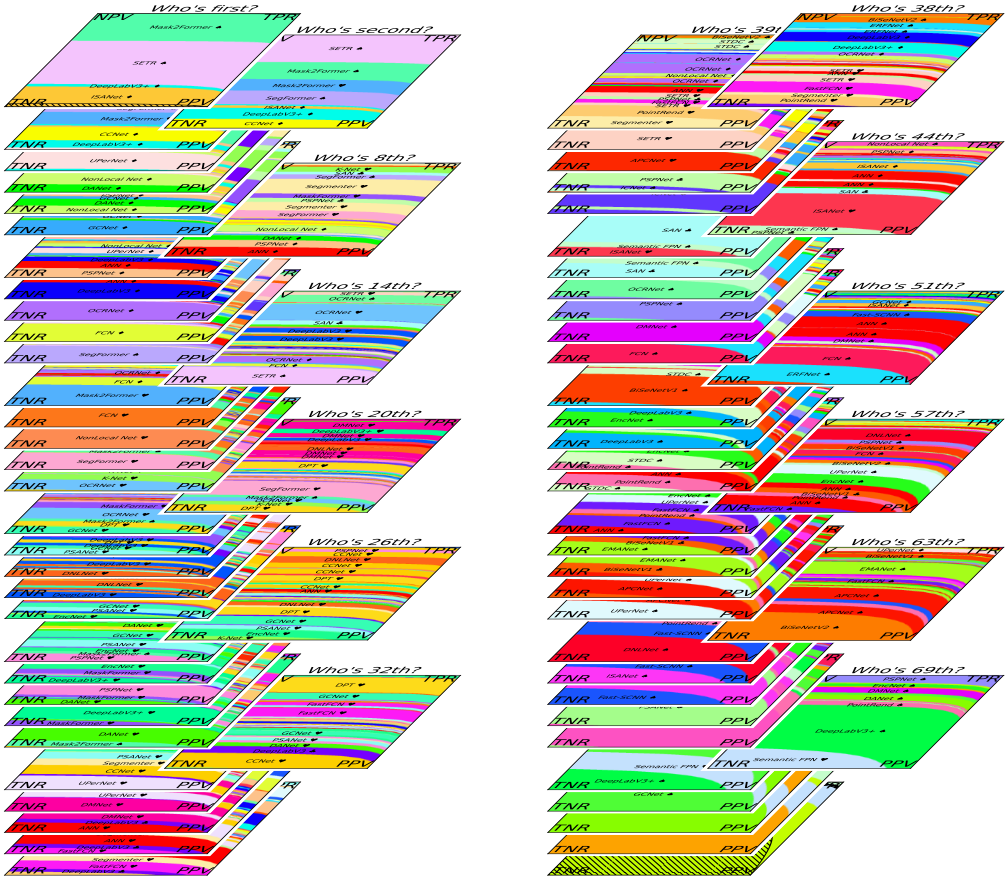


Figure 5.1.: This mille-feuille shows the stacking of all entity tiles, from the first rank (top) to the last one (bottom). The entities are the 74 semantic segmentation methods of [5].

Perspectives

The flavors introduced in this article are not peculiar to two-class crisp classification problems. The same principles could be used with any other task, but indeed it would not be possible to visualize the flavors on Tiles as the Tile is a graphical tool peculiar to two-class crisp classification.

As already said, the current practice is to rank entities based on an arbitrary score, for which the associated application preferences are unclear, and for which it is unknown whether the resulting classifications are meaningful. Without a solid theoretical foundation, these practices present ***methodological risks***, ***exaggerated confidence***, and ***impoverishment***, as there is a strong temptation to discard entities that are not top-ranked. However, these entities may have advantages that make them ideal for certain applications and user preferences. The flavors introduced in this paper have the potential to be impactful for the scientific community.

- The correlation flavor is a practical solution to the issue of ***methodological risks***. For a given distribution of performances (*e.g.*, those listed in Chapter 3), one can observe how some score X is rank-correlated with all ranking scores. If a perfect correlation is reached at some I, then it means that X can be used to establish rankings when the performances belong to the support of the studied distribution. Otherwise, it is a clear indicator that X cannot be used to rank without checking first that the induced performance ordering satisfies the axioms of Chapter 2.
- The value and ranking flavors are practical solutions to the issue of ***exaggerated confidence***. With them, one can observe the strengths and weaknesses of any entity.
- The ranking flavor is a practical solution to the issue of ***impoverishment***. Indeed, by observing the set of entities that can be ranked number one, one will be tempted to keep all of them.

Indeed, it is possible to define other flavors. See Chapters 6 and 7 for other examples.

Remarks

When we wrote this article, we already understood that there was a definite advantage to have Tiles displaying various types of information. These different Tiles were then seen as "Tile flavors". Since then, my conception has evolved slightly. I now see the Tile as a unique graphical tool that: (i) maps certain importances (*e.g.*, the canonical importances for which $I(tn) + I(tp) = I(fp) + I(fn)$) onto a square; (ii) possibly covers it with a texture in false colors; and (iii) finally draws annotations (*e.g.*, a grid, some entity names, the locus of scores putting all no-skill performances on an equal footing, the position of some scores, etc) on top. The information shown by the texture is not a purely graphical object but a mathematical object: it is a function of the importance I. This function can be either symbolic (*e.g.*, an *entity flavor*) or numeric (*e.g.*, a *correlation flavor*, a *value flavor*, or a *ranking flavor*). I therefore invite the reader to consider, for example, "correlation Tile" as a shortcut for "Tile showing a correlation flavor". It is useful to make this distinction because such a function is not intended solely for graphical representation. One may want to evaluate it, study its properties, or manipulate it in various ways. It is this vision that is currently being implemented in the Sorbetto library [67].

A Hitchhiker's Guide to Understanding Performances of Two-Class Classifiers

Anaïs Halin*, Sébastien Piérard*, Anthony Cioppa, and Marc Van Droogenbroeck
Montefiore Institute, University of Liège, Liège, Belgium
{Anais.Halin,S.Pierard,Anthony.Cioppa,M.VanDroogenbroeck}@uliege.be

Abstract

Properly understanding the performances of classifiers is essential in various scenarios. However, the literature often relies only on one or two standard scores to compare classifiers, which fails to capture the nuances of application-specific requirements. The Tile is a recently introduced visualization tool organizing an infinity of ranking scores into a 2D map. Thanks to the Tile, it is now possible to compare classifiers efficiently, displaying all possible application-specific preferences instead of having to rely on a pair of scores. This hitchhiker's guide to understanding the performances of two-class classifiers presents four scenarios showcasing different user profiles: a theoretical analyst, a method designer, a benchmarker, and an application developer. We introduce several interpretative flavors adapted to the user's needs by mapping different values on the Tile. We illustrate this guide by ranking and analyzing the performances of 74 state-of-the-art semantic segmentation models through the perspective of the four scenarios. Through these user profiles, we demonstrate that the Tile effectively captures the behavior of classifiers in a single visualization, while accommodating an infinite number of ranking scores. Code for mapping the different Tile flavors is available in supplementary material.

1. Introduction

As humans, performance and ranking are widespread in all aspects of our lives. For instance, at school, teachers evaluate tests using a score which reflects the performance of students. In some disciplines such as calculus, evaluation is straightforward as there are only two possible cases: either the answer is correct or wrong. The score can then be calculated as the ratio of correct answers to the total number of questions. Likewise, in most team sports, team A beats team B if they score more points. Even for reviewing papers, area chairs use scores provided by the reviewers to assess if a paper should be accepted or rejected [98].

*Equal contributions.

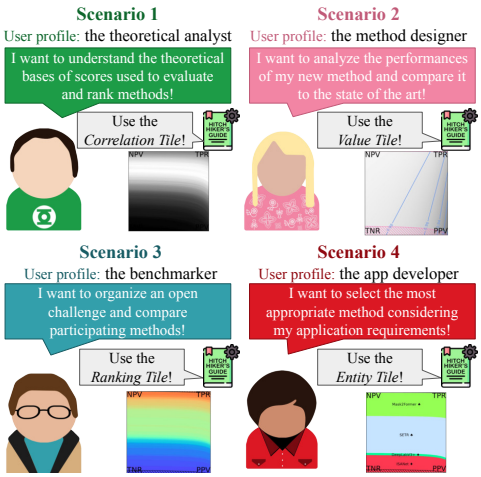


Figure 1. **Our hitchhiker's guide.** This hitchhiker's guide to understanding performances of two-class classifiers addresses four scenarios, answering specific requests from four user profiles: (1) *the theoretical analyst*, who is interested in understanding the theoretical relationship between different scores typically used for evaluating or ranking methods, (2) *the method designer*, who would like to analyze the performances of his/her new method and compare it to others, (3) *the benchmarker*, who organizes challenges for the scientific community and would like to know how to rank participating methods, and finally (4) *the application developer*, who wants to select the most appropriate method for his/her application. This guide provides specific tools and explains how to interpret them for each of those four scenarios.

However, not all evaluations are well-defined. For instance, when grocery shopping, consumers may choose product A over product B looking at different characteristics such as the price, the amount of sugar, or the packaging. In this case, the choice is based on several, sometimes contradictory scores. The question is, therefore, which score should the choice be based on? Similar questions arise in the field of machine learning: How can we determine

if a newly designed classifier outperforms existing ones? Which score(s) should we use to analyze its performance? How do we decide which score to consider for ranking the classifiers? Answers to these questions are not trivial, while potentially having a big impact on the development of a whole field of research.

In this paper, we propose a step-by-step hitchhiker's guide to help compare, analyze, and rank two-class classifiers. Throughout this guide, we study four scenarios, answering specific requests that four common user profiles may have, and provide visual tools adapted to their needs, as illustrated in Fig. 1. To do so, we rely on the infinite and parametric family of scores satisfying an axiomatic definition of the performance-based rankings introduced in [2] and the tool called the *Tile* introduced in [4]. More precisely, we leverage the newly introduced Tile tool to visualize various elements useful for understanding the performances of two-class classifiers from different perspective and considering an infinite number of scores. As such, we propose several ways to construct, use, and interpret different Tile flavors in the four scenarios, offering practical guidance to different user profiles with varied objectives, such as theoretically analyzing scores, developing new methods, benchmarking challenges, or designing applications.

We provide Python code that takes a list of two-class classification performances and generates a report (see supplementary material) from which most of the figures in this guide are issued. This report is intended for the theoretical analyst, the method designer, the benchmarker, and the application developer. It contains a recap of the compared performances (tables and plots), all flavors described in this guide (Correlation Tiles, Value Tiles, Baseline Value Tile, State-of-the-Art Value Tile, Ranking Tiles, and Entity Tile) as well as others (No-Skill Tile, Relative-Skill Tile), and gives several recommendations for the models to select (see scenario 4). Besides generating a report, our code also provides implementations of algorithms introduced in [4].

Contributions. We summarize our contributions as follows. **(i)** We provide the first hitchhiker's guide to understanding the performances of two-class classifiers, anchored in rigorous theoretical foundations. **(ii)** Throughout the guide, we answer the specific needs of four common user profiles via four scenarios by detailing which tool they should use, how they are constructed, and how they should interpret the results. **(iii)** We illustrate our guide for the computer vision community with an analysis and ranking of 74 state-of-the-art semantic segmentation models.

2. Related Work

For the analysis of performance measures, which is the focus of our guide, two strategies have emerged.

The first strategy consists in providing a series of performance scores that highlight the benefits and trade-offs in designing a classifier. Tharwat *et al*. [99] analyzed a series of scores for classification tasks that can be used alone or in combination. For example, the weighted harmonic mean between the recall TPR and the precision PPV, denoted by F_β, or simply by F_1 when weights are equal, is advocated by many authors. A score similar to F_1 is Jaccard's coefficient. Yet, it is known that Jaccard's coefficient and the F_1 score lead to the same ranking of classifiers, although Jaccard is probabilistic, unlike F_1 [29, 62]. However, the problem with the first strategy is that even though we can use some scores for analyzing the performances of classifiers and ranking them, it is not clear which ones to use.

The second strategy makes use of so-called evaluation spaces, such as the *Receiver Operating Characteristic* (ROC) [100] or *Precision-Recall* (PR) spaces, that combine two scores. In [30, 101], Fawcett explained how to use ROC graphs and avoid interpretation pitfalls. For instance, the area under the ROC curve (AUC-ROC) is commonly used as a performance indicator, as it has statistical significance [102]. Although several authors raised concerns about ROC-derived scores in decision-making [103, 104], recent works such as the one on contingency spaces [105] still build upon ROC graphs. More recently, the PR space has become the de facto replacement for the ROC space in the presence of imbalance [106], ignoring the fact that there is a bijection between these two spaces [24]. Moreover, the PR space suffers from the presence of an unachievable region [26], which is often overlooked in practice. The primary drawback of ROC- or PR-based curves analyzes is that, according to Menon *et al*. [107], the AUC-ROC and AUC-PR apply not to a classifier but to a scoring function, and that a scoring function yields a family of classifiers obtained for different thresholds. In other words, these curves consider a parametric family of classifiers rather than a specific classifier, and they are therefore inadequate to select classifiers when it comes to comparing unique instances from different families. As an example of these drawbacks, we show the performance of 74 two-class classifiers in the ROC space in Fig. 2. As can be seen, it is extremely challenging to uniquely determine which classifier is the best or the worst. Moreover, it is even harder to rank all 74 classifiers based only on this ROC space.

While these strategies provide a good insight from multiple perspectives, they only offer a scattered interpretation due to the inherent partial redundancy between scores, and ultimately only offer an incomplete look on the performances and ranking. In a recent attempt to formalize the notion of ranking, Nguyen *et al*. [45] proposed to impose three properties for ranking: (1) reliability, (2) meaningfulness (evaluated by humans), and (3) mathematical consistency. Pierard *et al*. [2] presents an alternative formal axiomatic definition of performance-based rankings, grounded in order theory and anchored within a probabilistic framework.

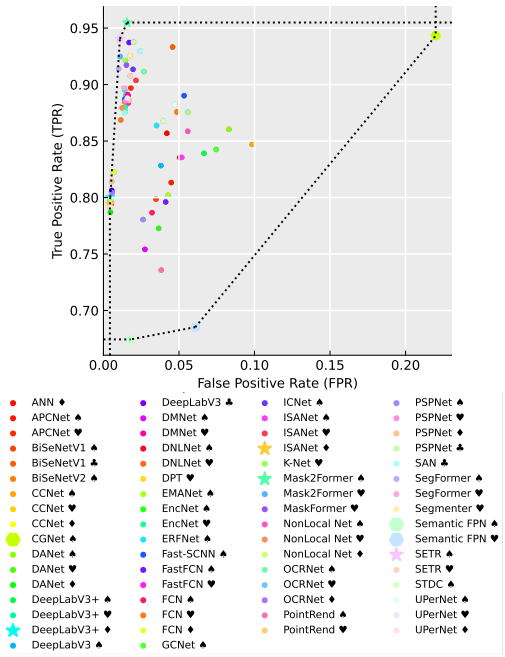


Figure 2. **How can we rank all these classification performances using the ROC space?** The classical ROC space does not provide an answer at a glance. The 74 performances shown here (for a positive prior equal to 0.124) serve as the showcase for our illustration in this guide. The dash lines correspond to the supremum and infimum of all achievable performances using a combination of the classifiers.

The axioms ensure the stability of rankings, *i.e.*, that if multiple entities, in our case two-class classifiers, are ranked, adding or removing an entity does not affect the relative order of the previously present entities. Additionally, it introduces ranking scores that satisfy these axioms, parameterized by a random variable I, called *importance*, which allows for consideration of application-specific preferences. Pierard *et al*. [4] proposes a spatial organization of the ranking scores in a 2D square map, called the Tile, for the particular case of two-class classification. It also studies properties of the Tile from a theoretical perspective. In this work, we show how both the axiomatic framework and the Tile can be used in practice and propose a hitchhiker's guide to use the Tile in various ways through four scenarios and user profiles.

3. Hitchhiker's Guide

This section presents our hitchhiker's guide to understanding performances of two-class classifiers. We begin by providing the recommended basic theoretical knowledge for using the guide, including a description of the terminology and notations, the general Tile tool, and the illustrative context used throughout. Next, we detail four scenarios, each aligned with the needs of a specific user profile. For each scenario, we highlight the relevant flavors of the Tile, and explain how to interpret the results.

3.1. Recommended Basic Theoretical Knowledge

Terminology and notations. To be consistent with [2] and [4], we use the same terminology and notations. Hence, an *entity* ϵ is a two-class classifier and the set of all entities of interest is denoted by $\mathbb{E}$. *Performances*, denoted by P, are probability measures, and the performance P_ϵ of an entity ϵ is the evaluation of this entity. *Scores* are functions associating a real value to performances, that is, $X : \mathrm{dom}(X) \to \mathbb{R} : P \mapsto X(P)$, with the domain of the score, $\mathrm{dom}(X)$, included in the set of all probability measures, $\mathbb{P}_{(\Omega,\Sigma)}$, on the measurable space (Ω, Σ), where Ω is the sample space (*i.e.*, the set of outcomes) and Σ is the event space (*i.e.*, a σ-algebra on Ω).

Construction and interpretation of the Tile. As explained in [4], the Tile for two-class classification is obtained by organizing ranking scores on a 2D square map, where the two axes, a and b, respectively represents the importance I given to true positive (tp) compared to true negative (tn), and false negative (fn) compared to false positive (fp):

$$a = I(tp) = 1 - I(tn)\,, \tag{1}$$

$$b = I(fn) = 1 - I(fp)\,. \tag{2}$$

The importance value can be arbitrarily chosen to reflect application-specific preferences. In the particular case of two-class classification, the ranking scores are given by

$$R_I(P) = \frac{I(tn)P(\{tn\}) + I(tp)P(\{tp\})}{I(tn)P(\{tn\}) + I(fp)P(\{fp\}) + I(fn)P(\{fn\}) + I(tp)P(\{tp\})}\,, \tag{3}$$

with $P(\{tn\})$ (resp. $P(\{fp\})$, $P(\{fn\})$, and $P(\{tp\})$) corresponding to the probability of the event $\{tn\}$ (resp. $\{fp\}$, $\{fn\}$, and $\{tp\}$). The Tile is then defined as follows:

Definition 1. *The Tile for two-class classification is the mapping*

$$[0,1]^2 \to \mathbb{X}_{(\Omega,\Sigma)} : (a,b) \mapsto R_I$$

with $\mathbb{X}_{(\Omega,\Sigma)}$ *denoting all possible scores on* (Ω, Σ) *and* $I(tn) = 1 - a$, $I(fp) = 1 - b$, $I(fn) = b$, $I(tp) = a$.

The layout of the Tile is shown in Fig. 3. By construction, the top-right corner of the Tile gives maximum importance values to both tp and fn (*i.e.*, $I(tp) = I(fn) = 1$).

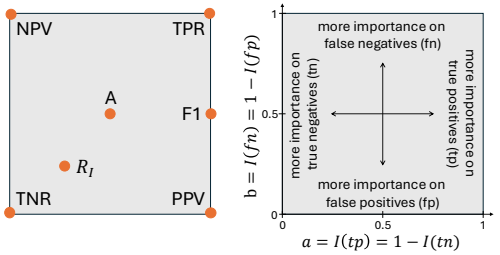


Figure 3. **Tile with ranking scores for two-class classification.** Each point of the Tile corresponds to a ranking score R_I which can be computed using Eq. (3). The Tile therefore contains an infinity of ranking scores, including the popular Accuracy (A), Negative Predictive Value (NPV), True Positive Rate (TPR), Positive Predictive Value (PPV), True Negative Rate (TNR), and F_1 scores. The location of a ranking score R_I in the Tile is defined by the values of a and b, which indicate the importance given to true positives (tp), true negatives (tn), false negatives (fn), and false positives (fp). In all following figures, the Tile flavors adopt the same layout as the one on the left hand side. For ease of reading, the names of the scores at the four corners are systematically written.

The ranking score in this corner is thus the True Positive Rate (TPR), also known as recall or sensitivity:

$$TPR = \frac{P(\{tp\})}{P(\{tp\})+P(\{fn\})} = P(Y = \hat{Y}|Y = c_+)\,, \quad (4)$$

where Y is the ground truth, $\hat{Y}$ the prediction, and c_+ the positive class. Conversely, the bottom-left corner gives minimum importance values to both tp and fn (*i.e.*, $I(tp) = I(fn) = 0$), leading to the ranking score named the True Negative Rate (TNR), also known as specificity:

$$TNR = \frac{P(\{tn\})}{P(\{tn\})+P(\{fp\})} = P(Y = \hat{Y}|Y = c_-)\,, \quad (5)$$

where c_- is the negative class. Similarly, the two other corners correspond to the Negative Predictive Value (NPV) and the Positive Predictive Value (PPV), also known as precision:

$$NPV = \frac{P(\{tn\})}{P(\{tn\})+P(\{fn\})} = P(Y = \hat{Y}|\hat{Y} = c_-)\,, \quad (6)$$

$$PPV = \frac{P(\{tp\})}{P(\{tp\})+P(\{fp\})} = P(Y = \hat{Y}|\hat{Y} = c_+)\,. \quad (7)$$

The score in the middle of the Tile, *i.e.*, giving the same importance value to tp, tn, fn, and fp (*i.e.*, $I(tp) = I(tn) = I(fn) = I(fp) = 0.5$) is the popular accuracy A, also defined as $P(Y = \hat{Y})$:

$$A = \frac{0.5\,P(\{tp\})+0.5\,P(\{tn\})}{0.5\,P(\{tp\})+0.5\,P(\{fn\})+0.5\,P(\{fp\})+0.5\,P(\{tn\})}\,. \quad (8)$$

Finally, F_β scores, with $\beta = \sqrt{b/1-b}$, are on the right-hand side that joins TPR to PPV, F_1 being located in the center of this side following:

$$F_1 = \frac{P(\{tp\})}{P(\{tp\}) + 0.5\,P(\{fp\}) + 0.5\,P(\{fn\})}\,. \quad (9)$$

As shown in this section, the Tile organizes an infinity of ranking scores into a 2D map through the concept of importance. Each point on the Tile corresponds to a score that assigns varying importance values to tp, tn, fp, and fn. This makes the Tile a useful visual tool for displaying the values these scores take for a given entity. Through our scenarios, we explain how to map different values on this tile for different user profiles.

Description of our illustration. In the following, we illustrate all our Tile flavors on a same use case related to computer vision for autonomous driving. We demonstrate that the applicability of the Tile and its flavors is far from being limited to two-class classification problems. Specifically, we compare and rank 74 state-of-the-art semantic segmentation models from the MMSegmentation toolbox [108]. What makes this use case challenging is that the sets of predicted classes differ from one model to another. This is because the models have been trained on several datasets with heterogeneous labels. As shown in Tab. 1, these include 31 models trained on ♠ Cityscapes [109], 27 on ♥ ADE20K [110], 12 on ♦ Pascal VOC 2012 [111], and 4 on ♣ COCO-Stuff 164k [112]. Motivated by the preponderant role played by the foreground-background differentiation in autonomous driving, we decided to group the semantic labels into a negative class, corresponding to background objects (*e.g.*, road, sidewalk, or building), and a positive class, corresponding to foreground, potentially moving, objects (*e.g.*, person, rider, or car). The grouping has been performed for both the learning sets and BDD100K [148], that we selected as testing set (see supplementary material for a detailed explanation on how it was performed). By doing so, the performances of the 74 models become directly comparable.

3.2. Four Scenarios of our Hitchhiker's Guide

This section provides a step-by-step guide for using the Tile to rank two-class classifiers (*i.e.*, entities), by illustrating its various potential uses through four scenarios. Each scenario corresponds to one user profile among (1) the theoretical analyst, who is interested in understanding the theoretical relationship between different scores typically used for evaluating or ranking methods, (2) the method designer, who needs to analyze the performances of his/her new method and compare it to others, (3) the benchmarker, who organizes challenges for the scientific community and has to rank participating methods, and finally (4) the application developer, who should be able to select the most appropriate method for his/her application. In the final scenario, we

Table 1. List of the models (columns) sorted w.r.t. the date of publication (from 2015 to 2023) from the MMSegmentation toolbox [108] trained on one or multiple datasets (rows). We respectively have 31 models trained on the ♠ Cityscapes [109] dataset, 27 on the ♥ ADE20K [110] dataset, 12 on the ♦ Pascal VOC 2012 [111] dataset, and 4 on the ♣ COCO-Stuff 164k [112] dataset, totaling 74 models.

	FCN [113, 114]	DeepLabV3 [115]	PSPNet [116]	ERFNet [38]	DeepLabV3+ [117]	EncNet [118]	NonLocal Net [119]	BiSeNetV1 [120]	ICNet [121]	PSANet [122]	UPerNet [123]	FastFCN [124]	Fast-SCNN [125]	ISANet [126]	APCNet [127]	DANet [128]	Semantic FPN [129]	ANN [130]	CCNet [131]	DMNet [132]	EMANet [133]	GCNet [134]	PointRend [135]	DNLNet [34]	OCRNet [136]	CGNet [137]	SETR [138]	BiSeNetV2 [139]	STDC [140]	DPT [141]	Segmenter [142]	K-Net [143]	MaskFormer [144]	SegFormer [145]	Mask2Former [146]	SAN [147]
♠ Cityscapes [109]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	
♥ ADE20K [110]	✓	✓	✓		✓	✓	✓			✓	✓	✓		✓	✓	✓		✓	✓	✓		✓	✓	✓	✓		✓									
♦ Pascal VOC 2012 [111]	✓	✓	✓		✓		✓				✓			✓				✓	✓			✓			✓											
♣ COCO-Stuff 164k [112]		✓	✓					✓																											✓

also discuss various strategies for selecting an entity based on the Tiles.

Scenario 1: The Theoretical Analyst

The theoretical analyst seeks to understand the foundational principles behind the scores used to evaluate and rank classifiers. Unlike other users, his/her focus is not on a specific application but rather on the theoretical relationships between various ranking scores. For this user, it is crucial to explore how different scores compare to one another, and how they correlate in terms of both value and ranking. The analyst aims to ensure that each selected score provides unique, non-redundant information, thereby enriching the evaluation process. Therefore, he/she requires a tool that can effectively illustrate the relationships between scores, helping discern whether the chosen scores are complementary or overlapping in the information they convey. To address these needs, the next section introduces the Correlation Tile, specifically designed for this type of analysis.

Correlation Tile. The Correlation Tile displays the correlation, using a linear (Pearson's r) or a rank (Spearman's ρ) correlation coefficient, between a score X, typically one used as a reference in a research field of interest, and the canonical ranking scores R_I across the Tile. The Correlation Tile is defined as:

Definition 2. *The Correlation Tile is the mapping*

$$corr_f : [0,1]^2 \to [-1,1] : (a,b) \mapsto f(X, R_I)\,,$$

where f is a correlation coefficient.

Figure 4 illustrates this correlation for the mean intersection over union (mIoU), a score commonly used to benchmark semantic segmentation models. A strong correlation between the mIoU and the Tile scores is observed in a horizontal band near the top. Hence, selecting a score in this band will lead to similar conclusions than the ones provided by the mIoU. However, selecting a score outside this band will most probably allow understanding the performances of semantic segmentation algorithms under a fresh angle. This Correlation Tile is therefore a powerful tool for understanding how to select complementary scores.

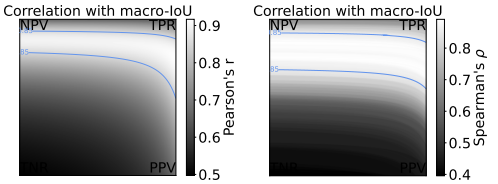


Figure 4. **Correlation Tile for linear (Pearson's r, left) and rank (Spearman's ρ, right) correlation coefficients.** These tiles show the linear and rank correlations between the canonical ranking scores and the macro-averaged IoU score that is usually taken as reference in the field of semantic segmentation, computed for the 31 models trained on ♠ Cityscapes. The blue lines delineate the area where the correlation coefficients are ≥ 0.85.

Scenario 2: The Method Designer

The method designer focuses on evaluating the performance of his/her newly developed method, comparing it against state-of-the-art and other baseline methods. This user is particularly interested in understanding how a method performs across different importances given to true positives (tp), false positives (fp), false negatives (fn), and true negatives (tn). In cases of a parametric method, the designer also seeks insights into the optimal hyper-parameter settings that maximize performance. Hereafter, we show how the Value Tile, Baseline Value Tile, and State-of-the-Art Value Tile can help the method designer understand the strengths and weaknesses of a method, providing insights needed to fine-tune performance and optimize hyper-parameters.

Value Tile. The Value Tile is a map displaying the value of each ranking score R_I across the Tile for a given entity and is defined as:

Definition 3. *For an entity ϵ, the* Value Tile *is the mapping*

$$V_\epsilon : [0,1]^2 \to [0,1] : (a,b) \mapsto R_I(P_\epsilon)\,.$$

In practice, several methods can be used to compute the *Value Tile*, which are discussed in supplementary material.

Figure 5 shows the Value Tile for the 4 entities ranked first among the 74 state-of-the-art semantic segmentation

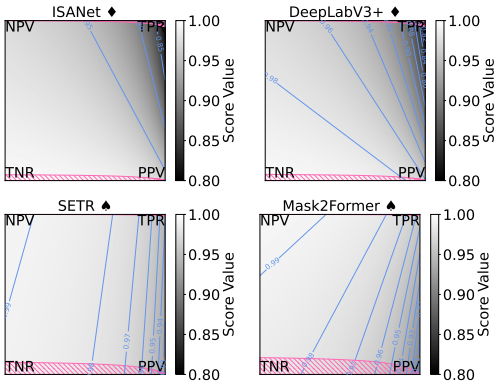


Figure 5. **Value Tile for the 4 entities ranked first (see Fig. 8).** The Value Tile shows the values of the canonical ranking scores. Blue lines are iso-value lines, *i.e.*, lines on which all scores have the same value. Hatched areas indicate regions where non-skilled performances (*i.e.*, when the ground truth Y and the prediction $\hat{Y}$ are independent) surpass that of the entity.

models (see Fig. 8). This Tile can be used to get in one look where the entity performs best, and therefore where improvements should come from the method designer.

The Value Tile shows a negative, zero, or positive horizontal gradient $\frac{\partial R_I}{\partial a}$ when, respectively, $P(\{tp\})$ is less than, equal to or greater than $P(\{tn\})$. Similarly, the vertical gradient $\frac{\partial R_I}{\partial b}$ is negative, zero, or positive when, respectively, $P(\{fn\})$ is less than, equal to, or greater than $P(\{fp\})$. As we have $P(\{tp\}) < P(\{tn\})$ for all the models in our particular illustrative case, it is enough to look at the direction of the iso-value lines in order to analyze the bias (in the sense of Byrt *et al.* [77]): a model is biased towards the negative class (see, *e.g.*, ISANet ♦ and DeepLabV3+ ♦ in Fig. 5), unbiased, or biased towards the positive class (see, *e.g.*, SETR ♠ and Mask2Former ♠ in Fig. 5) when, respectively, the iso-value lines are tilted to the left, vertical, or tilted to the right.

Baseline Value Tile and State-of-the-Art Value Tile. The Baseline Value Tile is a map displaying, for a given set of entities, the minimum value for each ranking score R_I across the Tile, while the State-of-the-Art Value Tile is a map displaying the maximum value for each ranking score R_I. They are respectively defined as follows.

Definition 4. *For a given set $\mathbb{E}$ of entities, the Baseline Value Tile is the mapping*

$$[0,1]^2 \to [0,1] : (a,b) \mapsto \min_{\epsilon \in \mathbb{E}} R_I(P_\epsilon)\,.$$

Definition 5. *For a given set $\mathbb{E}$ of entities, the State-of-the-Art Value Tile is the mapping*

$$[0,1]^2 \to [0,1] : (a,b) \mapsto \max_{\epsilon \in \mathbb{E}} R_I(P_\epsilon)\,.$$

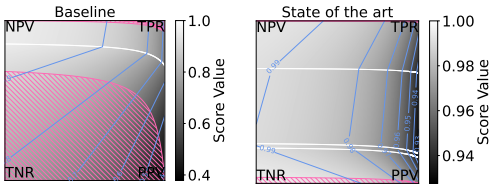


Figure 6. **Baseline Value Tile (left) and State-of-the-Art Value Tile (right).** These two tiles describe the current benchmark in semantic segmentation. The Baseline Value Tile and the State-of-the-Art Value Tile provide the *infimum* and the *supremum*, respectively, of the canonical ranking scores for the 74 semantic segmentation models. In other words, they give the value of the scores for the entities ranked last and first, respectively. The blue lines are iso-value lines, *i.e.*, lines on which all scores have the same value. The white lines mark the boundaries of these entities.

In other words, for a given set $\mathbb{E}$ of entities, the Baseline Value Tile gives the value of each score corresponding to the entity that is ranked last in each point of the Tile, while the State-of-the-Art Value Tile gives the value of the score corresponding to the entity that is ranked first.

The left-hand side of Fig. 6 illustrates the Baseline Value Tile for the 74 semantic segmentation models, showing the lowest canonical ranking score values in each point among all compared entities, while the right-hand side illustrates the State-of-the-Art Value Tile for the 74 semantic segmentation models, showing the highest canonical ranking score values in each point among all compared entities. The hatched areas indicate that, if the classifiers always predicting the positive class, c_+, or the negative class, c_-, were added to the set of entities to rank, they would occupy the top rank within these areas. These tiles are therefore interesting for the method designer when he/she compares them with the Value Tile of his/her entity. One can then easily compare where a method is close or above the State-of-the-Art Value Tile and where it is close or below the Baseline Value Tile. This indicates the advantages and drawbacks of a new method compared with previous works. Furthermore, these tiles may only consider a parametric family of the method, showing where different hyper-parameters perform well or bad, guiding the designer's choice.

Scenario 3: The Benchmarker

The benchmarker is focused on comparing methods from the literature to identify which ones outperform others. This user profile aims to establish a clear ranking among different methods, ensuring that their assessment is fair and consistent. Additionally, the benchmarker may be interested in

organizing open challenges, where the goal is to evaluate and rank participating methods to determine the top performer. In this context, it becomes crucial to accurately determine a winner by ranking the different entries based on their performance. To meet these needs, the benchmarker requires a tool that can generate reliable rankings.

Ranking Tile. The Ranking Tile maps, for a given entity ϵ, the rank of that entity across the tile. Following [2], for a given set of entities $\mathbb{E}$, rankings are based on an ordering $\lesssim_{R_I}$ of the performances induced by the ranking scores R_I. The Ranking Tile is defined as:

Definition 6. *The Ranking Tile is the mapping*

$$[0,1]^2 \to [1, |\mathbb{E}|] : (a,b) \mapsto \mathrm{rank}_{\mathbb{E}}(\epsilon)\,,$$

where $\mathrm{rank}_{\mathbb{E}}(\epsilon)$ *is computed according to* $\lesssim_{R_I}$.

Therefore, the Ranking Tile can be obtained by ordering the performances for each ranking score, *i.e.*, ordering in each point (a,b) the values of the ranking score for the Value Tile of each entity of the set.

Figure 7 shows the Ranking Tile for the 4 entities ranked first among the 74 state-of-the-art semantic segmentation models (see Fig. 8). The Value Tile (Fig. 5) and the Ranking Tile (Fig. 7) provide, at a glance, an overview of an entity's performance across the entire tile.

The Ranking Tile is therefore a great solution to help the benchmarker identify the best-performing methods and appropriately determine the winners in open challenges, as it provides a structured way to know in which case a method has a particular rank.

Scenario 4: The Application Developer 💂

The application developer is interested in selecting the best method that aligns with the specific requirements of his/her application. This developer may already know the importance of true positives, false positives, false negatives, and true negatives relevant to a particular use case. The goal is therefore to choose a method that best meets these predefined criteria. For this purpose, the application developer needs a tool that allows him/her to efficiently identify the most suitable method based on specific priorities.

Entity Tile. The Entity Tile maps entities that are at a given rank across the tile and is defined as follows.

Definition 7. *The Entity Tile, for a given rank* $r \in [1, |\mathbb{E}|]$, *is the mapping*

$$[0,1]^2 \to \mathbb{E} : (a,b) \mapsto \epsilon_r\,,$$

where ϵ_r *is the entity ranked* r*-th, according to* $\lesssim_{R_I}$.

The Entity Tile thus shows all entities that are at a given rank and is constructed similarly to the Ranking Tile. In the special case of $r = 1$, the Entity Tile shows the best entities among $\mathbb{E}$, while for $r = |\mathbb{E}|$, it shows the worst entities among $\mathbb{E}$. As illustrated in Fig. 8 for the semantic segmentation models, the first rank is shared by 4 entities, namely Mask2Former and SETR, both trained on ♠ Cityscapes, and DeepLabV3+ and ISANet, both trained on ♦ Pascal VOC 2012. These 4 entities also appear in the second and third ranks in other areas of the Ranking Tiles, *i.e.*, for other canonical ranking scores. Interestingly, the entities ranked first are the ones on the upper dashed broken line of Fig. 2 in the ROC space, and the ones ranked last are on the lower dashed broken line. This is a consequence of properties studied in [4].

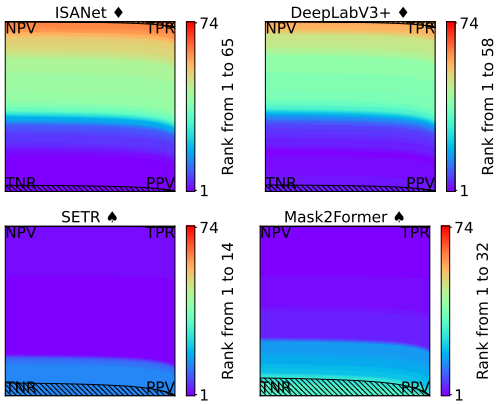


Figure 7. **Ranking Tile for the 4 entities that are ranked first, somewhere on the Tile.** The Ranking Tile shows the ranks of a given entity across the tile. Hatched areas highlight regions of the tile where no-skilled performances surpass that of the entity. Remarkably, we directly see that (1) ISANet ♦ and DeepLabV3+ ♦ have poor rankings in the upper part of the tile, while (2) SETR ♠ has the best overall performance and remains rather stable on the Tile, ranging from rank 1 to 14.

The application developer may choose an entity based on those tiles depending on 3 cases, based on the knowledge of importance values: (1) In the first case, the importance values are *known*. Hence, the application developer simply selects the method ranked number 1 at the corresponding place on the Tile. This is typically how most benchmarks rank methods, *i.e.*, based on a single score. (2) In the second case, the importance values are *unknown but can be determined* by analyzing the community practices. For example, in semantic segmentation, the "mean intersection over union" (mIoU) is often considered to be a good criterion. Although this score is not part of the Tile, the Correlation Tile provides a direct comparison between the orderings induced by all ranking scores and the macro-averaged

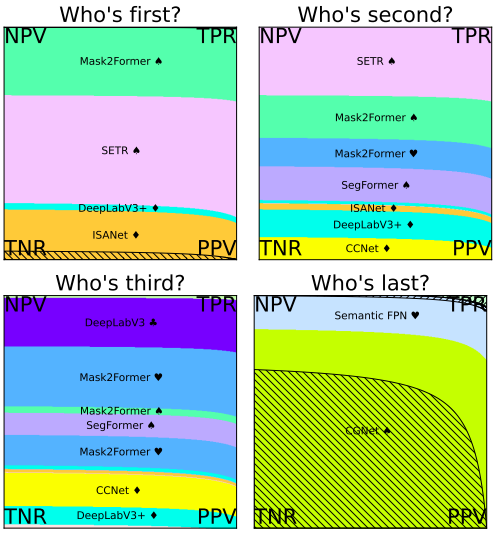


Figure 8. **Entity Tile showing the entities that are ranked at the first, second, third, and last positions.** Hatched areas in the "who's first?" and "who's last?" tiles highlight regions where no-skilled performances surpass that of the entities in the current map.

IoU. The application developer can then simply select the method ranked first in the area, where Spearman's ρ is the highest in the tile. (3) In the last case, there is *no information* on the importance values. One selection mechanism consists in minimizing the maximum rank over the Ranking Tile and, in case of ex aequo, minimizing the average rank.

Therefore, the Entity Tile enables the application developer to select the best method by visualizing how different options perform relative to the importance values, whether they are known or not. This allows to make informed decisions tailored to specific application requirements.

4. Discussion

Bands. Unlike the Value Tiles, all Correlation, Ranking and Entity Tiles illustrating this guide exhibit roughly horizontal bands. They can be explained as follows. All performances such that $P(\{tn\}) \propto P(\{tp\})$ are perfectly rank-correlated with all ranking scores on any given horizontal line of the Tile. In our illustration, the ratio $\frac{P(\{tp\})}{P(\{tn\})}$ is roughly constant ($\mu = 0.12$, $\sigma = 0.01$), which explains the roughly horizontal bands. Similarly, all performances such that $P(\{fp\}) \propto P(\{fn\})$ are perfectly rank-correlated with all ranking scores on any given vertical line of the Tile. This covers the case, not illustrated in this guide, in which all compared performances are unbiased, in the sense of Byrt *et al.* [77], *i.e.* $P(Y) = P(\hat{Y})$.

Multi-class classification extension. The universal theoretical foundations for performance-based ranking [2], on top of which the Tile was build [4], are indeed applicable to multi-class problems. With n classes, specifying the relative importance given to the samples ω such that $\hat{Y}(\omega) = Y(\omega)$ (as done on the horizontal axis of the Tile) requires $n-1$ dimensions, and specifying the relative importance given to the samples ω such that $\hat{Y}(\omega) \neq Y(\omega)$ (as done on the vertical axis of the Tile) requires n^2-n-1 dimensions. A "ranking space" similar to the Tile, but for n classes, would therefore require n^2-2 dimensions. This can be handled algorithmically, but it cannot be easily visualized. A typical alternative is to convert multi-class problems into two-class problems by micro- or macro-averaging [38]. While micro- and macro-averaged Value Tiles can be constructed and used to derive Entity Tiles, this approach is not necessarily advisable as it raises serious concerns. On the one hand, all micro-averaged scores are only functions of the multi-class accuracy, which leads to a loss of information, and, on the other hand, macro-averaging comes with interpretability issues [1]. We hope that this will motivate further studies on the topic.

Choice of a positive class. Once the labels have been grouped into two classes, there remains to choose which is the positive one, in order to comply with the classical convention. However, the opposite choice leads to identical Tiles, up to a central symmetry [4]. In particular, the list of entities (classifiers) identified as being the best (ranked first) does not depend on which class is chosen as positive.

5. Conclusion

In this paper, we introduced a hitchhiker's guide to understanding the performance of two-class classifiers. We organized our guide into four distinct scenarios, each corresponding to a specific user profile, covering the theoretical analyst, the method designer, the benchmarker, and the application developer. For each scenario, we provided examples of challenges each user may encounter when comparing classifier performances and identified the most suitable tools for their needs. Central to our approach is the idea of displaying different types of information, called *flavors*, on the Tile. By leveraging different mappings, we demonstrated the versatility of the Tile and its various flavors to address the diverse requirements of these user profiles. This guide offers a flexible and robust framework that enables users to effectively evaluate, rank, and interpret the performance of two-class classifiers, making it a valuable resource for researchers, practitioners, and developers alike.

References

[105] Azim Ahmadzadeh, Dustin J. Kempton, Petrus C. Martens, and Rafal A. Angryk. Contingency space: A semimetric space for classification evaluation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(2):1501–1513, 2023.

[102] Donald Bamber. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *J. Math. Psychol.*, 12(4):387–415, 1975.

[26] Kendrick Boyd, Victor Costa, Jesse Davis, and David Page. Unachievable region in precision-recall space and its effect on empirical evaluation. In *Int. Conf. Mach. Learn. (ICML)*, pages 639–646, Edinburgh, UK, 2012.

[103] William M. Briggs and Russell Zaretzki. The skill plot: A graphical technique for evaluating continuous diagnostic tests. *Biometrics*, 64(1):250–256, 2008.

[77] Ted Byrt, Janet Bishop, and John B. Carlin. Bias, prevalence and kappa. *J. Clin. Epidemiology*, 46(5):423–429, 1993.

[112] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. COCO-stuff: Thing and stuff classes in context. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 1209–1218, Salt Lake City, UT, USA, 2018. Inst. Electr. Electron. Eng. (IEEE).

[134] Yue Cao, Jiarui Xu, Stephen Lin, Fangyun Wei, and Han Hu. GCNet: Non-local networks meet squeeze-excitation networks and beyond. In *IEEE/CVF Int. Conf. Comput. Vis. Work. (ICCV Work.)*, pages 1971–1980, Seoul, South Korea, 2019. Inst. Electr. Electron. Eng. (IEEE).

[104] Andre M. Carrington, Douglas G. Manuel, Paul W. Fieguth, Tim Ramsay, Venet Osmani, Bernhard Wernly, Carol Bennett, Steven Hawken, Olivia Magwood, Yusuf Sheikh, Matthew McInnes, and Andreas Holzinger. Deep ROC analysis and AUC as balanced average accuracy, for improved classifier selection, audit and explanation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(1):329–341, 2023.

[115] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv*, abs/1706.05587, 2017.

[117] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Eur. Conf. Comput. Vis. (ECCV)*, pages 833–851, Munich, Germany, 2018. Springer Int. Publ.

[144] Bowen Cheng, Alexander G. Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. In *Adv. Neural Inf. Process. Syst. (NeurIPS)*, pages 17864–17875. Curran Assoc. Inc., 2021.

[146] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 1280–1289, New Orleans, LA, USA, 2022. Inst. Electr. Electron. Eng. (IEEE).

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

[109] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 3213–3223, Las Vegas, NV, USA, 2016. Inst. Electr. Electron. Eng. (IEEE).

[24] Jesse Davis and Mark Goadrich. The relationship between precision-recall and ROC curves. In *Int. Conf. Mach. Learn. (ICML)*, pages 233–240, Pittsburgh, Pennsylvania, 2006. ML Res. Press.

[111] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The PASCAL visual object classes (VOC) challenge. *Int. J. Comput. Vis.*, 88(2):303–338, 2010.

[140] Mingyuan Fan, Shenqi Lai, Junshi Huang, Xiaoming Wei, Zhenhua Chai, Junfeng Luo, and Xiaolin Wei. Rethinking BiSeNet for real-time semantic segmentation. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 9711–9720, Nashville, TN, USA, 2021. Inst. Electr. Electron. Eng. (IEEE).

[101] Tom Fawcett. ROC graphs: Notes and practical considerations for researchers. *Mach. Learn.*, 31:1–38, 2004.

[30] Tom Fawcett. An introduction to ROC analysis. *Pattern Recognit. Lett.*, 27(8):861–874, 2006.

[29] Peter A. Flach. The geometry of ROC space: Understanding machine learning metrics through ROC isometrics. In *Int. Conf. Mach. Learn. (ICML)*, pages 194–201, Washington, DC, USA, 2003. ML Res. Press.

[128] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 3141–3149, Long Beach, CA, USA, 2019. Inst. Electr. Electron. Eng. (IEEE).

[132] Junjun He, Zhongying Deng, and Yu Qiao. Dynamic multi-scale filters for semantic segmentation. In *IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pages 3561–3571, Seoul, South Korea, 2019. Inst. Electr. Electron. Eng. (IEEE).

[127] Junjun He, Zhongying Deng, Lei Zhou, Yali Wang, and Yu Qiao. Adaptive pyramid context network for semantic segmentation. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 7511–7520, Long Beach, CA, USA, 2019. Inst. Electr. Electron. Eng. (IEEE).

[126] Lang Huang, Yuhui Yuan, Jianyuan Guo, Chao Zhang, Xilin Chen, and Jingdong Wang. Interlaced sparse self-attention for semantic segmentation. *arXiv*, abs/1907.12273, 2019.

[131] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. CCNet: Criss-cross attention for semantic segmentation. In *IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pages 603–612, Seoul, South Korea, 2019. Inst. Electr. Electron. Eng. (IEEE).

[106] Jacqueline M. Hughes-Oliver. Population and empirical PR curves for assessment of ranking algorithms. *arXiv*, abs/1810.08635, 2018.

[129] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollar. Panoptic segmentation. In

IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pages 9396–9405, Long Beach, CA, USA, 2019. Inst. Electr. Electron. Eng. (IEEE).

[135] Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross Girshick. PointRend: Image segmentation as rendering. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 9796–9805, Seattle, WA, USA, 2020. Inst. Electr. Electron. Eng. (IEEE).

[133] Xia Li, Zhisheng Zhong, Jianlong Wu, Yibo Yang, Zhouchen Lin, and Hong Liu. Expectation-maximization attention networks for semantic segmentation. In *IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pages 9166–9175, Seoul, South Korea, 2019. Inst. Electr. Electron. Eng. (IEEE).

[98] Yusha Liu, Yichong Xu, Nihar B. Shah, and Aarti Singh. Integrating rankings into quantized scores in peer review. *Trans. Mach. Learn. Res.*, pages 1–28, 2022.

[113] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 3431–3440, Boston, MA, USA, 2015. Inst. Electr. Electron. Eng. (IEEE).

[107] Aditya Menon, Harikrishna Narasimhan, Shivani Agarwal, and Sanjay Chawla. On the statistical consistency of algorithms for binary classification under class imbalance. In *Int. Conf. Mach. Learn. (ICML)*, pages 603–611, Atlanta, GA, USA, 2013.

[100] Francis Sahngun Nahm. Receiver operating characteristic curve: overview and practical use for clinicians. *Korean Journal of Anesthesiology*, 75(1):25–36, 2022.

[45] Tran Thien Dat Nguyen, Hamid Rezatofighi, Ba-Ngu Vo, Ba-Tuong Vo, Silvio Savarese, and Ian Reid. How trustworthy are performance evaluations for basic vision tasks? *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(7):8538–8552, 2023.

[1] Sébastien Piérard and Marc Van Droogenbroeck. Summarizing the performances of a background subtraction algorithm measured on several videos. In *IEEE Int. Conf. Image Process. (ICIP)*, pages 3234–3238, Abu Dhabi, United Arab. Emir., 2020.

[4] 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.

[2] Sébastien Piérard, Anaïs Halin, Anthony Cioppa, Adrien Deliège, and Marc Van Droogenbroeck. Foundations of the theory of performance-based ranking. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 14293–14302, Nashville, TN, USA, 2025. IEEE.

[125] Rudra P. K. Poudel, Stephan Liwicki, and Roberto Cipolla. Fast-SCNN: Fast semantic segmentation network. *arXiv*, abs/1902.04502, 2019.

[62] David M. W. Powers. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness & correlation. *J. Mach. Learn. Technol.*, 2(1):37–63, 2011.

[141] Rene Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pages 12159–12168, Montréal, Can., 2021. Inst. Electr. Electron. Eng. (IEEE).

[35] Eduardo Romera, Jose M. Alvarez, Luis M. Bergasa, and Roberto Arroyo. ERFNet: Efficient residual factorized ConvNet for real-time semantic segmentation. *IEEE Trans. Intell. Transp. Syst.*, 19(1):263–272, 2018.

[114] Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(4):640–651, 2017.

[38] Marina Sokolova and Guy Lapalme. A systematic analysis of performance measures for classification tasks. *Inf. Process. & Manag.*, 45(4):427–437, 2009.

[142] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pages 7242–7252, Montréal, Can., 2021. Inst. Electr. Electron. Eng. (IEEE).

[99] Alaa Tharwat. Classification assessment methods. *Appl. Comput. Informatics*, 17(1):168–192, 2021.

[119] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 7794–7803, Salt Lake City, UT, USA, 2018. Inst. Electr. Electron. Eng. (IEEE).

[124] Huikai Wu, Junge Zhang, Kaiqi Huang, Kongming Liang, and Yizhou Yu. FastFCN: Rethinking dilated convolution in the backbone for semantic segmentation. *arXiv*, abs/1903.11816, 2019.

[137] Tianyi Wu, Sheng Tang, Rui Zhang, Juan Cao, and Yongdong Zhang. CGNet: A light-weight context guided network for semantic segmentation. *IEEE Trans. Image Process.*, 30:1169–1179, 2021.

[123] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Eur. Conf. Comput. Vis. (ECCV)*, pages 432–448. Springer Int. Publ., 2018.

[145] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. SegFormer: Simple and efficient design for semantic segmentation with transformers. In *Adv. Neural Inf. Process. Syst. (NeurIPS)*, pages 12077–12090, Virtual conference, 2021. Curran Assoc. Inc.

[147] Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. Side adapter network for open-vocabulary semantic segmentation. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 2945–2954, New Orleans, LA, USA, 2023. Inst. Electr. Electron. Eng. (IEEE).

[34] Minghao Yin, Zhuliang Yao, Yue Cao, Xiu Li, Zheng Zhang, Stephen Lin, and Han Hu. Disentangled non-local neural networks. In *Eur. Conf. Comput. Vis. (ECCV)*, pages 191–207. Springer Int. Publ., 2020.

[120] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. BiSeNet: Bilateral segmentation network for real-time semantic segmentation. In *Eur. Conf. Comput. Vis. (ECCV)*, pages 325–341. Springer Int. Publ., 2018.

[139] Changqian Yu, Changxin Gao, Jingbo Wang, Gang Yu, Chunhua Shen, and Nong Sang. BiSeNet V2: Bilateral net-

work with guided aggregation for real-time semantic segmentation. *Int. J. Comput. Vis.*, 129(11):3051–3068, 2021.

[148] 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 Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 2633–2642, Seattle, WA, USA, 2020. Inst. Electr. Electron. Eng. (IEEE).

[136] Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. In *Eur. Conf. Comput. Vis. (ECCV)*, pages 173–190. Springer Int. Publ., 2020.

[118] Hang Zhang, Kristin Dana, Jianping Shi, Zhongyue Zhang, Xiaogang Wang, Ambrish Tyagi, and Amit Agrawal. Context encoding for semantic segmentation. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 7151–7160, Salt Lake City, UT, USA, 2018. Inst. Electr. Electron. Eng. (IEEE).

[143] Wenwei Zhang, Jiangmiao Pang, Kai Chen, and Chen Change Loy. K-Net: Towards unified image segmentation. In *Adv. Neural Inf. Process. Syst. (NeurIPS)*, pages 10326–10338. Curran Assoc. Inc., 2021.

[116] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 6230–6239, Honolulu, HI, USA, 2017. Inst. Electr. Electron. Eng. (IEEE).

[121] Hengshuang Zhao, Xiaojuan Qi, Xiaoyong Shen, Jianping Shi, and Jiaya Jia. ICNet for real-time semantic segmentation on high-resolution images. In *Eur. Conf. Comput. Vis. (ECCV)*, pages 418–434. Springer Int. Publ., 2018.

[122] Hengshuang Zhao, Yi Zhang, Shu Liu, Jianping Shi, Chen Change Loy, Dahua Lin, and Jiaya Jia. PSANet: Point-wise spatial attention network for scene parsing. In *Eur. Conf. Comput. Vis. (ECCV)*, pages 270–286. Springer Int. Publ., 2018.

[138] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip H.S. Torr, and Li Zhang. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 6877–6886, Nashville, TN, USA, 2021. Inst. Electr. Electron. Eng. (IEEE).

[110] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ADE20K dataset. In *IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 5122–5130, Honolulu, HI, USA, 2017. Inst. Electr. Electron. Eng. (IEEE).

[130] Zhen Zhu, Mengdu Xu, Song Bai, Tengteng Huang, and Xiang Bai. Asymmetric non-local neural networks for semantic segmentation. In *IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pages 593–602, Seoul, South Korea, 2019. Inst. Electr. Electron. Eng. (IEEE).

6. Multi-domain performance analysis (6th article)

This chapter reproduces the following article [6]:

> S. Piérard, A. Deliège, and M. Van Droogenbroeck, "Multi-domain performance analysis with scores tailored to user preferences," *arXiv*, vol. abs/2512.08715, 2025.

It covers the notions of ***performance***, ***ranking*** and ***domain***.

Context

In this chapter, we look at how to analyze the performance of an entity across multiple domains. We already started this topic in Chapter 1, but it was before introducing the foundations of the theory of performance-based ranking in Chapter 2. We will now revisit and expand on this topic by proposing to relate the analysis to the applicative preferences.

Motivation

The analyses presented in this chapter allow us to identify the domains that are the strengths and weaknesses for a given entity, and to relate them to the applicative preferences.

Main contributions

While Chapter 1 focused on unconditional and conditional probabilistic scores, we now generalize our previous observations to scores that can be expressed as expected values and ratios of expected values. They include all the ranking scores R_I that have been introduced in Chapter 2. By doing so, we establish a link between the summarization and the importance values I or, in other words, the applicative preferences.

We add 4 new flavors to the catalog of flavors created in Chapter 5: the easiest, most difficult, preponderant, and bottleneck domains. The last two relate to the summarized (*i.e.*, averaged) performance. We illustrate these flavors on the Tile introduced in Chapter 4 for tasks similar to two-class crisp classification. This indicates that the Tile is a graphical tool that can be used for purposes other than ranking classifiers in a given domain.

In fact, what is presented in this chapter can be seen as domain ranking for a given classifier. The new flavors relate to domain rankings. The element that varies in the random experiment is not the entity but the data source. As the entity is fixed, the compared performances are not strictly speaking those of the entities but those in the domains. However, this does not change the suitability of the axioms given in Chapter 2. We can therefore use the ranking scores and the Tile to rank the domains. The easiest and the most difficult domains are those ranked, respectively, first and last when the data source used in the random evaluation experiment corresponds to the domain. To obtain a ranking relative to the bottleneck domain, we simply need to consider in the random experiment a hybrid data source corresponding to the set of all other domains.

Perspectives

Readers wishing to explore the subject further will realize that the principles introduced in this article for multi-domain and the summarized performance are similar to those that could be introduced for multi-class and the micro-averaged performance. In this case, we would have tiles showing the easiest, most difficult, preponderant, and bottleneck classes.

Remarks

In this chapter, I did not define an average domain. This will be done in Chapters 8 and 9. The summarized performance will be equivalent to the performance in the average domain, provided that the entity is fixed, or in other words, that it does not adapt in any way to the domain in which it is used.

Multi-domain performance analysis with scores tailored to user preferences

Sébastien Piérard, Adrien Deliège, and Marc Van Droogenbroeck *

Montefiore Institute, University of Liège, Liège, Belgium

Abstract. The performance of algorithms, methods, and models tends to depend heavily on the distribution of cases on which they are applied, this distribution being specific to the applicative domain. After performing an evaluation in several domains, it is highly informative to compute a (weighted) mean performance and, as shown in this paper, to scrutinize what happens during this averaging. To achieve this goal, we adopt a probabilistic framework and consider a performance as a probability measure (*e.g.*, a normalized confusion matrix for a classification task). It appears that the corresponding weighted mean is known to be the summarization, and that only some remarkable scores assign to the summarized performance a value equal to a weighted arithmetic mean of the values assigned to the domain-specific performances. These scores include the family of ranking scores, a continuum parameterized by user preferences, and that the weights to consider in the arithmetic mean depend on the user preferences. Based on this, we rigorously define four domains, named easiest, most difficult, preponderant, and bottleneck domains, as functions of user preferences. After establishing the theory in a general setting, regardless of the task, we develop new visual tools for two-class classification.

1 Introduction

The development of algorithms, methods, and models is often complicated by two types of uncertainty. First, the *applicative domain*, and thus the distribution of inputs, is rarely unique and precisely known during development. Second, *user preferences* regarding tradeoffs affecting the achievable performance (*e.g.*, to what degree do they prefer false negatives to false positives in classification?) are also generally not precisely known and can differ from one user to another.

Developers can, of course, implement various techniques (continual learning, test-time adaptation, domain adaptation [149], domain generalization [150], ...) to make their *entity* —the term used in this paper for algorithm, method, or model— usable in various domains, but the achievable performance still varies from one domain to another. After evaluating the developed entity in several domains, the developer is left with many questions, such as: Which domains are the strengths and weaknesses of the developed entities? Which domains have the greatest influence on average performance? Which domains constitute the bottlenecks for the average performance improvement? We need to address these questions from the end-user preferences perspective.

*Sébastien Piérard is funded by the Walloon region (Service Public de Wallonie Recherche, Belgium) under grant 2010235 - *ARIAC by DIGITALWALLONIA4.AI*. Adrien Deliège is a F.R.S.-FNRS postdoctoral researcher.

The objective of this paper is to rigorously demonstrate that, no matter what the task is, by taking an adequate way of averaging the domain-specific performances (*i.e.*, the *summarization* introduced in [1]) as well as an adequate family of scores (*i.e.*, the *ranking scores* introduced in [2]), the *easiest domain*, *most difficult domain*, *preponderant domain*, and *bottleneck domain* become all well-defined and relative to some easy to interpret user preferences.

Our contributions are threefold:

1. Theoretical contribution on scores. We highlight the fact that the *expected value scores* (including all *unconditional probabilistic scores*) and the *expected value ratio scores* (including all *probabilistic conditional scores* and all *ranking scores*) have the remarkable property that the value taken for the averaged performance obtained by the *summarization* technique [1] is equal to some weighted arithmetic mean of the values taken for the domain-specific performances, and provide closed-form formulas to compute these weights.

2. Contribution on user-centered multi-domain performance analysis. Capitalizing on the remarkable property and focusing on the family of *ranking scores* [2], we rigorously define the *easiest*, *most difficult*, *preponderant*, and *bottleneck* domains as functions of a random variable that can be used to encode some user preferences for any task.

3. Contribution on visualization tools. In the case of two-class crisp classification, we introduce new variants of the Tile, called *flavors* [4, 5], that allow us to visualize in a square how the *easiest*, *most difficult*, *preponderant*, and *bottleneck* domains depend on the user preferences.

Notations. Following [1, 2, 5], we consider a probabilistic framework. A performance is a probability measure (*e.g.*, a normalized confusion matrix for a classification task) on a measurable space (Ω, Σ), where Ω is the sample space and Σ the event space. The set of all performances is denoted by $\mathbb{P}_{(\Omega,\Sigma)}$. We denote the set of domains by $\mathbb{D}$, the performance in the domain d by P_d, and the weight arbitrarily given to the domain d by λ_d.

2 Averaging performances with the summarization

In this paper, we build upon the summarization [1], which is a probabilistic performance averaging technique that preserves the relationships between scores and preserves the probabilistic meaning of both unconditional and conditional probabilistic scores. It is applicable regardless of the task. The summarized performance $\overline{P}$ is the following mixture of performances:

$$\overline{P} = \frac{\sum_{d\in\mathbb{D}} \lambda_d P_d}{\sum_{d\in\mathbb{D}} \lambda_d} \qquad \text{with} \qquad \lambda_d \geq 0 \,\forall d \in \mathbb{D}\,. \tag{1}$$

As noticed in [1], for some remarkable scores X, the summarization leads to

$$X(\overline{P}) = \sum_{d\in\mathbb{D}} \omega_{X,d} X(P_d) \qquad \text{with} \qquad \begin{cases} \omega_{X,d} \geq 0 \,\forall d \in \mathbb{D} \\ \sum_{d\in\mathbb{D}} \omega_{X,d} = 1 \end{cases}. \tag{2}$$

We call λ_d the *domain weight* and $\omega_{X,d}$ the *summarization weight*.

Formulas for the weights for expected value scores. We parameterize these scores by a random variable V and define them as $X_V^{EV} : \mathbb{P}_{(\Omega,\Sigma)} \to \mathbb{R} : P \mapsto X_V^{EV}(P) = \mathbf{E}_P[V]$, where $\mathbf{E}$ is the mathematical expectation. With these scores, eq. (2) holds and the summarization weights do not depend on the performance:

$$\omega_{X_V^{EV},d} = \frac{\lambda_d}{\sum_{d\in\mathbb{D}} \lambda_d} \,. \tag{3}$$

Formulas for the weights for expected value ratio scores. We parameterize these scores by two random variables, V_1 and $V_2 \neq 0$, and define them as $X_{V_1,V_2}^{EVR} : \mathrm{dom}(X_{V_1,V_2}^{EVR}) \to \mathbb{R} : P \mapsto X_{V_1,V_2}^{EVR}(P) = \frac{\mathbf{E}_P[V_1]}{\mathbf{E}_P[V_2]}$, where $\mathrm{dom}(X_{V_1,V_2}^{EVR}) = \{P \in \mathbb{P}_{(\Omega,\Sigma)} : \mathbf{E}_P[V_2] \neq 0\}$. With these scores, eq. (2) also holds, but contrary to the expected value scores, the summarization weights depend on the performance:

$$\omega_{X_{V_1,V_2}^{EVR},d} = \frac{\lambda_d X_{V_2}^{EV}(P_d)}{\sum_{d\in\mathbb{D}} \lambda_d X_{V_2}^{EV}(P_d)} = \frac{\lambda_d}{\sum_{d\in\mathbb{D}} \lambda_d} \frac{X_{V_2}^{EV}(P_d)}{X_{V_2}^{EV}(\overline{P})} \,. \tag{4}$$

Equations (3) and (4) generalize the ones given in [1], in the particular case of *unconditional* and *conditional* probabilistic scores, respectively.

3 Taking user preferences into account through scores

In addition, we also build upon the ranking scores [2], which establish a connection between the performances P, the task (modeled through a random variable S called *satisfaction*), and some user preferences (modeled through a non-negative random variable I called *importance*). These scores induce meaningful performance orderings [2], in the sense that these orderings satisfy the fundamental axioms of performance-based ranking. They are defined as:

$$R_I : \mathrm{dom}(R_I) \to \mathbb{R} : P \mapsto R_I(P) = \frac{\mathbf{E}_P[IS]}{\mathbf{E}_P[I]} \,, \tag{5}$$

where $\mathrm{dom}(R_I) = \{P \in \mathbb{P}_{(\Omega,\Sigma)} : \mathbf{E}_P[I] \neq 0\}$. As $R_I = X_{IS,I}^{EVR}$, all ranking scores are particular cases of expected value ratio scores and eq. (4) also holds for them:

$$\omega_{R_I,d} = \frac{\lambda_d X_I^{EV}(P_d)}{\sum_{d\in\mathbb{D}} \lambda_d X_I^{EV}(P_d)} = \frac{\lambda_d \mathbf{E}_{P_d}[I]}{\sum_{d\in\mathbb{D}} \lambda_d \mathbf{E}_{P_d}[I]} = \frac{\lambda_d}{\sum_{d\in\mathbb{D}} \lambda_d} \frac{\mathbf{E}_{P_d}[I]}{\mathbf{E}_{\overline{P}}[I]} \,. \tag{6}$$

Thanks to these scores, we can now analyze, in the perspective of user preferences, the performance of an entity that has been evaluated in several domains.

For a given entity, we can rank the domains, from the easiest (with the best performance) to the most difficult (with the worst performance). This ranking depends on the user preferences as follows:

$$\text{easiest-domain}(I) = \arg\max_{d\in\mathbb{D}} R_I(P_d) \,, \tag{7}$$

$$\text{most-difficult-domain}(I) = \arg\min_{d\in\mathbb{D}} R_I(P_d)\,. \tag{8}$$

We can also determine the preponderant domain *w.r.t.* user preferences. It is the domain for which the summarization weight is the largest:

$$\text{preponderant-domain}(I) = \arg\max_{d\in\mathbb{D}} \frac{\lambda_d \mathbf{E}_{P_d}[I]}{\sum_{d\in\mathbb{D}} \lambda_d \mathbf{E}_{P_d}[I]} = \arg\max_{d\in\mathbb{D}} \lambda_d \mathbf{E}_{P_d}[I]\,. \tag{9}$$

A common practice in software engineering is to iteratively identify the parts that are bottlenecks and improve these parts. Similarly, if the objective is to design an entity that performs the best on average, one must be able to identify the domains that are bottlenecks. These are the domains for which the summarization weights are high and the values taken by the ranking scores are weak. We unify these two—sometimes contradictory—objectives by domain ablation: we define the bottleneck domain as the one that has to be removed from the set of domains to maximize a given ranking score:

$$\text{bottleneck-domain}(I) = \arg\max_{d\in\mathbb{D}} R_I\left(\frac{\sum_{d'\in\mathbb{D}\setminus\{d\}} \lambda_{d'} P_{d'}}{\sum_{d'\in\mathbb{D}\setminus\{d\}} \lambda_{d'}}\right)\,. \tag{10}$$

4 Visualization tools for two-class crisp classification

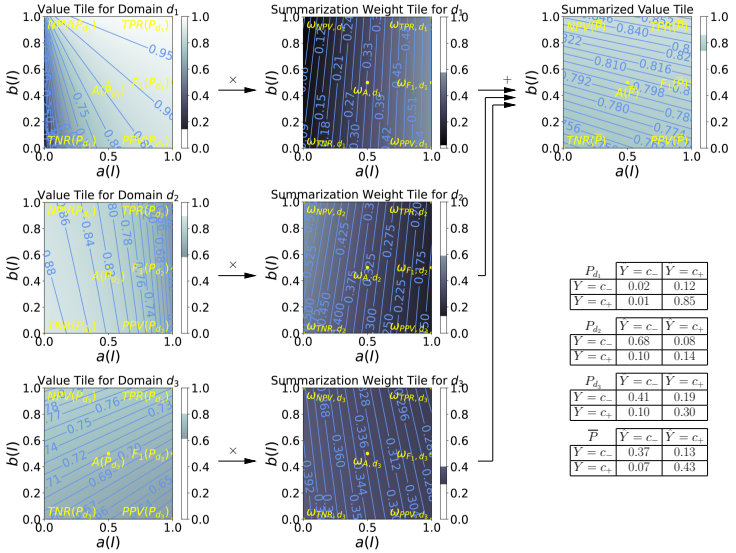
Lastly, we build upon the Tile [4], which is a tool for two-class crisp classification to visualize functions of the importance I on a square space. The horizontal axis indicates the relative importance considered by the user for the *true positives w.r.t.* the *true negatives*. The vertical axis indicates the relative importance for the *false negatives w.r.t.* the *false positives*.

Defining the task. From the performance perspective [2], two-class crisp classification is defined by the sample space $\Omega = \{tn, fp, fn, tp\}$ and the satisfaction $S = \mathbf{1}_{\{tn,tp\}}$. The samples tn, fp, fn, and tp are traditionally interpreted as a true negative, false positive, false negative, and true positive, respectively.

Setting the user preferences. In our setting, the user can freely specify some preferences through the importance values $I(tn)$, $I(fp)$, $I(fn)$, and $I(tp)$. The generic formula for the ranking scores given in eq. (5) is particularized as follows:

$$R_I(P) = \frac{P(\{tn\})\,I(tn) + P(\{tp\})\,I(tp)}{P(\{tn\})\,I(tn) + P(\{fp\})\,I(fp) + P(\{fn\})\,I(fn) + P(\{tp\})\,I(tp)}\,.$$

Using the Tile. As the ranking scores sharing the same values for $a(I) = \frac{I(tp)}{I(tn)+I(tp)}$ and $b(I) = \frac{I(fn)}{I(fp)+I(fn)}$ induce the same ranking, it has been proposed in [4] to focus on the *canonical ranking scores* such that $I(tn) + I(tp) = I(fp) + I(fn)$ and to map them on the $(a, b) \in [0, 1]^2$ space called *Tile*. So, $I(tn) = 1 - a$, $I(fp) = 1 - b$, $I(fn) = b$, and $I(tp) = a$. Note that the accuracy, the true negative rate, the true positive rate, the negative predictive value, the positive predictive value, and F_1 are particular cases of canonical ranking scores.



P_{d_1}

	$Y = c_-$	$Y = c_+$
$Y = c_-$	0.02	0.12
$Y = c_+$	0.01	0.85

P_{d_2}

	$Y = c_-$	$Y = c_+$
$Y = c_-$	0.68	0.08
$Y = c_+$	0.10	0.14

P_{d_3}

	$Y = c_-$	$Y = c_+$
$Y = c_-$	0.41	0.19
$Y = c_+$	0.10	0.30

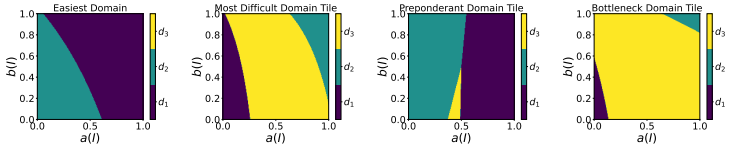
$\overline{P}$

	$Y = c_-$	$Y = c_+$
$Y = c_-$	0.37	0.13
$Y = c_+$	0.07	0.43

Fig. 1: Tiles showing the summarization on 3 domains: by multiplying each domain-specific *Value Tile* by the corresponding *Summarization Weight Tile* and adding the results together, one obtains the *Summarized Value Tile*, *i.e.* the *Value Tile* for $\overline{P}$. This Tile is exactly the same as the *Value Tile* that would be obtained from $\overline{P}$ after computing it with eq. (1). However, by scrutinizing what happens during the performance averaging, it becomes clear that the actual weights to consider strongly depend on the user preferences (*i.e.*, point in the Tile), which is something hidden in eq. (1).

Visualizing the summarization. Figure 1 depicts the summarization, *i.e.* eq. (1), with Tiles. The values taken by the canonical ranking scores and the weights $\omega_{R_I,d}$ are shown, respectively, on *Value Tiles* and *Summarization Weight Tiles*.

Introducing new flavors. We extend the catalog of flavors (*i.e.*, functions of the importance I) compiled in [5] and implemented in the Sorbetto library [151], and propose to depict on the Tile the easiest, the most difficult, the preponderant, and the bottleneck domains (see fig. 2). Practically, to improve on the worst case scenario, the developer should improve on the most difficult domain, and to improve on average, improve on the bottleneck domain. Currently, only the central point is generally considered (*i.e.* the accuracy), such as in [150], overlooking entirely the dimension of user preferences.



(a) easiest (b) most difficult (c) preponderant (d) bottleneck

Fig. 2: In this work, we propose four new flavors for the Tile, as visual tools to perform multi-domain performance analyses *w.r.t.* user preferences. These are the Tiles obtained for the example of fig. 1. Notably, this analysis covers the two sources of uncertainty met in a model development: the domain dimension (the colors on the Tiles), and the user preferences (the position on the Tiles).

5 Conclusion

By considering the technique for averaging domain-specific performances known as the *summarization* [1] and the family of scores known as the *ranking scores* [2], we rigorously defined the *easiest*, *most difficult*, *preponderant*, and *bottleneck* domains as functions of some user preferences. First, we did it in a way that makes it possible for our theoretical results to be used for any task. Then, we particularized our results to the task of two-class crisp classification, demonstrating how the dependence of the four defined domains regarding the user preferences can be visualized on a square space that is known as the Tile [4]. By doing so, we extended the catalog of Tile flavors compiled in [5] and currently implemented in the Sorbetto library [151].

References

[149] G. Wilson and D. J. Cook, "A survey of unsupervised deep domain adaptation," *ACM Trans. Intell. Syst. Technol.*, vol. 11, pp. 1–46, Jul. 2020.

[150] I. Gulrajani and D. Lopez-Paz, "In search of lost domain generalization," *arXiv*, vol. abs/2007.01434, 2020.

[1] S. Piérard and M. Van Droogenbroeck, "Summarizing the performances of a background subtraction algorithm measured on several videos," in *IEEE Int. Conf. Image Process. (ICIP)*, (Abu Dhabi, United Arab. Emir.), pp. 3234–3238, Oct. 2020.

[2] S. Piérard, A. Halin, A. Cioppa, A. Deliège, and M. Van Droogenbroeck, "Foundations of the theory of performance-based ranking," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, (Nashville, TN, USA), pp. 14293–14302, IEEE, Jun. 2025.

[4] S. Piérard, A. Halin, A. Cioppa, A. Deliège, and M. Van Droogenbroeck, "The Tile: A 2D map of ranking scores for two-class classification," *arXiv*, vol. abs/2412.04309, 2024.

[5] A. Halin, S. Piérard, A. Cioppa, and M. Van Droogenbroeck, "A hitchhiker's guide to understanding performances of two-class classifiers," *arXiv*, vol. abs/2412.04377, 2024.

[151] S. Piérard, A. Halin, F. Marelli, S. Pernas, and J. Pierre, "Sorbetto: a Python library for producing classification tiles," 2025. https://github.com/uliege-performance/sorbetto, DOI: 10.5281/zenodo.17591788.

7. Predicting rankings (7th article)

This chapter reproduces the following article [7]:

> S. Piérard, A. Deliège, A. Halin, and M. 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)*, (Nantes, France), pp. 1–6, IEEE, June 2025.

It covers the notions of ***performance***, ***ranking*** and ***domain***. The supplementary material of this article is reproduced in Chapter E.

Context

The scientific community is making significant efforts to carry out several evaluations for the algorithms (entities) developed, these evaluations being related to different domains (test sets). This trend can be seen both in publications (reviewers generally tend to request more and more evaluations) as well as in benchmarks that report the performance (scores) of the various algorithms for different (sub-)datasets. This is the case, for example, with CDnet, which we have already discussed in Chapters 1 and 3.

Motivation

The benchmarks can be used in several ways. We have already seen that they can be used to get an idea of the "average performance" of a method (see Chapter 1) or to determine the domains that are the easiest and most difficult to manage for a given entity, as well as the domains that are bottlenecks for the average performance (see Chapter 6). However, these analyses do not meet the needs of integrators who must select an entity for their application domain that differs from those represented in the benchmarks.

For entities that are deep learning models, model transferability estimation methods [152] attempt to predict how good the accuracy (and more rarely F_1 [153] or the balanced accuracy [154]) of the models can become after transferring (*i.e.*, adapting or fine-tuning) them for the applicative domain. But not all entities are fine-tunable deep learning models (*e.g.*, algorithms are not models). People who wish or need to keep an entity as is are left with their intuition when it comes to extrapolating the rankings present in the benchmarks to their own applicative domains.

Main contributions

In general, the relationship between the applicative domain and the domains considered in the available rankings is not obvious, and there is no reason to believe that the domains included in the benchmarks are sufficiently diverse for the applicative domain to be intermediate (this concept will be formally defined in Chapters 8 and 9). This raises the question of whether or not the existing benchmarks are valuable.

The article presented in this chapter explores this topic in the case of ranking background subtraction (BGS) algorithms (*i.e.*, pixelwise two-class crisp classifiers). After evaluating various intuitive ranking prediction strategies, I conclude that either these strategies are not very appropriate or that the benchmarks do not contain sufficient domain diversity.

Perspectives

We don't know yet how to make the best use of the existing benchmarks. Thanks to this work, we are now aware of this huge issue. Of course, the more diverse the domains to be considered, the more critical this problem becomes. Computer vision is undoubtedly one of the disciplines with the most varied applicative domains. From my point of view, more research on this topic will be needed in the future, especially for computer vision. The future research will benefit from my observation that the ranking prediction strategies should take the applicative preferences into account.

Remarks

It should be underlined that the task considered in this chapter consists in predicting rankings. Therefore, we need to compare and rank strategies that predict rankings. The theory presented in Chapter 2 applies to this task. The choice of Kendall's τ to induce the rankings is not insignificant, since this score belongs to the family of ranking scores with an appropriately chosen evaluation random experiment. More details are provided in Section A.2.6 of [2]. The Tile, developed for the ranking of two-class classifiers, proves nevertheless to be a well-suited visualization tool also in this case. However, I had to develop new flavors beyond those already presented in Chapters 5 and 6.

A Methodology to Evaluate Strategies Predicting Rankings on Unseen Domains

Sébastien Piérard, Adrien Deliège, Anaïs Halin, Marc Van Droogenbroeck
Montefiore Institute, University of Liège, Belgium
{S.Pierard,Adrien.Deliege,Anais.Halin,M.VanDroogenbroeck}@uliege.be

***Abstract*—Frequently, multiple entities (methods, algorithms, procedures, solutions, etc.) can be developed for a common task and applied across various domains that differ in the distribution of scenarios encountered. For example, in computer vision, the input data provided to image analysis methods depend on the type of sensor used, its location, and the scene content. However, a crucial difficulty remains: can we predict which entities will perform best in a new domain based on assessments on known domains, without having to carry out new and costly evaluations? This paper presents an original methodology to address this question, in a leave-one-domain-out fashion, for various application-specific preferences. We illustrate its use with 30 strategies to predict the rankings of 40 entities (unsupervised background subtraction methods) on 53 domains (videos).**

***Index Terms*—performance, multi-domain, ranking prediction, evaluation methodology, Tile, background subtraction**

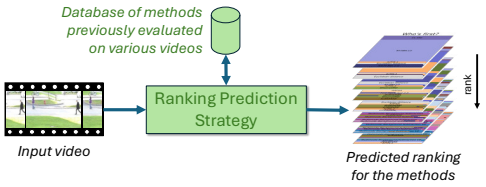


Fig. 1: In this paper, we explore the problem of predicting the rankings of computer vision methods (we take the particular case of background subtraction methods) on any new domain (video) based on a database storing the performances of these methods, previously evaluated on other domains (videos).

I. Introduction

Why do computer scientists rank entities (methods, algorithms, etc.)? Certainly, not just for picking a winner for a contest. The aim is rather to gain some knowledge that can help to select the most promising entity for a particular application. Predicting a ranking is a topic that has rarely been studied. As depicted in Fig. 1, we explore the case of a computer vision task and present an original methodology for investigating this theme, based on recent theory [2] and tools [4], [5] that are briefly recalled.

Background subtraction (BGS) is a computer-vision task for which the established rankings drastically differ from a video to another. Undoubtedly, this is due, on the one hand, to the variety of principles supporting these methods and, on the other hand, to the wide range of potential applications, and so to the wide variety of videos on which these methods can be applied. As a result, practitioners are frequently confronted with the selection of the most appropriate method for new video sequences, new camera placements, or video streams with characteristics changing unpredictably over time.

A first path consists in predicting the same ranking for all applicative domains, regardless of their characteristics. This is expected to be suboptimal, but it is worth trying such strategies for the sake of simplicity.

A second path consists in predicting the rankings based on the knowledge the practitioner has about his/her applicative domain. For example, he/she can specify a category to which all the inputs belong. The dataset *CDnet 2012* [21] defined 6 categories: *baseline*, *dynamic background*, *camera jitter*, *intermittent object motion*, *shadow*, and *thermal*, later complemented in *CDnet 2014* [155] by *bad weather*, *low framerate*, *night videos*, *PTZ*, and *turbulence*.

A third path consists in analyzing the input and predicting the ranking based on the perceived domain characteristics. Practitioners can collect representative images and analyze them to characterize their applicative domains as statistical distributions. For example, Halin *et al.* [9] have demonstrated the feasibility of characterizing domains, based on images, in a physically interpretable way with simulation-based inference [156] techniques. Alternatively, one could also classify pixels (*e.g.*, with semantic segmentation) and characterize the domain by the predicted distribution of classes.

In fact, our BGS example is just a particular case of a general and common problem that consists in predicting the rankings of entities for unseen domains based on performance evaluations previously performed for other domains, without having to carry out any new and costly evaluation. We go beyond just predicting the "best" entity as a user could be constrained to any subset of entities. To evaluate a predictive strategy, we compute the probability that, for any randomly chosen pair of entities (ϵ_a, ϵ_b), ϵ_a is predicted as worse or better than ϵ_b when it is actually the case.

Currently, scientists are not aware of how well simple and intuitive strategies perform. The ability to exploit existing rankings is all a blur. Our objective is to provide methodological elements and rigorous tools to clarify this. An originality of this work is that we do not rely on a small set of scores, but on a large family of scores, all of which have the required properties to induce meaningful rankings covering a diversity of application-specific preferences.

Our two main contributions are the following. (1) First, we establish a new and original methodology to evaluate and compare strategies predicting rankings on unseen domains. It leverages the recently introduced theoretical foundations for performance-based ranking [2] as well as a recently described tool (*i.e.*, the Tile [4], [5]). This framework allows us to cover an infinite and diversified family of meaningful performance-based rankings and to see how strategies perform according to application-specific preferences. (2) Second, we show a new usage for the Tile, originally introduced as a graphical tool to analyze and compare performances of two-class classifiers. In this work, the Tile appears to be the appropriate tool to depict the performance of strategies predicting rankings of two-class classifiers on unseen domains, and to compare them.

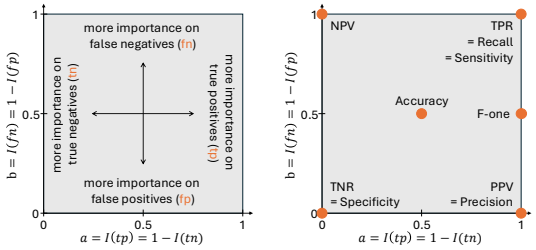


Fig. 2: Two equivalent readings of the Tile: a map of application-specific importances (left) and a map of scores to induce meaningful performance-based rankings (right).

II. Related Work

A. Background Subtraction

Most of the literature on BGS is about the methods [16], [157], [158], with some intended to be all-purpose (*i.e.*, universal) while others are developed for specific application domains [159]. Some consider constraints such as limited resources (time, memory, etc.) [10], [160], and some tackle multimodality [161].

A small part of the literature also concerns datasets, initially used to compare the BGS methods, then leveraged to develop supervised methods [16]. The best-known datasets are *CDnet 2012* [21] and its extension *CDnet 2014* [155], which are integrated into the changedetection.net platform.

Eventually, a part of the literature focuses on performance analysis. BGS is typically identified to a pixel-based two-class classification, where the negative and positive classes correspond to the background and the foreground (in the following, we denote a true negative, false positive, false negative, and true positive by tn, fp, fn, and tp). Jodoin *et al.* [18] and Piérard *et al.* [162] determined which performances are achievable by combining existing BGS methods. Braham *et al.* [163] introduced a generic technique to improve BGS methods performances. Piérard *et al.* [164] studied the limits of achievable performances with simple methods (pixelwise color comparisons), and how they vary with some properties (*e.g.*, amounts of noise and shadows) of the videos analyzed. Piérard *et al.* [1] highlighted the pitfalls of averaging values of scores measured on several videos, and provided a technique for summarizing the corresponding performances into a single interpretable one.

B. Performance-Based Ranking

Theoretical foundations for performance-based rankings, grounded in probability and order theories, have recently been introduced by Piérard *et al.* [2] through a rigorous axiomatic framework[1]. These axioms are guardrails to guarantee meaningful rankings while leaving the flexibility to adjust the rankings *w.r.t.* application-specific preferences.

When particularized to two-class classification (*e.g.*, between background and foreground, or between negatives and positives), it has been shown in [2], [4], [5] that one can fine-tune the relative importance given to the two types of correct classifications, as well as to the two types of erroneous classifications, by inducing the performance ordering from a parametric score R_I called *ranking score*:

$$R_I : P \mapsto R_I(P) = \frac{\sum_{\omega \in \{tn,tp\}} I(\omega) P(\{\omega\})}{\sum_{\omega \in \{tn,fp,fn,tp\}} I(\omega) P(\{\omega\})}, \quad (1)$$

where P denotes the performance, and I the importance, *i.e.*, the application-specific preferences. Ranking scores sharing the same values for $a = \frac{I(tp)}{I(tn)+I(tp)}$ and $b = \frac{I(fn)}{I(fp)+I(fn)}$ lead to the same rankings [4]. So, the diversity of rankings can be shown on the $(a, b) \in [0, 1]^2$ square, coined as the Tile [4], [5]. Interestingly, the family of ranking scores contains scores commonly used in the BGS community (see Fig. 2).

C. Ranking Prediction

Predicting a ranking of entities, sometimes called "learning to rank" in machine learning, is a task used for instance in recommendation systems, or action quality assessment (*e.g.*, in sports). The strategies generally fall into three categories: *pointwise*, where strategies predict the performance score of each entity individually, and derive a ranking from these scores [165], [166]; *pairwise*, where strategies predict the relative ordering between entity pairs, and derive a ranking from pairwise orderings [167]; and *listwise*, where strategies predict directly the entire ranked list [168]. Pointwise approaches are simple and efficient but ignore the relative nature of ranking tasks. Pairwise strategies better capture it and are

[1]Their 1st axiom states that any performance-based ranking should be derived from a preorder on performances, which ensures the stability of the rankings *w.r.t.* insertions and deletions of ranked entities. Their 2nd axiom gives compatibility conditions between the preorders and the considered task, modeled by a random variable called *satisfaction*. Their 3rd axiom gives compatibility conditions between the preorders and known properties about the evaluation (*i.e.*, the mapping of the entities to their performances).

widely used, though they struggle with global ranking consistency. Listwise approaches, by directly optimizing ranking scores, often achieve superior performance, but are typically more complex and computationally demanding [169], [170]. Compared to that literature, we do not aim to learn a model that ranks entities based on domain (videos) features; instead, we leverage pre-computed rankings across multiple domains to infer the ranking on a previously unseen one.

III. Methodology

We now describe our new methodology to evaluate and compare strategies predicting the rankings of methods in unseen domains. We focus on strategies able to predict a large family of meaningful rankings: all those induced by ranking scores [2] (*cf.* Eq. 1). For problems comparable to two-class classification, these rankings can be mapped on the Tile [4], [5], which leads to a *mille-feuille*[2], as shown in Fig. 3.

We propose to evaluate strategies, for any given domain, with correlation Tiles [5] giving the rank-correlation between the predicted and ground-truth rankings, *w.r.t.* the various application-specific preferences (specified by I, or a and b). We choose Kendall's τ as this score is coherent with the performance-based ranking framework[3]. Moreover, Kendall's τ can be nicely interpreted as the probability of observing, in two given rankings, the same relative order between two randomly chosen methods, this probability being linearly mapped into the $[-1, +1]$ interval (see Fig. 4).

Considering several domains can be done in a *leave-one-domain-out* fashion, predicting the ranking for each domain using information from the others. The expected performance of the strategy is given by the *mean correlation Tile*, whereas the worst-case is given by the *minimum correlation Tile*. This will be illustrated in Sec. IV.

Two natural baselines can be defined when the ground-truth rankings are known for several domains. (1) The mean of τ over all domain pairs estimates the expected agreement when generalizing a ranking from a randomly chosen reference domain to another random domain. (2) The minimum of τ across all domain pairs captures the worst-case agreement for any fixed domain and arbitrary reference.

To compare various strategies, it is essential to bear in mind that the best strategy might depend on the application-specific preferences I. This can be mapped on the Tile.

IV. Application

We now apply our methodology to the problem of predicting the performance-based rankings of 40 unsupervised BGS methods on the 53 videos of *CDnet 2014*. As said, we use only rankings, no implementation of these methods, ground-truth segmentation masks, nor any inference and evaluation step. In particular, the priors and the rates of predictions for the background and foreground are unknown. We compare strategies that have no knowledge about the input video (their expected distribution put aside) and strategies that leverage some domain characterization: a category (denoted by *), a distribution of semantic classes[4] (denoted by †), or both.

The information known about the test video (*i.e.*, for which we must predict the ranking) is its two-fold domain characterization: (1) its category in *CDnet 2014* and (2) an estimate of the statistical distribution of the pixel-based semantic labels. The information known for the others is three-fold: (1) its category, (2) the join distribution of the pixel-based semantic labels, ground-truth class, and estimated classes, and (3) the performance-based ranking of the 40 BGS methods. All distributions are for the publicly available subset of pixels annotated as background or foreground in *CDnet 2014*.

A. Baselines for Ranking Prediction Strategies

The two natural baselines defined in Sec. III are shown in Fig. 5. On the left-hand side, we see that there is only a little average correlation between rankings from one video to another. The mean value for τ around 0.25 means that there is about 62.5% chance of correctly ordering two methods if we refer to a predetermined ranking for any randomly chosen video. Moreover, we see that the difficulty of predicting rankings is not uniform: it seems to be a more difficult problem when $I(fn)$ is low *w.r.t.* $I(fp)$ (bottom of the Tile). On the right-hand side, we observe that, in the worst case, regardless of what the application-specific preferences I are, there can be a negative correlation between the rankings on two videos. All things considered, the problem of predicting the performance-based rankings of BGS methods is a difficult problem.

B. First Experiment

The aim of the 1st experiment is to compare a variety of 11 very intuitive strategies. [CDnet] is the ranking given on changedetection.net[5] at an arbitrarily chosen moment[6]. It is not sensitive to application-specific preferences. [mean-P] is the ranking based on the summarization [1] of the performances determined on the 52 other videos. [sem-P†] is the ranking based on the performance predicted from, on the one hand, the distribution of semantic labels in the input video and, on the other hand, the probabilities of false/true background/foreground conditionally to the semantic label for the other videos. [sem-d†] is the ranking known for the closest video in terms of semantic characterization (Bhattacharyya

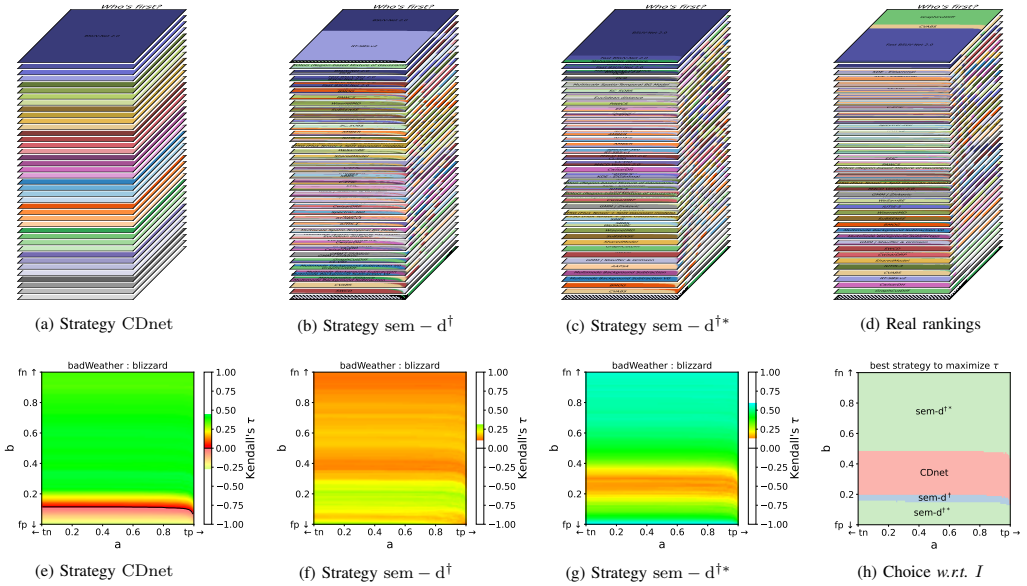
[2] *Mille-feuille* pastries are composed of several layers, with some made with puff pastry, resembling to a stacking of multiple sheets.

[3] It turns out that Kendall's τ is itself a ranking score for ranking problems when pairs of miss-ordered and well-ordered methods are assigned a satisfaction of -1 and +1, respectively. See the appendix of [2] for more information.

[4] The semantic labels are used only to cluster the pixels. Thus, we can rely on predicted (not perfect) segmentations. We selected a model from MMSegmentation [108] based on two criteria: semantic diversity and expected performance. For the former, we limited the choice to ADE20K [110] models and COCO-Stuff [112] models. Based on performance insights from *entity Tiles* [5], we selected Mask2Former [146] trained on ADE20K.

[5] This is our sole strategy that is not leave-one-video-out, as all our input videos are involved in this ranking. The related results could be optimistic.

[6] Version of March, 11th 2025. This remark is needed because the overall ranking of changedetection.net does not satisfy the 1st axiom of [2]: it is not stable. The relative order between the 40 methods, can be influenced by the presence of other methods (*e.g.*, the supervised ones) on the platform, and could change if a method was added or removed from the platform.



(a) Strategy CDnet (b) Strategy sem − d† (c) Strategy sem − d†* (d) Real rankings

(e) Strategy CDnet (f) Strategy sem − d† (g) Strategy sem − d†* (h) Choice *w.r.t. I*

Fig. 3: Illustration of our methodology for the evaluation and comparison of 3 strategies to predict the rankings of 40 BGS methods on the video "bad weather: blizzard". We use Tiles [4] to cover the different application-specific preferences (a, b). The upper row shows: (a) the predicted ranking based just on the global ranking given on changedetection.net (strategy CDnet), (b) the one based on the performance measured on another semantically close video (strategy sem − d†), (c) the one based on the performance measured on another semantically close video in the same category (strategy sem − d†*), and (d) the actual ranking we would like to predict (our ground truth). These "mille-feuilles" are stackings of entity Tiles [5]: the k-th layer shows the methods ranked k-th, the worst methods being at the base of the mille-feuille, and the best ones on its top. The lower row shows the correlation Tiles [5] between these rankings: (e) the correlation between (a) and (d), (f) the correlation between (b) and (d), and (g) the correlation between (c) and (d). The Tile (h) shows which strategy gives the best correlation. This methodology is applied in Sec. IV to compare many more strategies on a diversified set of 53 videos.

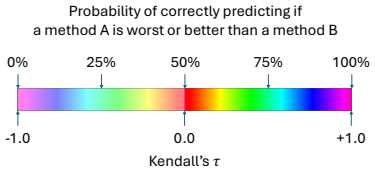


Fig. 4: Interpretation of the rank correlation τ (Kendall). Note the color code for displaying the value of τ, as is used in other figures.

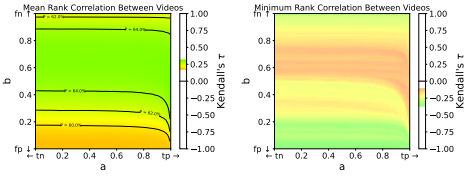


Fig. 5: The two natural baselines for ranking prediction strategies (as defined in our methodology) in the particular case of 40 unsupervised BGS algorithms ranked on 53 videos.

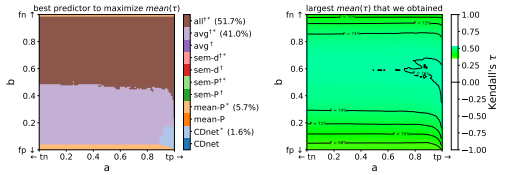
distance between distributions of semantic labels). [avg†] is the ranking based on the summarization [1] of the performances predicted for CDnet[7], mean-P, sem-P†, and sem-d†.

[7]Because the summarization only accepts performances as input, and not real values, we need a solution to convert a value into a performance. For any value $v \in [l, u]$, we assign the performance P such that $P(\{tn\}) = P(\{tp\}) = \frac{1}{2}\frac{v-l}{u-l}$ and $P(\{fp\}) = P(\{fn\}) = \frac{1}{2}\frac{u-v}{u-l}$, as $R_I(P)$ is strictly monotonously increasing with v for all ranking scores R_I.

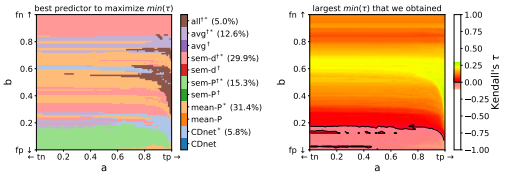
[...*]: each of these five strategies comes with a variant that is category-specific and that is denoted by *. [all†] is the ranking based on the summarization [1] of the performances predicted for avg† and avg†*.

The results are shown in Fig. 6. We see that, when we want to maximize $mean(\tau)$, it is clearly useful to consider hybrid strategies, choosing one of the basic strategies according to

ART. 7

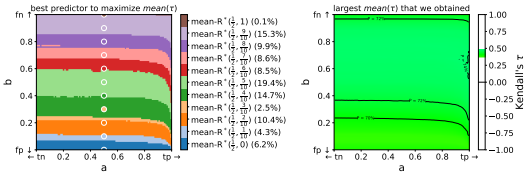


(a) Hybrid strategy maximizing $mean(\tau)$.

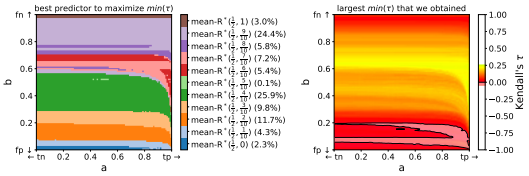


(b) Hybrid strategy maximizing $min(\tau)$.

Fig. 6: Results of our 1st experiment.



(a) Hybrid strategy maximizing $mean(\tau)$.



(b) Hybrid strategy maximizing $min(\tau)$.

Fig. 7: Results of our 2nd experiment.

the application-specific preferences: none of the 11 strategies tested in this 1st experiment is the best in more than 51.7% of the Tile. Moreover, we see that for the vast part of the Tile, it is best to opt for strategies based on the summarization of performances predicted in various ways. The maximum $mean(\tau)$ obtained is relatively independent of application-specific preferences (the Tile is all in green shades). When we want to maximize $min(\tau)$, the strategies to select are, over the majority of the Tile, strategies that exploit knowledge of the category to which the input video belongs.

C. Second Experiment

Now, we aim to establish something simpler, and easier to interpret, than the first experiment. We study the vertical behavior of the Tile through a parametric family of strategies. We focus on strategies that exploit the knowledge of the category, but put semantics aside. This experiment involves 11 strategies to predict rankings. [mean-R$^*(\frac{1}{2}, x)$] is the category-specific weighted[8] arithmetic mean of the ranks induced by the canonical ranking scores $R_{I_{a,b}}$ with $a = \frac{1}{2}$ and $b = x$.

Results are shown in Fig. 7 for the discrete set of values $x \in \{0, \frac{1}{10}, \frac{2}{10} \dots 1\}$. When we want to maximize $mean(\tau)$, we see that all 11 strategies are useful. Moreover, the point $(a, b) = (\frac{1}{2}, x)$ lies in, or is not far from, the area of the Tile where the strategy mean-R$^*(\frac{1}{2}, x)$ is the best (see the colored points). The average τ achievable with these strategies is close to the average τ obtained with the strategies of the first experiment. When we want to maximize $min(\tau)$, our conclusions are similar to those given for the maximization of $mean(\tau)$. This paves the way for an even more refined hybrid strategy, combining not only 11 strategies as in this experiment, but an infinite number of them (one for each a and b), which is done in the 3rd experiment.

[8]We weight videos as in changedetection.net: all categories have the same weight, and all videos in a given category have also the same weight. The video on which the rankings are predicted has a zero weight.

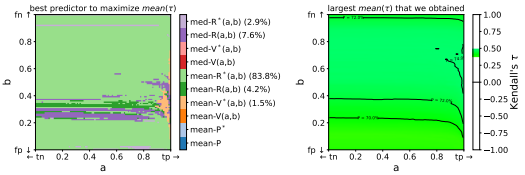
D. Third Experiment

Now, we aim to explore two questions: (1) is it better to compute average performances, average values, or average ranks? (2) is it more suitable to compute means or medians? This experiment involves 10 strategies. [mean-P] is the ranking based on the summarization [1] of the performances determined on the 52 other videos. [mean-V] is the weighted arithmetic mean of the values taken by the canonical ranking score, corresponding to the application-specific preferences. [mean-R] is the weighted arithmetic mean of the ranks induced by the canonical ranking score, corresponding to the application-specific preferences. [med-V] is similar to mean-V, except that we take the median instead of the mean. [med-R] is similar to mean-R, except that we take the median instead of the mean. [. . .*]: as in the 1st experiment, each of these five strategies comes with a variant that is category-specific and that is denoted by *.

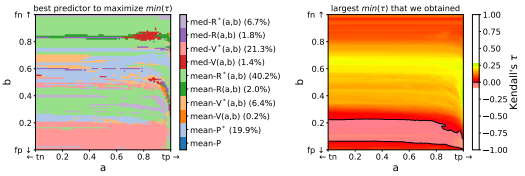
The results of the 3rd experiment are displayed in Fig. 8. When we want to maximize $mean(\tau)$, in 83.8% of the Tile, mean-R* is the best strategy among the 10. When we want to maximize $min(\tau)$, for most of the Tile, it is better to take only strategies that exploit videos within the same category. The $mean(\tau)$ and $min(\tau)$ Tiles are very similar to the ones we obtained in our first two experiments.

V. Conclusion

This work introduces a methodology for comparing strategies for the prediction of rankings of computer vision methods (*e.g.*, BGS methods) on new domains (*e.g.*, videos), *w.r.t.* application-specific preferences, by exploiting the knowledge that can be found in publicly available multi-domain rankings (*e.g.*, changedetection.net). The methodology, as presented for a problem similar to the two-class classification, takes



(a) Hybrid strategy maximizing $mean(\tau)$.



(b) Hybrid strategy maximizing $min(\tau)$.

Fig. 8: Results of our 3rd experiment.

advantage of a recently introduced visualization tool called "Tile" [4], [5], which is based on a theoretical framework for performances [2]. In the case that we studied, we have shown that the performance-based rankings of computer vision methods on new domains is far from being an easy or solved problem. Because of its practical importance, we hope that this work will stimulate the community to investigate and develop more powerful strategies, and that our proposed methodology will help to establish their effectiveness.

Acknowledgment

S. Piérard, A. Deliège, and A. Halin are funded respectively by (1) the Service Public de Wallonie (SPW) Recherche (grant 8573, ReconnAIssance project), (2) ULiège (project DESTINA), and (3) the SPW EER, Wallonia, Belgium (grant 2010235, ARIAC by DIGITALWALLONIA4.AI).

References

[2] Sébastien Piérard, Anaïs Halin, Anthony Cioppa, Adrien Deliège, and Marc Van Droogenbroeck, "Foundations of the theory of performance-based ranking," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Nashville, TN, USA, Jun. 2025, pp. 14293–14302, IEEE.

[4] 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*, vol. abs/2412.04309, 2024.

[5] Anaïs Halin, Sébastien Piérard, Anthony Cioppa, and Marc Van Droogenbroeck, "A hitchhiker's guide to understanding performances of two-class classifiers," *arXiv*, vol. abs/2412.04377, 2024.

[21] Nil Goyette, Pierre-Marc Jodoin, Fatih Porikli, Janusz Konrad, and Prakash Ishwar, "Changedetection.net: A new change detection benchmark dataset," in *IEEE Int. Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW)*, Providence, RI, USA, Jun. 2012, pp. 1–8, Inst. Electr. Electron. Eng. (IEEE).

[155] 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 Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW)*, Columbus, OH, USA, Jun. 2014, pp. 393–400, Inst. Electr. Electron. Eng. (IEEE).

[9] Anaïs Halin, Sébastien Piérard, Renaud Vandeghen, Benoît Gérin, Maxime Zanella, Martin Colot, Jan Held, Anthony Cioppa, Emmanuel Jean, Gianluca Bontempi, Saïd Mahmoudi, Benoît Macq, and Marc Van Droogenbroeck, "Physically interpretable probabilistic domain characterization," in *Asian Conf. Comput. Vis. Work. (ACCV Work.)*, Hanoi, Vietnam., 2025, vol. 15482 of *Lect. Notes Comput. Sci.*, pp. 17–35, Springer Nat. Singap.

[156] Kyle Cranmer, Johann Brehmer, and Gilles Louppe, "The frontier of simulation-based inference," *Proc. National Acad. Sci. (PNAS)*, vol. 117, no. 48, pp. 30055–30062, May 2020.

[16] 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*, vol. 117, pp. 8–66, Sept. 2019.

[157] 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, Oct. 2019.

[158] 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, Oct. 2019.

[159] Marc Braham and Marc Van Droogenbroeck, "Deep background subtraction with scene-specific convolutional neural networks," in *IEEE Int. Conf. Syst. Signals Image Process. (IWSSIP)*, Bratislava, Slovakia, May 2016, pp. 1–4.

[10] Olivier Barnich and Marc Van Droogenbroeck, "ViBe: A universal background subtraction algorithm for video sequences," *IEEE Trans. Image Process.*, vol. 20, no. 6, pp. 1709–1724, Jun. 2011.

[160] Anthony Cioppa, Marc Van Droogenbroeck, and Marc Braham, "Real-time semantic background subtraction," in *IEEE Int. Conf. Image Process. (ICIP)*, Abu Dhabi, United Arab. Emir., Oct. 2020, pp. 3214–3218, Inst. Electr. Electron. Eng. (IEEE).

[161] Jérôme Leens, Sébastien Piérard, Olivier Barnich, Marc Van Droogenbroeck, and Jean-Marc Wagner, "Combining color, depth, and motion for video segmentation," in *Computer Vision Systems*. 2009, vol. 5815 of *Lect. Notes Comput. Sci.*, pp. 104–113, Springer.

[18] Pierre-Marc Jodoin, Sébastien Piérard, Yi Wang, and Marc Van Droogenbroeck, "Overview and benchmarking of motion detection methods," in *Background Modeling and Foreground Detection for Video Surveillance*, chapter 24. Chapman and Hall/CRC, Jul. 2014.

[162] Sébastien Piérard, Marc Braham, and Marc Van Droogenbroeck, "An exploration of the performances achievable by combining unsupervised background subtraction algorithms," *arXiv*, vol. abs/2202.12563, 2022.

[163] Marc Braham, Sébastien Piérard, and Marc Van Droogenbroeck, "Semantic background subtraction," in *IEEE Int. Conf. Image Process. (ICIP)*, Beijing, China, Sept. 2017, pp. 4552–4556.

[164] Sébastien Piérard and Marc Van Droogenbroeck, "A perfect estimation of a background image does not lead to a perfect background subtraction: analysis of the upper bound on the performance," in *Int. Conf. Image Anal. Process. (ICIAP), Work. Scene Backgr. Model. Initial. (SBMI)*. 2015, vol. 9281 of *Lect. Notes Comput. Sci.*, pp. 527–534, Springer.

[1] Sébastien Piérard and Marc Van Droogenbroeck, "Summarizing the performances of a background subtraction algorithm measured on several videos," in *IEEE Int. Conf. Image Process. (ICIP)*, Abu Dhabi, United Arab. Emir., Oct. 2020, pp. 3234–3238.

[165] David Cossock and Tong Zhang, "Subset ranking using regression," in *Learn. Theory*. 2006, vol. 4005 of *Lect. Notes Comput. Sci.*, pp. 605–619, Springer Berl. Heidelb.

[166] Ping Li, Qiang Wu, and Christopher Burges, "McRank: Learning to rank using multiple classification and gradient boosting," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, Vancouver, Can., Dec. 2007, vol. 20, pp. 1–8, Curran Assoc. Inc.

[167] Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender, "Learning to rank using gradient descent," in *Int. Conf. Mach. Learn. (ICML)*, Bonn, Germany, 2005, pp. 89–96, ACM Press.

[168] Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li, "Learning to rank: from pairwise approach to listwise approach," in *Int. Conf. Mach. Learn. (ICML)*, Corvalis, OR, USA, Jun. 2007, pp. 129–136, ACM.

[169] T. Liu, "Learning to rank for ir," *Foundation and Trends in IR*, vol. 3, no. 3, pp. 225–331, 2009.

[170] Niek Tax, Sander Bockting, and Djoerd Hiemstra, "A cross-benchmark comparison of 87 learning to rank methods," *Inf. Process. & Manag.*, vol. 51, no. 6, pp. 757–772, Nov. 2015.

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

[110] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba, "Scene parsing through ADE20K dataset," in *IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 5122–5130, Inst. Electr. Electron. Eng. (IEEE).

[112] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari, "COCO-stuff: Thing and stuff classes in context," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, Jun. 2018, pp. 1209–1218, Inst. Electr. Electron. Eng. (IEEE).

[146] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar, "Masked-attention mask transformer for universal image segmentation," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 1280–1289, Inst. Electr. Electron. Eng. (IEEE).

ART. 7

8. Mixture domain adaptation (8th article)

This chapter reproduces the following article [8]:

> S. Piérard, A. Cioppa, A. Halin, R. Vandeghen, M. Zanella, B. Macq, S. Mahmoudi, and M. Van Droogenbroeck, "Mixture domain adaptation to improve semantic segmentation in real-world surveillance," in *IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, (Waikoloa, Hawaii, USA), pp. 22–31, IEEE, January 2023.

It covers both the notions of ***performance*** and ***domain***. The supplementary material of this article is reproduced in Chapter F. This article results from a collaboration initiated on the occasion of the TRAIL summer workshop 2022. I came up with the idea for this research project and focused mainly on the theoretical aspects as well as on the analysis of the results.

Context

In Chapter 7, we have seen that the ranking depends not only on the applicative preferences but also on the domain. The same goes for the best entity. In practice, the domain is not necessarily known in advance and may even change over time and space. Instead of being limited to a finite set of entities and switching abruptly from one entity to another when the domain changes, we would like to be able to combine the (outputs of the) available entities to obtain a hybrid entity capable of continuously adapting to the current domain.

Motivation

Let us take the example of an autonomous car. The distribution of situations it is likely to encounter in the near future may depend on the environment (urban, rural, etc.), weather conditions, traffic density, etc. Given the wide variety of situations an autonomous vehicle may encounter, it seems rather utopian to believe that it can be taught to behave correctly in all situations at once and that it can be tested in all situations to guarantee its correctness. In theory, optimal decisions depend on the statistical distribution of situations that may arise. And this distribution may differ from one vehicle to another, depending on where they are located. It may also change over time, sometimes even rapidly. This means that an intelligent vehicle must have the ability to adapt on the fly

to the current domain. Of course, we cannot periodically stop the vehicle, collect data, have it annotated by human experts to obtain ground truths, and learn a new predictive model (entity) or fine-tune an existing model based on the new data and annotations before restarting the vehicle.

Main contributions

Being able to perform real-time adaptation without new data and annotations is a real challenge. In this chapter, I assume that one knows the relationship that links the current (applicative) domain to the domains for which the available entities are optimal.

In line with the idea presented in Chapter 1 and extended in Chapter 6, I consider that the different domains are weighted. More precisely, I study a particular type of relationship between the target and the source domains: a mixture (convex combination) of probability measures. In Chapter 9, I propose a methodology to estimate the mixture weights. But in this chapter, I simply consider that they are given. In the classical terminology of domain adaptation, the current domain is the target domain, while the domains that are associated with the available entities are called source domains. I present the theory and an unsupervised algorithm called *Mixture Domain Adaptation* (MDA) to determine how entities optimized for the source domains can be combined to obtain an entity optimized for the target domain. Experimental results are reported for soft semantic segmentation models with 19 strongly imbalanced semantic classes.

The key idea behind this work is to note that the probability theory can be used to reason in this context if we associate a joint distribution on the input (the evidence E, sometimes also called attributes, features, or observations X) and the expected output (the hypothesis H to be chosen, or the ground truth Y). More precisely, we consider a measurable space, associate probability measures to the domains, and define the evidences and hypotheses as events. For a given set of domains, we can define the intermediate domains as those associated with the probability measures that are mixtures (*i.e.*, convex combinations) of those associated with the base domains.

Perspectives

The algorithm presented in Fig. 4 of the article can be viewed as a software architecture where the different blocks correspond to distinct software modules. This original software architecture enables the adaptability, modularity, and reliability. Modularity is necessary for the development of complex systems. With our approach, to manage a larger set of domains, it is sufficient to add pre-trained models for additional source domains and update the domain discriminator model, which requires only unlabeled data. Reliability stems partly from the ability to test the specialized modules one by one rather than having to test an entire "back box" system, and also partly from the way of gluing these modules that is theoretically guaranteed to be correct. In the next chapter, I will discuss how to complete this architecture to further improve its reliability and to address the interpretability.

It should be stressed that the illustration of the intelligent vehicle chosen in this article is far from being the only scenario in which MDA could be used. In the medical field, it is well known that the prevalence and the form of a given disease can vary from one hospital to another and from one region of the world to another. This is due to different environmental characteristics, to different lifestyles, and to populations with different genetic characteristics. There is therefore a clear advantage in being able to obtain models optimized for each hospital. Some already have them; others do not. For the latter, instead of undertaking a long and costly campaign to collect and annotate data, it would be desirable to be able to leverage the models that have already been optimized in other hospitals. I assume that, based on demographic and environmental data, it would be possible to express the distribution of cases that can arise at a given hospital as a mixture of what can arise at others. If the various training datasets were available, they could be combined, weighted appropriately, and used for further training. Unfortunately, sharing annotated data between hospitals is a sensitive issue. The MDA technique presented in this paper avoids the need to share annotated data by requiring only the sharing of trained models.

Remarks

- Readers wishing to gain an in-depth understanding of the link between MDA and summarization are invited to take a careful look to Fig. 6 in the article. The curves show the accuracies obtained with various entities in intermediate (target) domains between two source domains (they are at the left and right extremities). The blue curve corresponds to one *fixed* entity (*i.e.*, without any domain adaptation), the orange curve corresponds to another *fixed* entity, and each point on the other curves to *adapted* entities. With any fixed entity, the performance in an intermediate domain is the summarized performance, with the performance weights corresponding to the domain weights. However, we saw in Chapters 1 and 6 that the accuracy of the summarized performance is the arithmetic mean of the accuracies with the weights of the values corresponding to the weights given to the performances. We can immediately conclude that the accuracy is the arithmetic mean of the accuracies with the weights of the values corresponding to the weights given to the domains. This is why the blue and orange curves should theoretically be straight lines. If one wants to change the accuracy to any other score, the blue and orange curves can be computed from the performances determined in the source domains, without performing any test in intermediate domains, by applying the chosen score on summarized performances.

- The attentive reader will note the similarity of reasoning with the performance summarization introduced in Chapters 1 and 6. We saw that the weights to be used for arithmetically averaging conditional probability scores, expected ratio scores, and ranking scores are not the weights given to the domains. Similarly, we see here that the weights to be used for arithmetically averaging the posteriors are not the mixture weights. To estimate them, we propose using a domain discriminator

model. This is a first example of *game of weights*. A second example is the effect of a prior shift operation. It appears in this article and is the same as the one already encountered in Chapter 4.

Mixture Domain Adaptation to Improve Semantic Segmentation in Real-World Surveillance

Sébastien Piérard
University of Liège, Belgium
S.Pierard@uliege.be

Anthony Cioppa
University of Liège, Belgium
Anthony.Cioppa@uliege.be

Anaïs Halin
University of Liège, Belgium
Anais.Halin@uliege.be

Renaud Vandeghen
University of Liège, Belgium
R.Vandeghen@uliege.be

Maxime Zanella
University of Louvain-la-Neuve, Belgium
Maxime.Zanella@uclouvain.be

Benoît Macq
University of Louvain-la-Neuve, Belgium
Benoit.Macq@uclouvain.be

Saïd Mahmoudi
University of Mons, Belgium
Said.Mahmoudi@umons.ac.be

Marc Van Droogenbroeck
University of Liège, Belgium
M.VanDroogenbroeck@uliege.be

Abstract

Various tasks encountered in real-world surveillance can be addressed by determining posteriors (e.g. by Bayesian inference or machine learning), based on which critical decisions must be taken. However, the surveillance domain (acquisition device, operating conditions, etc.) is often unknown, which prevents any possibility of scene-specific optimization. In this paper, we define a probabilistic framework and present a formal proof of an algorithm for the unsupervised many-to-infinity domain adaptation of posteriors. Our proposed algorithm is applicable when the probability measure associated with the target domain is a convex combination of the probability measures of the source domains. It makes use of source models and a domain discriminator model trained off-line to compute posteriors adapted on the fly to the target domain. Finally, we show the effectiveness of our algorithm for the task of semantic segmentation in real-world surveillance. The code is publicly available at `https://github.com/rvandeghen/MDA`.

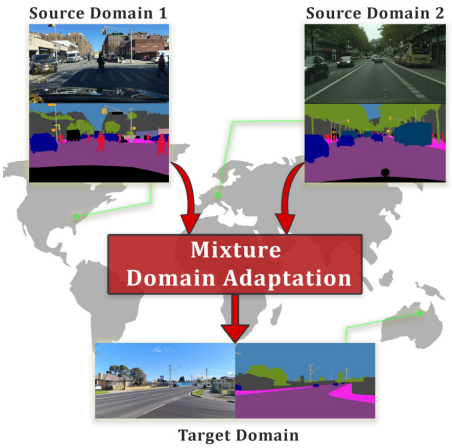


Figure 1. We present a theoretically motivated algorithm for unsupervised domain adaptation of posteriors on the fly. In particular, we express the target domain as a convex combination of the source domains. We show the effectiveness of our algorithm for the task of semantic segmentation in real-world surveillance.

1. Introduction

Applying artificial intelligence methods to real-world surveillance requires taking into account the specificity of the considered scene. In this paper, we address this issue by computing posteriors, which are probabilities related to the content of the scene. For example, for the task of semantic segmentation, the posteriors are the conditional probabilities of the semantic classes (the *hypotheses*) given a pixel from an image (an *evidence*).

There are three main reasons for focusing on posteriors: (1) they help interpret results compared to hard decisions (*e.g*. classifying an object with some degree of confidence rather than simply claiming it belongs to a class), (2) they enable optimal decisions for most common criteria (*e.g*. maximizing the posteriors for the accuracy or maximizing the likelihoods for the balanced accuracy), and (3) they allow to adapt to the specific domain of a scene, as demonstrated in this paper.

In general, the domains are rarely known in advance, which prevents any possibility of scene-specific optimization. Indeed, the domains depend on the choices made for the data acquisition (*e.g*. where the camera is installed or when the videos or images are captured) as well as on the distribution of scenarios that can occur (*e.g*. varying weather or lighting conditions). Furthermore, in some applications, the domains may change over time, making it even more challenging to take into account the specificity of the scene. This typically occurs when the surveillance cameras are onboard of moving vehicles.

When dealing with multiple domains, a training dataset can be collected in each domain. Common but oversimplistic assumptions are that the testing distribution (referred to as the *target domain*) is a fixed mixture of the various training datasets (referred to as the *source domains*) and that the mixture parameters are known in advance. Under these assumptions, it is reasonable to mix the various training datasets before training the models. This, however, is impossible in most real-world cases as the mixture parameters depend on the target domain which is unknown at learning time. An on-the-fly and unsupervised domain adaptation is therefore required for the sake of flexibility.

In this paper, we propose an algorithm that makes use of *source models* and a *domain discriminator model* trained off-line to obtain posteriors adapted on the fly to the target domain. We present a formal proof of this algorithm that is interpretable and runs in real time. Its effectiveness is shown experimentally for the task of semantic segmentation in real-world surveillance.

Contributions. Our contributions are as follows. **(i)** We define a probabilistic framework for a particular unsupervised mixture domain adaptation problem. **(ii)** Based on this framework, we present a formal proof of an unsupervised algorithm to adapt the posteriors on the fly for a changing target domain. **(iii)** We compare our algorithm with common posterior combination heuristics and show its superiority on several real-world surveillance datasets.

2. Related work

Hereafter, we present the related work first on unsupervised domain adaptation and second on semantic segmentation, with some references specific to the field of real-world surveillance.

Unsupervised domain adaptation. Domain adaptation aims at transferring knowledge from source domains to unseen target domains and alleviating the impact of domain shift in data distributions [171]. An even more challenging scenario arises when no label is available in the target domains; this is called unsupervised domain adaptation [172]. Domain shifts, which are defined as a change in the data distribution between the training distribution of an algorithm and the inference distribution, can be categorized into three main categories [171]: (1) *prior shift* (*a.k.a. target shift*), when the priors are different but the likelihoods are identical [94]; (2) *covariate shift*, when the marginal probability distributions differ between the source and target domains, but not the posteriors [173]; and (3) *concept shift*, when data distributions remain unchanged across source and target domains, while the posteriors differ between domains [174].

A common solution to unsupervised domain adaptation consists in adding a domain discriminator to learn domain-independent features [175]. This approach needs a new training phase, making real-time predictions unfeasible. Moreover, aligning the features can reduce their discriminative power [176]. A more restrictive scenario can be investigated, in which the source domain data cannot be accessed at test-time. Liang *et al*. [177] freeze the classifier and learn representations from the target domains aligned to the source hypothesis. Wang *et al*. [178] adapt the normalization layers while minimizing the prediction entropy. Boudiaf *et al*. [179] propose to keep the parameters of the model unchanged to avoid collapse in case of poorly diversified batches and only correct the predictions by encouraging neighboring representations to have consistent predictions.

Other works tackle the problem of domain shifts in the case of semantic segmentation. Zhang *et al*. [180] suggest to use a curriculum learning approach to learn more transferable global and local properties across domains. Li *et al*. [181] propose a bidirectional learning framework in which a translation module and a segmentation model are trained alternatively.

In the context of multi-source semantic segmentation, unsupervised domain adaptation has been studied with some approaches generating adapted domains [182, 183] or aligning their features distribution by matching their moments [184]. He *et al*. [185] propose to reduce the discrepancy between domains by aligning their pixel distribution before training models in a collaborative way to share knowledge between sources. One of the closest works to ours for binary classification is presented in [186]. The authors derive posteriors from Bayes' theorem by approximating priors with distances to each point in each source domain and likelihoods from nearest neighbor points. In our work, we address a multi-class case in a dense task, *i.e*. semantic segmentation. We further characterize the target

domain as a mixture of the source domains.

Semantic segmentation. For the past decade, semantic segmentation has proven to be a powerful tool for global scene understanding [187, 188, 189]. To support the development of semantic segmentation networks, many annotated datasets emerged such as ADE20K [110, 190], PascalVOC2012 [111], COCO [191], or CityScapes [109]. Their availability has boosted the performance of algorithms over the years, leading to algorithms such as PSPNet [116], Mask R-CNN [192], PointRend [135], or SETR [138].

In the field of real-world surveillance, semantic segmentation can form the basis of complex downstream tasks such as urban scene characterization [193], maritime surveillance [194], water level estimation [195], low-light video enhancement [196], or super-resolution [197]. Due to the variety of scenes and downstream tasks, it is difficult to develop semantic segmentation algorithms that remain competitive without further scene-specific tuning. For instance, when placing a new surveillance camera, we need to ensure that the segmentation network will be able to perform well in this novel environment. Furthermore, the network must be robust to dynamic domain changes. Short-term changes may include illumination changes, sudden heavy rains, or different occupancy during rush hours, while long-term changes may include day-lasting heavy snow during winter or road constructions that last for months. If the network has not been trained on all those particular domains, it may simply fail to predict the correct labels and therefore lead to critical failures. Unfortunately, the networks that are robust to a wide variety of domains are often too large, may not fit in memory, may not be fast enough for real-time processing, and may require high power consumption.

One naive scene-specific adaptation consists in training a small real-time network on scene-specific data in a supervised fashion. However, this requires collecting a dataset prior to installation, which may be impractical. Also, if the domain changes dynamically, the network may be unable to adapt. As a solution, Cioppa *et al*. [198] propose the online distillation framework, in which a large network is used to train a small real-time network on the fly. This allows the real-time network to adapt to dynamically changing domains without requiring any manual annotation. We propose an alternative for dynamic domain adaptation by leveraging real-time networks trained on various domains.

3. Method

3.1. Probabilistic framework

Similarities and differences between domains. By assumption, all domains share a measurable space (Ω, Σ), where Ω is a non-empty set (the *sample space*, or *universe*) and $\Sigma \subseteq \Omega^2$ is a σ-algebra on it (the *event space*). They also share a non-empty set of *evidences*, $\mathbb{E} \subseteq \Sigma$, and a non-empty set of *hypotheses*, $\mathbb{H} \subseteq \Sigma$. The differences between the various domains stand in the probability measures. We denote the one associated with the domain d by P_d.

Bayesian inference. This task consists in determining the posterior $P(H \mid E)$ for any given hypothesis $H \in \mathbb{H}$ and any given evidence $E \in \mathbb{E}$. In the following, entities performing this task are called *models*. In some cases, *exact models* $f_H : \mathbb{E} \to \mathbb{R} : E \mapsto P(H \mid E)$ can be established theoretically. In Section 3, we relax this constraint and assume that the models are *exact up to a target shift*, *i.e.* $f^*_H : \mathbb{E} \to \mathbb{R} : E \mapsto P^*(H \mid E)$, the probability measures P and P^* having equal likelihoods:

$$P(E \mid H) = P^*(E \mid H) \qquad \forall E \in \mathbb{E}, \forall H \in \mathbb{H}. \quad (1)$$

In case of non-zero priors, one can recover an exact model by applying a correction, called *target shift*, to the output $P^*(H \mid E)$. When $\mathbb{H}$ forms a partition of Ω, this correction [94, 199] can be written as

$$P(H \mid E) = \frac{\frac{P(H)}{P^*(H)} P^*(H \mid E)}{\sum_{H \in \mathbb{H}} \frac{P(H)}{P^*(H)} P^*(H \mid E)}. \quad (2)$$

Geometric representation. When $\mathbb{H}$ is finite and forms a partition of Ω, priors and posteriors can be represented by points in any non-degenerated Euclidean simplex in $|\mathbb{H}| - 1$ dimensions (*e.g.* a point, a segment, a triangle, a tetrahedron). After establishing a bijection between its vertices and $\mathbb{H}$, the probability relative to the hypothesis H (either the prior or the posterior) can be obtained by projecting the point on the axis that passes through the vertex H and that is orthogonal to its opposite face, the probability being equal to 0 on the face and to 1 at the vertex.

Decision-making. In any domain d, making a decision for a given evidence E comes to choose a hypothesis based on the posteriors $P_d(H \mid E)\ \forall H \in \mathbb{H}$. Together, the Bayesian inference and the decision-making form a function $\mathbb{E} \to \mathbb{H}$. We recall two standard decision-making strategies.

- MAP (*maximum a posteriori*) selects the hypothesis with the highest posterior. MAP has a discontinuity where the hypotheses are equally likely given the evidence. It maximizes the probability of making a correct decision, *i.e*. the accuracy.
- MLE (*maximum likelihood estimation*) selects the hypothesis with the highest likelihood (or, equivalently, the highest posterior to prior ratio). MLE has a discontinuity where posteriors and priors are equal, *i.e*. when the evidence provides no information about the hypotheses. It maximizes the balanced accuracy.

As shown in Figure 2, both strategies partition the Euclidean simplex into $|\mathbb{H}|$ convex parts. These two strategies are

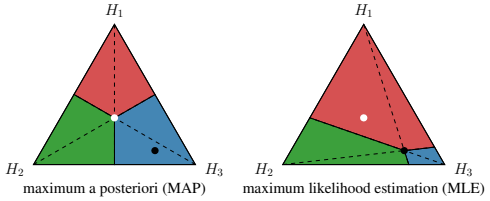


Figure 2. Comparison between two decision-making strategies: maximizing the accuracy (MAP, left) and maximizing the balanced accuracy (MLE, right). Here, the three colors represent the decisions for the three hypotheses $\mathbb{H} = \{H_1, H_2, H_3\}$ forming a partition of Ω. The black dot represents the priors, arbitrarily taken as $P(H_1) = 0.1$, $P(H_2) = 0.2$, and $P(H_3) = 0.7$.

related to each others by the fact that the balanced accuracy corresponds to the accuracy after shifting the priors to equally likely hypotheses.

3.2. Problem statement

Let us now focus on real-world problems in which there are two kinds of domains: the *source* and *target* domains, given respectively by the sets $\mathbb{D}_S$ and $\mathbb{D}_T$. In all source domains $d_S \in \mathbb{D}_S$, we assume that it is possible to obtain models computing or estimating the posteriors $P_{d_S}(H \mid E)$, possibly after gathering annotated data following the distribution $P_{d_S}(E, H)$, to the contrary of any "new" target domain $d_T \in \mathbb{D}_T \setminus \mathbb{D}_S$. The computation of $P_{d_T}(H \mid E)$ is an *unsupervised domain adaptation* problem.

We consider now the particular many-to-infinity domain adaptation problem in which the probability measure of any target domain $d_T \in \mathbb{D}_T$ can be obtained as a convex combination of the probability measures of the source domains in the non-empty set $\mathbb{D}_S = \{d_{S1}, d_{S2}, \dots d_{Sn}\}$ as follows

$$P_{d_T} = \sum_{k=1}^{n} \lambda_k P_{d_{S_k}}\,, \tag{3}$$

with $\lambda = (\lambda_1, \lambda_2 \dots \lambda_n) \in \mathbf{\Delta}^{n-1}$.[1] This problem is complex as the distribution of evidences $P(E)$, the priors $P(H)$, and the likelihoods (*a.k.a.* appearance models) $P(E \mid H)$ may vary from one domain to another.

From Equation 3, it results that the distribution of evidences in the target domain is a mixture of the distributions of evidences in the source domains: $P_{d_T}(E) = \sum_{k=1}^{n} \lambda_k P_{d_{S_k}}(E)$, $\forall E \in \mathbb{E}$. Therefore, the domain adaptation problem studied in this paper is a particular case of the more general *mixture adaptation problem* introduced by Mansour *et al*. in [200].

[1] We use the notation $\mathbf{\Delta}^{n-1}$ to denote the $(n-1)$-probabilistic simplex, that is $\lambda \in \mathbf{\Delta}^{n-1}$ if and only if $\sum_{k=1}^{n} \lambda_k = 1$ and $\lambda_k \geq 0$, $\forall k$.

By construction, we consider that the priors can be pre-computed for source domains. Consequently, the priors in the target domain can be computed by $P_{d_T}(H) = \sum_{k=1}^{n} \lambda_k P_{d_{S_k}}(H)$, $\forall H \in \mathbb{H}$. Moreover, the likelihoods in the target domain are mixtures of the corresponding likelihoods from the source domains as $P_{d_T}(E \mid H) = \sum_{k=1}^{n} \omega_{k,H} P_{d_{S_k}}(E \mid H)$ with $\omega_{k,H} = \lambda_k {}^{P_{d_{S_k}}(H)}/_{P_{d_T}(H)}$, $\forall E \in \mathbb{E}$ and $\forall H \in \mathbb{H}$. Therefore, the domain adaptation problem studied here is a sub-case of a more general one in which the target likelihoods are mixtures of fixed components (in our case, the likelihoods in the source domains), the component weights and the target priors being known only at runtime. Some theoretical and experimental results for that on-the-fly domain adaptation problem can be found in [199], in the very specific case of two-class classifiers.

3.3. Problem analysis and theoretical solution

Let us consider any evidence E such that $P_{d_T}(E) \neq 0$. From Equation 3 and using probability theory, we can derive an exact formula to compute $P_{d_T}(H \mid E)$ in any target domain $d_T \in \mathbb{D}_T$. We have, $\forall H \in \mathbb{H}$,

$$P_{d_T}(H \mid E) = \sum_{d_{S_k} \in \Phi(E)} \omega_{k,E} P_{d_{S_k}}(H \mid E) \tag{4}$$

with

$$\Phi(E) = \{d_S \in \mathbb{D}_S : P_{d_S}(E) \neq 0\} \neq \emptyset \tag{5}$$

and

$$\omega_{k,E} = \lambda_k \frac{P_{d_{S_k}}(E)}{P_{d_T}(E)} = \frac{\lambda_k P_{d_{S_k}}(E)}{\sum_{k'=1}^{n} \lambda_{k'} P_{d_{S_{k'}}}(E)}\,. \tag{6}$$

In the following, we pose $\omega_E = (\omega_{1,E}, \omega_{2,E} \dots \omega_{n,E})$. From Equation 6, we see that $\omega_E \in \mathbf{\Delta}^{n-1}$.

One recognizes in Equation 4 the *distribution weighted combining rule* proposed by Mansour *et al*. in [200]. The theoretical analysis performed in that paper shows that this combination rule behaves well when the posteriors in the source domains are affected by a bounded uncertainty.

3.3.1 Motivation for determining ω_E

In general, $P_{d_T}(H \mid E)$ cannot be determined based only on λ_k and $P_{d_{S_k}}(H \mid E)$, $\forall i \in \{1, 2, \dots n\}$. This is because, unless some λ_k are null, the vector ω_E can sweep the complete probabilistic simplex $\mathbf{\Delta}^{n-1}$. Therefore, the posteriors for the target domain can be anywhere within the convex hull of the posteriors for the source domains d_{Sk} such that $d_{Sk} \in \Phi(E)$ and $\lambda_k > 0$. Figure 3 shows this uncertainty for $|\mathbb{H}| = 3$.

Since MAP is independent of the priors, one can show, by a convexity argument, that, if the decisions are the same in all source domains, then it is also the same in the target

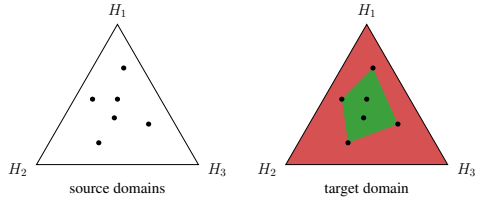


Figure 3. Given the vectors of posteriors in the source domains (black dots), there might be an important uncertainty on the vector of posteriors in the target domain (green area), as explained in Section 3.3.1. A solution is given in Section 3.3.2. In this figure, there are 6 source domains and $\mathbb{H} = \{H_1, H_2, H_3\}$.

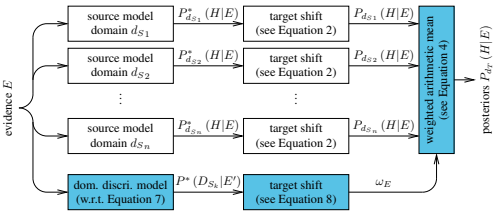


Figure 4. Our unsupervised domain adaptation algorithm. The boxes with a white background are used for determining the posteriors in the source domains. Those with a blue background depict the elements that we have added to obtain, on the fly, the posteriors in the target domain.

domain. Computing ω_E is necessary only when the decisions taken in the source domains are contradictory.

On the contrary, since MLE depends on the priors, the decision made in the target domain can be other than the decisions made in the source domains. In particular, if the decisions are the same in all source domains, then the decision in the target domain could be different.

3.3.2 Determination of ω_E

In order to establish a way of computing ω_E, we start by giving a probabilistic meaning to it. To this aim, we introduce the overall measurable space (Ω', Σ') with $\Omega' = \Omega \times \mathbb{D}_S$ and $\Sigma' = \Sigma \times 2^{\mathbb{D}_S}$. We define the event $D_{S_k} = \Omega \times \{d_{S}k\}$ for the k-th source domain and, for any event $X \in \Sigma$, we define the corresponding event $X' = X \times \mathbb{D}_S \in \Sigma'$.

All probabilities considered until now can be written in terms of a unique probability measure P on (Ω', Σ'). It is such that $P(X \times \{d_{S}k\}) = \lambda_k P_{d_{S}k}(X)$ for all $X \in \Sigma$ and for all $d_{S}k \in \mathbb{D}_S$. For example, $P_{d_{S}k}(X) = P(X' \mid D_{S_k})$ and $P_{d_T}(X) = P(X')$, $\forall X \in \Sigma$. Moreover, λ_k and ω_E acquire a probabilistic meaning as $\lambda_k = P(D_{S_k})$ and $\omega_{k,E} = P(D_{S_k} \mid E')$.

Unfortunately, P cannot be used during the off-line stage because it depends on λ. Instead, we choose to work with the following probability measure P^* on (Ω', Σ'):

$$P^*(X \times \{d_{S}k\}) = \kappa_k P_{d_{S}k}(X) \ , \tag{7}$$

$\forall X \in \Sigma$ and $\forall d_{S}k \in \mathbb{D}_S$, with any arbitrarily chosen vector of strictly positive weights $\kappa = (\kappa_1, \kappa_2 \ldots \kappa_n) \in \mathring{\mathbf{\Delta}}^{n-1}$.[2]

In the same way that we assumed the existence of a model determining $P_{d_S}(H \mid E)$ for any $E \in \mathbb{E}$ and any $H \in \mathbb{H}$, we can also assume that it is possible to obtain a model for $P^*(D_{S_k} \mid E')$, for any given $E \in \mathbb{E}$ and any given source domain. As $P^*(D_{S_k}) = \kappa_k$ and $P^*(E' \mid D_{S_k}) = P_{d_{S}k}(E)\ \forall E \in \mathbb{E}$, Equation 6 leads to

$$\omega_{k,E} = \frac{\frac{\lambda_k}{\kappa_k} P^*(D_{S_k} \mid E')}{\sum_{k'=1}^{n} \frac{\lambda_{k'}}{\kappa_{k'}} P^*\left(D_{S_{k'}} \mid E'\right)} \ . \tag{8}$$

This equation is similar to Equation 2 and can be interpreted as a prior shift for priors on the n source domains instead of priors on the $|\mathbb{H}|$ hypotheses.

3.4. Our domain adaptation algorithm

Figure 4 shows the algorithm that we propose to compute $P_{d_T}(H \mid E)$ on the fly for any hypothesis $H \in \mathbb{H}$, any evidence $E \in \mathbb{E}$, and any target domain $d_T \in \mathbb{D}_T$.

The off-line step. Two types of models are obtained off-line, before knowing the target domain: $|\mathbb{D}_S|$ *source models* and 1 *domain discriminator model*. They are used to determine $P^*_{d_{S}k}(H \mid E)$ and $P^*(D_{S_k} \mid E')$, respectively. Training the source models requires single-source labeled data. Training the domain discriminator model requires multiple-source unlabeled data. Equations 1 and 7 specify the distributions that have to be assumed by the learner.

The on-the-fly step. Equation 2 is applied to obtain $P_{d_{S}k}(H \mid E)$ from $P^*_{d_{S}k}(H \mid E)$. Moreover, knowing the target domain (*i.e.* the vector λ), Equation 8 is applied to obtain $\omega_{k,E}$ from $P^*(D_{S_k} \mid E')$. Finally, Equation 4 leads to the posterior $P_{d_T}(H \mid E)$ in the target domain. The overall algorithm gives the flexibility to adapt the posteriors on the fly to an evolving target domain.

Discussion. If all the source models and the domain discriminator model are *exact up to a target shift*, then our algorithm behaves as an *exact model* for the target domain. In practice, however, models are rarely exact (*i.e.* they have noisy outputs). Taking the L_1 distance to measure the errors, it can be shown that: if the error on ω_E (over $\mathbb{D}_S$) is bounded by ϵ_{ω_E} and if the error on $P_{d_S}(H \mid E)$ (over $\mathbb{H}$, a partition of Ω) is bounded by ϵ_{d_S} for all $d_S \in \mathbb{D}_S$, then the error on $P_{d_T}(H \mid E)$ is bounded by $\epsilon_{d_T} \leq \epsilon_{d_S} + \epsilon_{\omega_E}$.

[2] We use the notation $\mathring{\mathbf{\Delta}}^{n-1}$ to denote the interior of $\mathbf{\Delta}^{n-1}$, that is $\kappa \in \mathring{\mathbf{\Delta}}^{n-1}$ if and only if $\sum_{k=1}^{n} \kappa_k = 1$ and $\kappa_k > 0\ \forall k$.

In other words, the last step of the algorithm behaves well when facing noisy inputs. However, theoretically, the upstream target shift operations can amplify the noise affecting the outputs of the various models, especially when priors are low. Therefore, in the next section, we will demonstrate the practical suitability of this algorithm, in the challenging case in which some priors are low. Also, we validate the predicted posteriors after the target shifts.

4. Experiments

4.1. Experimental Setup

We evaluate and compare our domain adaptation algorithm on the semantic segmentation task, which consists in predicting, for each pixel of an image, the semantic class of its enclosing object or region in the scene.

In accordance with the previous section, the notations are as follows. We denote the set of images by $\mathbb{I}$, the set of pixels by $\mathbb{P}$, and the set of semantic classes by $\mathbb{S}$. All these sets being finite in our experimental setup, we take $\Omega = \mathbb{I}\times\mathbb{P}\times\mathbb{S}$ and $\Sigma = 2^{\Omega}$. The evidence $E_{i,p}$ for the pixel p in image i is $\{(i,p,s) \mid s \in \mathbb{S}\}$ and the hypothesis H_s for the groundtruth semantic class s is $\{(i,p,s) \mid i \in \mathbb{I},\, p \in \mathbb{P}\}$. Note that $\mathbb{H}$ is a partition of Ω. With our notations, the semantic segmentation task aims at choosing an element of $\mathbb{S}$, the predicted semantic class, based on the posteriors $P_d\left(H_s \mid E_{i,p}\right)$.

4.1.1 Off-line choices

Our algorithm requires two types of off-line models: (1) source models and (2) a domain discriminator model. In the following, we describe each choice for these models.

Choice of models. Various machine learning techniques can be used as posterior estimators. In general, deep learning is suitable for predicting posteriors when the loss minimized during the training stage is carefully chosen. In particular, the cross-entropy reaches a local minimum when the output of the network corresponds to the posteriors $P_d\left(H \mid E\right)\, \forall H \in \mathbb{H}$ [201, 202]. For the architecture of the source models and the domain discriminator model, we chose the TinyNet [198, 203] segmentation network, which is a lightweight architecture that only needs a few training samples and is fast to train.

Training the source models. For all our experiments, we train each source model on its own source dataset separately. The models are trained with batches of 12 images randomly sampled from the training set of the source domain. We use a learning rate of 10^{-4} with a reduce-on-plateau-scheduling strategy, a patience of 10 and a reduction factor of 0.1. We use the Adam optimizer with default parameters and no weight decay [204]. To avoid typical problems when dealing with unbalanced distributions of classes, we use the weighted cross-entropy loss, for which the weighting factor is estimated using the priors of the source domain. These choices are motivated by the wish to obtain source models *exact up to a target shift*. We aim at satisfying Equation 1 with $P^*_{d_S}\left(H_s\right) = \left|\mathbb{S}\right|^{-1}\, \forall H_s \in \mathbb{H}$. Indeed, the chosen loss achieves a local minimum for a pixel p in an image i when the source model $f^*_{sm} : \mathbb{E} \rightarrow \mathbf{\Delta}^{|\mathbb{S}|-1}$ gives $f^*_{sm}(i,p)_s = P^*_{d_S}\left(H_s \mid E_{i,p}\right)$ for all semantic classes (indexed by s here).

Training the domain discriminator model. To estimate $\omega_{k,E}$, we train a soft discriminator to recognize the corresponding domain of each pixel. We use the same TinyNet architecture, training procedure, and hyperparameters than for training the source models. Nevertheless, we replace the number of semantic classes with the number of source domains in the output layer and the batch size by 4.

Regarding training samples, we propose a way to generate multi-domain images, inspired by the mosaic transformation presented in [205]. We combine four patches cropped from four randomly drawn images to create new training images. These choices are motivated by the wish to satisfy Equation 7 with $\kappa_k = \left|\mathbb{D}_S\right|^{-1}$. Indeed, the cross-entropy loss achieves a local minimum for a pixel p in an image i when the domain discriminator model $f^*_{ddm} : \mathbb{E} \rightarrow \mathbf{\Delta}^{n-1}$ gives $f^*_{ddm}(i,p)_k = P^*\left(D_{S_k} \mid E'_{i,p}\right)$ for all source domains (indexed by k here).

Validation of predicted posteriors. The source models and the domain discriminator model are expected to be exact up to a target shift. Throughout our experiments, we systematically perform two tests to establish their trustablility. First, we compute the posteriors estimated by these models and verify (after the necessary target shifts) that these predictions correspond, in expectation, to the respective priors. Second, we verify by visual inspection that these models are well calibrated using calibration plots following [206, 207].

4.1.2 Evaluation and comparison with heuristics

For the sake of evaluation, we simulate various target domains by mixing the test sets of the different source domains. This is achieved by weighting the images. In our experiments, all images have the same size and all test sets contain the same amount of images. Thus, we weight the images of the k-th test set by λ_k to satisfy Equation 3.

We use 4 common performance scores to evaluate the quality of the semantic segmentations: the accuracy, balanced accuracy, mean IoU (macro-averaging), and balanced mean IoU. All reported results are for the decision-making strategy MAP. We compare the scores obtained with our algorithm to those obtained with the 3 following heuristics.

1. **Source models.** The first heuristic consists in using each source model separately on the target domains. For a fair comparison, we apply a target shift on the

Figure 5. Examples of images (left) from the datasets Cityscapes (above) and BDD100K (below), with the corresponding groundtruth images (right).

posteriors corresponding to the target domain priors following Equation 2.

2. **Random selection of source models.** The second heuristic randomly selects the decision derived from the k-th source model with a probability given by λ_k.

3. **Linear combination of posteriors.** The third heuristic combines linearly the posteriors provided by the source models as follows: $\sum_{k=1}^{n} \lambda_k \, P_{d_{S_k}}(H_s \mid E_{i,p})$.

4.2. Experiment with 2 source domains

In the first experiment, we consider the semantic segmentation of high-definition (1280×720) color images acquired by cameras installed in moving vehicles, behind the windshield. In this experiment, there are 2 source domains and 19 strongly imbalanced semantic classes.

Motivation. The semantic segmentation of such images is reputed to be a preliminary step towards autonomous vehicles. There is a native need for considering multiple domains as we expect differences in the appearance of roads, traffic signs, and cars from one country to another. Moreover, the sensors and their positioning can differ from one car to another. And, last but not least, the weather and traffic conditions can vary continuously. The authors of [148] noted a dramatic domain shift between two datasets (BDD100K [148] and Cityscapes [109]). According to their results, the semantic segmentation models perform much worse when tested on a different dataset.

Data. The source domains are represented by the datasets CityScapes (data acquired in European cities, mostly in Germany) and BDD100K (data acquired in the USA). These datasets gather color images with their respective semantic segmentation groundtruths (see Figure 5). We resized and cropped all images to 1280×720. We also randomly split each dataset into a training set (2726 images per dataset), a validation set (250 images per dataset), and a test set (500 images per dataset).

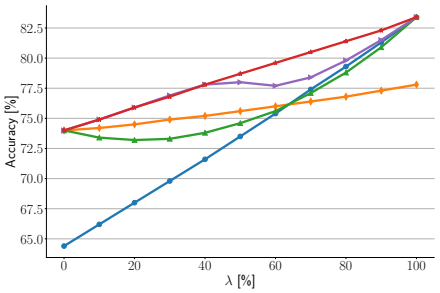


Figure 6. Results of our first experiment (all decisions made with the MAP strategy) showing the accuracy with the **source model trained on CityScapes (in blue)**, the **source model trained on BDD100K (in orange)**, the **random selection of source models (in green)**, the **linear combination of posteriors (in purple)**, and **our algorithm (in red)**. $\lambda = 0$ (resp. $= 100$) corresponds to BDD100K (resp. CityScapes) as target domain.

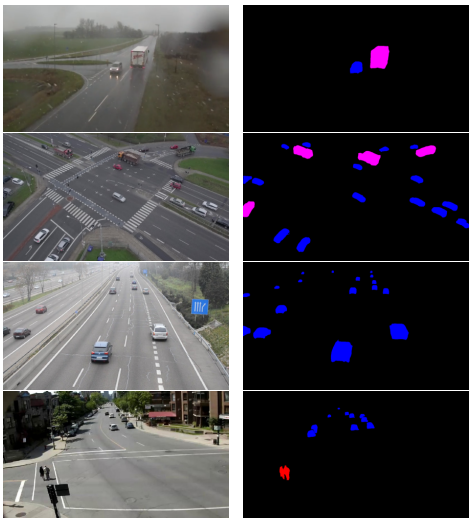


Figure 7. Examples of images (left) from the surveillance datasets (from top to bottom) RainSnow, MTID/Drone, GRAM-RTM/M-30-HD and UT/Sherbrooke, with the corresponding pseudo-groundtruth images (right).

Results. We compare the accuracy of our algorithm with the 3 heuristics in Figure 6.

The first heuristic (source models) leads to the orange (model trained on BDD100K) and blue (model trained on CityScapes) curves. When the target domain coincides with one of the source domains (both sides of the graph), the corresponding source model behaves much better than the

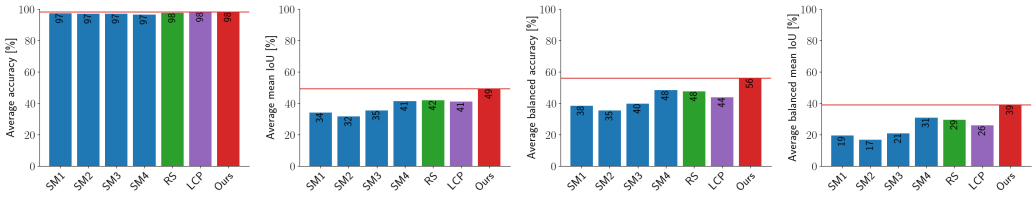


Figure 8. Results of our second experiment (all decisions made with the MAP strategy) showing the mean behaviour over 15 diversified target domains. We compare the different **source models (SM1-4, in blue)** with the **random selection of source models (RS, in green)**, **linear combination of posteriors (LCP, in purple)**, and **our algorithm (in red)**. Four performance scores are averaged over the 15 target domains: the accuracy, mean IoU, balanced accuracy, and balanced mean IoU (from left to right).

other one. This confirms the observation made in [148] and the need for domain adaptation, as simply choosing a source model and applying it directly to a different target domain leads to a drastic decrease in performance.

The second heuristic (random selection of source models, in green) is clearly suboptimal compared to the best source model for the current target domain. It is therefore not an effective way to combine the source models.

The third heuristic (linear combination of the posteriors, in purple) reaches better performances than the two previous heuristics. This is a surprise as this simple heuristic is known to be inappropriate, by theoretical arguments [200].

Our domain adaptation algorithm (red curve) outperforms all heuristics, especially when the weight of CityScapes is above 40% in the target domain mixture.

4.3. Experiment with 4 source domains

Let us consider a second semantic segmentation task, with images acquired by fixed video surveillance traffic cameras. In this experiment, there are 4 source domains and 4 strongly imbalanced semantic classes.

Data. The four source domains are represented by the following datasets. (1) AAU RainSnow Traffic Surveillance Dataset [208] (called *RainSnow*) contains traffic surveillance videos in rainfall and snowfall from seven different intersections. (2) Multi-view Traffic Intersection Dataset (MTID) [209] (called *MTID/Drone*) contains footage of the same intersection from two different points of view. We only use the images recorded with a drone from one point of view. (3) GRAM Road-Traffic Monitoring (GRAM-RTM) dataset [210] (called *GRAM-RTM/M-30-HD*) consists of three videos recorded under different conditions and with different platforms. We only use one video, named M-30-HD. (4) The dataset from the Urban Tracker [211] project (called *UT/Sherbrooke*) contains also several videos. We only select the Sherbrooke video, filmed at the Sherbrooke/Amherst intersection in Montreal.

For each dataset, we randomly select 1299 images for the training set, 300 images for the validation set and 300 images for the test set. All images have been resized (*e.g.* by upsampling, downsampling, or cropping) to 1280×720.

Since no segmentation groundtruth is available for all of these datasets and since we need all datasets to share the same semantic classes, we use the PointRend [135] algorithm trained on the COCO [191] dataset to obtain pseudo-groundtruths. For this experiment, we consider 4 semantic classes: background, person, two wheels (bicycle and motorcycle), and four wheels (car, bus, and truck). An example of an image and its pseudo-groundtruth for each dataset is illustrated in Figure 7.

Performance analysis. Since we have 4 source domains in this experiment, it is impossible to depict the performances as we did in the first experiment. Instead, Figure 8 shows the mean behaviour over several target domains. Bar plots are provided for 4 performance scores. They are obtained by averaging the scores of our algorithm and of the heuristics over 15 target domains obtained by mixing either 1, 2, 3, or 4 source domains with equal weights. As can be seen, on average, our algorithm outperforms all heuristics, with a large margin for three of the four scores. This demonstrates the superiority of our algorithm.

5. Conclusion

In this paper, we focus on the domain adaptation problem in which the probability measure of the target domain is a convex combination of the probability measures of the source domains. We define a probabilistic framework and show that the posteriors in the target domain do not depend solely on posteriors, priors, and relative weights of the source domains. From there, we provide a theoretical proof of an algorithm to compute the posteriors for the target domain in an unsupervised way. This is particularly valuable when the target domain can change on the fly. Interestingly, this flexibility is achieved by keeping the annotated data collected in different source domains separate and by sharing only unlabeled data. Finally, we test our unsupervised domain adaptation algorithm on a semantic segmentation task in real-world surveillance, and show its superiority compared to common heuristics.

Acknowledgements

This work has been made possible thanks to the *TRAIL* initiative (`https://trail.ac`). Part of it was supported by the Walloon region (Service Public de Wallonie Recherche, Belgium) under grant n°2010235 – *ARIAC by DIGITALWALLONIA4.AI*. Anthony Cioppa is funded by the FNRS (`https://www.frs-fnrs.be/en/`).

References

[171] A. Farahani, S. Voghoei, K. Rasheed, and H. R. Arabnia, "A brief review of domain adaptation," *Adv. Data Sci. Inf. Eng.*, pp. 877–894, 2021.

[172] X. Liu, C. Yoo, F. Xing, H. Oh, G. El Fakhri, J.-W. Kang, and J. Woo, "Deep unsupervised domain adaptation: A review of recent advances and perspectives," *APSIPA Trans. Signal Inf. Process.*, vol. 11, no. 1, pp. 1–51, Aug. 2022.

[94] T. Sipka, M. Sulc, and J. Matas, "The hitchhiker's guide to prior-shift adaptation," in *IEEE Winter Conf. Appl. Comput. Vis. (WACV)*. Waikoloa, HI, USA: IEEE, Jan. 2022, pp. 2031–2039.

[173] M. Kirchmeyer, A. Rakotomamonjy, E. de Bezenac, and P. Gallinari, "Mapping conditional distributions for domain adaptation under generalized target shift," in *Int. Conf. Learn. Represent. (ICLR)*, Virtual conference, Sept. 2022, pp. 1–22.

[174] I. Redko, E. Morvant, A. Habrard, M. Sebban, and Y. Bennani, *Advances in domain adaptation theory*. Elsevier, 2019.

[175] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Int. Conf. Mach. Learn. (ICML)*, ser. Proc. Mach. Learn. Res., vol. 37. Lille, France: ML Res. Press, Jul. 2015, pp. 1180–1189.

[176] K. Saito, K. Watanabe, Y. Ushiku, and T. Harada, "Maximum classifier discrepancy for unsupervised domain adaptation," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*. Salt Lake City, UT, USA: IEEE, Jun. 2018, pp. 3723–3732.

[177] J. Liang, D. Hu, and J. Feng, "Do we really need to access the source data? Source hypothesis transfer for unsupervised domain adaptation," in *Int. Conf. Mach. Learn. (ICML)*, ser. Proc. Mach. Learn. Res., vol. 119. ML Res. Press, Jul. 2020, pp. 6028–6039.

[178] D. Wang, E. Shelhamer, S. Liu, B. A. Olshausen, and T. Darrell, "Tent: Fully test-time adaptation by entropy minimization," in *Int. Conf. Learn. Represent. (ICLR)*. Virtual conference: Int. Conf. Learn. Represent. (ICLR), May 2021, pp. 1–15.

[179] M. Boudiaf, R. Mueller, I. B. Ayed, and L. Bertinetto, "Parameter-free online test-time adaptation," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*. New Orleans, LA, USA: Inst. Electr. Electron. Eng. (IEEE), Jun. 2022, pp. 8334–8343.

[180] Y. Zhang, P. David, and B. Gong, "Curriculum domain adaptation for semantic segmentation of urban scenes," in *IEEE Int. Conf. Comput. Vis. (ICCV)*. Venice, Italy: IEEE, Oct. 2017, pp. 2039–2049.

[181] Y. Li, L. Yuan, and N. Vasconcelos, "Bidirectional learning for domain adaptation of semantic segmentation," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*. Long Beach, CA, USA: IEEE, Jun. 2019, pp. 6929–6938.

[182] S. Zhao, B. Li, X. Yue, Y. Gu, P. Xu, R. Hu, H. Chai, and K. Keutzer, "Multi-source domain adaptation for semantic segmentation," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 32. Vancouver, Can.: Curran Assoc. Inc., Dec. 2019, pp. 1–14.

[183] S. Zhao, B. Li, P. Xu, X. Yue, G. Ding, and K. Keutzer, "MADAN: Multi-source adversarial domain aggregation network for domain adaptation," *Int. J. Comput. Vis.*, vol. 129, no. 8, pp. 2399–2424, May 2021.

[184] X. Peng, Q. Bai, X. Xia, Z. Huang, K. Saenko, and B. Wang, "Moment matching for multi-source domain adaptation," in *IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*. Seoul, South Korea: IEEE, Oct. 2019, pp. 1406–1415.

[185] J. He, X. Jia, S. Chen, and J. Liu, "Multi-source domain adaptation with collaborative learning for semantic segmentation," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*. Nashville, TN, USA: IEEE, Jun. 2021, pp. 11 003–11 012.

[186] S.-L. Sun and H.-L. Shi, "Bayesian multi-source domain adaptation," in *Int. Conf. Mach. Learn. Cybern. (ICMLC)*. Tianjin, China: IEEE, Jul. 2013, pp. 24–28.

[187] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, and J. Garcia-Rodriguez, "A review on deep learning techniques applied to semantic segmentation," *arXiv*, vol. abs/1704.06857, Apr. 2017.

[188] X. Liu, Z. Deng, and Y. Yang, "Recent progress in semantic image segmentation," *Artificial Intelligence Review*, vol. 52, pp. 1089–1106, Jun. 2018.

[189] S. Minaee, Y. Y. Boykov, F. Porikli, A. J. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 7, pp. 3523–3542, Jul. 2022.

[110] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ADE20K dataset," in *IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*. Honolulu, HI, USA: Inst. Electr. Electron. Eng. (IEEE), Jul. 2017, pp. 5122–5130.

[190] B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, and A. Torralba, "Semantic understanding of scenes through the ADE20K dataset," *Int. J. Comput. Vis.*, vol. 127, no. 3, pp. 302–321, Dec. 2018.

[111] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Jun. 2010.

[191] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common Objects in Context," in *Eur. Conf. Comput. Vis. (ECCV)*, ser. Lect. Notes Comput. Sci., vol. 8693. Springer, Sept. 2014, pp. 740–755.

[109] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*. Las Vegas, NV, USA: Inst. Electr. Electron. Eng. (IEEE), Jun. 2016, pp. 3213–3223.

[116] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*. Honolulu, HI, USA: Inst. Electr. Electron. Eng. (IEEE), Jul. 2017, pp. 6230–6239.

[192] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," in *IEEE Int. Conf. Comput. Vis. (ICCV)*. Venice, Italy: IEEE, Oct. 2017, pp. 2980–2988.

[135] A. Kirillov, Y. Wu, K. He, and R. Girshick, "PointRend: Image segmentation as rendering," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*. Seattle, WA, USA: Inst. Electr. Electron. Eng. (IEEE), Jun. 2020, pp. 9796–9805.

[138] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. Torr, and L. Zhang, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*. Nashville, TN, USA: Inst. Electr. Electron. Eng. (IEEE), Jun. 2021, pp. 6877–6886.

[193] W. Zhou, J. Liu, J. Lei, L. Yu, and J.-N. Hwang, "GMNet: Graded-feature multilabel-learning network for RGB-thermal urban scene semantic segmentation," *IEEE Trans. Image Process.*, vol. 30, pp. 7790–7802, 2021.

[194] T. Cane and J. Ferryman, "Evaluating deep semantic segmentation networks for object detection in maritime surveillance," in *IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*. Auckland, New Zealand: IEEE, Nov. 2018, pp. 1–6.

[195] N. A. Muhadi, A. F. Abdullah, S. K. Bejo, M. R. Mahadi, and A. Mijic, "Deep learning semantic segmentation for water level estimation using surveillance camera," *Appl. Sci.*, vol. 11, no. 20, pp. 1–25, Oct. 2021.

[196] S. Zheng and G. Gupta, "Semantic-guided zero-shot learning for low-light image/video enhancement," in *IEEE/CVF Winter Conf. Appl. Comput. Vis. Work. (WACVW)*. Waikoloa, HI, USA: IEEE, Jan. 2022, pp. 581–590.

[197] A. Aakerberg, A. S. Johansen, K. Nasrollahi, and T. B. Moeslund, "Semantic segmentation guided real-world super-resolution," in *IEEE/CVF Winter Conf. Appl. Comput. Vis. Work. (WACVW)*. Waikoloa, HI, USA: IEEE, Jan. 2022, pp. 449–458.

[198] A. Cioppa, A. Deliège, M. Istasse, C. De Vleeschouwer, and M. Van Droogenbroeck, "ARTHuS: Adaptive real-time human segmentation in sports through online distillation," in *IEEE Int. Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW), CVsports*. Long Beach, CA, USA: Inst. Electr. Electron. Eng. (IEEE), Jun. 2019, pp. 2505–2514.

[199] S. Piérard, A. Marcos Alvarez, A. Lejeune, and M. Van Droogenbroeck, "On-the-fly domain adaptation of binary classifiers," in *Proceedings of the BENELEARN Conference*. Brussels, Belgium: Belgian-Dutch Conference on Machine Learning (BENELEARN), Jun. 2014, pp. 20–28.

[200] Y. Mansour, M. Mohri, and R. Rostamizadeh, "Domain adaptation with multiple sources," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 21. Vancouver, Canada: Curran Assoc. Inc., Dec. 2008, pp. 1041–1048.

[201] J. W. Miller, R. Goodman, and P. Smyth, "On loss functions which minimize to conditional expected values and posterior probabilities," *IEEE Trans. Inf. Theory*, vol. 39, no. 4, pp. 1404–1408, Jul. 1993.

[202] D. Ramos, J. Franco-Pedroso, A. Lozano-Diez, and J. Gonzalez-Rodriguez, "Deconstructing cross-entropy for probabilistic binary classifiers," *Entropy*, vol. 20, no. 3, pp. 1–20, Mar. 2018.

[203] A. Cioppa, A. Deliège, and M. Van Droogenbroeck, "A bottom-up approach based on semantics for the interpretation of the main camera stream in soccer games," in *IEEE Int. Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW), CVsports*, Salt Lake City, UT, USA, Jun. 2018, pp. 1846–1855.

[204] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Int. Conf. Learn. Represent. (ICLR)*, San Diego, CA, USA, May 2015, pp. 1–15.

[205] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," *arXiv*, vol. abs/2004.10934, 2020.

[206] A. Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," in *Int. Conf. Mach. Learn. (ICML)*. Bonn, Germany: ACM Press, 2005, pp. 625–632.

[207] 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," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Nov. 2011.

[148] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell, "BDD100K: A diverse driving dataset for heterogeneous multitask learning," in *IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*. Seattle, WA, USA: Inst. Electr. Electron. Eng. (IEEE), Jun. 2020, pp. 2633–2642.

[208] C. H. Bahnsen and T. B. Moeslund, "Rain removal in traffic surveillance: Does it matter?" *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 8, pp. 2802–2819, Aug. 2019.

[209] M. B. Jensen, A. Mogelmose, and T. B. Moeslund, "Presenting the multi-view traffic intersection dataset (MTID): A detailed traffic-surveillance dataset," in *IEEE Int. Conf. Intell. Transp. Syst. (ITSC)*. Rhodes, Greece: IEEE, Sept. 2020, pp. 1–6.

[210] R. Guerrero-Gómez-Olmedo, R. J. López-Sastre, S. Maldonado-Bascón, and A. Fernández-Caballero, "Vehicle tracking by simultaneous detection and viewpoint estimation," in *Natural and Artificial Computation in Engineering and Medical Applications*, ser. Lect. Notes Comput. Sci., vol. 7931. Springer Berl. Heidelb., 2013, pp. 306–316.

[211] J.-P. Jodoin, G.-A. Bilodeau, and N. Saunier, "Urban tracker: Multiple object tracking in urban mixed traffic," in *IEEE Winter Conf. Appl. Comput. Vis. (WACV)*. Steamboat Springs, Colorado, USA: IEEE, Mar. 2014, pp. 885–892.

ART. 8

9. Domain characterization (9th article)

This chapter reproduces the following article [9]:

> A. Halin*, S. Piérard*, R. Vandeghen, B. Gérin, M. Zanella, M. Colot, J. Held, A. Cioppa, E. Jean, G. Bontempi, S. Mahmoudi, B. Macq, and M. Van Droogenbroeck, "Physically interpretable probabilistic domain characterization," in *Asian Conference on Computer Vision Workshops (ACCV Workshops)*, vol. 15482 of *Lecture Notes in Computer Science*, (Hanoi, Vietnam), pp. 17–35, Springer Nature Singapore, 2025.

It is focused on the notion of ***domain***. This article results from a collaboration initiated on the occasion of the TRAIL summer workshop 2024. I came up with the idea for this research project, for which I was the team leader, and I focused mainly on the theoretical aspects as well as on the analysis of the results.

Context

In the previous chapter, we saw that specifying the relationship between an applicative domain (*i.e.*, the target domain) and the chosen reference domains (*i.e.*, the source domains) is useful for determining how models optimized for the reference domains can be combined to obtain a model optimized for the applicative domain. We saw how to implement this in practice in cases where the probability measure associated with the applicative domain is a mixture of the probability measures associated with the reference domains, even when the weights of the mixture can change in real time. However, we assumed that these weights were known.

Motivation

Let us consider a (semi-)autonomous vehicle to illustrate current needs. It is useful to keep track of the events that just happened (the *what*) and of the reasons that led to what happened (the *why*). For example, the EDR (Event Data Recorder) and DSSAS (Data Storage System for Automated Driving) are standard types of recording systems respectively for the *what* and for the *why*. As for the why, given that the optimal decisions depend on the domain, it is clear that a (semi-)autonomous vehicle must be able to store the knowledge it has of the domain in which it operates, in a form that is understandable to a human expert. From my point of view, a distribution of physical parameters is a suitable form.

Furthermore, it also seems clear to me that we should be able to explain how the vehicle adapts its decisions to the current domain and to justify this adaptation. From my point of view, the weights given to probability measures in the mixture are easily interpretable, as is the probabilistic reasoning used in the mixture domain adaptation (MDA). This part has already been discussed in Chapter 8.

Finally, the conditions under which a (semi-)autonomous vehicle is capable of driving must be specified. These are the conditions for which its operation and safety are guaranteed. They form the ODD (Operational Design Domain) [212]. Using the terminology used in this thesis, it must be possible to determine if the current domain is part of the ODD. Therefore, the ODD is a set of domains. In the case of MDA, if a domain is part of the ODD, then the probability measure associated with that domain is a convex combination of those associated with the reference domains.

Main contributions

In this chapter, we explore the possibility for a system to become aware of the domain in which it operates by characterizing it as a joint distribution of physical parameters. We also explore the possibility of automatically determining whether the applicative domain is possibly a mixture of the chosen reference domains, as well as estimating the weights of this mixture. All of this is made possible thanks to methods that encode distributions (*e.g.*, normalizing flows) and methods that estimate them (*e.g.*, simulation-based inference).

Perspectives

The perspectives given in Chapter 8 include a software architecture for which I discussed the adaptability, modularity, and reliability. Combined with the techniques presented in this article, its reliability improves and some interpretability is gained.

- Regarding reliability, it is better to do nothing if the system is incapable of doing it safely. In this sense, being able to detect when the applicative domain is definitely not part of the ODD is very useful. A (semi-)autonomous vehicle could detect that it is operating in a domain that it does not support, warn its passengers, or automatically park on the side of the road. This greatly improves the reliability of the driving process.
- Regarding the interpretability, the vehicle can now record a description of what is perceived about the domain in which it operates, and provide it to human experts as needed. Moreover, one can easily understand how the mixture weights used for the adaptation are computed based on what is perceived.

Physically Interpretable Probabilistic Domain Characterization

Anaïs Halin[⋆1], Sébastien Piérard[⋆1], Renaud Vandeghen[1], Benoît Gérin[2], Maxime Zanella[2,4], Martin Colot[3], Jan Held[1], Anthony Cioppa[1], Emmanuel Jean[5], Gianluca Bontempi[3], Saïd Mahmoudi[4], Benoît Macq[2], and Marc Van Droogenbroeck[1]

[1] Montefiore Institute, University of Liège (ULiège), Liège, Belgium
[2] Catholic University of Louvain (UCLouvain), Louvain-la-Neuve, Belgium
[3] Université Libre de Bruxelles (ULB), Brussels, Belgium
[4] University of Mons (UMons), Mons, Belgium
[5] Multitel research & innovation centre, Mons, Belgium

Abstract. Characterizing domains is essential for models analyzing dynamic environments, as it allows them to adapt to evolving conditions or to hand the task over to backup systems when facing conditions outside their operational domain. Existing solutions typically characterize a domain by solving a regression or classification problem, which limits their applicability as they only provide a limited summarized description of the domain. In this paper, we present a novel approach to domain characterization by characterizing domains as probability distributions. Particularly, we develop a method to predict the likelihood of different weather conditions from images captured by vehicle-mounted cameras by estimating distributions of physical parameters using normalizing flows. To validate our proposed approach, we conduct experiments within the context of autonomous vehicles, focusing on predicting the distribution of weather parameters to characterize the operational domain. This domain is characterized by physical parameters (absolute characterization) and arbitrarily predefined domains (relative characterization). Finally, we evaluate whether a system can safely operate in a target domain by comparing it to multiple source domains where safety has already been established. This approach holds significant potential, as accurate weather prediction and effective domain adaptation are crucial for autonomous systems to adjust to dynamic environmental conditions.

Keywords: Domain Characterization · Distribution Prediction · Normalizing Flows · Simulation-Based Inference · Domain Adaptation · Weather · Autonomous Vehicles · Operational Design Domain

1 Introduction

Advances in computer vision allow widespread camera monitoring, but diverse weather conditions lead to visually different data, sometimes impacting the per-

⋆ Equal contributions.

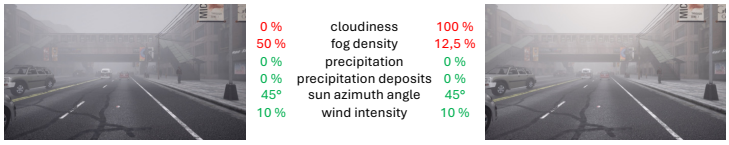


Fig. 1: The two synthetic images, A and C, generated using the *CARLA* software, appear almost identical, despite being acquired under very different weather conditions. This highlights the challenge of information loss from sensors like cameras when dealing with weather-related physical parameters. As a result, predictions that diverge from the ground truth in such ambiguous cases should not be penalized during evaluation.

formance of high-level tasks like object detection and surveillance. Current solutions often lack generalizability across multi-weather scenarios, highlighting the need for adaptive methods that can process visual data under diverse conditions.

More specifically, weather conditions significantly affect the perception capabilities of autonomous driving systems [213], particularly under harsh conditions, such as heavy rain or fog, which can compromise their ability to operate safely. Therefore, it is essential to develop reliable approaches to analyze the environment, irrespective of the weather conditions. Detecting critical circumstances such as extreme weather events allow the system to respond appropriately within and outside its *Operational Design Domain* (*ODD*), defined by SAE International [212] as the "operating conditions under which a given driving automation system, or feature thereof, is specifically designed to function, including, but not limited to, environmental, geographical, and time-of-day restrictions, and/or the requisite presence or absence of certain traffic or roadway characteristics".

However, predicting exact weather conditions is challenging due to the many ambiguous cases. A single observation of the environment (*e.g.*, recorded by a vehicle-mounted camera) could be characterized by multiple sets of weather parameters. Fig. 1 presents two synthetic images, generated using the *CARLA* software [214], that appear nearly identical but were captured under different weather conditions. This shows the ambiguities inherent to characterizing the domain based on a single observation (*i.e.*, image).

Current methods predict weather conditions through regression or classification [215–217]. However, those approaches only produce a single crisp answer, lacking the insight about the ambiguities of characterizing the weather conditions. Fig. 2 illustrates how the two scenarios from Fig. 1, when processed with an intermediate set of values, result in very different visual representations.

In this work, we propose a novel solution based on a probabilistic characterization of the weather. Historically, statistical inference relying on likelihood estimation was intractable when dealing with high dimensional data [156]. Recently, one class of density estimation techniques based on neural networks called normalizing flows [218–220] has become quite popular to solve this kind of problem.

Fig. 2: Considering the images A and C from Fig. 1, generating an image B for the arithmetically averaged parameters, leads to an image very different from A and C. In other words, there are images i such that the probability $P(I = i, W = \hat{w}(i))$ of the pair (image, estimated weather) is zero when the estimated weather is the expected value of the weather knowing the image, $\hat{w}(i) = E[W|I = i]$, as done in regression. Working with distributions of weather parameters (as shown using a contour plot on the right-hand side) as proposed in this work, rather than predicting specific values for each parameter, *i.e.*, regression (as shown on the left-hand side), avoids this problem.

These models learn invertible transformations to go from complex distributions to more handy ones, and can therefore be used for modeling weather parameters from highly complex image and weather distributions.

These novel techniques allow us to express the domain of an image by a probabilistic distribution, *e.g.*, weather conditions, compared to current deterministic approaches. More specifically, we show on three consecutive tasks how various weather conditions can be predicted based (1) on a single color image acquired in front of a vehicle, (2) on a bag of color images (*absolute characterization*), and (3) how the current domain is related to arbitrarily chosen source domains (*relative characterization*).

Notions similar to *ODD* are very common in other fields where equivalent terms are used, such as *Operational Envelope* for maritime and *Operational Context* for railroad [221]. Additionally, other fields are facing domain shifts, such as in medical imagery from different acquisition devices [222]. We argue that our approach could serve a large range of practical applications. For example, determining the most suitable model for a given scenario within a fleet of lightweight AI models able to analyze the environment, as proposed in previous works [8, 200, 223], detecting significant domain transitions to collect new data for domain adaptation methods that rely on buffers or adaptable internal statistics [178, 224–228] or activating adaptation mechanisms based on clustering [179, 229–231].

We summarize our contributions as follows: (1) We propose a novel probabilistic methodology to characterize domains in the case of autonomous vehicles driving in various weather conditions. (2) We demonstrate that simulation-based inference (normalizing flows) is adequate to obtain distribution for weather parameters that are used to characterize the domain and compare different back-

bones for features extraction. (3) Based on this weather domain characterization, we show how to characterize a new target domain as a mixture of source models.

2 Related Work

2.1 Predicting Weather Conditions from Images

Prediction of weather conditions from images was first formulated as a single-label classification task (*e.g.*, sunny, cloudy, or foggy). In 2014, Lu *et al.* [232] proposed a binary classification task between sunny and cloudy weathers, using features extracted from visual cues such as the sky, shadows, reflections, contrast, and haze. Later, Guerra *et al.* [233] proposed a multi-class dataset extending the scope of the classification task to rain, snow, and fog. Recent works focus on optimizing *Convolutional Neural Network* (*CNN*) to obtain strong features for common and uncommon weather conditions [234, 235]. However, weather conditions can hardly be represented by crisp classification due to its continuous nature.

To reach a more realistic description of intricate relations between weather conditions, recent methods simultaneously regress several physical parameters [215, 216]. To increase interpretability, Li *et al.* [217] also assign cues of weather characteristics to each pixel. However, these methods still lack the ability to represent ambiguous scenarios (see Fig. 2). In this work, we predict the joint distribution of weather parameters by proposing a novel method based on normalizing flows.

2.2 Handling Weather Conditions for Autonomous Driving

Autonomous car driving systems need to be efficient under all weather conditions. Some methods propose to integrate a generalization step to the model to remove the environmental influences on the acquired images, by introducing a style layer inspired by images style-transfer neural networks [236] or through adversarial training [237]. Generalized features obtained from alignment of source and target domains tend to be suboptimal, as they do not consider the task. To improve the estimation, a solution is to add a domain adaptation step [238]. Jeon *et al.* [239] further improved their estimation by adding an unsupervised domain adaptation step after domain alignment.

Many studies have also highlighted the importance of a strong *ODD* definition to properly assess the ability of automatic driving systems to work in given conditions regarding weather, location, other vehicles on the road, state of the car sensors and many other environmental parameters [240–242]. Many strategies have been proposed to evaluate different situations according to specific evaluations of potential damage cost [243, 244] and define the boundaries of the *ODD*. In this work, we characterize the domain by a probability distribution focusing on weather parameters in autonomous driving environment.

2.3 Probabilistic Modeling of Parameters Distributions

We propose to leverage recent observations in the *Simulation-Based Inference* (*SBI*) literature [156,245] for domain characterization. This literature has seen a rapid expansion thanks to new density estimation techniques in problems where likelihood estimation was often intractable, especially for high dimensional data. One class of these density estimation techniques based on neural networks is normalizing flows [219]. The principle consists in transforming an arbitrarily chosen distribution (*e.g.*, a Gaussian) into the desired distribution. Different types exist, such as the *Neural Spline Flow* (*NSF*) type [246] that can be used in two ways: (1) to determine the value of the *Probability Density Function* (*PDF*) at a given point and (2) to draw samples at random. Different techniques can be used to learn them, *e.g.* the *Neural Posterior Estimation* (*NPE*) technique [247, 248].

There are a few techniques to analyze the performance of models predicting distributions. *Coverage Plots* [249] show, objectively and quantitatively, whether the distribution prediction models are underconfident (*i.e.*, conservative), calibrated, or overconfident. Another technique, specific for parameter distributions (*e.g.*, weather parameters) predicted from an observation (*e.g.*, an image) and widespread in the field of *SBI*, is known as *Posterior Predictive Check* (*PPC*) [250]. It consists in drawing parameters at random from a predicted distribution, injecting these parameters into a simulator or physical system and comparing the resulting observations with the one from which the distribution was predicted. In this work, we leverage those analysis techniques for assessing the quality of our weather characterization models.

3 The Three Fundamental Tasks Behind the Physically Interpretable Probabilistic Domain Characterization

Our experiments are organized around three different tasks, involving a prediction of the distribution of weather conditions given some images acquired by color cameras placed in front of vehicles. Before elaborating on these tasks in Sec. 3.2, 3.3, and 3.4, we briefly introduce our framework in Sec. 3.1.

3.1 Framework

Mathematical Modeling. We denote the set of all possible values for the physical parameters (*e.g.* weather conditions) by $\mathbb{W}$ and the set of all observations of interest (*e.g.* images) by $\mathbb{I}$. We adopt the probability theory of Kolmogorov [32, 33] and consider a measurable space (Ω, Σ) as well as the (generalized) random variables $W : \Omega \to \mathbb{W}$ for the physical parameters and $I : \Omega \to \mathbb{I}$ for the observation. Following the mathematical modeling introduced in [8], we consider the set $\mathbb{D}_{(\Omega,\Sigma)}$ of domains d in which there is a probability measure P_d on (Ω, Σ). We see the *ODD* of a given autonomous system as the set of domains in which it can be used safely, no matter if this has been established by design or by testing. Thus, $ODD \subseteq \mathbb{D}_{(\Omega,\Sigma)}$.

Table 1: In *CARLA*, the weather is controlled through 13 physical parameters. This table shows the range of values that we consider in our experiments and indicates, for each of them, if they are considered in our predictions.

parameter	range	predicted
`cloudiness`	0 to 100%	yes
`fog density`	0 to 100%	yes
`precipitation`	0 to 100%	yes
`sun azimuth angle`	0° to 360°	no
`sun altitude angle`	−90° to 90°	yes
`wind intensity`	0 to 100%	yes
`precipitation deposits`	0 to 100%	yes

parameter	range	predicted
`fog distance`	fixed to 0.75	no
`fog falloff`	fixed to 0.1	no
`mie scattering scale`	fixed to 0.03	no
`rayleigh scattering scale`	fixed to 0.033	no
`scattering intensity`	fixed to 1.0	no
`wetness`	fixed to 0.0	no

Data. All our experiments are performed on data generated using the *CARLA* software, an open-source simulator for autonomous driving research [214]. All images are acquired by a simulated camera placed in front of a vehicle called *ego vehicle*. The weather is controlled through 13 real-valued parameters (see Table 1). There are 6 parameters for which we make predictions. Thus, $\mathbb{W} \subseteq \mathbb{R}^6$.

3.2 Task I: Predicting Distributions of Physical Parameters

The first task (see Fig. 3) consists in predicting how likely various physical parameters are, jointly. As these parameters cannot, in general, be measured directly, it is necessary to estimate the likelihood, *i.e.* the distribution of plausible values, based on an indirect observation.

Case Study. We study here the particular case in which the physical parameters are relative to the weather conditions and the observations are color images $i \in \mathbb{I}$ acquired by cameras placed in front of vehicles. We aim at learning, offline, a deep learning model $i \mapsto \hat{P}_{d_R}(W|I = i)$, where d_R denotes a domain of reference

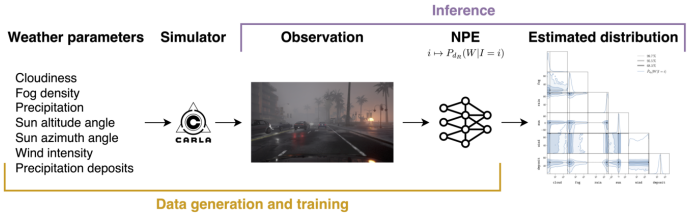


Fig. 3: Task I. The aim of the first task is, based on an image, to predict the joint distribution of the weather parameters. For this purpose, (1) we generate data using the *CARLA* software for uniformly distributed weather parameters (offline), (2) we train a *NPE* model using the learning set (LS) of our generated data (offline), and (3) we infer, given an image from the test set (TS), the estimated weather distribution (normalizing flow) and show the result on a corner plot.

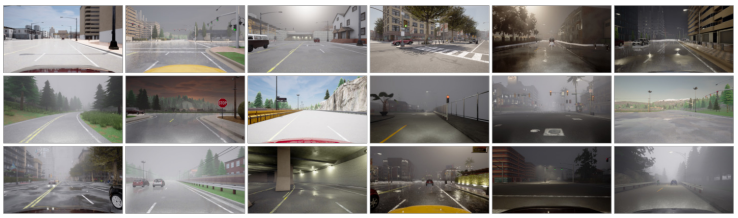


Fig. 4: Excerpt of the images in our dataset generated with the *CARLA* simulator. The ground-truth weather parameters are drawn at random for each image, following a uniform distribution with the bounds given in Table 1.

in which the probability measure on (Ω, Σ) is P_{d_R}. This domain is arbitrarily chosen in such a way that P_{d_R} has a large support.

Data. We use *CARLA* to generate a dataset with 635k images and the corresponding ground-truth values for the weather parameters. The dataset (Fig. 4) is split into a learning set (*LS*) with 600k images (500k for the training set and 100k for the validation set) and a test set (*TS*) with 35k images. Letting the model of the ego vehicle, the map, the number of pedestrians, and the number of vehicles vary brings a touch of diversity in our data.

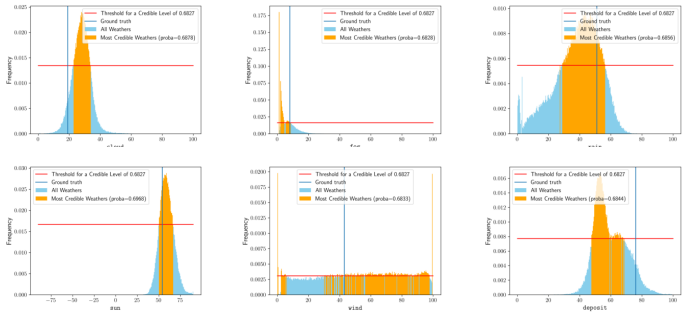
Method. We consider three different (frozen) backbones to extract features from the input images: *ResNet-50* [251], *DINOv2* [252], and *CLIP* [253]. We use the libraries *LAMPE* [254] and *ZUKO* [220] to learn a model (of type *NPE*) and to manipulate the normalizing flows (of type *NSF*), respectively. All predicted weather distributions are posteriors relative to the weather priors in the *LS*. Note that *LAMPE* and *ZUKO* were not developed for domain characterization, but rather for *SBI*. Also, to the best of our knowledge, these libraries have only been used once with very high dimensional input data [245].

Evaluation and Results. Five different analyses are carried out.

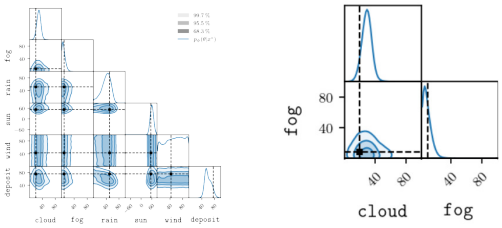
1st analysis: histograms (Fig. 5b). Due to their physical meaning, the weather parameters are easily interpretable. This paves the way to a first, subjective, evaluation. We draw samples at random out of the predicted weather distributions $\hat{P}_{d_R}(W|I = i)$, for some images i arbitrarily chosen in *TS* and conduct a visual inspection of the histograms for the 6 marginals. The most credible weather parameters (*a.k.a.* highest density credibility sets, highest density regions, plausible sets, *etc.*) are highlighted for a credible level $l = 68.27\%$. For any input image i, these weathers are those such that the predicted *PDF* is above some threshold $t(i)$ and the predicted probability of the set is l [255].

weather parameter	ground-truth value
`cloudiness (cloud)`	19
`fog density (fog)`	8
`precipitation (rain)`	51
`sun azimuth angle`	90
`sun altitude angle (sun)`	54
`wind intensity (wind)`	43
`precipitation deposits (deposit)`	76

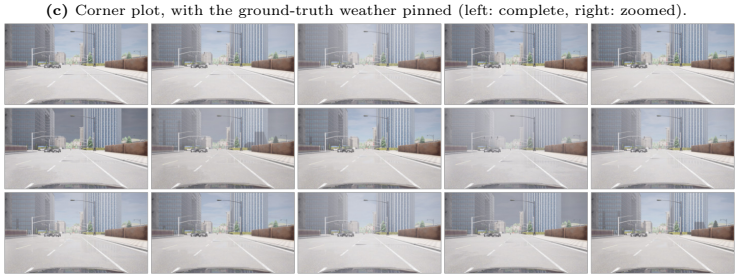
(a) An input image i and the corresponding ground-truth values.



(b) Histograms for the 6 marginals of the predicted weather distribution for i.



(c) Corner plot, with the ground-truth weather pinned (left: complete, right: zoomed).



(d) Posterior predictive check.

Fig. 5: Task I: results obtained, with the model learned with the *ResNet-50* backbone on 500k learning samples, for an arbitrarily chosen input image in the *TS*.

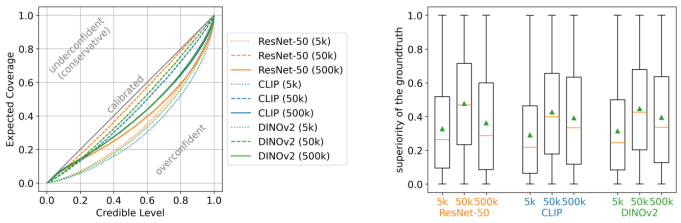


Fig. 6: Task I: comparison of our 9 models on the overall *TS*. Left: coverage plot. Right: box-and-whisker plots for the π statistic. Both analyses agree that the model learned with the *ResNet-50* backbone on 50k learning samples is preferable.

2nd analysis: corner plots (Fig. 5c). A corner plot is a triangular array of plots. Those on the diagonal show the n marginals of a n-D distribution. Those below the diagonal depict the highest density credibility regions delimited at some arbitrarily chosen credibility levels (in this paper: 68.27%, 95.45%, and 99.73%), for each pair of parameters. We observe on the corner plots wide distributions, meaning that there is a large uncertainty for the weather parameters given an image. However, when analyzing the results, this uncertainty is explainable and the following analysis shows that this uncertainty is not excessive, *i.e.*, our models are not underconfident.

3rd analysis: posterior predictive checks (Fig. 5d). Once a weather distribution is predicted for an observation (input image), it is possible to (1) draw weather parameters vectors at random from it and then (2) to inject these vectors in *CARLA*, keeping all other parameters unchanged and immobilizing the vehicles and pedestrians, to obtain new images in order to finally (3) compare those with the initial input image. One cannot expect to obtain identical images as, in *CARLA*, the traffic lights continue to run, the rain continues to fall, the plants move with the wind, and pedestrians' poses are not perfectly frozen. Putting this aside, we observe that most retrieved images are very similar to the input image. We conclude that our models are not underconfident.

4th analysis: coverage plots (Fig. 6). Coverage plots show how the expected coverage varies with the credible level [249]. By definition, the expected coverage, at the credible level $l \in [0, 1]$, is the probability that the ground-truth weather belongs to the most credible weathers at level l. A method estimating the distribution of weathers based on an image is said underconfident (*i.e.*, conservative), calibrated, or overconfident at level l when the expected coverage at level l is, respectively, $> l$, $= l$, or $< l$. We observe that our models are all overconfident. The best calibrated model is the one that we obtained with 50k learning samples and *ResNet-50* as backbone.

5th analysis: superiority of the ground truth (Fig. 6). We also determine to what extent the ground truth is more credible than the other weathers. This is the proportion π of weathers that can be drawn at random from the normalizing flow and that are predicted as having a *PDF* value (*i.e.*, likelihood) lower or equal than the one from the ground truth. The higher π is, the better it is, but approaching 1.0 is notably very challenging. We estimated π for each image of the *TS* and made box-and-whisker plots for the 9 models. This analysis is complementary to the coverage plot in the sense that a model can be perfectly calibrated while still presenting room for improvement. That being said, we are in the particular case in which this analysis leads to the same conclusion as the coverage plot: the model based on the backbone *ResNet-50* and learned from 50k samples is preferable to the others.

3.3 Task II: Absolute Domain Characterization

The second task (see Fig. 7) consists in characterizing a domain in an easy-to-interpret way. For this, we opt for a distribution of physical parameters estimated based on a bag (*a.k.a.* multiset) of observations.

Case Study. We characterize a domain $d \in \mathbb{D}_{(\Omega,\Sigma)}$ by the estimated distribution $\hat{P}_d(W)$ of weather conditions based on a real-valued bag b (multiset) of arbitrarily weighted images acquired by cameras placed in front of vehicles. In the following, we denote the weight (multiplicity) of the image $i \in b$ by $\omega(i)$. The images can either originate from a unique vehicle or from several, in case of vehicle-to-vehicle (*V2V*) communications. We want to establish $b \mapsto \hat{P}_d(W|B = b)$.

Data. We consider a bag b of 1,000 equally weighted images generated in maps already used in Task I. The weather parameters have been drawn at random

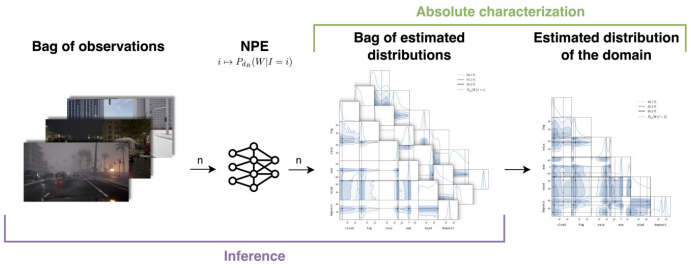


Fig. 7: Task II. The aim of the second task is to obtain an absolute characterization of a domain of interest. We process with the *NPE* each observation of a bag of observations of the domain of interest to obtain the bag of estimated distributions corresponding to the observations, then we obtain the distribution of weather parameters for the domain of interest by averaging the individual distributions of the bag.

Fig. 8: Task II: example of result. Left: corner plot showing an arbitrarily chosen distribution of weather conditions, $P_d(W)$ (ground truth). Center: a bag b of images generated using *CARLA* with weather conditions drawn from $P_d(W)$. Right: corner plot showing the estimated likelihood of the weather conditions $\hat{P}_d(W)$ based on b.

from an arbitrary distribution $P_d(W)$ for which we fixed all parameters but two, `fog density` and `precipitation`, that follow a mixture of Gaussians.

Method. We use the *NPE* model developed for task I and build our solution for task II on top of it. We implement the following estimator for $P_d(W)$: $\hat{P}_d(W) = \sum_{i \in b} \omega(i) \hat{P}_{d_R}(W|I = i)$. The motivation for this estimator is threefold. (1) It is straightforward to evaluate the probability density function of $\hat{P}_d(W)$ and to draw weather conditions $w \in \mathbb{W}$ at random, following $\hat{P}_d(W)$, as we can do it with the normalizing flow $P_{d_R}(W|I = i)$. This will be valuable in our third task. (2) This estimator is useful for linear temporal filtering, *e.g.*, when one wants to weight more the recently acquired images than the old ones. (3) Finally, this estimator is fully justifiable under the assumptions that $P_d(W|I) = P_{d_R}(W|I)$ and $\omega(i) = P_d(I = i)$ as, in this case, $P_d(W) = \int_i \omega(i) P_{d_R}(W|I = i)di$.

Evaluation and Results. Our result is shown in Fig. 8. We observe that the three modes of the ground-truth weather distribution $P_d(W)$ are within the highest density regions of the prediction $\hat{P}_d(W)$. We made the same observation for many other ground-truth distributions (results omitted due to the limited space). We stress the fact that $\hat{P}_d(W)$ differs significantly from $P_d(W)$ can be explained by the inherent loss of information resulting from the use of color cameras, as already discussed in the introduction (*cf*. Fig. 1).

3.4 Task III: Relative Domain Characterization

The third task (see Fig. 9) consists in characterizing any target domain $d_T \in \mathbb{D}_{(\Omega,\Sigma)}$, relatively, w.r.t. to some arbitrarily chosen source domains d_{S_1}, d_{S_2}, $\ldots d_{S_{n_S}} \in \mathbb{D}_{(\Omega,\Sigma)}$. Our motivation for this task originates from the importance of knowing if a system implementing a given many-to-one domain adaptation method can operate safely in the target domain d_T when it is known to operate safely in the source domains $d_{S_1}, d_{S_2}, \ldots d_{S_{n_S}}$.

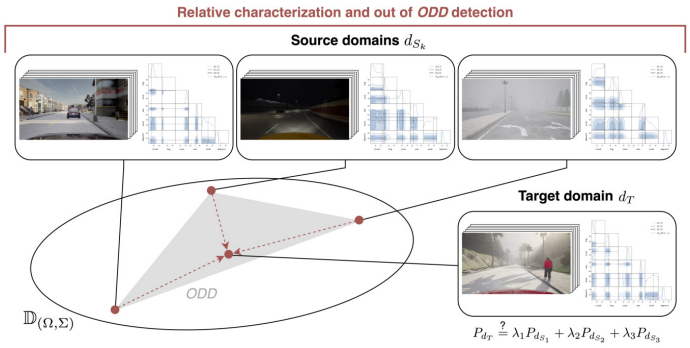


Fig. 9: Task III. The aim of the third task is twofold: (1) having a relative characterization of a target domain d_T based on source domains d_{S_k} and (2) detecting when this target domain d_T is out of the *ODD*.

Case Study. We discuss the case of *Mixture Domain Adaptation* (*MDA*) in which the system adapts to any target domain for which $P_{d_T} = \sum_{k=1}^{n_S} \lambda_k P_{d_{S_k}}$, with $\sum_{k=1}^{n_S} \lambda_k = 1$ and $\lambda_k \geq 0 \, \forall k$ (this is the *mixture assumption*). In this framework, the *ODD* is the convex hull of $\{P_{d_{S_1}}, P_{d_{S_2}}, \dots P_{d_{S_{n_S}}}\}$. An autonomous car implementing *MDA* is expected to drive safely when $P_{d_T} \in ODD$. So, our goal in this third task is to determine if the mixture assumption holds and if so, what the values of the mixture weights λ_k are.

MDA supports many potential applications in the high-level tasks of the *Sense* pillar of the *Sense-Plan-Act* model [256]. While Mansour studied it generically [200], the application to the two-class classification task has been studied in [199], and the application to the semantic segmentation task of images acquired by vehicle-mounted cameras has been studied in [8]. These latter two works put an important emphasis on on-the-fly applicability and provide mathematically proven exact solutions. However, a critical limitation of these works is the need for the mixture weights to be known at adaptation time. Here, we remove this limitation by introducing a method that determines these weights automatically.

Data. We consider 4 subsets of the *TS* that we created for Task I: b_0, b_1, b_2, and b_3, containing 176, 27, 21, and 26 images, respectively. In the bag b_k, the distribution of the ground-truth weather parameters follows a uniform distribution on $\mathbb{W}_k \subsetneq \mathbb{W}$. These sets are such that $\mathbb{W}_0 \cap (\mathbb{W}_1 \cup \mathbb{W}_2 \cup \mathbb{W}_3) = \emptyset$.

Method. We characterize the source and target domains with the method presented for Task II, using the same weather distribution predictive model for all

domains. We define the mean squared gap between the target domain and the mixture of the source domains as:

$$\delta(\hat{\lambda}_1, \dots \hat{\lambda}_{n_S}) = \int_{\mathbb{W}} \left[\hat{P}_{d_T}(W = w) - \sum_{k=1}^{n_S} \hat{\lambda}_k \hat{P}_{d_{S_k}}(W = w)\right]^2 \hat{P}_{d_T}(W = w) dw$$
$$\simeq \frac{1}{n_W} \sum_i^{n_W} \left[\hat{P}_{d_T}(W = w_i) - \sum_{k=1}^{n_S} \hat{\lambda}_k \hat{P}_{d_{S_k}}(W = w_i)\right]^2 \quad (1)$$

with $\{w_i\}_{i=1}^{n_W} \sim \hat{P}_{d_T}(W)$. We aim at finding the values of $\hat{\lambda}_1, \dots \hat{\lambda}_{n_S}$ that minimize $\delta(\hat{\lambda}_1, \dots \hat{\lambda}_{n_S})$. This is a constrained least squares problem in which the constraints are $\sum_{k=1}^{n_S} \hat{\lambda}_k = 1$ and $\hat{\lambda}_k \geq 0 \, \forall k$. In order to use the *CVXPY* library [257, 258] to solve our problem, we converted our original problem into a convex quadratic programming problem, with the same constraints.

Evaluation and Results. The goal of this experiment is twofold.

1. To show that it is possible to recover the mixture weights needed for the *MDA* technique presented in [8]. If the probability measure in the target domain is a mixture of the probability measures in the source domains, then we expect our characterization of the target domain to be a mixture of our characterizations for the source domains, with the same weights, as $P_{d_T} = \sum_{k=1}^{n_S} \lambda_k P_{d_{S_k}} \Rightarrow P_{d_T}(W) = \sum_{k=1}^{n_S} \lambda_k P_{d_{S_k}}(W)$.
2. To show that it is possible to detect when the target domain is not a mixture of the source domains, which means that the target domain is out of the *ODD* for the *MDA* technique presented in [8]. If our characterization of the target domain is not a mixture of our characterizations for the source domains, then we expect that the probability measure in the target domain is not a mixture of the probability measures in the source domains, as we have $P_{d_T}(W) \neq \sum_{k=1}^{n_S} \lambda_k P_{d_{S_k}}(W) \Rightarrow P_{d_T} \neq \sum_{k=1}^{n_S} \lambda_k P_{d_{S_k}}$.

To achieve this goal, we perform experiments in which we characterize target domains relatively to $n_S = 3$ source domains. However, to study the robustness of our method, we let the probability measures in the target domains d_T be mixtures not only of the probability measures in the source domains, but also in some domains that are out of the *ODD*. We use the bags b_1, b_2, and b_3 for the source domains d_{S_1}, d_{S_2}, and d_{S_3}, respectively. We also use the bag b_0 for domains $d_{\notin ODD}$ that are, by construction, guaranteed to be out of the *ODD*. Putting this into equations, we have $P_{d_T} = (1 - \eta) P_{d_{\in ODD}} + \eta P_{d_{\notin ODD}}$ with $P_{d_{\in ODD}} = \lambda_1 P_{d_{S_1}} + \lambda_2 P_{d_{S_2}} + \lambda_3 P_{d_{S_3}}$. The quantity $\eta \in [0, 1]$ is interpreted as a proportion of noise, which is the keystone to assess the robustness of our method.

In each experiment, the images in b_1, b_2, and b_3 are equally weighted, whereas those in b_0 are randomly weighted following a uniform distribution. The method introduced in Task II for the absolute characterization of domains is applied for d_{S_1}, d_{S_2}, d_{S_3}, and d_T. We arbitrarily chose $(\lambda_1, \lambda_2, \lambda_3) = (0.2, 0.3, 0.5)$. The mixture weights are optimized on $n_W = 16$ points.

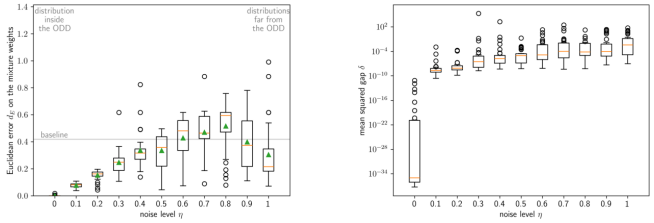


Fig. 10: Task III: results obtained when the noise level η sweeps the $[0, 1]$ interval. Left: box-and-whisker plots showing the distributions of d_E. Right: box-and-whisker plots showing the distributions of $\delta(\hat{\lambda}_1, \dots \hat{\lambda}_{n_S})$. These plots show (1) that we can recover the relationship between the target domain and the source domains, under the mixture assumption, when the target domain belongs to the *ODD* and (2) that the $\delta(\hat{\lambda}_1, \dots \hat{\lambda}_{n_S})$ can be used to determine when the target domain belongs to the *ODD*.

The overall experiment consists in (1) choosing a target domain at random, as described here-above, (2) computing its absolute characterization with the method of Task II, (3) executing the weight estimation algorithm, and (4) reporting both the mean square gap $\delta(\hat{\lambda}_1, \dots \hat{\lambda}_{n_S})$ and the Euclidean distance d_E between $(\hat{\lambda}_1, \hat{\lambda}_2, \hat{\lambda}_3)$ and $(\lambda_1, \lambda_2, \lambda_3)$. This experiment has been performed 330 times (30 times for 11 values of η). The results are shown in Fig. 10. The mean square gap achievable by chance is shown (baseline). Note that the target domain belongs to the *ODD* only when $\eta = 0$. We observe that d_E is negligible when $d_T \in$ *ODD* and that the mean square gap is effective to detect when $d_T \notin$ *ODD*.

4 Conclusion

In this work, we present a novel approach to characterize domains by estimating the distribution of physical parameters. Our method enhances interpretability, facilitates domain adaptation, and provides safeguards for systems operating outside their *ODD*. Our experiments, organized around three tasks, are performed in the particular, but very important, case of autonomous vehicles and demonstrate how to obtain an absolute characterization of a domain by predicting a distribution of weather parameters (1) using a single image acquired by a vehicle-mounted camera and (2) using a bag of images. Our experiments also demonstrate (3) how to obtain a relative characterization of a target domain based on arbitrarily chosen source domains. Our approach includes two types of domain characterization: absolute and relative. The relative characterization is particularly valuable for domain adaptation, allowing the expression of a target domain in terms of source domains and verifying whether the current domain is part of the *ODD*. This is important for autonomous driving systems as well as for other fields requiring interpretable and trustworthy domain adaptation.

Acknowledgements

This work has been made possible thanks to the *TRAIL* initiative (https://trail.ac). Part of it was supported by the Walloon region (Service Public de Wallonie Recherche, Belgium) under grant n°2010235 (ARIAC by DIGITALWALLONIA4.AI). J. Held and A. Cioppa are funded by the F.R.S.-FNRS (https://www.frs-fnrs.be/en/).

References

258. Agrawal, A., Verschueren, R., Diamond, S., Boyd, S.: A rewriting system for convex optimization problems. Journal of Control and Decision **5**(1), 42–60 (January 2018). https://doi.org/10.1080/23307706.2017.1397554
179. Boudiaf, M., Mueller, R., Ayed, I.B., Bertinetto, L.: Parameter-free online test-time adaptation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8334–8343 (June 2022), https://doi.org/10.1109/CVPR52688.2022.00816
242. Colwell, I., Phan, B., Saleem, S., Salay, R., Czarnecki, K.: An automated vehicle safety concept based on runtime restriction of the operational design domain. In: IEEE Intelligent Vehicles Symposium. pp. 1910–1917 (June 2018), https://doi.org/10.1109/IVS.2018.8500530
156. Cranmer, K., Brehmer, J., Louppe, G.: The frontier of simulation-based inference. Proceedings of the National Academy of Sciences (PNAS) **117**(48), 30055–30062 (May 2020). https://doi.org/10.1073/pnas.1912789117
257. Diamond, S., Boyd, S.: CVXPY: A Python-embedded modeling language for convex optimization. Journal of Machine Learning Research **17**(83), 1–5 (2016)
214. Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., Koltun, V.: CARLA: An open urban driving simulator. In: Annual Conference on Robot Learning. Proceedings of Machine Learning Research, vol. 78, pp. 1–16 (November 2017), https://proceedings.mlr.press/v78/dosovitskiy17a.html
246. Durkan, C., Bekasov, A., Murray, I., Papamakarios, G.: Neural spline flows. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 32, pp. 1–12 (December 2019), https://papers.nips.cc/paper_files/paper/2019/hash/7ac71d433f282034e088473244df8c02-Abstract.html
223. Gérin, B., Halin, A., Cioppa, A., Henry, M., Ghanem, B., Macq, B., De Vleeschouwer, C., Van Droogenbroeck, M.: Multi-stream cellular test-time adaptation of real-time models evolving in dynamic environments. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). vol. 33, pp. 4472–4482 (June 2024), https://doi.org/10.1109/CVPRW63382.2024.00450
222. Gérin, B., Zanella, M., Wynen, M., Mahmoudi, S., Macq, B., De Vleeschouwer, C.: Exploring viability of test-time training: Application to 3D segmentation in multiple sclerosis. In: IEEE Conference on Artificial Intelligence (CAI). vol. 34, pp. 557–562 (June 2024), https://doi.org/10.1109/CAI59869.2024.00110
225. Gong, T., Jeong, J., Kim, T., Kim, Y., Shin, J., Lee, S.J.: NOTE: Robust continual test-time adaptation against temporal correlation. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 35, pp. 27253–27266 (November 2022), https://openreview.net/forum?id=E9HNxrCFZPV

248. Greenberg, D., Nonnenmacher, M., Macke, J.H.: Automatic posterior transformation for likelihood-free inference. In: International Conference on Machine Learning (ICML). Proceedings of Machine Learning Research, vol. 97, pp. 2404–2414 (2019), https://proceedings.mlr.press/v97/greenberg19a.html
241. Gyllenhammar, M., Johansson, R., Warg, F., Chen, D., Heyn, H.M., Sanfridson, M., Söderberg, J., Thorsén, A., Ursing, S.: Towards an operational design domain that supports the safety argumentation of an automated driving system. In: European congress on embedded real time systems (ERTS). pp. 1–10 (January 2020)
251. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (June 2016), https://doi.org/10.1109/CVPR.2016.90
249. Hermans, J., Delaunoy, A., Rozet, F., Wehenkel, A., Begy, V., Louppe, G.: A crisis in simulation-based inference? beware, your posterior approximations can be unfaithful. Transactions on Machine Learning Research **11**, 2835–8856 (November 2022)
228. Houyon, J., Cioppa, A., Ghunaim, Y., Alfarra, M., Halin, A., Henry, M., Ghanem, B., Van Droogenbroeck, M.: Online distillation with continual learning for cyclic domain shifts. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). vol. abs 2211 16234, pp. 2437–2446 (June 2023), https://doi.org/10.1109/CVPRW59228.2023.00242
255. Hyndman, R.J.: Computing and graphing highest density regions. The American Statistician **50**(2), 120–126 (May 1996). https://doi.org/10.1080/00031305.1996.10474359
215. Ibrahim, M., Haworth, J., Cheng, T.: WeatherNet: Recognising weather and visual conditions from street-level images using deep residual learning. ISPRS International Journal of Geo-Information **8**(12), 1–18 (November 2019). https://doi.org/10.3390/ijgi8120549
216. Introvigne, M., Ramazzina, A., Walz, S., Scheuble, D., Bijelic, M.: Real-time environment condition classification for autonomous vehicles. In: IEEE Intelligent Vehicles Symposium (IV). pp. 1527–1533 (June 2024), https://doi.org/10.1109/iv55156.2024.10588692
239. Jeon, M., Seo, J., Min, J.: DA-RAW: Domain adaptive object detection for real-world adverse weather conditions. In: IEEE International Conference on Robotics and Automation (ICRA). vol. 17, pp. 2013–2020 (May 2024), https://doi.org/10.1109/ICRA57147.2024.10611219
218. Kobyzev, I., Prince, S.J.D., Brubaker, M.A.: Normalizing flows: An introduction and review of current methods. IEEE Transactions on Pattern Analysis and Machine Intelligence **43**(11), 3964–3979 (November 2021). https://doi.org/10.1109/tpami.2020.2992934
32. Kolmogorov, A.N.: Grundbegriffe der Wahrscheinlichkeitsrechnung, Ergebnisse der Mathematik und Ihrer Grenzgebiete, vol. 2. Springer Berlin Heidelberg, Berlin, Germany (1933). https://doi.org/10.1007/978-3-642-49888-6, https://doi.org/10.1007/978-3-642-49888-6
33. Kolmogorov, A.N.: Foundations of the theory of probability. Chelsea Publishing Company, New York City, New York, USA (1950), https://archive.org/details/foundationsofthe00kolm
244. Lee, C.W., Nayeer, N., Garcia, D.E., Agrawal, A., Liu, B.: Identifying the operational design domain for an automated driving system through assessed risk. In: IEEE Intelligent Vehicles Symposium. vol. IV, pp. 1317–1322 (October 2020), https://doi.org/10.1109/IV47402.2020.9304552

238. Lee, S., Kim, D., Kim, N., Jeong, S.G.: Drop to adapt: Learning discriminative features for unsupervised domain adaptation. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 91–100 (October 2019), https://doi.org/10.1109/ICCV.2019.00018
237. Li, J., Xu, R., Ma, J., Zou, Q., Ma, J., Yu, H.: Domain adaptive object detection for autonomous driving under foggy weather. In: IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 612–622 (January 2023), https://doi.org/10.1109/WACV56688.2023.00068
217. Li, X., Wang, Z., Lu, X.: A multi-task framework for weather recognition. In: ACM International Conference on Multimedia (MM). pp. 1318–1326 (October 2017), https://doi.org/10.1145/3123266.3123382
234. Li, Z., Li, Y., Zhong, J., Chen, Y.: Multi-class weather classification based on multi-feature weighted fusion method. In: Earth and Environmental Science. Journal of Physics: Conference Series, vol. 558 (August 2020), https://doi.org/10.1088/1755-1315/558/4/042038
232. Lu, C., Lin, D., Jia, J., Tang, C.K.: Two-class weather classification. IEEE Transactions on Pattern Analysis and Machine Intelligence **39**(12), 2510–2524 (December 2017). https://doi.org/10.1109/tpami.2016.2640295
247. Lueckmann, J., Gonçalves, P.J., Bassetto, G., Öcal, K., Nonnenmacher, M., Macke, J.H.: Flexible statistical inference for mechanistic models of neural dynamics. In: Advances in Neural Information Processing Systems (NeurIPS). pp. 1289–1299 (2017), https://papers.nips.cc/paper/6728-flexible-statistical-inference-for-mechanistic-models-of-neural-dynamics
200. Mansour, Y., Mohri, M., Rostamizadeh, R.: Domain adaptation with multiple sources. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 21, pp. 1041–1048 (December 2008), https://papers.nips.cc/paper_files/paper/2008/hash/0e65972dce68dad4d52d063967f0a705-Abstract.html
252. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning robust visual features without supervision. Transactions on Machine Learning Research **1**, 1–32 (2024)
219. Papamakarios, G., Nalisnick, E., Rezende, D.J., Mohamed, S., Lakshminarayanan, B.: Normalizing flows for probabilistic modeling and inference. Journal of Machine Learning Research **22**(57), 1–64 (2021)
240. Pappalardo, G., Caponetto, R., Varrica, R., Cafiso, S.: Assessing the operational design domain of lane support system for automated vehicles in different weather and road conditions. Journal of Traffic and Transportation Engineering **9**(4), 631–644 (August 2022). https://doi.org/10.1016/j.jtte.2021.12.002
213. Perrels, A., Votsis, A., Nurmi, V., Pilli-Sihvola, K.: Weather conditions, weather information and car crashes. ISPRS International Journal of Geo-Information **4**(4), 1–23 (November 2015). https://doi.org/10.3390/ijgi4042681
8. Piérard, S., Cioppa, A., Halin, A., Vandeghen, R., Zanella, M., Macq, B., Mahmoudi, S., Van Droogenbroeck, M.: Mixture domain adaptation to improve semantic segmentation in real-world surveillance. In: IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW). pp. 22–31 (January 2023), https://doi.org/10.1109/WACVW58289.2023.00007
199. Piérard, S., Marcos Alvarez, A., Lejeune, A., Van Droogenbroeck, M.: On-the-fly domain adaptation of binary classifiers. In: Proceedings of the BENELEARN Conference. pp. 20–28 (June 2014)

253. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning (ICML). Proceedings of Machine Learning Research, vol. 139, pp. 8748–8763 (July 2021), https://proceedings.mlr.press/v139/radford21a.html
236. Rebut, J., Bursuc, A., Perez, P.: StyleLess layer: Improving robustness for real-world driving. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 8992–8999 (September 2021), https://doi.org/10.1109/IROS51168.2021.9636204
250. Robertson, C., Long, J.A., Nathoo, F.S., Nelson, T.A., Plouffe, C.C.F.: Assessing quality of spatial models using the structural similarity index and posterior predictive checks. Geographical Analysis **46**(1), 53–74 (January 2014). https://doi.org/10.1111/gean.12028
220. Rozet, F., Divo, F., Schnake, S.: Zuko: Normalizing flows in PyTorch. Software (2022). https://doi.org/10.5281/ZENODO.7625672, https://pypi.org/project/zuko
254. Rozet, F., Miller, B.K., Delaunoy, A.: LAMPE: Likelihood-free amortized posterior estimation. Software (2021). https://doi.org/10.5281/ZENODO.8405782, https://pypi.org/project/lampe
212. SAE International: Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles. Tech. Rep. SAE Standard J3016 202104, Society of Automobile Engineers, Warrendale, PA, USA (April 2021). https://doi.org/10.4271/J3016_202104, https://doi.org/10.4271/J3016_202104
256. Shoker, A., Yasmin, R., Esteves-Verissimo, P.: WIP: Savvy: Trustworthy autonomous vehicles architecture. In: Symposium on Vehicle Security Privacy (VehicleSec). pp. 1–7 (Feb 2024), https://doi.org/10.14722/vehiclesec.2024.23058
243. Sun, C., Deng, Z., Chu, W., Li, S., Cao, D.: Acclimatizing the operational design domain for autonomous driving systems. IEEE Intelligent Transportation Systems Magazine **14**(2), 10–24 (March 2022). https://doi.org/10.1109/mits.2021.3070651
229. Tang, H., Chen, K., Jia, K.: Unsupervised domain adaptation via structurally regularized deep clustering. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8722–8732 (June 2020), https://doi.org/10.1109/CVPR42600.2020.00875
221. Tonk, A., Boussif, A.: Operational design domain or operational envelope; seeking a suitable concept for autonomous railway systems. In: Proceedings of the European Safety and Reliability Conference (ESREL). pp. 2104–2111 (August 2022), https://doi.org/10.3850/978-981-18-5183-4_S06-08-245-cd
245. Vasist, M., Rozet, F., Absil, O., Mollière, P., Nasedkin, E., Louppe, G.: Neural posterior estimation for exoplanetary atmospheric retrieval. Astronomy & Astrophysics **672**, 1–10– (April 2023). https://doi.org/10.1051/0004-6361/202245263
233. Villarreal Guerra, J.C., Khanam, Z., Ehsan, S., Stolkin, R., McDonald-Maier, K.: Weather classification: A new multi-class dataset, data augmentation approach and comprehensive evaluations of convolutional neural networks. In: NASA/ESA Conference on Adaptive Hardware and Systems (AHS). pp. 305–310 (August 2018), https://doi.org/10.1109/AHS.2018.8541482

178. Wang, D., Shelhamer, E., Liu, S., Olshausen, B.A., Darrell, T.: Tent: Fully test-time adaptation by entropy minimization. In: International Conference on Learning Representations (ICLR). pp. 1–15 (May 2021), https://openreview.net/forum?id=uXl3bZLkr3c
224. Wang, Q., Fink, O., Van Gool, L., Dai, D.: Continual test-time domain adaptation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7191–7201 (June 2022), https://doi.org/10.1109/CVPR52688.2022.00706
227. Wang, W., Zhong, Z., Wang, W., Chen, X., Ling, C., Wang, B., Sebe, N.: Dynamically instance-guided adaptation: A backward-free approach for test-time domain adaptive semantic segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 24090–24099 (June 2023), https://doi.org/10.1109/CVPR52729.2023.02307
226. Yuan, L., Xie, B., Li, S.: Robust test-time adaptation in dynamic scenarios. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15922–15932 (June 2023), https://doi.org/10.1109/CVPR52729.2023.01528
230. Zanella, M., Ayed, I.B.: On the test-time zero-shot generalization of vision-language models: Do we really need prompt learning? In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 23783–23793 (June 2024), https://doi.org/10.1109/CVPR52733.2024.02245
231. Zanella, M., Gérin, B., Ayed, I.B.: Boosting vision-language models with transduction. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 37, pp. 62223–62256 (December 2024)
235. Zhang, Y., Sun, J., Chen, M., Wang, Q., Yuan, Y., Ma, R.: Multi-weather classification using evolutionary algorithm on EfficientNet. In: Proceedings of the IEEE International Conference on Pervasive Computing and Communications Workshops and Other Affiliated Events (PerCom Workshops). pp. 546–551 (March 2021), https://doi.org/10.1109/PerComWorkshops51409.2021.9430939

ART. 9

Conclusion

This thesis addresses the ***evaluation***, ***comparison***, and ***adaptation*** of entities that could be algorithms, devices, methods, models, procedures, strategies, systems, solutions, etc. Through nine articles, it shows evidence that a probabilistic approach is well suited to evaluate entities and define their ***performances***, compare entities and establish meaningful ***rankings***, and adapt entities to applicative ***domains***. Although a particular focus is placed on theory, most results are illustrated using concrete examples encountered in computer vision. Tasks similar to classification and segmentation have been extensively studied. In this case, the entities are classifiers (as considered in Chapters 3 to 5), change detection algorithms (as for the case studies of Chapters 1 and 7), or semantic segmentation methods (as for the case studies of Chapter 8), and the domains are represented by sets of images, videos, or categories of videos. Nevertheless, none of the theoretical results given in this thesis is peculiar to computer vision.

Summary of my contributions

Contributions to the evaluation of entities and to the definition of their performances

To compare entities that have the potential to be used in various domains, we need to get view of an entity's reliability. Until now, the standard way to obtain it was to average performance scores for some given set of domains. These scores were chosen arbitrarily, and the choices could differ from one study to another, leading to inconsistent results. For example, in classification, users of the ROC space naturally tend to average the false positive rate and the true positive rate, while users of the PR space naturally tend to average the precision and the recall. Unfortunately, these two approaches lead to different conclusions (*e.g.*, when it comes to determining a mean across domains). The biases are obvious, and the conclusions are difficult to interpret safely.

Motivated by this observation, I began by establishing the need to consider the performance as a mathematical object rather than as a list of numbers ("measures" or score values). I have shown it in the case of pixel-based two-class classification applied to videos. The expected relationships between scores (*e.g.*, F_1 is the harmonic mean of the precision and recall) are broken when one averages, score by score, the values obtained on different videos. In other words, averaging score values for different domains leads to inconsistent values. This problem is solved by viewing the performance as a mathematical object, defining a way to average these objects through a convex combination (the summarization), considering scores as real functions of performance, and calculating the values taken by the scores on the average performance rather than the averages of the

values taken by the scores on all performances.

The topic of evaluation has been a recurring theme throughout this thesis. With probabilistic reasoning, we know that it is very important to specify the random experiment to avoid issues highlighted by Bertrand's paradox [28]. I propose to see the evaluation as a (thought) random experiment involving the entity to evaluate and the performance as the resulting distribution of outcomes. When several entities are evaluated with the same experiment, their performances are comparable. The outcomes of the evaluation experiment are such that one could specify how satisfied he/she is with each of them. The *satisfaction* is a random variable that describes the task. Another random variable, positive, can be used to describe the applicative preferences with the *importance* given to each possible outcome. It is useful to study the properties of the evaluation experiments. For example, if the evaluated entity is used only once when the experiment is realized, then the experiment has a *linear* behavior: the performance of a hybrid entity selecting one of the combined entities at random before running it is the mixture of the performances corresponding to the combined entities, with the mixture weights corresponding to the probabilities of selection. It is also useful to study transformations of the evaluation experiments. Combining them non-deterministically leads to the *summarization* (*cf* Chapter 1). Restarting them non-deterministically based on what the outcome is leads to the *filtering* operation and to the *ranking scores* (*cf* Chapter 2).

Contributions to the comparison of entities and to their performance-based rankings

To the best of my knowledge, until now, there was no general mathematical theory for the performance-based ranking of entities. In Chapter 2, I developed the first one on top of both the probability and order theories: (1) I introduced a rigorous mathematical framework, (2) I proposed an axiomatic definition of the performance-based ranking of entities, and (3) I gave a generic family of scores that can be used to induce performance orderings satisfying my axioms: the *ranking scores* R_I. They form a continuum and are parameterized by the importance I. In the particular case of two-class crisp classification, this family encompasses well-known performance scores, including the accuracy, the true positive rate (*a.k.a.* recall, sensitivity), the true negative rate (*a.k.a.* specificity), the positive predictive value (*a.k.a.* precision), and the F-scores F_β.

My approach is generic in the sense that this theory can be used with any type of task, as long as one agrees with the probabilistic interpretation of the evaluation. My approach is also integrated in the sense that the rankings are related to the performances P, to the applicative preferences (with the importance I), and to the task (with the satisfaction S).

The manifold of performance orderings induced by the ranking scores R_I is introduced in Chapter 3 and is a cornerstone for several results presented in this thesis. It turns out that the distances between orderings are closely related to rank correlations (Kendall's τ along the manifold and Spearman's ρ in straight lines). These distances depend on the considered distribution of performances. I have motivated the definition of the optimal trade-off between some ranking scores as the Karcher-Fréchet mean (*i.e.*, the ranking

score that minimizes the Fréchet variance) relative to the Kendall distance (*a.k.a.* bubble-sort distance). I performed an extensive study of what the optimal trade-off between the precision and the recall is. It appears that the F-score F_1, often considered for the trade-off, is often far from being optimal. In some cases, the ranking induced by F_1 perfectly mimics the ranking induced by the precision or by the recall and completely ignores the other one. I gave all the theoretical and algorithmic tools necessary to determine what the optimal trade-off is, for any reference distribution of performance or any finite set of entities to be ranked.

For the two-class crisp classification, my theory leads to innovative practical tools: the Tile (see Chapter 4) and its flavors (see Chapter 5). It turns out that, with this task, the performance orderings induced by the ranking scores form a 2D manifold that can be mapped on a 2D square (the Tile). Its axes can be interpreted as the relative importance between satisfying cases (true positive *vs* true negative) and between unsatisfying cases (false negative *vs* false positive). This space can be used to depict different types of information (the flavors), such as the best entity or the rank of a given entity *w.r.t.* the application- or user-related preferences. Rank correlations can also be depicted on the Tile, which allows one to visualize the behavior of any score.

My theory has the potential to be particularized to any task. Studying the particularization to tasks other than the two-class crisp classification is left for future research.

Contributions to the adaptation of entities *w.r.t.* the applicative domains, and their characterization

In addition to the summarization, I presented four main contributions on domains.

The first, given in Chapter 6, is to provide a means to perform multi-domain performance analyses. I introduced the easiest and most difficult domains, the preponderant and bottleneck domains, the latter two being rigorously defined through the summarization. These four domains depend on the applicative preferences, and are therefore functions of the importance I. In other words, they are what I have called flavors and can be depicted on the Tile.

The second, given in Chapter 7 is a methodology to evaluate strategies predicting rankings on unseen domains. Applying it on a computer vision problem with well diversified domains, it appears that we don't know yet how to make the best use of the existing benchmarks. This issue seems to be overlooked in the literature.

The third, given in Chapter 8, is a method that can be used to dynamically adapt entities to a target domain in which the distribution of cases that can arise is a mixture (convex combination) of the distribution in source domains. I presented the theory and an unsupervised algorithm called *Mixture Domain Adaptation* (MDA) to determine how entities optimized for the source domains can be combined to obtain an entity optimized for the target domain.

The last, given in Chapter 9, is an ensemble of techniques to characterize domains in a physically interpretable way and to analyze the relationship between them.

Perspectives

Global directions for future research

In the VIULaB (formerly TELIM) laboratory where I completed this thesis, we have developed the practice of structuring our developments around several pillars: a *data* pillar, a *method* pillar, and *performance* pillar. As a matter of fact, most of the work published in the literature in my core discipline —computer vision— focuses heavily on the *methods* pillar and only deals superficially with the other two. From my point of view, there is currently a significant imbalance in the research being carried out. The work presented in this thesis clearly shows that there are still many important unanswered questions in the *performance* pillar and that there is a definite need to continue researching in this direction.

All disciplines can benefit from the theoretical foundations of performance-based ranking

All the components of my theoretical foundations are intrinsically interdisciplinary (in particular the entities, performances, tasks, applicative importances, as well as the need to establish rankings). My approach, based on the probability and order theories, promotes the use of the mathematical language, which is universal, and the level of abstraction is high. My theoretical foundations thus have a beautiful potential for wide applicability. The only limitation that I see is that it is only for those who agree with the probabilistic nature of the performance and with a quantitative approach, describing numerically the performances, the tasks, and the applicative preferences.

This thesis has a particular focus on tasks similar to crisp classifications with two classes. The theoretical results presented and the practical tools developed for them (the Tile and its flavors) have the potential to be used in many disciplines. Indeed, this kind of task occurs in everyday life, when one has to distinguish between junk and desirable mails; in medicine, when one has to decide whether a patient has a certain disease or a certain syndrome (*e.g.*, [259]); in change detection, when one has to decide if a given pixel belongs to the foreground or to the background of the scene (*e.g.*, [260, 161, 18, 164, 163, 261]); in video surveillance, when one has to determine whether there are people in a filmed scene (*e.g.*, [262, 263, 264]) or if some video clip corresponds to some given activity (*e.g.*, [265, 266]); or in astronomy, when one has to tell whether there is one (or several) exoplanet(s) near a star (*e.g.*, [267]). There are yet other problems for which the evaluation leads to two satisfying ($S = 1$) and two unsatisfying ($S = 0$) cases: for the prediction of software defects [268], for the evaluation of rules in logic [269], for the evaluation of clustering methods [54], etc. All researchers, method developers, and benchmarkers facing this type of problem can now benefit from my developments.

Beyond two-class crisp classification, the theoretical foundations of performance-based ranking could be particularized to any other task (see Fig. 2.4, page 31). For all of them, it should be particularly informative to (1) identify the measurable space (Ω, Σ) and the random variables that one needs to make connections with the literature, (2) determine the consensus on the associated satisfaction, (3) study how the performance space is

structured, (4) determine the applicative preferences associated with the scores previously defined in the literature, (5) study how the manifold of meaningful performance orderings is structured, and (6) explore the innovative tools and algorithms that can be developed.

Stop loosing our works!

The best entity for a given application and a given user is not, in general, the same for others. A single ranking cannot therefore suit everyone.

Unfortunately, the current practice consists in standardizing the way in which the developed methods are ranked, on the pretext that the results reported in the literature are then comparable. This is often done by opting for a reference score specific for a given community and a given task. However, this way of working leads to a unique ranking. But the developed methods are difficult to publish if they are not at the top of this unique ranking. And as soon as they are no longer top ranked, the community tends to forget about them. This is unfortunate because these methods may be precisely those that best meet the needs of certain applications and users.

Instead, it would be preferable to consider all possible rankings for all possible applicative importances. For tasks similar to two-class crisp classification, we have seen that the Tile allows to display the full spectrum of rankings. For other tasks, we can imagine creating benchmarks in the form of dynamic web pages. The performances of all available entities would be saved, and the benchmark visitors would be asked to specify their importances I to show them a ranking tailored to their needs. This would be the ranking induced by the ranking score R_I. In this way, no potentially useful entity would be lost.

Towards a crystal clear understanding of the performances and adaptation capabilities of entities

Most systems designed in computer science (and more generally in engineering) are successful. However, to trust the reliability of these systems, it is essential to have a crystal-clear understanding of their performances and their ability to adapt not only to a diversity of applicative domains but also to various application- and user-related preferences.

Unfortunately, the current understanding is harmed by *incorrect beliefs* on which the scientific community relies and that are widely accepted. Let me list a few that were rebutted in this thesis: "The operations on real numbers that we learned in elementary school can be applied without caution to the values taken by scores." (rebutted in Chapter 1); "We have long been well aware of how to evaluate, compare, and analyze the performances of the methods that we develop." (rebutted in Chapter 2); "Intuition is enough to establish that one's favorite score can be used to establish a meaningful ranking." (rebutted in Chapter 2); "To obtain trade-offs between scores, one can simply average values or ranks." (rebutted in Chapters 2 and 3); "Only a few scores are needed to analyze performances: a space based on two scores (*e.g.*, the ROC or Precision-Recall spaces) or a table of values are the most appropriate ways to compare performances." (rebutted in Chapters 4 and 5); "The weights given to domains can be used as is to average

score values." (rebutted in Chapters 1 and 6); "The available benchmarks are relevant and can easily be extrapolated." (rebutted in Chapters 2 and 7); "To have systems that make good decisions, all one needs to do is to put together all the available data and to train a single large predictive model." (rebutted in Chapter 8); and "A system does not need to be aware of the domain in which it operates to behave reliably." (rebutted in Chapter 9). A large community of researchers can now benefit from these general clarifications.

Final words

I hope that evaluation, comparison, and adaptation of methods are now less of a mystery to you, and that you now find this field of research as fascinating as I do. I also hope that this thesis has helped you to better understand the generic concepts of performance, ranking associated with specific application preferences, and domain, as well as the links between these concepts. Finally, I hope that this thesis has convinced you of the need for solid theoretical foundations, and that these foundations can be established on top of the probability theory.

A. Supplementary material for the 2nd article

This appendix contains the supplementary material provided with my 2nd selected article [2], presented in Chapter 2.

A. Supplementary Material

Contents

A.1. List of Symbols . . . 149
A.1.1 Mathematical Symbols . . . 149
A.1.2 Symbols Related to Our Mathematical Framework . . . 149
A.1.3 Symbols used for Operations on Performances . . . 150
A.1.4 Symbols used in the Performance Ordering and Performance-Based Ranking Theory . . . 150
A.1.5 Symbols used for the Particular Case of Two-Class Crisp Classifications . . . 150
A.2 How to Use our Framework: a Little Catalog of Problems . . . 151
A.2.1 Multi-Class Classification (with a Note on Micro- and Macro-Averaging) . . . 151
A.2.2 Regression (with a Note on the Mean Squared Error and the Mean Absolute Error) . . . 152
A.2.3 Information Retrieval . . . 152
A.2.4 Detection (with a Note about the Intersection-over-Union and the F-Score) . . . 153
A.2.5 Clustering (with a Note about Fowlkes-Mallows Index) . . . 154
A.2.6 Ranking (with a Note about Kendall's τ) . . . 154
A.3. Supplementary Material about Sec. 3.2 . . . 156
A.3.1 Reminders of Classical Definitions. . . . 156
A.3.2 The 4 Cases in the Comparison of Two Performances with a Preorder $\lesssim$. . . . 156
A.3.3 Implications of the Transitivity of $\lesssim$. . . . 156
A.3.4 Properties of $\sim$, $>$, $<$, $\not\lesssim$, $\lesssim$, and $\gtrsim$. . . . 157
A.4. Supplementary Material about Sec. 3.6 . . . 159
A.4.1 The Visual Inspection of Formulas, to Determine the Importance given by Scores, can be Misleading! 159
A.4.2 So, How can we Determine the Importance given by Scores? . . . 159
A.5. Supplementary Material about Sec. 4.2 . . . 160
A.5.1 Proof of Theorem 1. . . . 160
A.5.2 Proof of Theorem 2 . . . 160
A.5.3 Proof of Theorem 3 . . . 161
A.6. Supplementary Material about Sec. 4.3 . . . 163
A.6.1 All Ranking Scores can be Used to Rank Performances (for $\Phi = \mathrm{conv}$) . . . 163
A.6.2 On the Properties of Ranking Scores. . . . 165
A.7. Supplementary Material about Sec. 5.2 . . . 168
A.7.1 Link between Classical Formulation and Ours. . . . 168
A.7.2 Custom Optimization Algorithm to Estimate Kendall's τ. . . . 168
A.7.3 Scores Perfectly Correlated with a Ranking Score, for all Performances . . . 168
A.7.4 Scores Perfectly Correlated with a Ranking Score, for the Performances Corresponding to Given Class Priors $\pi_- \neq 0$ and $\pi_+ \neq 0$ . . . 168

A.1. List of Symbols

A.1.1 Mathematical Symbols

- $\mathbf{1}_U$: the 0-1 indicator function of subset U
- $\mathbb{R}$: the real numbers
- $\mathcal{R}$: a relation
- conv: the set of convex combinations
- $\vee$: the *inclusive disjunction* (*i.e.*, logical or)
- $\wedge$: the *conjunction* (*i.e.*, logical and)
- $\circ$: the composition of functions, *i.e.* $(g \circ f)(x) = g(f(x))$
- $\mathbf{E}$: the mathematical expectation

A.1.2 Symbols Related to Our Mathematical Framework

We organize these symbols according to the 6 pillars depicted in Fig. 1, which correspond to the 6 subsections of Sec. 3.

Symbols related to the 1st pillar (Sec. 3.1)
- Ω: the sample space (universe)
- ω: a sample (*i.e.*, an element of Ω)
- Σ: the event space (a σ-algebra on Ω, *e.g.* 2^Ω)
- E: an event (*i.e.*, an element of Σ)
- (Ω, Σ): the measurable space
- $\mathbb{P}_{(\Omega,\Sigma)}$: all performances on (Ω, Σ)
- Π: a set of performances ($\Pi \subseteq \mathbb{P}_{(\Omega,\Sigma)}$)
- P: a performance (*i.e.*, an element of $\mathbb{P}_{(\Omega,\Sigma)}$)

Symbols related to the 2nd pillar (Sec. 3.2)
- $\lesssim$: binary relation *worse or equivalent* on $\mathbb{P}_{(\Omega,\Sigma)}$
- $\gtrsim$: binary relation *better or equivalent* on $\mathbb{P}_{(\Omega,\Sigma)}$
- $\sim$: binary relation *equivalent* on $\mathbb{P}_{(\Omega,\Sigma)}$
- $>$: binary relation *better* on $\mathbb{P}_{(\Omega,\Sigma)}$
- $<$: binary relation *worse* on $\mathbb{P}_{(\Omega,\Sigma)}$
- $\lesssim \atop \gtrsim$: binary relation *incomparable* on $\mathbb{P}_{(\Omega,\Sigma)}$

Symbols related to the 3rd pillar (Sec. 3.3)
- S: the random variable *Satisfaction*
- $s_{min,\Omega}$: the minimum satisfaction value
- $s_{max,\Omega}$: the maximum satisfaction value

Symbols related to the 4th pillar (Sec. 3.4)
- $\mathbb{E}$: the set of entities to rank
- ϵ: an entity, *i.e.* an element of $\mathbb{E}$
- eval: the performance *evaluation* function
- Φ: some performances that are for sure achievable

Symbols related to the 5th pillar (Sec. 3.5)
- X: a score
- $\mathrm{dom}(X)$: the domain of the score X
- X^E_V: the *expected value score* parameterized by the random variable V
- $X^P_{E_1|E_2}$: the *probabilistic score* parameterized by the events E_1 and E_2

Symbols related to the 6th pillar (Sec. 3.6)
- I: the random variable *Importance*

A.1.3 Symbols used for Operations on Performances

- filter_I: the *filtering* operation

A.1.4 Symbols used in the Performance Ordering and Performance-Based Ranking Theory

- $\mathrm{rank}_{\mathbb{E}}$: the *ranking* function, w.r.t. the set of entities $\mathbb{E}$
- $\lesssim_X$: the ordering induced by the score X (*cf*. Theorem 1)
- R_I: the *ranking score* parameterized by the importance I
- τ: the rank correlation coefficient of Kendall [50]

A.1.5 Symbols used for the Particular Case of Two-Class Crisp Classifications

Particularization of the mathematical framework
- tn: the sample *true negative*
- fp: the sample *false positive*, *a.k.a.* type I error
- fn: the sample *false negative*, *a.k.a.* type II error
- tp: the sample *true positive*

Extensions to the mathematical framework
- ROC: the *Receiver Operating Characteristic* space, *i.e.* $FPR \times TPR$
- PR: the *Precision-Recall* space, *i.e.* $TPR \times PPV$
- Y: the random variable for the ground truth
- $\hat{Y}$: the random variable for the prediction
- $\mathbb{C}$: the set of classes
- c: a class (*i.e.*, an element of $\mathbb{C}$)
- c_-: the negative class
- c_+: the positive class

Scores
- A: the *accuracy*
- TNR: the *true negative rate*
- FPR: the *false positive rate*
- TPR: the *true positive rate*
- PPV: the *positive predictive value*
- F_β: the F-scores
- π_+: the *prior of the positive class*
- π_-: the *prior of the negative class*

A.2. How to Use our Framework: a Little Catalog of Problems

Throughout the paper, we have exemplified our theory with the problem of two-class classification. This section aims at showing the universality of our theory. It presents a little catalog of other problems, together with discussions on how to use our framework for them. These discussions are introductions. As shown by Sec. 5 and two recent works [4, 5], an in-depth analysis and particularization of our theory to the various problems (*e.g.*, to highlight their distinctive features, to review the current ranking practices from the literature and the consistency of popular scores, and to establish practical tools tailored to different user needs) may require substantial work that is out of the scope of this supplementary material.

In the following, we adopt a systematic approach: for each problem, we start by specifying a thought random experiment for the evaluation (this is an arbitrary choice since the random experiment is not unique for each problem; do not hesitate to use a different random experiment, the important is to specify it explicitly!), then we discuss the possible choices for the sample space Ω (and what the set of all performances is), for the modeling of tasks with the satisfaction S, for the modeling of the knowledge we have about the evaluation with the function Φ, and for the modeling of the application-specific preferences with the importance I.

A.2.1 Multi-Class Classification (with a Note on Micro- and Macro-Averaging)

Let us consider the following thought experiment to evaluate classifiers predicting classes in a finite and non-empty set $\mathbb{C}$.

Random Experiment 1. *(1) Draw a sample $s \in \mathbb{S}$ at random from a given* source $\mathcal{S}$. *(2) Apply the* oracle $\mathcal{O}$ *on s to obtain the ground-truth class $y(s) \in \mathbb{C}$. (3) Apply a* descriptor $\mathcal{D}$ *on s to obtain the features (*a.k.a. *attributes) $x(s) \in \mathbb{X}$. (4) Feed the* classifier $\mathcal{C}$ *with $x(s)$ to obtain the predicted class $\hat{y}(x(s)) \in \mathbb{C}$. (5) Set the outcome of the experiment to the pair $(y, \hat{y}) \in \mathbb{C}^2$.*

Choice for Ω and $\mathbb{P}_{(\Omega,\Sigma)}$. Our theory applies with $\Omega = \mathbb{C}^2$. By definition, the function eval gives, for any classifier $\mathcal{C} : \mathbb{X} \to \mathbb{C}$ (the evaluated entity, either deterministic or not), the distribution of outcomes resulting from this random experiment: $P_{\mathcal{C}} = \mathrm{eval}(\mathcal{C})$. Note that it is implicit that the performances are specific for some given source $\mathcal{S}$ (*e.g.*, evaluation dataset), oracle $\mathcal{O}$, and descriptor $\mathcal{D}$. By convenience, one can manipulate the ground-truth and predicted classes with, respectively, the random variables Y and $\hat{Y}$ defined in such a way that $\omega = (Y(\omega), \hat{Y}(\omega))\ \forall \omega \in \Omega$.

Choice for S. Several classification tasks can be distinguished, as the following two examples show. (1) One can consider that all erroneous classifications are unsatisfactory and that correct classifications are satisfactory. For this task, the satisfaction is then binary and given by $S = \mathbf{1}_{Y=\hat{Y}}$. The expected value of the satisfaction, which is a particular case of ranking scores, is then equal to the multi-class accuracy. (2) One can also consider the similarity $\mathrm{sim} : \mathbb{C}^2 \to \mathbb{R}$ between classes, and choose the satisfaction accordingly: $S(\omega) = \mathrm{sim}(Y(\omega); \hat{Y}(\omega))$. A wide variety of tasks can be considered by tuning sim. In general, the expected value of the satisfaction is different from the multi-class accuracy.

Choice for Φ. As the classifier is used once and only once during the execution of the evaluation, we know that if the performances P_1 and P_2 are achievable by some classifiers $\mathcal{C}_1$ and $\mathcal{C}_2$, then any performance $\overline{P} = \lambda_1 P_1 + \lambda_2 P_2$ (with $\lambda_1 \geq 0$, $\lambda_2 \geq 0$, and $\lambda_1 + \lambda_2 = 1$) is achievable by a classifier $\overline{\mathcal{C}}$ obtained by a non-deterministic combination of $\mathcal{C}_1$ and $\mathcal{C}_2$ that chooses them with respective probabilities λ_1 and λ_2. Thus, $\Phi = \mathrm{conv}$ makes sense, and all ranking scores can be used to rank classifiers. However, it would be possible to go further, by considering other functions Φ that would include the knowledge that we can predict the performance achievable by composing the classifier with any of the $|\mathbb{C}|^{|\mathbb{C}|}$ functions $f : \mathbb{C} \to \mathbb{C}$. This would lead to other performance orderings suitable for ranking classifiers.

Choice for I. When $\Phi = \mathrm{conv}$, we have demonstrated that all rankings induced by the ranking scores R_I satisfy all our three axioms. This leaves a great flexibility for the users to fine-tune the ranking w.r.t. their application-specific preferences through the random variable I. As we have seen, the only constraints are that $I \neq 0$ and $I(\omega) \geq 0\ \forall \omega \in \Omega$.

Note on micro- and macro-averaging. Micro- and macro-averaging are commonly used techniques to build scores for multi-class classification from scores for two-class classification [38]. We warn that they have pitfalls. In general, micro- and macro-averaging scores suitable for ranking two-class classifiers do not lead to scores suitable for ranking multi-class classifiers. The accuracy put aside, the performance orderings induced from the micro-averaged versions of the scores put in green in Tab. 1 are incompatible with $S = \mathbf{1}_{Y=\hat{Y}}$: our 2nd axiom is not satisfied. Moreover, the accuracy put again aside, the performance orderings induced from the macro-averaged versions of the scores put in green in Tab. 1 are incompatible with $\Phi = \mathrm{conv}$: our 3rd axiom is not satisfied. A solution consists in using directly ranking scores defined for multi-class classification.

A.2.2 Regression (with a Note on the Mean Squared Error and the Mean Absolute Error)

Let us consider the following thought experiment to evaluate regressors.

Random Experiment 2. *(1) Draw a sample $s \in \mathbb{S}$ at random from a given* source $\mathcal{S}$. *(2) Apply the* oracle $\mathcal{O}$ *on s to obtain the ground-truth value $y(s) \in \mathbb{R}$. (3) Apply a* descriptor $\mathcal{D}$ *on s to obtain the features (*a.k.a. *attributes) $x(s) \in \mathbb{X}$. (4) Feed the* regressor $\mathcal{R}$ *with $x(s)$ to obtain the predicted value $\hat{y}(x(s)) \in \mathbb{R}$. (5) Set the outcome of the experiment to the pair $(y, \hat{y}) \in \mathbb{R}^2$.*

Choice for Ω and $\mathbb{P}_{(\Omega,\Sigma)}$. Our theory applies with $\Omega = \mathbb{R}^2$. By definition, the function eval gives, for any regressor $\mathcal{R} : \mathbb{X} \to \mathbb{R}$ (the evaluated entity, either deterministic or not), the distribution of outcomes resulting from this random experiment. It is the performance $P_{\mathcal{R}} = \text{eval}(\mathcal{R})$ of $\mathcal{R}$. By convenience, one can manipulate the ground-truth and predicted values with, respectively, the random variables Y and $\hat{Y}$ defined in such a way that $\omega = (Y(\omega), \hat{Y}(\omega)) \, \forall \omega \in \Omega$.

Choice for S. Clearly, it is not a good idea to choose $S = \mathbf{1}_{Y=\hat{Y}}$, similarly as one can do in classification. In practice, a regressor as no chance to predict the same value as the oracle, so this unfortunate choice for S would lead to performances P such that $P(S = 0) = 1$. In other words, all performances observed about real regressors would belong to the set of the worst performances (see Corollary 2), and their ranking would be of little interest (see Corollary 1). A better option consists in specifying a tolerance $\epsilon > 0$ and choosing $S = \mathbf{1}_{|Y-\hat{Y}|\leq\epsilon}$. An even more flexible option, which takes advantage of the fact that satisfaction values do not necessarily have to be positive, is to choose $S = f(|Y - \hat{Y}|)$ with any arbitrarily chosen monotonically decreasing function f. The plethora of choices that can be made for S makes it clear that there is an infinity of tasks related to the regression problem.

Choice for Φ. As the regressor is used once and only once during the execution of the evaluation, we know that if the performances P_1 and P_2 are achievable by some regressors $\mathcal{R}_1$ and $\mathcal{R}_2$, then any performance $\overline{P} = \lambda_1 P_1 + \lambda_2 P_2$ (with $\lambda_1 \geq 0$, $\lambda_2 \geq 0$, and $\lambda_1 + \lambda_2 = 1$) is achievable by a regressor $\overline{\mathcal{R}}$ obtained by a non-deterministic combination of $\mathcal{R}_1$ and $\mathcal{R}_2$ that chooses them with respective probabilities λ_1 and λ_2. Thus, $\Phi = \text{conv}$ makes sense, and all ranking scores can be used to rank regressors. However, it would be possible to go further, by considering other functions Φ that would include the knowledge that we can predict the performance achievable by adding noise, or applying a transformation on the output of the regressor. This would lead to other performance orderings suitable for ranking regressors.

Choice for I. When $\Phi = \text{conv}$, we have demonstrated that all rankings induced by the ranking scores R_I satisfy all our three axioms. This leaves a great flexibility for the users to fine-tune the ranking w.r.t. their application-specific preferences through the random variable I. As we have seen, the only constraints are that $I \neq 0$ and $I(\omega) \geq 0 \, \forall \omega \in \Omega$.

Note on the mean squared error and the mean absolute error. If we choose $S = -|Y - \hat{Y}|^2$, the ranking score R_I corresponding to uniform importance values, *i.e.*, the expected satisfaction X_S^E, yields a ranking that minimizes the *mean squared error* (MSE). If we choose $S = -|Y - \hat{Y}|$, the ranking score R_I corresponding to uniform importance values, *i.e.*, the expected satisfaction X_S^E, yields a ranking that minimizes the *mean absolute error* (MAE).

A.2.3 Information Retrieval

Let us consider the following thought experiment to evaluate information retrieval systems. We denote by $\mathbb{Q}$ the set of all possible queries.

Random Experiment 3. *(1) Draw a query $q \in \mathbb{Q}$ at random from a given source $\mathcal{S}$. (2) Apply the oracle $\mathcal{O}$ on q to obtain the ground-truth set of results $\mathbb{Y}$. (3) Apply the evaluated information retrieval system $\mathcal{S}$ on q to obtain the predicted set of results $\hat{\mathbb{Y}}$. (4) If $\mathbb{Y} = \emptyset$ and $\hat{\mathbb{Y}} = \emptyset$, restart the experiment, otherwise draw a result r at random in $\mathbb{Y} \cup \hat{\mathbb{Y}}$. (5) Choose the outcome as follows: fp if $r \notin \mathbb{Y}$ and $r \in \hat{\mathbb{Y}}$, fn if $r \in \mathbb{Y}$ and $r \notin \hat{\mathbb{Y}}$, and tp if $r \in \mathbb{Y}$ and $r \in \hat{\mathbb{Y}}$.*

Choice for Ω and $\mathbb{P}_{(\Omega,\Sigma)}$. Our theory applies with $\Omega = \{fp, fn, tp\}$. By definition, the function eval gives, for any retrieval system $\mathcal{S}$ defined on $\mathbb{Q}$ (the evaluated entity, either deterministic or not), the distribution of outcomes resulting from this random experiment: $P_{\mathcal{S}} = \text{eval}(\mathcal{S})$.

Choice for S. Intuitively, everyone certainly agrees that $S(fp) < S(tp)$ and $S(fn) < S(tp)$. But we expect different opinions regarding whether the outcome (sample) fp gives less, equal, or more satisfaction than fn.

Choice for Φ. This random experiment is very interesting as, during its execution, the evaluated entity (the information retrieval system $\mathcal{S}$) can be used multiple times. In such a case, we have to discuss whether $\Phi = \mathrm{conv}$ is adequate. Let us consider two systems $\mathcal{S}_1, \mathcal{S}_2$ and their respective performances $P_1 = \mathrm{eval}(\mathcal{S}_1)$, $P_2 = \mathrm{eval}(\mathcal{S}_2)$. It is possible to show that the performance of a retrieval system $\overline{\mathcal{S}}$ obtained by a non-deterministic combination of $\mathcal{S}_1$ and $\mathcal{S}_2$, that chooses them with respective probabilities λ_1 and λ_2, is some interpolated performance $\overline{P} = \mu_1 P_1 + \mu_2 P_2$ (with $\mu_1 \geq 0$, $\mu_2 \geq 0$, and $\mu_1 + \mu_2 = 1$). Unless being in very particular cases, $\mu \neq \lambda$. In other words, we know that the performances that are convex combinations of achievable performances are also achievable (for any λ, there exists μ), but we do not know in general how to achieve them (for most μ it is not possible to determine λ). This contrasts with the other kinds of problems discussed in this catalog. In fact, the question we raise here is not specific to the information retrieval problem: it is peculiar to the random experiment that we have chosen for it. By slightly modifying the thought experiment, the question vanishes: instead of restarting the experiment when $\mathbb{Y} \cup \hat{\mathbb{Y}} = \emptyset$, one could yield a fourth outcome (and add it to the sample space Ω). By doing so, the evaluated entity $\mathcal{S}$ is used only once and $\Phi = \mathrm{conv}$ makes sense for sure.

Choice for I. If $\Phi = \mathrm{conv}$ is considered as adequate, then we have demonstrated that all rankings induced by the ranking scores R_I satisfy all our three axioms. This leaves a great flexibility for the users to fine-tune the ranking w.r.t. their application-specific preferences through the random variable I. As we have seen, the only constraints are that $I \neq 0$ and $I(\omega) \geq 0 \, \forall \omega \in \Omega$.

A.2.4 Detection (with a Note about the Intersection-over-Union and the F-Score)

Different types of detections are present in the literature. An example of spatial detection aims at predicting the axis-aligned bounding boxes around all the objects that match some given properties (*i.e.*, a semantic class) in input images. Examples of temporal detections include the detection of events in video streams and in audio recordings. By definition, such detection problems are called *action spotting* when the temporal window is small, and *activity detection* otherwise. Let us consider the following, generic, thought experiment to evaluate detectors.

Random Experiment 4. *(1) Draw an input at random from a given* source $\mathcal{S}$ *(*e.g.*, dataset). (2) Apply the* oracle $\mathcal{O}$ *on it to obtain a set $\mathbb{Y}$ of ground-truth detections. (3) Also apply the* detector $\mathcal{D}$ *on it to obtain a set $\hat{\mathbb{Y}}$ of predicted detections. (4) If $\mathbb{Y} = \emptyset$ and $\hat{\mathbb{Y}} = \emptyset$, then end the experiment with the outcome $\circledast$. Otherwise: (5) Apply a matching criterion between $\mathbb{Y}$ and $\hat{\mathbb{Y}}$ such that to any detection in $\mathbb{Y}$ should be associated at most a detection in $\hat{\mathbb{Y}}$ and vice versa. (6) Draw a detection d at random in $\mathbb{Y} \cup \hat{\mathbb{Y}}$. (7) Give as outcome fp, fn or tp depending on whether d is a prediction, a ground truth, or both (*i.e.*, a match).*

Choice for Ω and $\mathbb{P}_{(\Omega,\Sigma)}$. Our theory applies with $\Omega = \{\circledast, fp, fn, tp\}$. By definition, the function eval gives, for any detector (the evaluated entity), the distribution of outcomes resulting from this random experiment. It is the performance $P_{\mathcal{D}} = \mathrm{eval}(\mathcal{D})$ of $\mathcal{D}$.

Choice for S. Intuitively, everyone certainly agrees that $\circledast$ and tp give entire satisfaction. Moreover, we expect agreement on $S(fp) < S(tp)$ and $S(fn) < S(tp)$. But we expect different opinions regarding whether fp gives less, equal, or more satisfaction than fn.

Choice for Φ. As the detector is used once and only once during the execution of the evaluation, we know that if the performances P_1 and P_2 are achievable by some detectors $\mathcal{D}_1$ and $\mathcal{D}_2$, then any performance $\overline{P} = \lambda_1 P_1 + \lambda_2 P_2$ (with $\lambda_1 \geq 0$, $\lambda_2 \geq 0$, and $\lambda_1 + \lambda_2 = 1$) is achievable by a detector $\overline{\mathcal{D}}$ obtained by a non-deterministic combination of $\mathcal{D}_1$ and $\mathcal{D}_2$ that chooses them with respective probabilities λ_1 and λ_2. Thus, $\Phi = \mathrm{conv}$ makes sense, and all ranking scores can be used to rank detectors.

Choice for I. If $\Phi = \mathrm{conv}$ is considered as adequate, then we have demonstrated that all rankings induced by the ranking scores R_I satisfy all our three axioms. This leaves a great flexibility for the users to fine-tune the ranking w.r.t. their application-specific preferences through the random variable I. As we have seen, the only constraints are that $I \neq 0$ and $I(\omega) \geq 0 \, \forall \omega \in \Omega$.

Note about the Intersection-over-Union and the F-score. Traditionally, in the literature, as soon as one has symbols fp, fn, and tp, regardless of their very fine meaning, one defines quantities $IoU = \frac{P(tp)}{P(fp)+P(fn)+P(tp)}$ and $F_1 = \frac{2P(tp)}{P(fp)+P(fn)+2P(tp)}$, and name them *Intersection-over-Union* (or *Jaccard*) and *F-one*, respectively. The exact meaning

of these quantities is not well standardized. In particular, the random experiment supporting the evaluation, if it exists, is rarely specified explicitly. For this reason, we cannot give the guarantee that these quantities are suitable to rank detectors in all works in which they have been used. However, with the random experiment given here-above, with $\Phi = \text{conv}$, and with $S = \mathbf{1}_{\{\circledast,tp\}}$, we can guarantee that IoU and F_1 are suitable to rank detectors because they are equal to the ranking scores with, respectively, $I = \mathbf{1}_{\{fp,fn,tp\}}$ and $I = \mathbf{1}_{\{fp,tp\}} + \mathbf{1}_{\{fn,tp\}}$. Thus, the performance orderings induced by them fulfill our three axioms.

A.2.5 Clustering (with a Note about Fowlkes-Mallows Index)

Let us consider the following thought experiment to evaluate clustering methods. These methods aim to place in different clusters (groups) dissimilar objects and in the same cluster (group) objects that are similar to each other. We denote by $\mathbb{E}$ the set of elements that these methods have to deal with. For the sake of simplicity, we do not consider hierarchical clustering.

Random Experiment 5. *(1) Apply both the* clustering method $\mathcal{C}$ *and the* oracle $\mathcal{O}$ *on* $\mathbb{E}$ *to obtain, respectively, the predicted and ground-truth clusterings. (2) Randomly draw two distinct elements,* ϵ_1 *and* ϵ_2*, from* $\mathbb{E}$*. (3) Consider that the pair* $(\epsilon_1;\epsilon_2)$ *is a negative or a positive, in a given clustering, when* ϵ_1 *and* ϵ_2 *are in different clusters or in the same cluster, respectively. (4) Choose the outcome as follows:* tn *when* $(\epsilon_1;\epsilon_2)$ *is negative in both the predicted and ground-truth clusterings,* fp *when* $(\epsilon_1;\epsilon_2)$ *is negative in the ground-truth clustering and positive in the predicted clustering,* fn *when* $(\epsilon_1;\epsilon_2)$ *is positive in the ground-truth clustering and negative in the predicted clustering, and* tp *when* $(\epsilon_1;\epsilon_2)$ *is positive in both the predicted and ground-truth clusterings.*

Choice for Ω and $\mathbb{P}_{(\Omega,\Sigma)}$. Our theory applies with $\Omega = \{tn, fp, fn, tp\}$. By definition, the function eval gives, for any clustering method (the evaluated entity), the distribution of outcomes resulting from this random experiment. It is the performance $P_{\mathcal{C}} = \text{eval}(\mathcal{C})$ of $\mathcal{C}$.

Choice for S. When S is chosen such that $S(fp) = S(fn) = 0$ and $S(tn) = S(tp) = 1$, we are in the same setting as the one we studied for the two-class classification in Sec. 5. This is indeed not because we use the same symbols for the elements of Ω —this is just a convention—, but because, in both settings, we have $|\Omega| = 4$, $|S = 0| = 2$, and $|S = 1| = 2$. This implies that the performance orderings that satisfy our three axioms for ranking two-class classifiers can also be used for ranking clustering methods, and vice versa.

Choice for Φ. As the clustering method is used once and only once during the execution of the evaluation, we know that if the performances P_1 and P_2 are achievable by some clustering methods $\mathcal{C}_1$ and $\mathcal{C}_2$, then any performance $\overline{P} = \lambda_1 P_1 + \lambda_2 P_2$ (with $\lambda_1 \geq 0$, $\lambda_2 \geq 0$, and $\lambda_1 + \lambda_2 = 1$) is achievable by a clustering method $\overline{\mathcal{C}}$ obtained by a non-deterministic combination of $\mathcal{C}_1$ and $\mathcal{C}_2$ that chooses them with respective probabilities λ_1 and λ_2. Thus, $\Phi = \text{conv}$ makes sense, and all ranking scores can be used to rank clustering methods.

Choice for I. When $\Phi = \text{conv}$, we have demonstrated that all rankings induced by the ranking scores R_I satisfy all our three axioms. This leaves a great flexibility for the users to fine-tune the ranking w.r.t. their application-specific preferences through the random variable I. As we have seen, the only constraints are that $I \neq 0$ and $I(\omega) \geq 0\, \forall \omega \in \Omega$.

Note about Fowlkes-Mallows index. The score FMI known as *Fowlkes-Mallows index* [54] and *cosine coefficient* [55], which is commonly used for clustering methods and defined as the geometric mean of the positive predictive value PPV (*a.k.a.* precision) and true positive rate TPR (*a.k.a.* sensitivity and recall), does not satisfy our 3rd axiom: the performance ordering induced by FMI is incompatible with $\Phi = \text{conv}$. More precisely, the clustering method $\mathcal{C}$ obtained by randomly choosing between some methods $\mathcal{C}_1$ or $\mathcal{C}_2$ can be such that $FMI(\text{eval}(\mathcal{C})) < \min(FMI(\text{eval}(\mathcal{C}_1)), FMI(\text{eval}(\mathcal{C}_2)))$, while it makes no sense to say that $\mathcal{C}$ can be worse than $\mathcal{C}_1$ or $\mathcal{C}_2$. From this perspective, it is advisable to use any ranking score instead of FMI, for example, those in green in Tab. 1.

A.2.6 Ranking (with a Note about Kendall's τ)

Let us consider the following thought experiment to evaluate ranking methods. We denote by $\mathbb{E}$ the set of elements that these methods have to rank. For the sake of simplicity, we prefer to deal only with the case with no tie hereafter.

Random Experiment 6. *(1) Apply both the* ranking method $\mathcal{R}$ *and the* oracle $\mathcal{O}$ *on* $\mathbb{E}$ *to obtain, respectively, the predicted and ground-truth sequences of elements. (2) Randomly draw two distinct elements,* ϵ_1 *and* ϵ_2*, from* $\mathbb{E}$*. (3) Four cases can*

occur depending on whether ϵ_1 is before or after ϵ_2 in the predicted sequence and whether ϵ_1 is before or after ϵ_2 in the ground-truth sequence. Nevertheless, two outcomes are enough: choose ☺ if ϵ_1 and ϵ_2 appear in the same order in both sequences, ☹ otherwise.

Choice for Ω and $\mathbb{P}_{(\Omega,\Sigma)}$. Our theory applies with $\Omega = \{☺, ☹\}$. By definition, the function eval gives, for any ranking method (the evaluated entity), the distribution of outcomes resulting from this random experiment. It is the performance $P_{\mathcal{R}} = \mathrm{eval}(\mathcal{R})$ of $\mathcal{R}$. The probability of drawing a *discordant pair* is given by $P(\{☹\})$, and the probability of drawing a *concordant pair* is given by $P(\{☺\})$.

Choice for S. Clearly, $S(☹) < S(☺)$ is wanted.

Choice for Φ. As the ranking method is used once and only once during the execution of the evaluation, we know that if the performances P_1 and P_2 are achievable by some ranking methods $\mathcal{R}_1$ and $\mathcal{R}_2$, then any performance $\overline{P} = \lambda_1 P_1 + \lambda_2 P_2$ (with $\lambda_1 \geq 0$, $\lambda_2 \geq 0$, and $\lambda_1 + \lambda_2 = 1$) is achievable by a ranking method $\overline{\mathcal{R}}$ obtained by a non-deterministic combination of $\mathcal{R}_1$ and $\mathcal{R}_2$ that chooses them with respective probabilities λ_1 and λ_2. Thus, $\Phi = \mathrm{conv}$ makes sense, and all ranking scores can be used to rank ranking methods.

Choice for I. Because $|\Omega| = 2$ and $S(☹) \neq S(☺)$, we are in a particular case in which all ranking scores rank the ranking methods in the same way. From this point of view, fine-tuning I is useless.

Note about Kendall's τ. When $S(☹) = -1$ and $S(☺) = 1$, the expected value of the satisfaction is given by $X_S^E(P) = 1 - 2P(\{☹\}) = P(\{☺\}) - P(\{☹\}) = \tau(P)$. In other words, with the task corresponding to this choice for the satisfaction, Kendall's correlation coefficient τ [50] is the ranking score corresponding to uniform importance values.

A.3. Supplementary Material about Sec. 3.2

This section is devoted to reminders about the order theory.

A.3.1 Reminders of Classical Definitions.

Let $\mathcal{R}$ be a homogeneous binary relation on $\mathbb{P}_{(\Omega,\Sigma)}$. It is said:
- *reflexive iif* $P\mathcal{R}P\,\forall P$;
- *irreflexive iif* $\nexists P : P\mathcal{R}P$;
- *transitive iif* $P_1\mathcal{R}P_2 \wedge P_2\mathcal{R}P_3 \Rightarrow P_1\mathcal{R}P_3\,\forall P_1, P_2, P_3$;
- *symmetric iif* $P_1\mathcal{R}P_2 \Leftrightarrow P_2\mathcal{R}P_1\,\forall P_1, P_2$;
- *asymmetric iif* $\nexists (P_1, P_2) : P_1\mathcal{R}P_2 \wedge P_2\mathcal{R}P_1$;
- and *antisymmetric iif* $P_1\mathcal{R}P_2 \wedge P_2\mathcal{R}P_1 \Rightarrow P_1 = P_2$.

Two homogeneous binary relations $\mathcal{R}_a$ and $\mathcal{R}_b$ on $\mathbb{P}_{(\Omega,\Sigma)}$ are said *converse iif* $P_1\mathcal{R}_aP_2 \Leftrightarrow P_2\mathcal{R}_bP_1\,\forall P_1, P_2$.

A relation $\mathcal{R}$ is:
- an *equivalence iif* it is reflexive, transitive, and symmetric;
- a *preorder iif* it is reflexive and transitive;
- and an *order iif* it is reflexive, transitive, and antisymmetric.

An order $\mathcal{R}$ is said *total iif* $\nexists (P_1, P_2) : P_1 \not\mathcal{R} P_2 \wedge P_2 \not\mathcal{R} P_1$. It is said *partial* otherwise.

A.3.2 The 4 Cases in the Comparison of Two Performances with a Preorder $\lesssim$.

Let us now consider a preorder $\lesssim$ and derive the homogeneous binary relations $\sim$, $>$, $<$, $\lessgtr$ as follows:

$$P_1 \sim P_2 \Leftrightarrow P_1 \lesssim P_2 \wedge P_2 \lesssim P_1 \tag{1}$$
$$P_1 > P_2 \Leftrightarrow P_1 \not\lesssim P_2 \wedge P_2 \lesssim P_1 \tag{2}$$
$$P_1 < P_2 \Leftrightarrow P_1 \lesssim P_2 \wedge P_2 \not\lesssim P_1 \tag{3}$$
$$P_1 \lessgtr P_2 \Leftrightarrow P_1 \not\lesssim P_2 \wedge P_2 \not\lesssim P_1\,. \tag{4}$$

Indeed, we have:

$$P_1 \lesssim P_2 \Leftrightarrow P_1 < P_2 \vee P_1 \sim P_2\,. \tag{5}$$

Similarly, one can derive other binary relations taking unions of $\sim$, $>$, $<$, or $\lessgtr$. For example,

$$P_1 \gtrsim P_2 \Leftrightarrow P_1 > P_2 \vee P_1 \sim P_2\,. \tag{6}$$

A.3.3 Implications of the Transitivity of $\lesssim$.

We can easily check, for each $\mathcal{R}_{ab} \in \{\lesssim, \not\lesssim\}$, each $\mathcal{R}_{ba} \in \{\lesssim, \not\lesssim\}$, each $\mathcal{R}_{bc} \in \{\lesssim, \not\lesssim\}$, each $\mathcal{R}_{cb} \in \{\lesssim, \not\lesssim\}$, each $\mathcal{R}_{ca} \in \{\lesssim, \not\lesssim\}$, and each $\mathcal{R}_{ac} \in \{\lesssim, \not\lesssim\}$, if there exists (P_a, P_b, P_c) such that $P_a\mathcal{R}_{ab}P_b$, $P_b\mathcal{R}_{ba}P_a$, $P_b\mathcal{R}_{bc}P_c$, $P_c\mathcal{R}_{cb}P_b$, $P_c\mathcal{R}_{ca}P_a$, and $P_a\mathcal{R}_{ac}P_c$. Because of the assumed transitivity of $\lesssim$, there are only 29 possible cases out of the 2^6:

1. $P_a \lessgtr P_b$, $P_b \lessgtr P_c$, $P_a \lessgtr P_c$
2. $P_a \lessgtr P_b$, $P_b \lessgtr P_c$, $P_a > P_c$
3. $P_a \lessgtr P_b$, $P_b \lessgtr P_c$, $P_a < P_c$
4. $P_a \lessgtr P_b$, $P_b \lessgtr P_c$, $P_a \sim P_c$
5. $P_a \lessgtr P_b$, $P_b > P_c$, $P_a \lessgtr P_c$
6. $P_a \lessgtr P_b$, $P_b > P_c$, $P_a > P_c$
7. $P_a \lessgtr P_b$, $P_b < P_c$, $P_a \lessgtr P_c$
8. $P_a \lessgtr P_b$, $P_b < P_c$, $P_a < P_c$
9. $P_a \lessgtr P_b$, $P_b \sim P_c$, $P_a \lessgtr P_c$
10. $P_a > P_b$, $P_b \lessgtr P_c$, $P_a \lessgtr P_c$
11. $P_a > P_b$, $P_b \lessgtr P_c$, $P_a > P_c$
12. $P_a > P_b$, $P_b > P_c$, $P_a > P_c$
13. $P_a > P_b$, $P_b < P_c$, $P_a \lessgtr P_c$
14. $P_a > P_b$, $P_b < P_c$, $P_a > P_c$
15. $P_a > P_b$, $P_b < P_c$, $P_a < P_c$

16. $P_a > P_b, P_b < P_c, P_a \sim P_c$
17. $P_a > P_b, P_b \sim P_c, P_a > P_c$
18. $P_a < P_b, P_b \lneqq P_c, P_a \lneqq P_c$
19. $P_a < P_b, P_b \lneqq P_c, P_a < P_c$
20. $P_a < P_b, P_b > P_c, P_a \lneqq P_c$
21. $P_a < P_b, P_b > P_c, P_a > P_c$
22. $P_a < P_b, P_b > P_c, P_a < P_c$
23. $P_a < P_b, P_b > P_c, P_a \sim P_c$
24. $P_a < P_b, P_b < P_c, P_a < P_c$
25. $P_a < P_b, P_b \sim P_c, P_a < P_c$
26. $P_a \sim P_b, P_b \lneqq P_c, P_a \lneqq P_c$
27. $P_a \sim P_b, P_b > P_c, P_a > P_c$
28. $P_a \sim P_b, P_b < P_c, P_a < P_c$
29. $P_a \sim P_b, P_b \sim P_c, P_a \sim P_c$

From this list, we can derive some rules for manipulating the binary relations $\sim$, $>$, $<$, and $\lneqq$. First, we can see that $\sim$, $>$, and $<$ are transitive:

$$P_1 \sim P_2 \wedge P_2 \sim P_3 \Rightarrow P_1 \sim P_3 \tag{7}$$
$$P_1 > P_2 \wedge P_2 > P_3 \Rightarrow P_1 > P_3 \tag{8}$$
$$P_1 < P_2 \wedge P_2 < P_3 \Rightarrow P_1 < P_3\,. \tag{9}$$

Second, we can also see how $\sim$ can be combined with the other 3 relations:

$$(P_1 \sim P_2 \wedge P_2 > P_3) \vee (P_1 > P_2 \wedge P_2 \sim P_3) \Rightarrow P_1 > P_3 \tag{10}$$
$$(P_1 \sim P_2 \wedge P_2 < P_3) \vee (P_1 < P_2 \wedge P_2 \sim P_3) \Rightarrow P_1 < P_3 \tag{11}$$
$$(P_1 \sim P_2 \wedge P_2 \lneqq P_3) \vee (P_1 \lneqq P_2 \wedge P_2 \sim P_3) \Rightarrow P_1 \lneqq P_3\,. \tag{12}$$

And, third, we can see how $\lneqq$ can be combined with $>$ and $<$:

$$(P_1 \lneqq P_2 \wedge P_2 > P_3) \vee (P_1 > P_2 \wedge P_2 \lneqq P_3) \Rightarrow P_1 > P_3 \vee P_1 \lneqq P_3 \tag{13}$$
$$(P_1 \lneqq P_2 \wedge P_2 < P_3) \vee (P_1 < P_2 \wedge P_2 \lneqq P_3) \Rightarrow P_1 < P_3 \vee P_1 \lneqq P_3\,. \tag{14}$$

A.3.4 Properties of $\sim$, $>$, $<$, $\lneqq$, $\lesssim$, and $\gtrsim$.

Lemma 1. *When $\lesssim$ is a preorder, $\sim$ is reflexive.*

Proof. This results from the reflexivity of $\lesssim$ and from Eq. (1): $P \sim P \Leftrightarrow P \lesssim P \wedge P \lesssim P \Leftrightarrow true$. □

Lemma 2. *When $\lesssim$ is a preorder, $\sim$ in transitive.*

Proof. This results from the transitivity of $\lesssim$ (*cf*. Eq. (7)). □

Lemma 3. *When $\lesssim$ is a preorder, $\sim$ is symmetric.*

Proof. This results from the fact that the conjunction is symmetric and from Eq. (1): $P_1 \sim P_2 \Leftrightarrow P_1 \lesssim P_2 \wedge P_2 \lesssim P_1 \Leftrightarrow P_2 \lesssim P_1 \wedge P_1 \lesssim P_2 \Leftrightarrow P_2 \sim P_1$. □

Lemma 4. *When $\lesssim$ is a preorder, $>$ and $<$ are converse.*

Proof. This results from the fact that the conjunction is symmetric and from Eqs. (2) and (3): $P_1 > P_2 \Leftrightarrow P_1 \not\lesssim P_2 \wedge P_2 \lesssim P_1 \Leftrightarrow P_2 \lesssim P_1 \wedge P_1 \not\lesssim P_2 \Leftrightarrow P_2 < P_1$. □

Lemma 5. *When $\lesssim$ is a preorder, $>$ and $<$ are irreflexive.*

Proof. For $>$, this results from the reflexivity of $\lesssim$ and from Eq. (2): $P > P \Leftrightarrow P \not\lesssim P \wedge P \lesssim P \Leftrightarrow false \wedge true = false$. For $<$, the proof is similar. □

Lemma 6. *When $\lesssim$ is a preorder, $>$ and $<$ are asymmetric.*

Proof. For $>$, this is because $P_1 > P_2 \wedge P_2 > P_1 \Leftrightarrow (P_1 \not\lesssim P_2 \wedge P_2 \lesssim P_1) \wedge (P_2 \not\lesssim P_1 \wedge P_1 \lesssim P_2) \Leftrightarrow (P_1 \not\lesssim P_2 \wedge P_1 \lesssim P_2) \wedge (P_2 \not\lesssim P_1 \wedge P_2 \lesssim P_1) \Leftrightarrow false \wedge false \Leftrightarrow false$. For $<$, the proof is similar. □

Lemma 7. *When $\lesssim$ is a preorder, $>$ and $<$ are transitive.*

Proof. This results from the transitivity of $\lesssim$ (*cf*. Eqs. (8) and (9)). □

Lemma 8. *When $\lesssim$ is a preorder, $\lessgtr$ is irreflexive.*

Proof. This results from the reflexivity of $\lesssim$ and from Eq. (4): $P \lessgtr P \Leftrightarrow P \not\lesssim P \wedge P \not\lesssim P \Leftrightarrow false \wedge false \Leftrightarrow false$. □

Lemma 9. *When $\lesssim$ is a preorder, $\lessgtr$ is symmetric.*

Proof. This results from the fact that the conjunction is symmetric and from Eq. (4): $P_1 \lessgtr P_2 \Leftrightarrow P_1 \not\lesssim P_2 \wedge P_2 \not\lesssim P_1 \Leftrightarrow P_2 \not\lesssim P_1 \wedge P_1 \not\lesssim P_2 \Leftrightarrow P_2 \lessgtr P_1$. □

Lemma 10. *When $\lesssim$ is a preorder, $\lesssim$ and $\gtrsim$ are converse.*

Proof. From Eqs. (5) and (6), as $>$ and $<$ are converse, we have $P_1 \gtrsim P_2 \Leftrightarrow P_1 > P_2 \vee P_1 \sim P_2 \Leftrightarrow P_2 < P_1 \vee P_2 \sim P_1 \Leftrightarrow P_2 \lesssim P_1$. □

Lemma 11. *When $\lesssim$ is a preorder, $\lesssim$ and $\gtrsim$ are reflexive.*

Proof. For $\lesssim$, it is by definition of preorders. For $\gtrsim$, from Eqs. (1), (2) and (6), we have $P \gtrsim P \Leftrightarrow P > P \vee P \sim P \Leftrightarrow (P \not\lesssim P \wedge P \lesssim P) \vee (P \lesssim P \wedge P \lesssim P) \Leftrightarrow (false \wedge true) \vee (true \wedge true) \Leftrightarrow true$. □

Lemma 12. *When $\lesssim$ is a preorder, $\lesssim$ and $\gtrsim$ are transitive.*

Proof. For $\lesssim$, it is by definition of preorders. For $\gtrsim$, as $\lesssim$ and $\gtrsim$ are converse, $P_1 \gtrsim P_2 \wedge P_2 \gtrsim P_3 \Leftrightarrow P_3 \lesssim P_2 \wedge P_2 \lesssim P_1 \Rightarrow P_3 \lesssim P_1 \Leftrightarrow P_1 \gtrsim P_3$. □

A.4. Supplementary Material about Sec. 3.6

A.4.1 The Visual Inspection of Formulas, to Determine the Importance given by Scores, can be Misleading!

Let us consider the example of two-class classification, with $P(\{tn\})$, $P(\{fp\})$, $P(\{fn\})$, and $P(\{tp\})$ denoting, respectively, the probability (or proportion) of true negatives, false positives, false negatives, and true positives. Here are two classical scores, the accuracy and the true positive rate:

$$A = P(\{tn\}) + P(\{tp\}) \qquad TPR = \frac{P(\{tp\})}{P(\{fn\}) + P(\{tp\})}$$

The formula for the accuracy gives the illusion that the same importance is given to $\{tn\}$ and $\{tp\}$ and that no importance at all is given to $\{fp\}$ and $\{fn\}$. For the true positive rate, the formula might give the impression that the same importance is given to $\{fn\}$ and $\{tp\}$ and that no importance at all is given to $\{tn\}$ and $\{fp\}$. In fact, the visual inspection of formulas like these is not reliable at all to judge the importance given by a score to the various events. To see it, consider rewriting the previous equations as

$$A = (1-\alpha)P(\{tn\}) - \alpha P(\{fp\}) - \alpha P(\{fn\}) + (1-\alpha)P(\{tp\}) + \alpha \qquad \forall \alpha$$
$$TPR = \frac{-\alpha P(\{tn\}) - \alpha P(\{fp\}) - \alpha P(\{fn\}) + (1-\alpha)P(\{tp\}) + \alpha}{-\beta P(\{tn\}) - \beta P(\{fp\}) + (1-\beta)P(\{fn\}) + (1-\beta)P(\{tp\}) + \beta} \qquad \forall \alpha, \beta$$

A visual inspection of such formulas would lead, indeed, to other illusions about the events that have no importance. This observation, however, should not stop us from thinking in terms of importance. In fact, we do it in this paper, but we do it in a mathematical framework that allows to do it rigorously.

A.4.2 So, How can we Determine the Importance given by Scores?

One cannot determine the application-specific preferences implicitly considered by any score X. However, one can determine those that are consistent, or have the best consistency, with X. We detail these notions hereafter.

For a given set of performances. Consider a score X and a ranking score R_I. If, for a given set $\Pi \subseteq \mathbb{P}_{(\Omega,\Sigma)}$, the scores X and R_I are linked by a strict monotonic relationship on $\Pi \cap \mathrm{dom}(X) \cap \mathrm{dom}(R_I)$ and if $\Pi \cap \mathrm{dom}(X) = \Pi \cap \mathrm{dom}(R_I)$, then the performance orderings $\lesssim_X$ and $\lesssim_{R_I}$ (induced by X and R_I in the way specified in Theorem 1) are identical on Π. We say that the score X *is consistent with* the application-specific preferences I, on this set. A score can be consistent with different importance values (*e.g.*, as consequence of Properties 3 and 4).

For a given distribution of performances. Let $\mathring{\mathbb{P}} = \{P \in \mathbb{P}_{(\Omega,\Sigma)} : P(\{\omega\}) > 0 \forall \omega \in \Omega)\}$. All ranking scores are defined on this set. Consider a score X and the set $\Pi = \mathrm{dom}(X) \cap \mathring{\mathbb{P}}$. We say that the score X *has the best consistency with* the application-specific preferences I when I maximizes the rank correlation between X and R_I, on Π, for the given distribution of performances. See Sec. A.7.2 for computational details.

A.5. Supplementary Material about Sec. 4.2

A.5.1 Proof of Theorem 1.

For convenience, we provide a reminder of Theorem 1 and Axiom 1 below.

Theorem 1 (Sufficient condition for Axiom 1)**.** *A binary relation $\lesssim_X$ on $\mathbb{P}_{(\Omega,\Sigma)}$ induced by a score X as $P_1 \lesssim_X P_2$ iif either $P_1 = P_2$ or $P_1 \in \mathrm{dom}(X)$ and $P_2 \in \mathrm{dom}(X)$ and $X(P_1) \leq X(P_2)$, is a preorder satisfying Axiom 1.*

Axiom 1. *The ranking function $\mathrm{rank}_{\mathbb{E}} : \mathbb{E} \to [1, |\mathbb{E}|] : \epsilon \mapsto \mathrm{rank}_{\mathbb{E}}(\epsilon)$ satisfies $|\{\epsilon' \in \mathbb{E} : \mathrm{eval}(\epsilon) < \mathrm{eval}(\epsilon')\}| + 1 \leq \mathrm{rank}_{\mathbb{E}}(\epsilon) \leq |\{\epsilon' \in \mathbb{E} : \mathrm{eval}(\epsilon) \lesssim \mathrm{eval}(\epsilon')\}|$, where $\lesssim$ is a preorder on $\mathbb{P}_{(\Omega,\Sigma)}$.*

Proof. To establish that $\lesssim_X$ is a preorder, we have to show that it is (1) reflexive and (2) transitive.

(1) The reflexivity of $\lesssim_X$ is trivial to establish, since $P_1 = P_2 \Rightarrow P_1 \lesssim_X P_2$.

(2) The transitivity of $\lesssim_X$ can be shown as follows. $P_1 \lesssim_X P_2 \wedge P_2 \lesssim_X P_3$ implies that:

- either $P_1 \in \mathrm{dom}(X)$, $P_2 \in \mathrm{dom}(X)$, $P_3 \in \mathrm{dom}(X)$, and $X(P_1) \leq X(P_2) \wedge X(P_2) \leq X(P_3) \Rightarrow X(P_1) \leq X(P_3) \Rightarrow P_1 \lesssim_X P_3$;
- or $P_1 \notin \mathrm{dom}(X)$, $P_2 \notin \mathrm{dom}(X)$, $P_3 \notin \mathrm{dom}(X)$, and $P_1 = P_2 \wedge P_2 = P_3 \Rightarrow P_1 = P_3 \Rightarrow P_1 \lesssim_X P_3$.

We conclude that, in all cases, $P \lesssim_X P$ and $P_1 \lesssim_X P_2 \wedge P_2 \lesssim_X P_3 \Rightarrow P_1 \lesssim_X P_3$. The orderings $\lesssim_X$ induced by scores X are thus preorders. □

Summary. If the homogeneous binary relations $\sim$, $>$, $<$, and $\lesseqgtr$ on $\mathbb{P}_{(\Omega,\Sigma)}$ are derived from the ordering $\lesssim_X$ as explained above, and if the $\lesssim_X$ is derived from the score X, then the comparison between performances P_1 and P_2 can be summarized as follows.

	$P_1 \in \mathrm{dom}(X)$	$P_1 \notin \mathrm{dom}(X)$
$P_2 \in \mathrm{dom}(X)$	$X(P_1) < X(P_2) \Leftrightarrow P_1 < P_2$ $X(P_1) = X(P_2) \Leftrightarrow P_1 \sim P_2$ $X(P_1) > X(P_2) \Leftrightarrow P_1 > P_2$	$P_1 \lesseqgtr P_2$
$P_2 \notin \mathrm{dom}(X)$	$P_1 \lesseqgtr P_2$	$P_1 = P_2 \Leftrightarrow P_1 \sim P_2$ $P_1 \neq P_2 \Leftrightarrow P_1 \lesseqgtr P_2$

A.5.2 Proof of Theorem 2

For convenience, we provide a reminder of Theorem 2 and Axiom 2 below.

Theorem 2 (Sufficient condition for Axiom 2)**.** *If a score X satisfies $\min_{\omega \in E} S(\omega) \leq X(P) \leq \max_{\omega \in E} S(\omega)$ for all events $E \in \Sigma$ and all performances $P \in \mathrm{dom}(X)$ such that $P(E) = 1$, then the ordering $\lesssim_X$ satisfies Axiom 2.*

Axiom 2. *For $P_1, P_2 \in \mathbb{P}_{(\Omega,\Sigma)}$ such that $P_1(S \leq s) = 1$ and $P_2(S \geq s) = 1$ for some s, then $P_1 \lesssim P_2$ or $P_1 \lesseqgtr P_2$.*

Proof. Axiom 2 is satisfied when $P_1 \notin \mathrm{dom}(X)$ or $P_2 \notin \mathrm{dom}(X)$.

- Either $P_1 = P_2 \Leftrightarrow P_1 \sim P_2 \Rightarrow P_1 \lesssim P_2$,
- or $P_1 \neq P_2 \Leftrightarrow P_1 \lesseqgtr P_2$.

Axiom 2 is also satisfied when $P_1 \in \mathrm{dom}(X)$ and $P_2 \in \mathrm{dom}(X)$.

- On the one hand, the axiom stipulates that the event $E_1 = \{\omega \in \Omega : S(\omega) \leq s\}$ and the performance P_1 are such that $P_1(E_1) = 1$. Trivially, we have $\max_{\omega \in E_1} S(\omega) \leq s$. On the other hand, the theorem stipulates that, as $P_1(E_1) = 1$, $X(P_1) \leq \max_{\omega \in E_1} S(\omega)$. Putting all together, we have $X(P_1) \leq s$.
- On the one hand, the axiom stipulates that the event $E_2 = \{\omega \in \Omega : S(\omega) \geq s\}$ and the performance P_2 are such that $P_2(E_2) = 1$. Trivially, we have $s \leq \min_{\omega \in E_2} S(\omega)$. On the other hand, the theorem stipulates that, as $P_2(E_2) = 1$, $\min_{\omega \in E_2} S(\omega) \leq X(P_2)$. Putting all together, we have $s \leq X(P_2)$.
- As we have established that $X(P_1) \leq s$ and $s \leq X(P_2)$, we have $X(P_1) \leq X(P_2) \Leftrightarrow P_1 \lesssim P_2$.

□

A.5.3 Proof of Theorem 3

For convenience, we provide a reminder of Theorem 3 and Axiom 3 below.

Theorem 3 (Sufficient condition for Axiom 3)**.** *If a score X is such that $\Pi \subseteq \mathrm{dom}(X) \Rightarrow \Phi(\Pi) \subseteq \mathrm{dom}(X)$ and $\min_{P\in\Pi} X(P) \leq X(\overline{P}) \leq \max_{P\in\Pi} X(P)$ for all $\Pi \subseteq \mathrm{dom}(X)$ and all $\overline{P} \in \Phi(\Pi)$, then the ordering $\lesssim_X$ satisfies Axiom 3.*

Axiom 3. *Let P be a performance, and Π be a set of performances on $\mathbb{P}_{(\Omega,\Sigma)}$ such that $P' \lesssim P \vee P \lesssim P' \forall P' \in \Pi$.*
- *$P' \lesssim P \forall P' \in \Pi \Rightarrow \overline{P} \lesssim P \forall \overline{P} \in \Phi(\Pi)$;*
- *$P' \not\lesssim P \forall P' \in \Pi \Rightarrow \overline{P} \not\lesssim P \forall \overline{P} \in \Phi(\Pi)$;*
- *$P \lesssim P' \forall P' \in \Pi \Rightarrow P \lesssim \overline{P} \forall \overline{P} \in \Phi(\Pi)$;*
- *and $P \not\lesssim P' \forall P' \in \Pi \Rightarrow P \not\lesssim \overline{P} \forall \overline{P} \in \Phi(\Pi)$.*

Proof. We take $\lesssim = \lesssim_X$ and $\Pi \neq \emptyset$.
Remainder of the conditions. The first condition of Theorem 3 is:

$$\Pi \subseteq \mathrm{dom}(X) \Rightarrow \Phi(\Pi) \subseteq \mathrm{dom}(X)\,. \tag{15}$$

The second condition of Theorem 3 is:

$$\min_{P\in\Pi} X(P) \leq X(\overline{P}) \leq \max_{P\in\Pi} X(P) \qquad \forall \Pi \subseteq \mathrm{dom}(X) \qquad \forall \overline{P} \in \Phi(\Pi)\,. \tag{16}$$

The condition of Axiom 3 is that P is comparable to all performances in the set Π:

$$P' \lesssim P \vee P \lesssim P' \qquad \forall P' \in \Pi\,. \tag{17}$$

On the domain of X. By Theorem 1, this last condition implies that

$$P \in \mathrm{dom}(X)\,, \tag{18}$$

and

$$\Pi \subseteq \mathrm{dom}(X)\,. \tag{19}$$

Taking Eq. (15) and Eq. (19) together, we have

$$\Phi(\Pi) \subseteq \mathrm{dom}(X) \qquad \Leftrightarrow \qquad \overline{P} \in \mathrm{dom}(X) \qquad \forall \overline{P} \in \Phi(\Pi)\,. \tag{20}$$

Proof that $P' \lesssim P \forall P' \in \Pi \Rightarrow \overline{P} \lesssim P \forall \overline{P} \in \Phi(\Pi)$. On the one hand, we have, by Theorem 1,

$$\begin{aligned} P' \lesssim P \forall P' \in \Pi \Leftrightarrow & X(P') \leq X(P) \qquad \forall P' \in \Pi \\ \Leftrightarrow & \max_{P'\in\Pi} X(P') \leq X(P)\,. \end{aligned}$$

On the other hand, Eq. (16) implies that

$$X(\overline{P}) \leq \max_{P'\in\Pi} X(P') \qquad \forall \overline{P} \in \Phi(\Pi)\,.$$

Considering the last two equations together, we obtain

$$\begin{aligned} & X(\overline{P}) \leq \max_{P'\in\Pi} X(P') \leq X(P) \qquad \forall \overline{P} \in \Phi(\Pi) \\ \Rightarrow & X(\overline{P}) \leq X(P) \qquad \forall \overline{P} \in \Phi(\Pi)\,. \end{aligned}$$

By Theorem 1, we have thus $\overline{P} \lesssim P$.
Proof that $P' \not\lesssim P \forall P' \in \Pi \Rightarrow \overline{P} \not\lesssim P \forall \overline{P} \in \Phi(\Pi)$. On the one hand, we have, by Eq. (17) and Theorem 1,

$$\begin{aligned} P' \not\lesssim P \forall P' \in \Pi \Rightarrow & P < P' \forall P' \in \Pi \\ \Leftrightarrow & X(P) < X(P') \qquad \forall P' \in \Pi \\ \Leftrightarrow & X(P) < \min_{P'\in\Pi} X(P') \end{aligned}$$

On the other hand, Eq. (16) implies that

$$\min_{P' \in \Pi} X(P') \leq X(\overline{P}) \qquad \forall\, \overline{P} \in \Phi(\Pi)$$

Considering the last two equations together, we obtain

$$\begin{aligned} & X(P) < \min_{P' \in \Pi} X(P') \leq X(\overline{P}) \qquad \forall\, \overline{P} \in \Phi(\Pi) \\ \Rightarrow & X(P) < X(\overline{P}) \qquad \forall\, \overline{P} \in \Phi(\Pi)\,. \end{aligned}$$

By Theorem 1, we have thus $P < \overline{P}$, and by Eq. (17), $\overline{P} \not\lesssim P$.

Proof that $P \lesssim P' \,\forall P' \in \Pi \Rightarrow P \lesssim \overline{P} \,\forall \overline{P} \in \Phi(\Pi)$**.** On the one hand, we have, by Theorem 1,

$$\begin{aligned} P \lesssim P' \,\forall P' \in \Pi \Leftrightarrow & X(P) \leq X(P') \qquad \forall\, P' \in \Pi \\ \Leftrightarrow & X(P) \leq \min_{P' \in \Pi} X(P')\,. \end{aligned}$$

On the other hand, Eq. (16) implies that

$$\min_{P' \in \Pi} X(P') \leq X(\overline{P}) \qquad \forall\, \overline{P} \in \Phi(\Pi)\,.$$

Considering the last two equations together, we obtain

$$\begin{aligned} & X(P) \leq \min_{P' \in \Pi} X(P') \leq X(\overline{P}) \qquad \forall\, \overline{P} \in \Phi(\Pi) \\ \Rightarrow & X(P) \leq X(\overline{P}) \qquad \forall\, \overline{P} \in \Phi(\Pi)\,. \end{aligned}$$

By Theorem 1, we have thus $P \lesssim \overline{P}$.

Proof that $P \not\lesssim P' \,\forall P' \in \Pi \Rightarrow P \not\lesssim \overline{P} \,\forall \overline{P} \in \Phi(\Pi)$**.** On the one hand, we have, by Eq. (17) and Theorem 1,

$$\begin{aligned} P \not\lesssim P' \,\forall\, P' \in \Pi \Rightarrow & P' < P \,\forall\, P' \in \Pi \\ \Leftrightarrow & X(P') < X(P) \qquad \forall\, P' \in \Pi \\ \Leftrightarrow & \max_{P' \in \Pi} X(P') < X(P) \end{aligned}$$

On the other hand, Eq. (16) implies that

$$X(\overline{P}) \leq \max_{P' \in \Pi} X(P') \qquad \forall\, \overline{P} \in \Phi(\Pi)$$

Considering the last two equations together, we obtain

$$\begin{aligned} & X(\overline{P}) \leq \max_{P' \in \Pi} X(P') < X(P) \qquad \forall\, \overline{P} \in \Phi(\Pi) \\ \Rightarrow & X(\overline{P}) < X(P) \qquad \forall\, \overline{P} \in \Phi(\Pi)\,. \end{aligned}$$

By Theorem 1, we have thus $\overline{P} < P$, and by Eq. (17), $P \not\lesssim \overline{P}$.

□

A.6. Supplementary Material about Sec. 4.3

A.6.1 All Ranking Scores can be Used to Rank Performances (for $\Phi = \text{conv}$)

To show that all ranking scores can be used to rank performances, for $\Phi = \text{conv}$, we show that these scores satisfy the conditions of Theorems 1, 2, and 3.

All ranking scores satisfy the conditions of Theorem 1, and thus Axiom 1. For convenience, we provide a reminder of Theorem 1 and Axiom 1 below.

Theorem 1 (Sufficient condition for Axiom 1)**.** *A binary relation $\lesssim_X$ on $\mathbb{P}_{(\Omega,\Sigma)}$ induced by a score X as $P_1 \lesssim_X P_2$ iif either $P_1 = P_2$ or $P_1 \in \text{dom}(X)$ and $P_2 \in \text{dom}(X)$ and $X(P_1) \leq X(P_2)$, is a preorder satisfying Axiom 1.*

Axiom 1. *The ranking function $\text{rank}_{\mathbb{E}} : \mathbb{E} \to [1, |\mathbb{E}|] : \epsilon \mapsto \text{rank}_{\mathbb{E}}(\epsilon)$ satisfies $|\{\epsilon' \in \mathbb{E} : \text{eval}(\epsilon) < \text{eval}(\epsilon')\}| + 1 \leq \text{rank}_{\mathbb{E}}(\epsilon) \leq |\{\epsilon' \in \mathbb{E} : \text{eval}(\epsilon) \lesssim \text{eval}(\epsilon')\}|$, where $\lesssim$ is a preorder on $\mathbb{P}_{(\Omega,\Sigma)}$.*

Theorem 4. *All ranking scores satisfy the conditions of Theorem 1.*

Proof. For all ranking scores R_I, it is possible to induce an ordering $\lesssim_{R_I}$ satisfying the requirements of Theorem 1. □

All ranking scores satisfy the conditions of Theorem 2, and thus Axiom 2. For convenience, we provide a reminder of Theorem 2 and Axiom 2 below.

Theorem 2 (Sufficient condition for Axiom 2)**.** *If a score X satisfies $\min_{\omega \in E} S(\omega) \leq X(P) \leq \max_{\omega \in E} S(\omega)$ for all events $E \in \Sigma$ and all performances $P \in \text{dom}(X)$ such that $P(E) = 1$, then the ordering $\lesssim_X$ satisfies Axiom 2.*

Axiom 2. *For $P_1, P_2 \in \mathbb{P}_{(\Omega,\Sigma)}$ such that $P_1(S \leq s) = 1$ and $P_2(S \geq s) = 1$ for some s, then $P_1 \lesssim P_2$ or $P_1 \precnsim P_2$.*

Theorem 5. *All ranking scores satisfy the conditions of Theorem 2.*

Proof. We take $X = R_I$. When $P(E) = 1$, we have

$$R_I(P) = \frac{\sum_{\omega \in \Omega} I(\omega) S(\omega) P(\{\omega\})}{\sum_{\omega \in \Omega} I(\omega) P(\{\omega\})} = \frac{\sum_{\omega \in E} I(\omega) S(\omega) P(\{\omega\})}{\sum_{\omega \in E} I(\omega) P(\{\omega\})}$$

with $\sum_{\omega \in \Omega} I(\omega) P(\{\omega\}) > 0$ when $P \in \text{dom}(R_I)$.

- Let $M = \max_{\omega \in E} S(\omega)$. We have

$$\begin{aligned}
& S(\omega) \leq M \qquad \forall \omega \in E \\
\Leftrightarrow & S(\omega) - M \leq 0 \qquad \forall \omega \in E \\
\Rightarrow & \sum_{\omega \in E} I(\omega) \left[S(\omega) - M\right] P(\{\omega\}) \leq 0 \qquad \text{as } I(\omega) \geq 0 \text{ and } P(\{\omega\}) \geq 0 \\
\Leftrightarrow & \sum_{\omega \in E} I(\omega) S(\omega) P(\{\omega\}) \leq \sum_{\omega \in E} I(\omega) M P(\{\omega\}) \\
\Leftrightarrow & \sum_{\omega \in E} I(\omega) S(\omega) P(\{\omega\}) \leq M \underbrace{\sum_{\omega \in E} I(\omega) P(\{\omega\})}_{>0} \\
\Leftrightarrow & \frac{\sum_{\omega \in E} I(\omega) S(\omega) P(\{\omega\})}{\sum_{\omega \in E} I(\omega) P(\{\omega\})} \leq M \\
\Leftrightarrow & R_I(P) \leq M
\end{aligned}$$

- Let $m = \min_{\omega \in E} S(\omega)$. We have

$$\begin{aligned}
& S(\omega) \geq m \qquad \forall \omega \in E \\
\Leftrightarrow & S(\omega) - m \geq 0 \qquad \forall \omega \in E \\
\Rightarrow & \sum_{\omega \in E} I(\omega) \left[S(\omega) - m\right] P(\{\omega\}) \geq 0 \qquad \text{as } I(\omega) \geq 0 \text{ and } P(\{\omega\}) \geq 0 \\
\Leftrightarrow & \sum_{\omega \in E} I(\omega) S(\omega) P(\{\omega\}) \geq \sum_{\omega \in E} I(\omega) m P(\{\omega\}) \\
\Leftrightarrow & \sum_{\omega \in E} I(\omega) S(\omega) P(\{\omega\}) \geq m \underbrace{\sum_{\omega \in E} I(\omega) P(\{\omega\})}_{>0} \\
\Leftrightarrow & \frac{\sum_{\omega \in E} I(\omega) S(\omega) P(\{\omega\})}{\sum_{\omega \in E} I(\omega) P(\{\omega\})} \geq m \\
\Leftrightarrow & R_I(P) \geq m
\end{aligned}$$

- Putting all together, when $P(E) = 1$, we have $m \leq R_I(P) \leq M$, and so,

$$\min_{\omega \in E} S(\omega) \leq R_I(P) \leq \max_{\omega \in E} S(\omega)\,. \tag{21}$$

□

All ranking scores satisfy the conditions of Theorem 3, and thus Axiom 3 (for $\Phi = \mathrm{conv}$). For convenience, we provide a reminder of Theorem 3 and Axiom 3 below.

Theorem 3 (Sufficient condition for Axiom 3)**.** *If a score X is such that $\Pi \subseteq \mathrm{dom}(X) \Rightarrow \Phi(\Pi) \subseteq \mathrm{dom}(X)$ and $\min_{P \in \Pi} X(P) \leq X(\overline{P}) \leq \max_{P \in \Pi} X(P)$ for all $\Pi \subseteq \mathrm{dom}(X)$ and all $\overline{P} \in \Phi(\Pi)$, then the ordering $\lesssim_X$ satisfies Axiom 3.*

Axiom 3. *Let P be a performance, and Π be a set of performances on $\mathbb{P}_{(\Omega,\Sigma)}$ such that $P' \lesssim P \vee P \lesssim P' \forall P' \in \Pi$.*

- *$P' \lesssim P \forall P' \in \Pi \Rightarrow \overline{P} \lesssim P \forall \overline{P} \in \Phi(\Pi)$;*
- *$P' \not\lesssim P \forall P' \in \Pi \Rightarrow \overline{P} \not\lesssim P \forall \overline{P} \in \Phi(\Pi)$;*
- *$P \lesssim P' \forall P' \in \Pi \Rightarrow P \lesssim \overline{P} \forall \overline{P} \in \Phi(\Pi)$;*
- *and $P \not\lesssim P' \forall P' \in \Pi \Rightarrow P \not\lesssim \overline{P} \forall \overline{P} \in \Phi(\Pi)$.*

Theorem 6. *All ranking scores satisfy the conditions of Theorem 3 (for $\Phi = \mathrm{conv}$).*

Proof. The proof is in two parts.

- First, let us show that $\Pi \subseteq \mathrm{dom}(R_I) \Rightarrow \mathrm{conv}(\Pi) \subseteq \mathrm{dom}(R_I)$. For any $\overline{P} \in \mathrm{conv}(\Pi)$ there exists a weighting function $\lambda_{\Pi,\overline{P}} : \Pi \to \mathbb{R}_{\geq 0} : P \mapsto \lambda_{\Pi,\overline{P}}(P)$ such that $\sum_{P \in \Pi} \lambda_{\Pi,\overline{P}}(P) = 1$ and $\sum_{P \in \Pi} \lambda_{\Pi,\overline{P}}(P) P = \overline{P}$. For all $\overline{P} \in \mathrm{conv}(\Pi)$, we have:

$$\begin{aligned}
\Pi \subseteq \mathrm{dom}(R_I) & \Leftrightarrow \sum_{\omega \in \Omega} I(\omega) P(\{\omega\}) \neq 0 \qquad \forall P \in \Pi \\
& \Rightarrow \sum_{P \in \Pi} \lambda_{\Pi,\overline{P}}(P) \sum_{\omega \in \Omega} I(\omega) P(\{\omega\}) \neq 0 \\
& \Leftrightarrow \sum_{\omega \in \Omega} I(\omega) \sum_{P \in \Pi} \lambda_{\Pi,\overline{P}} P(\{\omega\}) \neq 0 \\
& \Leftrightarrow \sum_{\omega \in \Omega} I(\omega) \overline{P}(\{\omega\}) \neq 0 \\
& \Leftrightarrow \overline{P} \in \mathrm{dom}(R_I)\,.
\end{aligned}$$

- Second, let us show that, for all $\overline{P} \in \text{conv}(\Pi)$, $\min_{P\in\Pi} R_I(P) \leq R_I(\overline{P}) \leq \max_{P\in\Pi} R_I(P)$. Let us pose $l = \min_{P\in\Pi} R_I(P)$ and $u = \max_{P\in\Pi} R_I(P)$. We have:

$$
\begin{aligned}
& l \leq R_I(P) \leq u \qquad \forall P \in \Pi \\
\Leftrightarrow & l \leq \frac{\sum_{\omega\in\Omega} I(\omega)S(\omega)P(\{\omega\})}{\sum_{\omega\in\Omega} I(\omega)P(\{\omega\})} \leq u \qquad \forall P \in \Pi \\
\Leftrightarrow & l \sum_{\omega\in\Omega} I(\omega)P(\{\omega\}) \leq \sum_{\omega\in\Omega} I(\omega)S(\omega)P(\{\omega\}) \leq u \sum_{\omega\in\Omega} I(\omega)P(\{\omega\}) \qquad \forall P \in \Pi \\
\Rightarrow & l \sum_{\omega\in\Omega} I(\omega)\overline{P}(\{\omega\}) \leq \sum_{\omega\in\Omega} I(\omega)S(\omega)\overline{P}(\{\omega\}) \leq u \sum_{\omega\in\Omega} I(\omega)\overline{P}(\{\omega\}) \\
\Leftrightarrow & l \leq \frac{\sum_{\omega\in\Omega} I(\omega)S(\omega)\overline{P}(\{\omega\})}{\sum_{\omega\in\Omega} I(\omega)\overline{P}(\{\omega\})} \leq u \\
\Leftrightarrow & l \leq R_I(\overline{P}) \leq u
\end{aligned}
$$

□

A.6.2 On the Properties of Ranking Scores.

Proof of Property 1. Let us demonstrate that we have $R_I(P) = X_S^E(P')$ with $P' = \text{filter}_I(P)$. For all $\omega \in \Omega$, we have:

$$P'(\{\omega\}) = \frac{P(\{\omega\})I(\omega)}{\sum_{\omega'\in\Omega} P(\{\omega'\})I(\omega')}$$

Thus,

$$
\begin{aligned}
X_S^E(P') &= \sum_{\omega\in\Omega} P'(\{\omega\})S(\omega) \\
&= \sum_{\omega\in\Omega} \frac{P(\{\omega\})I(\omega)}{\sum_{\omega'\in\Omega} P(\{\omega'\})I(\omega')} S(\omega) \\
&= \frac{\sum_{\omega\in\Omega} P(\{\omega\})I(\omega)S(\omega)}{\sum_{\omega'\in\Omega} P(\{\omega'\})I(\omega')} \\
&= R_I(P)
\end{aligned}
$$

□

Proof of Property 2. Let $S' = \alpha S + \beta$ with $\alpha, \beta \in \mathbb{R}$.

$$\frac{\sum_{\omega\in\Omega} P(\{\omega\})S'(\omega)I(\omega)}{\sum_{\omega\in\Omega} P(\{\omega\})I(\omega)} = \alpha \frac{\sum_{\omega\in\Omega} P(\{\omega\})S(\omega)I(\omega)}{\sum_{\omega\in\Omega} P(\{\omega\})I(\omega)} + \beta$$

□

Proof of Property 3. $R_{kI} = \frac{\sum_{\omega\in\Omega} kI(\omega)S(\omega)P(\{\omega\})}{\sum_{\omega\in\Omega} kI(\omega)P(\{\omega\})} = \frac{\sum_{\omega\in\Omega} I(\omega)S(\omega)P(\{\omega\})}{\sum_{\omega\in\Omega} I(\omega)P(\{\omega\})} = R_I$ □

Proof of Property 4. Let us consider a binary satisfaction, that is $S(\omega) \in \{0, 1\}\, \forall \omega \in \Omega$. Let us define the events $E_0 = \{\omega \in \Omega : S(\omega) = 0\}$ and $E_1 = \{\omega \in \Omega : S(\omega) = 1\}$. If $I' = (\mathbf{1}_{S=0}\alpha_0 + \mathbf{1}_{S=1}\alpha_1)I$ with $\alpha_0 > 0$ and $\alpha_1 > 0$, then

$$
\begin{aligned}
R_I(P) &= \frac{\sum_{\omega\in\Omega} I(\omega)S(\omega)P(\{\omega\})}{\sum_{\omega\in\Omega} I(\omega)P(\{\omega\})} \\
&= \frac{\sum_{\omega\in E_1} I(\omega)P(\{\omega\})}{\sum_{\omega\in E_0} I(\omega)P(\{\omega\}) + \sum_{\omega\in E_1} I(\omega)P(\{\omega\})}
\end{aligned}
$$

and

$$\begin{aligned}
R_{I'}(P) &= \frac{\sum_{\omega\in\Omega} I'(\omega)S(\omega)P(\{\omega\})}{\sum_{\omega\in\Omega} I'(\omega)P(\{\omega\})} \\
&= \frac{\sum_{\omega\in E_1} I'(\omega)P(\{\omega\})}{\sum_{\omega\in E_0} I'(\omega)P(\{\omega\}) + \sum_{\omega\in E_1} I'(\omega)P(\{\omega\})} \\
&= \frac{\sum_{\omega\in E_1} \alpha_1 I(\omega)P(\{\omega\})}{\sum_{\omega\in E_0} \alpha_0 I(\omega)P(\{\omega\}) + \sum_{\omega\in E_1} \alpha_1 I(\omega)P(\{\omega\})} \\
&= \frac{\alpha_1 \sum_{\omega\in E_1} I(\omega)P(\{\omega\})}{\alpha_0 \sum_{\omega\in E_0} I(\omega)P(\{\omega\}) + \alpha_1 \sum_{\omega\in E_1} I(\omega)P(\{\omega\})}
\end{aligned}$$

Thus, $R_{I'} = \frac{\alpha_1 R_I}{\alpha_0(1-R_I)\alpha_1 R_I}$ and $\frac{\partial R_{I'}}{\partial R_I} = \frac{\alpha_0\alpha_1}{(\alpha_0(1-R_I)\alpha_1 R_I)^2} > 0$. This leads immediately to the conclusion that $\lesssim_{R_{I'}} = \lesssim_{R_I}$. □

Proof of Properties 5 and 6. Let us consider a binary satisfaction and the events $E_0 = \{\omega \in \Omega : S(\omega) = 0\}$ and $E_1 = \{\omega \in \Omega : S(\omega) = 1\}$. Let I_1 and I_2 be two random variables and $I = \lambda_1 I_1 + \lambda_2 I_2$ with $\lambda_1, \lambda_2 \in \mathbb{R}$ such that $\lambda_1 + \lambda_2 = 1$.

- When the random variables I_1 and I_2 are such that $I_1(\omega) = I_2(\omega) = I(\omega)\ \forall\omega \in E_1$, if we take $f : x \mapsto x^{-1}$,

$$\begin{aligned}
&\lambda_1 f\left(R_{I_1}(P)\right) + \lambda_2 f\left(R_{I_1}(P)\right) \\
=&\lambda_1 \left(1 + \frac{\sum_{\omega\in E_0} I_1(\omega)P(\{\omega\})}{\sum_{\omega\in E_1} I_1(\omega)P(\{\omega\})}\right) + \lambda_2 \left(1 + \frac{\sum_{\omega\in E_0} I_2(\omega)P(\{\omega\})}{\sum_{\omega\in E_1} I_2(\omega)P(\{\omega\})}\right) \\
=&\lambda_1 \left(1 + \frac{\sum_{\omega\in E_0} I_1(\omega)P(\{\omega\})}{\sum_{\omega\in E_1} I(\omega)P(\{\omega\})}\right) + \lambda_2 \left(1 + \frac{\sum_{\omega\in E_0} I_2(\omega)P(\{\omega\})}{\sum_{\omega\in E_1} I(\omega)P(\{\omega\})}\right) \\
=&(\lambda_1 + \lambda_2) + \frac{\lambda_1 \sum_{\omega\in E_0} I_1(\omega)P(\{\omega\}) + \lambda_2 \sum_{\omega\in E_0} I_2(\omega)P(\{\omega\})}{\sum_{\omega\in E_1} I(\omega)P(\{\omega\})} \\
=&1 + \frac{\sum_{\omega\in E_0} (\lambda_1 I_1 + \lambda_2 I_2)(\omega)P(\{\omega\})}{\sum_{\omega\in E_1} I(\omega)P(\{\omega\})} \\
=&1 + \frac{\sum_{\omega\in E_0} I(\omega)P(\{\omega\})}{\sum_{\omega\in E_1} I(\omega)P(\{\omega\})} \\
=&f\left(R_I(P)\right)
\end{aligned}$$

- When the random variables I_1 and I_2 are such that $I_1(\omega) = I_2(\omega) = I(\omega)\ \forall\omega \in E_0$, if we take $f : x \mapsto (1-x)^{-1}$,

$$\begin{aligned}
&\lambda_1 f\left(R_{I_1}(P)\right) + \lambda_2 f\left(R_{I_1}(P)\right) \\
=&\lambda_1 \left(1 + \frac{\sum_{\omega\in E_1} I_1(\omega)P(\{\omega\})}{\sum_{\omega\in E_0} I_1(\omega)P(\{\omega\})}\right) + \lambda_2 \left(1 + \frac{\sum_{\omega\in E_1} I_2(\omega)P(\{\omega\})}{\sum_{\omega\in E_0} I_2(\omega)P(\{\omega\})}\right) \\
=&\lambda_1 \left(1 + \frac{\sum_{\omega\in E_1} I_1(\omega)P(\{\omega\})}{\sum_{\omega\in E_0} I(\omega)P(\{\omega\})}\right) + \lambda_2 \left(1 + \frac{\sum_{\omega\in E_1} I_2(\omega)P(\{\omega\})}{\sum_{\omega\in E_0} I(\omega)P(\{\omega\})}\right) \\
=&(\lambda_1 + \lambda_2) + \frac{\lambda_1 \sum_{\omega\in E_1} I_1(\omega)P(\{\omega\}) + \lambda_2 \sum_{\omega\in E_1} I_2(\omega)P(\{\omega\})}{\sum_{\omega\in E_0} I(\omega)P(\{\omega\})} \\
=&1 + \frac{\sum_{\omega\in E_1} (\lambda_1 I_1 + \lambda_2 I_2)(\omega)P(\{\omega\})}{\sum_{\omega\in E_0} I(\omega)P(\{\omega\})} \\
=&1 + \frac{\sum_{\omega\in E_1} I(\omega)P(\{\omega\})}{\sum_{\omega\in E_0} I(\omega)P(\{\omega\})} \\
=&f\left(R_I(P)\right)
\end{aligned}$$

□

Proof of Property 7. Let $\mathcal{R} \in \{<, \leq, =, \geq, >\}$ and $v = R_I(P)$. For all $P' \in \mathbb{P}_{(\Omega,\Sigma)}$, we have:

$$
\begin{aligned}
& R_I(P')\mathcal{R}R_I(P) && (22)\\
\Leftrightarrow & R_I(P')\mathcal{R}v && (23)\\
\Leftrightarrow & \frac{\sum_{\omega\in\Omega} I(\omega)S(\omega)P'(\{\omega\})}{\sum_{\omega\in\Omega} I(\omega)P'(\{\omega\})}\mathcal{R}v && (24)\\
\Leftrightarrow & \left[\sum_{\omega\in\Omega} I(\omega)S(\omega)P'(\{\omega\})\right]\mathcal{R}\left[v\sum_{\omega\in\Omega} I(\omega)P'(\{\omega\})\right] && (25)\\
\Leftrightarrow & \sum_{\omega\in\Omega} I(\omega)\left[S(\omega)-v\right]P'(\{\omega\})\mathcal{R}0 && (26)
\end{aligned}
$$

This is either a linear equality or a linear inequality constraint. Thus,

$$
\begin{aligned}
\phi_{\mathcal{R}}(P) &= \left\{P' \in \mathbb{P}_{(\Omega,\Sigma)} : R_I(P')\mathcal{R}R_I(P)\right\} && (27)\\
&= \left\{P' \in \mathbb{P}_{(\Omega,\Sigma)} : \sum_{\omega\in\Omega} I(\omega)\left[S(\omega)-v\right]P'(\{\omega\})\mathcal{R}0\right\} && (28)
\end{aligned}
$$

is a convex subset of $\mathbb{P}_{(\Omega,\Sigma)}$. □

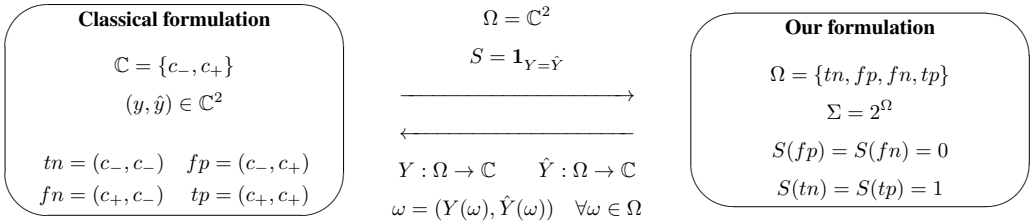


Figure A.7.1. Passages between two formulations (left: classical, right: ours) for the performance analysis of two-class classification problems.

A.7. Supplementary Material about Sec. 5.2

A.7.1 Link between Classical Formulation and Ours.

Fig. A.7.1 shows the connections between the classical formulation of the two-class classification task and our formulation, as explained in Sec. 5.

A.7.2 Custom Optimization Algorithm to Estimate Kendall's τ.

For any score X, our algorithm aims at determining the minimum and maximum values that a rank correlation between X and our ranking scores R_I can take over all possible importances I. Note that this algorithm is not specific to Kendall's τ [50] and could also be used with any other rank correlation, for example Spearman's ρ [51].

Variables. Leveraging Properties 3 and 4, we know that the rank-correlation between X and a ranking score R_{I_1} is equal to the rank-correlation between X and another ranking score R_{I_2} if $\frac{I_1(tp)}{I_1(tn)+I_1(tp)} = \frac{I_2(tp)}{I_2(tn)+I_2(tp)}$ and $\frac{I_1(fn)}{I_1(fp)+I_1(fn)} = \frac{I_2(fn)}{I_2(fp)+I_2(fn)}$. For this reason, we consider only two variables: $a = \frac{I(tp)}{I(tn)+I(tp)} \in [0,1]$ and $b = \frac{I(fn)}{I(fp)+I(fn)} \in [0,1]$.

Objective function. We optimize the function $\tau(a,b)$ that gives the rank correlation between X and R_{I^*} with $I^*(tp) = 1-a$, $I^*(tp) = 1-b$, $I^*(tp) = b$, and $I^*(tp) = a$. In practice, this is an estimation based on a finite set of performances on which X and R_{I^*} are applied. Note that $\tau(a,b)$ is not a continuous function when estimated on a finite set of performances. The chosen optimization technique circumvents the difficulties related to that.

Optimization technique. We implemented a custom coarse-to-fine grid-based direct search [56]: we compute $\tau(a,b)$ on a coarse grid over the unit square, locate the maximum on the grid, center a smaller square and a finer grid around that point, and iterate until the square is small enough.

A.7.3 Scores Perfectly Correlated with a Ranking Score, for all Performances

- **The accuracy:** $A = R_I$ with $I(tn) = 1/2$, $I(fp) = 1/2$, $I(fn) = 1/2$, and $I(tp) = 1/2$.
- **The F-score with** $\beta = 0.5$**:** $F_{0.5} = R_I$ with $I(tn) = 0$, $I(fp) = 4/5$, $I(fn) = 1/5$, and $I(tp) = 1$.
- **The F-score with** $\beta = 1.0$**:** $F_1 = R_I$ with $I(tn) = 0$, $I(fp) = 1/2$, $I(fn) = 1/2$, and $I(tp) = 1$.
- **The F-score with** $\beta = 2.0$**:** $F_2 = R_I$ with $I(tn) = 0$, $I(fp) = 1/5$, $I(fn) = 4/5$, and $I(tp) = 1$.
- **The negative predictive value:** $NPV = R_I$ with $I(tn) = 1$, $I(fp) = 0$, $I(fn) = 1$, and $I(tp) = 0$.
- **The positive predictive value:** $PPV = R_I$ with $I(tn) = 0$, $I(fp) = 1$, $I(fn) = 0$, and $I(tp) = 1$.
- **The true negative rate:** $TNR = R_I$ with $I(tn) = 1$, $I(fp) = 1$, $I(fn) = 0$, and $I(tp) = 0$.
- **The true positive rate:** $TPR = R_I$ with $I(tn) = 0$, $I(fp) = 0$, $I(fn) = 1$, and $I(tp) = 1$.

A.7.4 Scores Perfectly Correlated with a Ranking Score, for the Performances Corresponding to Given Class Priors $\pi_- \neq 0$ and $\pi_+ \neq 0$

- **The balanced accuracy:** $BA = R_I$ with $I(tn) = \pi_+$, $I(fp) = \pi_+$, $I(fn) = \pi_-$, and $I(tp) = \pi_-$.
- **Cohen's kappa:** $\kappa = \frac{R_I - 2\pi_-\pi_+}{\pi_-^2+\pi_+^2}$ with $I(tn) = \frac{\pi_+^2}{\pi_-^2+\pi_+^2}$, $I(fp) = \frac{1}{2}$, $I(fn) = \frac{1}{2}$, and $I(tp) = \frac{\pi_-^2}{\pi_-^2+\pi_+^2}$. Thus, $\frac{\partial \kappa}{R_I} > 0$.

- **The informedness (*a.k.a.* Youden's J):** $J_Y = 2R_I - 1$ with $I(tn) = \pi_+$, $I(fp) = \pi_+$, $I(fn) = \pi_-$, and $I(tp) = \pi_-$. Thus, $\frac{\partial J_Y}{R_I} > 0$.
- **The negative likelihood ratio:** $NLR = \frac{1-R_I}{R_I}$ with $I(tn) = 1$, $I(fp) = 0$, $I(fn) = 1$, and $I(tp) = 0$. Thus, $\frac{\partial NLR}{R_I} < 0$.
- **The positive likelihood ratio:** $PLR = \frac{R_I}{1-R_I}$ with $I(tn) = 0$, $I(fp) = 1$, $I(fn) = 0$, and $I(tp) = 1$. Thus, $\frac{\partial PLR}{R_I} > 0$.
- **The probability of the elementary event *true negative*:** $PTN = \pi_- R_I$ with $I(tn) = 1$, $I(fp) = 1$, $I(fn) = 0$, and $I(tp) = 0$. Thus, $\frac{\partial PTN}{R_I} > 0$.
- **The probability of the elementary event *true positive*:** $PTP = \pi_+ R_I$ with $I(tn) = 0$, $I(fp) = 0$, $I(fn) = 1$, and $I(tp) = 1$. Thus, $\frac{\partial PTP}{R_I} > 0$.

B. Supplementary material for the 3rd article

This appendix contains the supplementary material provided with my 3rd selected article [3], presented in Chapter 3.

A. Supplementary Material

This is the supplementary material for paper *What Is the Optimal Ranking Score Between Precision and Recall? We Can Always Find It and It Is Rarely F_1.*

Contents

A.1. Supplementary Material About Sec. 2.2 (Building Upon the Related Work) . . . 173
A.1.1. All F_β Lead to Meaningful Performance Orderings . . . 173
A.1.2. $SIVF$ Leads to a Meaningful Performance Ordering When the Class Priors Are Fixed . . . 173
A.1.3. $SIVF$ Leads to the Same Ranking as F_β with $\beta^2 = \frac{\pi_-}{\pi_+}$ . . . 173
A.2. Supplementary Material About Sec. 3 (Theory) . . . 175
A.2.1. On the Minimization of Fréchet Variance (Eq. (11)) . . . 175
A.2.2. On the Formulas for Degree of Optimality of Some F_β . . . 175
A.3. Supplementary Material About Sec. 4 (Case Studies) . . . 178
A.3.1. Principle of Analytical Computations for Kendall's rank correlations τ with Distributions of Performances 178
A.3.2. Detailed Results for the Uniform Distribution Over All Performances . . . 179
A.3.3. Detailed Results for the Uniform Distributions With Fixed Probability of True Negatives . . . 180
A.3.4. Detailed Results for the Uniform Distributions With Fixed Class Priors . . . 182
A.3.5. Detailed Results for the Uniform Distributions With Fixed Class Priors, Above No-Skill . . . 185
A.3.6. Detailed Results for the Uniform Distributions With Fixed Class Priors, Close to Oracle . . . 188
A.3.7. All Results for Some Real Sets of Performances . . . 191

A.1. Supplementary Material About Sec. 2.2 (Building Upon the Related Work)

A.1.1. All F_β Lead to Meaningful Performance Orderings

It has been shown in [2] that all performance orderings derived from some score of the form

$$R_I(P) = \frac{I(tn)P(\{tn\}) + I(tp)P(\{tp\})}{I(tn)P(\{tn\}) + I(fp)P(\{fp\}) + I(fn)P(\{fn\}) + I(tp)P(\{tp\})}\,, \tag{31}$$

where I denotes a positive random variable called *Importance*, satisfy all the axioms of the theory of performance-based ranking. These scores are known as *ranking scores*.

In order to prove that all F_β lead to meaningful performance orderings, we show here that the F_β are particular cases of R_I. The proofs for $\beta = 1/2$, $\beta = 1$, and $\beta = 2$ were already given in [2], so here we give a generalization. Our proof is based on the well known fact that F_β has two nearly equivalent definitions. The first one, that we have used in the introduction (Eq. (1)), is as follows:

$$F_\beta : \left\{P \in \mathbb{P}_{(\Omega,\Sigma)} : P(\{tp\}) \neq 0\right\} \to [0,1] : P \mapsto \left(\frac{1}{1+\beta^2} Pr^{-1}(P) + \frac{\beta^2}{1+\beta^2} Re^{-1}(P)\right)^{-1}\,. \tag{32}$$

The second one is

$$F_\beta : \left\{P \in \mathbb{P}_{(\Omega,\Sigma)} : P(\{tn\}) \neq 1\right\} \to [0,1] : P \mapsto \frac{(1+\beta^2)\,P(\{tp\})}{1\,P(\{fp\}) + \beta^2\,P(\{fn\}) + (1+\beta^2)\,P(\{tp\})}\,. \tag{33}$$

It turns out that the first definition is a restriction of the second one: $P(\{tp\}) \neq 0 \Rightarrow P(\{tn\}) \neq 1$ and both F_β are equal on $\left\{P \in \mathbb{P}_{(\Omega,\Sigma)} : P(\{tp\}) \neq 0\right\}$. It is now possible to compare Eq. (33) with Eq. (31) and to see that, for any given β, F_β is a particular case of ranking score with

$$(I(tn), I(fp), I(fn), I(tp)) \propto \left(0, 1, \beta^2, 1+\beta^2\right)\,, \tag{34}$$

or equivalently

$$(I(tn), I(fp), I(fn), I(tp)) \propto (0, 1-b, b, 1)\,, \tag{35}$$

with $b = \beta^2/1+\beta^2$.

A.1.2. $SIVF$ Leads to a Meaningful Performance Ordering When the Class Priors Are Fixed

The skew-insensitive version of F_1 as been defined in [29] as

$$SIVF = \frac{2TPR}{TPR + FPR + 1}\,. \tag{36}$$

When the class priors are fixed and such that $\pi_- \neq 0$ and $\pi_+ \neq 0$,

$$\begin{aligned} SIVF(P) &= \frac{2\frac{P(\{tp\})}{\pi_+}}{\frac{P(\{tp\})}{\pi_+} + \frac{P(\{fp\})}{\pi_-} + 1} = \frac{2\,\pi_-\,P(\{tp\})}{\pi_-\,P(\{tp\}) + \pi_+\,P(\{fp\}) + \pi_-\,\pi_+} \\ &= \frac{2\,\pi_-\,P(\{tp\})}{\pi_-\,P(\{tp\}) + \pi_+\,P(\{fp\}) + \pi_-\,(P(\{fn\}) + P(\{tp\}))} = \frac{2\,\pi_-\,P(\{tp\})}{\pi_+\,P(\{fp\}) + \pi_-\,P(\{fn\}) + 2\,\pi_-\,P(\{tp\})}\,. \end{aligned} \tag{37}$$

The comparison of Eq. (37) with Eq. (31) shows that $SIVF$ is a particular case of ranking score with

$$(I(tn), I(fp), I(fn), I(tp)) \propto \left(0, \pi_+, \pi_-, 2\,\pi_-\right)\,. \tag{38}$$

A.1.3. $SIVF$ Leads to the Same Ranking as F_β with $\beta^2 = \frac{\pi_-}{\pi_+}$

A first demonstration, based on the relationship between $SIVF$ and F_β. From [29], it is already known that $SIVF = F_1$ when $\pi_- = \pi_+ = 1/2$. Here, we give a generalization for $\pi_- \neq 0$ and $\pi_+ \neq 0$. We noticed that

$$\beta^2 = \frac{\pi_-}{\pi_+} \Rightarrow F_\beta = \frac{SIVF}{(\pi_+ - \pi_-)\,SIVF + 2\pi_-}\,. \tag{39}$$

When $\pi_- \neq 0$, we have

$$\beta^2 = \frac{\pi_-}{\pi_+} \Rightarrow \frac{\partial F_\beta}{\partial SIVF} > 0\,. \tag{40}$$

Therefore, when the priors are fixed, $SIVF$ leads to the same performance ordering as F_β with $\beta^2 = \frac{\pi_-}{\pi_+}$.

A second demonstration, based on the importance values. Property 4 of [2], when particularized for two-class classification, states that the performance ordering induced by a ranking score is insensitive to the uniform scaling of the importance given to the unsatisfying samples (fp and fn) or to the satisfying (tn and tp) samples. The comparison between Eq. (34) and Eq. (38) shows that:

- for the unsatisfying samples, we have $\left(\bullet, 1, \beta^2, \bullet\right) \propto \left(\bullet, \pi_+, \pi_-, \bullet\right)$ when $\beta^2 = \frac{\pi_-}{\pi_+}$;
- for the satisfying samples, we have always $\left(0, \bullet, \bullet, 1 + \beta^2\right) \propto \left(0, \bullet, \bullet, 2\,\pi_-\right)$.

Therefore, when the priors are fixed, $SIVF$ leads to the same performance ordering as F_β with $\beta^2 = \frac{\pi_-}{\pi_+}$.

A.2. Supplementary Material About Sec. 3 (Theory)

A.2.1. On the Minimization of Fréchet Variance (Eq. (11))

In Sec. 3.4.1, we say that the optimal tradeoffs are the F_β scores that minimize the Fréchet variance [64]:

$$\sigma^2(\beta) = d_\tau^2(Pr; F_\beta) + d_\tau^2(F_\beta; Re)\,.$$

We also say that the solutions, known as the Karcher means [13] are those that are equidistant of Pr and Re, *i.e.* such that

$$\begin{aligned} & d_\tau(Pr; F_*) = d_\tau(F_*; Re) = \frac{d_\tau(Pr; Re)}{2} \\ \Leftrightarrow\ & \tau(Pr; F_*) = \tau(F_*; Re) = \frac{1 + \tau(Pr; Re)}{2}\,. \end{aligned}$$

Hereafter, we discuss here the case of a continuum of rankings, as shown in Fig. A.2.1a. The case of a finite amount of rankings is depicted in Fig. A.2.1b and Fig. A.2.1c for, respectively, an odd and an even amount.

By restricting to the family of F_β scores, the solution is actually unique and is thus a Fréchet mean. Indeed, since Eq. (9) states that

$$d_\tau(Pr; Re) = d_\tau(Pr; F_\beta) + d_\tau(F_\beta; Re) \qquad \forall \beta \geq 0$$

we have

$$\sigma^2(\beta) = d_\tau^2(Pr; F_\beta) + (d_\tau(Pr; Re) - d_\tau(Pr; F_\beta))^2 \tag{41}$$

$$= 2d_\tau^2(Pr; F_\beta) - 2d_\tau(Pr; Re)d_\tau(Pr; F_\beta) + d_\tau^2(Pr; Re) \tag{42}$$

which is quadratic in $d_\tau(Pr; F_\beta)$ and thus uniquely minimized for β such that

$$d_\tau(Pr; F_\beta) = \frac{d_\tau(Pr; Re)}{2}\,. \tag{43}$$

This incidentally yields the same equality for $d_\tau(F_\beta; Re)$. The result in terms of correlation stems from the relation $\tau = 1 - 2d_\tau$.

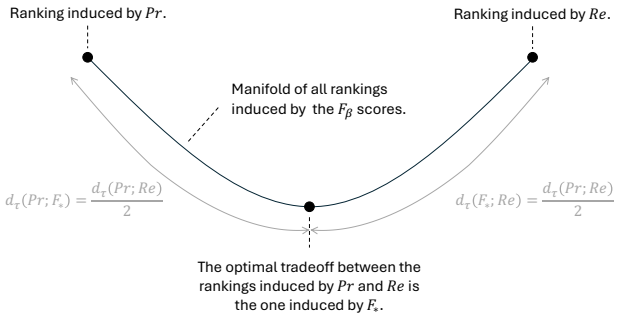
A.2.2. On the Formulas for Degree of Optimality of Some F_β

We provide here a few explanations about how Eqs. (16) to (18) cen be derived. For that, let us remind three interpretations for Kendall's distance $d_\tau(X_1; X_2) \in [0, 1]$.

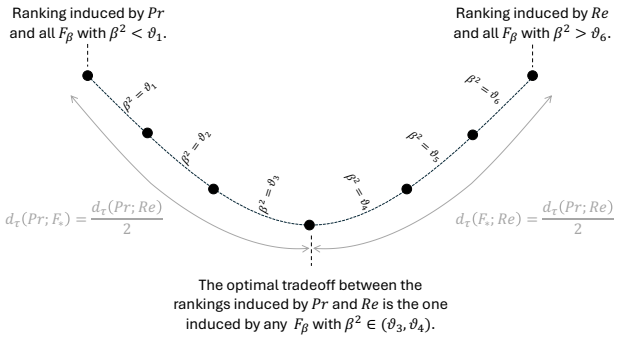
- First, Eq. (8) shows that $d_\tau(X_1; X_2)$ is equal to the proportion of pairs of elements for which X_1 and X_2 do not agree on the relative order. This interpretation is the cornerstone for probabilistic reasoning.
- Second, $d_\tau(X_1; X_2)$ is equal to the minimum number of swaps of consecutive elements that are needed to transform the ranking induced by X_1 into the ranking induced by X_2 and vice versa. This interpretation is the cornerstone for geometric reasoning, as it makes the connection with the distance along the manifold (or path graph) of rankings.
- Third, $d_\tau(X_1; X_2)$ is linearly related to the rank correlation $\tau(X_1; X_2)$ as $d_\tau = \frac{1-\tau}{2}$.

Keeping these three interpretations in mind is helpful to derive Eqs. (16) to (18).

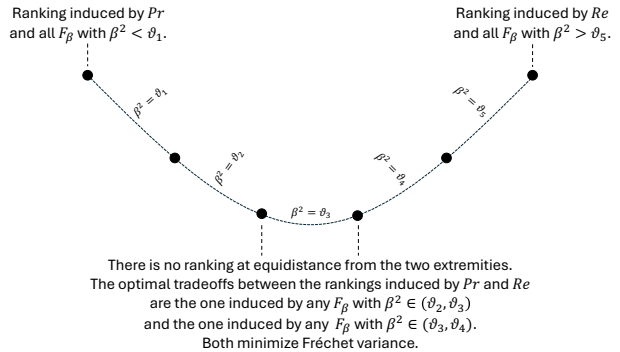
Let us consider Eq. (17), for example. We are interested in the probability $P(\text{✘})$. The first interpretation allows us to express it as $d_\tau(F_\beta; F_*)$. Then, it is possible to perform some geometric reasoning based on the second interpretation. As shown in Fig. A.2.2, $d_\tau(F_\beta; F_*) = 1/2 \left| d_\tau(F_\beta; Re) - d_\tau(Pr; F_\beta) \right|$. Finally, using the third interpretation, we find that $1/2 \left| d_\tau(F_\beta; Re) - d_\tau(Pr; F_\beta) \right| = 1/4 \left| \tau(Pr; F_\beta) - \tau(F_\beta; Re) \right|$.



(a) **Case in which there is a bijection between the values of β and the rankings induced by F_β.** The rankings form a continuum (manifold) and the minimization of Fréchet variance leads to the midpoint where $d_\tau(Pr; F_*) = d_\tau(F_*; Re)$.

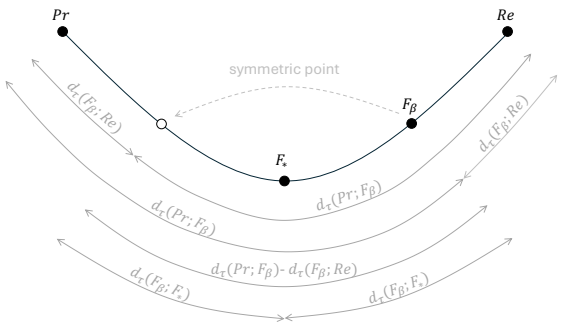


(b) **Case in which the number of rankings induced by the F_β scores is odd.** The rankings form a path graph. The nodes correspond to the rankings, each of them being induced by a range of values for β. The edges correspond to the swaps that occur between the consecutive rankings, at some given values ϑ for the β parameter. Assuming there is no group of at least three co-aligned performances, the median of all these ϑ values belongs to the range of βs for which the ranking is at the midpoint where $d_\tau(Pr; F_*) = d_\tau(F_*; Re)$.

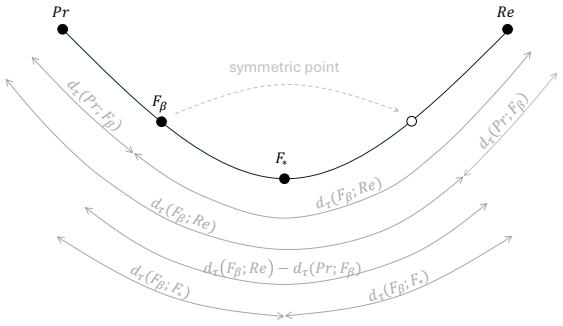


(c) **Case in which the number of rankings induced by the F_β scores is even.** Assuming there is no group of at least three co-aligned performances, the median of all the ϑ values corresponds to the swap between the two rankings minimizing Fréchet variance.

Figure A.2.1. The optimal tradeoff as the solution of the minimization of Fréchet variance $\sigma^2(\beta)$.



(a) Case in which the ranking induced by F_β is closer to the ranking induced by Re than to the ranking induced by Pr. We see that $d_\tau(F_\beta; F_*) = \frac{d_\tau(Pr;F_\beta) - d_\tau(F_\beta;Re)}{2}$.



(b) Case in which the ranking induced by F_β is closer to the ranking induced by Pr than to the ranking induced by Re. We see that $d_\tau(F_\beta; F_*) = \frac{d_\tau(F_\beta;Re) - d_\tau(Pr;F_\beta)}{2}$.

Figure A.2.2. Kendall distances on the manifold of rankings induced by the F_β scores. We see that $d_\tau(F_\beta; F_*) = \frac{\left|d_\tau(F_\beta;Re) - d_\tau(Pr;F_\beta)\right|}{2}$.

A.3. Supplementary Material About Sec. 4 (Case Studies)

A.3.1. Principle of Analytical Computations for Kendall's rank correlations τ with Distributions of Performances

Let us consider two scores X_1, X_2. We denote the probability for them to order differently two (different) performances P_A and P_B drawn at random, independently, by $d_\tau(X_1; X_2)$. Kendall's rank correlation is given by $\tau(X_1; X_2) = 1 - 2d_\tau(X_1; X_2)$. When we consider a distribution of performances $\mathcal{P}$, instead of a finite set Π, we cannot compute $d_\tau(X_1; X_2)$ with Eq. (8). Instead, we can compute analytically the following probabilities:

$$\text{Proba}\left[X_1(P_A) < X_1(P_B), X_2(P_A) < X_2(P_B)\right] = \int_{P_A \in \mathbb{P}_{(\Omega,\Sigma)}} \int_{P_B \in \mathbb{P}_{(\Omega,\Sigma)}} \mathbf{1}_{X_1(P_A)<X_1(P_B), X_2(P_A)<X_2(P_B)} f_{\mathcal{P}}(P_A) f_{\mathcal{P}}(P_B) dP_A dP_B \,, \tag{44}$$

$$\text{Proba}\left[X_1(P_A) < X_1(P_B), X_2(P_A) > X_2(P_B)\right] = \int_{P_A \in \mathbb{P}_{(\Omega,\Sigma)}} \int_{P_B \in \mathbb{P}_{(\Omega,\Sigma)}} \mathbf{1}_{X_1(P_A)<X_1(P_B), X_2(P_A)>X_2(P_B)} f_{\mathcal{P}}(P_A) f_{\mathcal{P}}(P_B) dP_A dP_B \,, \tag{45}$$

$$\text{Proba}\left[X_1(P_A) > X_1(P_B), X_2(P_A) < X_2(P_B)\right] = \int_{P_A \in \mathbb{P}_{(\Omega,\Sigma)}} \int_{P_B \in \mathbb{P}_{(\Omega,\Sigma)}} \mathbf{1}_{X_1(P_A)>X_1(P_B), X_2(P_A)<X_2(P_B)} f_{\mathcal{P}}(P_A) f_{\mathcal{P}}(P_B) dP_A dP_B \,, \tag{46}$$

$$\text{Proba}\left[X_1(P_A) > X_1(P_B), X_2(P_A) > X_2(P_B)\right] = \int_{P_A \in \mathbb{P}_{(\Omega,\Sigma)}} \int_{P_B \in \mathbb{P}_{(\Omega,\Sigma)}} \mathbf{1}_{X_1(P_A)>X_1(P_B), X_2(P_A)>X_2(P_B)} f_{\mathcal{P}}(P_A) f_{\mathcal{P}}(P_B) dP_A dP_B \,, \tag{47}$$

where the symbol $\mathbf{1}$ denotes the indicator and $f_{\mathcal{P}}$ the probability density function related to the distribution $\mathcal{P}$.

Kendall's rank correlation τ is then given by

$$\tau(X_1; X_2) = 1 - 2\left(\text{Proba}\left[X_1(P_A) < X_1(P_B), X_2(P_A) > X_2(P_B)\right] + \text{Proba}\left[X_1(P_A) > X_1(P_B), X_2(P_A) < X_2(P_B)\right]\right) . \tag{48}$$

Note that P_A and P_B are exchangeable in these equalities since they are drawn independently from the same distribution. Thus, by symmetry, we have

$$\text{Proba}\left[X_1(P_A) < X_1(P_B), X_2(P_A) < X_2(P_B)\right] = \text{Proba}\left[X_1(P_A) > X_1(P_B), X_2(P_A) > X_2(P_B)\right] , \tag{49}$$

and

$$\text{Proba}\left[X_1(P_A) < X_1(P_B), X_2(P_A) > X_2(P_B)\right] = \text{Proba}\left[X_1(P_A) > X_1(P_B), X_2(P_A) < X_2(P_B)\right] . \tag{50}$$

Moreover, the four probabilities sum to one. Thus, computing one of these four probabilities suffices to be able to determine τ, and if we know τ, then we are able to recover the four probabilities:

$$\tau(X_1; X_2) = 4\,\text{Proba}\left[X_1(P_A) < X_1(P_B), X_2(P_A) < X_2(P_B)\right] - 1 \,, \tag{51}$$

$$\tau(X_1; X_2) = 1 - 4\,\text{Proba}\left[X_1(P_A) < X_1(P_B), X_2(P_A) > X_2(P_B)\right] , \tag{52}$$

$$\tau(X_1; X_2) = 1 - 4\,\text{Proba}\left[X_1(P_A) > X_1(P_B), X_2(P_A) < X_2(P_B)\right] , \tag{53}$$

$$\tau(X_1; X_2) = 4\,\text{Proba}\left[X_1(P_A) > X_1(P_B), X_2(P_A) > X_2(P_B)\right] - 1 \,. \tag{54}$$

The source codes for the analytical results provided hereafter are for *Wolfram 14.2* (*a.k.a. Mathematica*).

A.3.2. Detailed Results for the Uniform Distribution Over All Performances

To find the probabilities necessary to compute τ, we integrate over the tetrahedron. We start with a few initializations.

```
Tetra = Simplex[{{1, 0, 0, 0}, {0, 1, 0, 0}, {0, 0, 1, 0}, {0, 0, 0, 1}}]
Ppv[ptn_, pfp_, pfn_, ptp_] := ptp/(pfp + ptp)
Tpr[ptn_, pfp_, pfn_, ptp_] := ptp/(pfn + ptp)
Fone[ptn_, pfp_, pfn_, ptp_] := 2*ptp/(pfp + pfn + 2*ptp)

tot = Integrate[
    1,
    {ptn1, pfp1, pfn1, ptp1} \[Element] Tetra,
    {ptn2, pfp2, pfn2, ptp2} \[Element] Tetra
]
```

We have

$$\tau(Pr; Re) = \frac{1}{3}, \tag{55}$$

as shown with the code:

```
pLtLt = Integrate[
    Boole[
        Ppv[ptn1, pfp1, pfn1, ptp1] < Ppv[ptn2, pfp2, pfn2, ptp2] &&
        Tpr[ptn1, pfp1, pfn1, ptp1] < Tpr[ptn2, pfp2, pfn2, ptp2]
    ],
    {ptn1, pfp1, pfn1, ptp1} \[Element] Tetra,
    {ptn2, pfp2, pfn2, ptp2} \[Element] Tetra
] / tot
tau = 4 * pLtLt - 1
```

We have

$$\tau(Pr; F_1) = \frac{2}{3}, \tag{56}$$

as shown with the code:

```
pLtGt = Integrate[
    Boole[
        Ppv[ptn1, pfp1, pfn1, ptp1] < Ppv[ptn2, pfp2, pfn2, ptp2] &&
        Fone[ptn1, pfp1, pfn1, ptp1] > Fone[ptn2, pfp2, pfn2, ptp2]
    ],
    {ptn1, pfp1, pfn1, ptp1} \[Element] Tetra,
    {ptn2, pfp2, pfn2, ptp2} \[Element] Tetra
] / tot
tau = 1 - 4 * pLtGt
```

And we have

$$\tau(F_1; Re) = \frac{2}{3}, \tag{57}$$

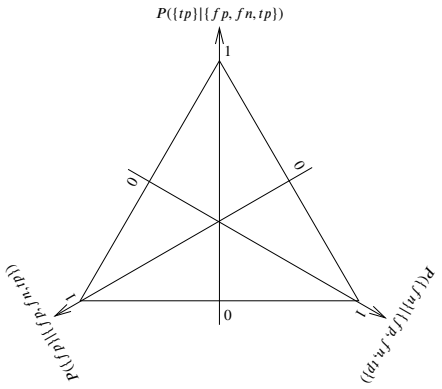
as shown with the code:

```
pLtGt = Integrate[
    Boole[
        Fone[ptn1, pfp1, pfn1, ptp1] < Fone[ptn2, pfp2, pfn2, ptp2] &&
        Tpr[ptn1, pfp1, pfn1, ptp1] > Tpr[ptn2, pfp2, pfn2, ptp2]
    ],
    {ptn1, pfp1, pfn1, ptp1} \[Element] Tetra,
    {ptn2, pfp2, pfn2, ptp2} \[Element] Tetra
] / tot
tau = 1 - 4 * pLtGt
```

A.3.3. Detailed Results for the Uniform Distributions With Fixed Probability of True Negatives

The demonstration could be provided similarly as we did for the previous case study, taking a simplex in two dimensions instead of a simplex in three dimensions. However, we would like to show a geometrical reasoning instead.

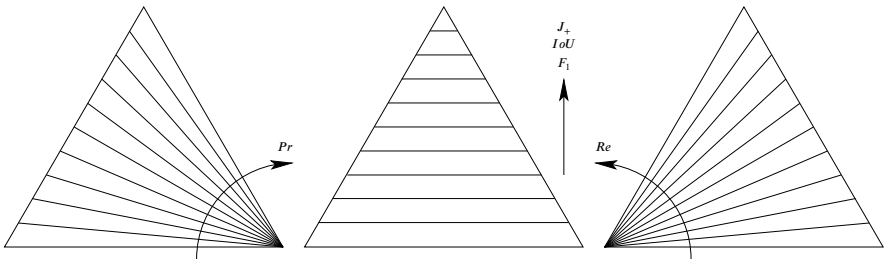
The equilateral triangles depicted in Fig. 3b form a "triangular space" in which the performances are assumed to be uniformly distributed. This "space" is parameterized as follows.



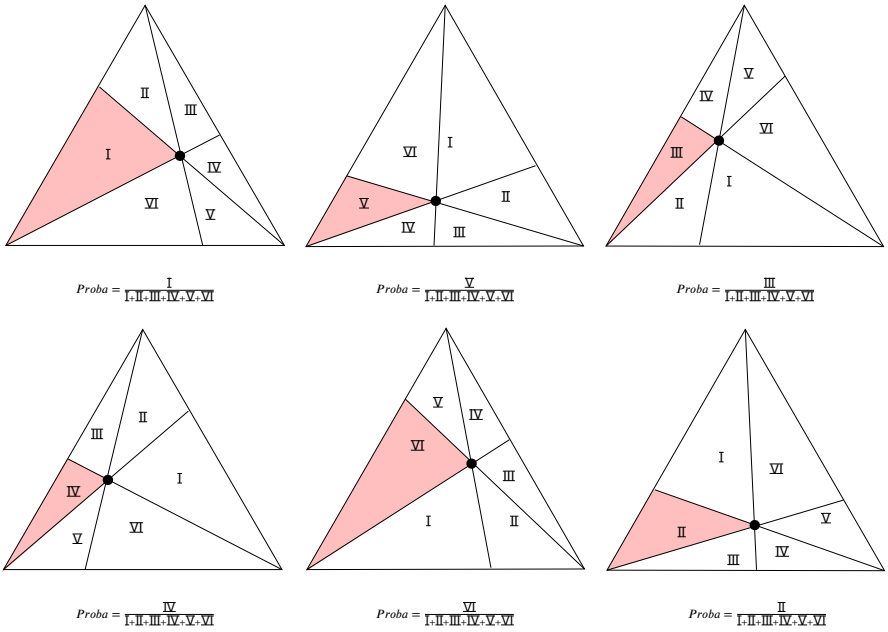
In fact, in the drawing shown here-above, the vertical axis corresponds to a score known as the *Intersection over Union* IoU or *Jaccard Coefficient* J_+. It is monotonically increasing with F_1:

$$F_1 = \frac{2IoU}{1 + IoU} \Rightarrow \frac{\partial F_1}{\partial IoU} > 0 \tag{58}$$

The isometrics of Pr form a pencil of lines whose vertex is located at the bottom-right corner. The isometrics of F_1 are horizontal lines. The isometrics of Re form a pencil of lines whose vertex is located at the bottom-left corner.



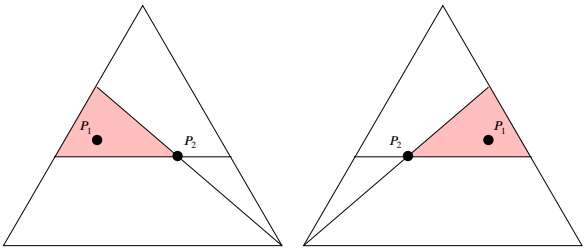
We now show that $\tau(Pr; Re) = 1/3$. In the following drawings, we consider a performance P_1 that can be anywhere in the "triangular space" and we choose one particular performance P_2, shown as a black point. The pink area represents the probability that $Pr(P_1) < Pr(P_2)$ and $Re(P_1) > Re(P_2)$. The six cases are symmetrical: one can obtain them by mirroring or rotating.



We see that, in average (that is, if we let P_2 be anywhere in the "triangular space"), the probability that $Pr(P_1) < Pr(P_2)$ and $Re(P_1) > Re(P_2)$ is $1/6$. Using Eq. (52), we find the announced result:

$$\tau(Pr; Re) = 1/3 \tag{59}$$

We now show that $\tau(Pr; F_1) = \tau(Re; F_1)$. Using Eq. (52), we know that it is the case if and only if the probability that $Pr(P_1) < Pr(P_2)$ and $F_1(P_1) > F_1(P_2)$ is equal to the probability that $Re(P_1) < Re(P_2)$ and $F_1(P_1) > F_1(P_2)$. These two probabilities are shown as pink areas, for some arbitrarily fixed P_2 (the black dot), in the following drawings.



We see that, by symmetry, the two probabilities are equal. If, furthermore, we do not fix anymore the performance P_2 but let it be anywhere in the "triangular space", the two probabilities remain equal. Using Eq. (52), we find the announced result:

$$\tau(Pr; F_1) = \tau(Re; F_1) \tag{60}$$

A.3.4. Detailed Results for the Uniform Distributions With Fixed Class Priors

To find the probabilities necessary to compute τ, we integrate over the whole ROC space. Let us denote by $(FPR, TPR) = (x, y)$ the coordinates of a performance in this space. We have

$$Pr(P_1) < Pr(P_2) \Leftrightarrow \frac{y_1}{x_1} < \frac{y_2}{x_2} \tag{61}$$

$$Re(P_1) < Re(P_2) \Leftrightarrow y_1 < y_2 \tag{62}$$

$$F_\beta(P_1) < F_\beta(P_2) \Leftrightarrow \frac{y_1}{x_1 + \ell} < \frac{y_2}{x_2 + \ell} \tag{63}$$

We start with a few initializations.

```
tot = Integrate[
    1,
    {x1, 0, 1}, {y1, 0, 1},
    {x2, 0, 1}, {y2, 0, 1}
]
```

We have

$$\tau(Pr; Re) = \frac{1}{2}, \tag{64}$$

as shown with the code (see Fig. A.3.3a):

```
pLtGt = Integrate[
        Boole[(y1 x2 < y2 x1) && (y1 > y2)],
        {x1, 0, 1}, {y1, 0, 1},
        {x2, 0, 1}, {y2, 0, 1}
] / tot
tau = 1 - 4 * pLtGt
```

We have

$$\tau(Pr; F_\beta) = 1 - \ell \left(\ell \log \left(\frac{\ell}{\ell + 1} \right) + 1 \right), \tag{65}$$

as shown with the code (see Fig. A.3.3b):

```
pLtGt = Integrate[
        Boole[(y1 x2 < y2 x1) && (y1 (x2 + l) > y2 (x1 + l))],
        {x1, 0, 1}, {y1, 0, 1},
        {x2, 0, 1}, {y2, 0, 1},
        Assumptions -> l > 0
] / tot
pLtGt = FullSimplify[pLtGt, Assumptions -> l > 0]
tau = 1 - 4 * pLtGt
```

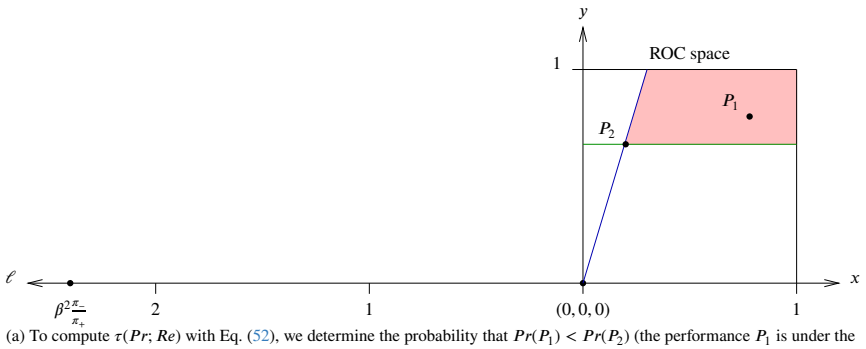
And we have

$$\tau(F_\beta; Re) = \ell^2 \left(-\log \left(\frac{1}{\ell} + 1 \right) \right) + \ell + \frac{1}{2}, \tag{66}$$

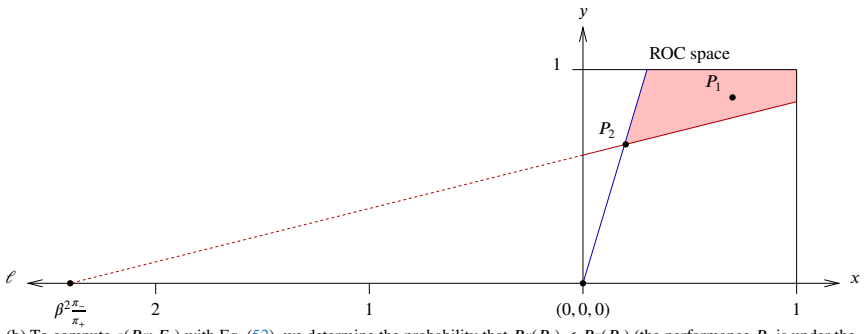
as shown with the code (see Fig. A.3.3c):

```
pLtGt = Integrate[
        Boole[(y1 (x2 + l) < y2 (x1 + l)) && (y1 > y2)],
        {x1, 0, 1}, {y1, 0, 1},
        {x2, 0, 1}, {y2, 0, 1},
        Assumptions -> l > 0
] / tot
tau = 1 - 4 * pLtGt
tau = FullSimplify[tau, Assumptions -> l > 0]
```

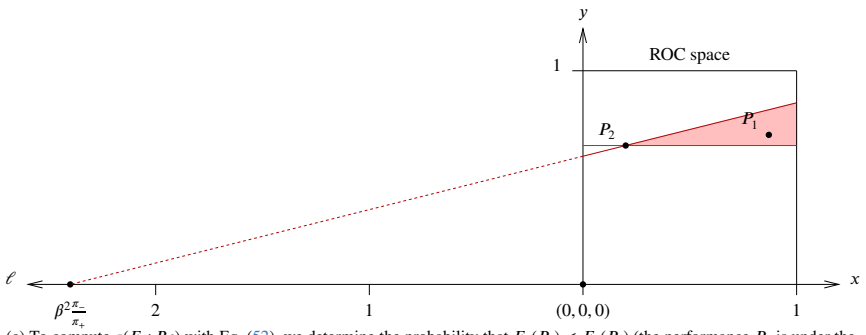
The plots are provided in Fig. A.3.4.



(a) To compute $\tau(Pr; Re)$ with Eq. (52), we determine the probability that $Pr(P_1) < Pr(P_2)$ (the performance P_1 is under the blue line) and $Re(P_1) > Re(P_2)$ (the performance P_1 is above the green line), which is the pink area averaged for all positions of P_2.

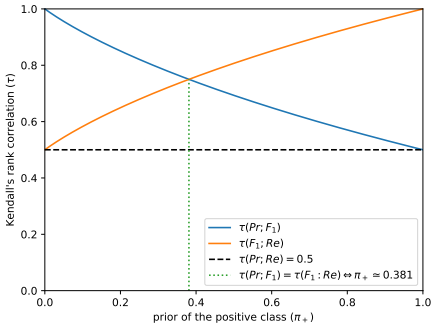


(b) To compute $\tau(Pr; F_\beta)$ with Eq. (52), we determine the probability that $Pr(P_1) < Pr(P_2)$ (the performance P_1 is under the blue line) and $F_\beta(P_1) > F_\beta(P_2)$ (the performance P_1 is above the red line), which is the pink area averaged for all positions of P_2.

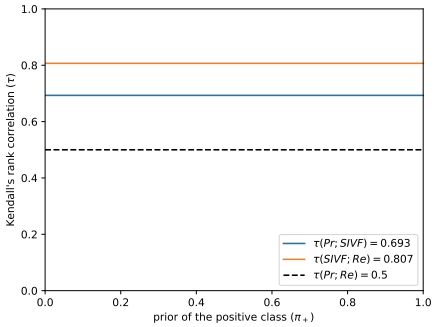


(c) To compute $\tau(F_\beta; Re)$ with Eq. (52), we determine the probability that $F_\beta(P_1) < F_\beta(P_2)$ (the performance P_1 is under the red line) and $Re(P_1) > Re(P_2)$ (the performance P_1 is above the green line), which is the pink area averaged for all positions of P_2.

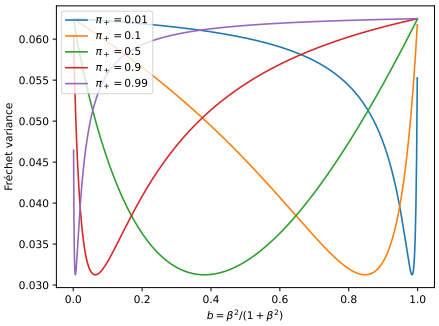
Figure A.3.3. Graphical representation, in and around the ROC space, of the principle that we use to derive the analytical expression for the optimal tradeoff between precision and recall for the purpose of ranking, in the case of uniform distributions with fixed class priors. Note that the ROC space is a linear mapping of the rectangles depicted in Fig. 3c.



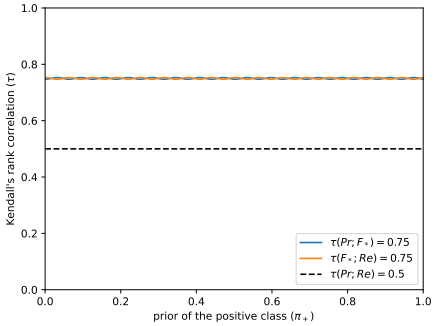
(a) F_1, the traditional (balanced) F-score, is not the optimal tradeoff: $\tau(Pr; F_1) \neq \tau(F_1; Re)$.



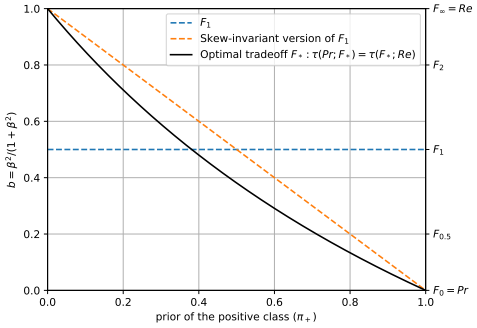
(b) $SIVF$, the skew-insensitive version of F_1 [29], is not the optimal tradeoff: $\tau(Pr; SIVF) \neq \tau(SIVF; Re)$.



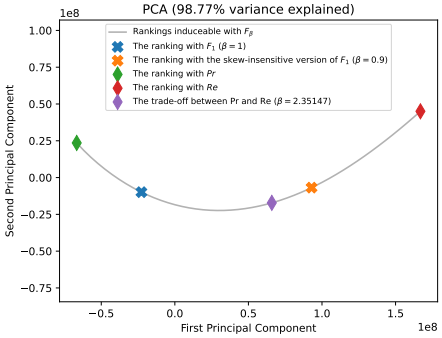
(c) Fréchet variance for various priors. It is defined at Eq. (11) and should be minimized to obtain the optimal tradeoff.



(d) F_* if the optimal tradeoff: $\tau(Pr; F_*) = \tau(F_*; Re)$.



(e) Adaptation of β w.r.t. class priors.



(f) PCA of the manifold (for $\pi_+ = 0.1$).

Figure A.3.4. Results for uniform distributions over the performances with fixed class priors, *i.e.* $\Pi_3(\pi_+)$.

A.3.5. Detailed Results for the Uniform Distributions With Fixed Class Priors, Above No-Skill

To find the probabilities necessary to compute τ, we integrate over half of the ROC space, the half above the raising diagonal. Let us denote by $(FPR, TPR) = (x, y)$ the coordinates of a performance in this space. We have

$$Pr(P_1) < Pr(P_2) \Leftrightarrow \frac{y_1}{x_1} < \frac{y_2}{x_2} \tag{67}$$

$$Re(P_1) < Re(P_2) \Leftrightarrow y_1 < y_2 \tag{68}$$

$$F_\beta(P_1) < F_\beta(P_2) \Leftrightarrow \frac{y_1}{x_1 + \ell} < \frac{y_2}{x_2 + \ell} \tag{69}$$

We start with a few initializations.

```
tot = Integrate[
    1,
    {x1, 0, 1}, {y1, x1, 1},
    {x2, 0, 1}, {y2, x2, 1}
]
```

We have

$$\tau(Pr; Re) = 0\,, \tag{70}$$

as shown with the code (see Fig. A.3.5a):

```
pLtGt = Integrate[
        Boole[(y1 x2 < y2 x1) && (y1 > y2)],
    {x1, 0, 1}, {y1, x1, 1},
    {x2, 0, 1}, {y2, x2, 1}
] / tot
tau = 1 - 4 * pLtGt
```

We have

$$\tau(Pr; F_\beta) = 1 - \frac{2}{3}\ell\left(-6\ell^2 - 6\left(\ell^2 - 1\right)\ell \log\left(\frac{\ell}{\ell + 1}\right) + 3\ell + 4\right)\,, \tag{71}$$

as shown with the code (see Fig. A.3.5b):

```
pLtGt = Integrate[
        Boole[(y1 x2 < y2 x1) && (y1 (x2 + 1) > y2 (x1 + 1))],
    {x1, 0, 1}, {y1, x1, 1},
    {x2, 0, 1}, {y2, x2, 1},
        Assumptions -> l > 0
] / tot
pLtGt = FullSimplify[pLtGt, Assumptions -> l > 0]
tau = 1 - 4 * pLtGt
```

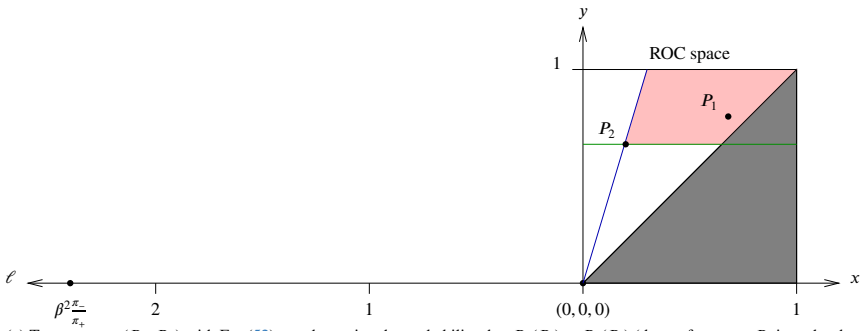
And we have

$$\tau(F_\beta; Re) = \frac{2}{3}\ell\left(-6\ell^2 + 6\left(\ell^2 - 1\right)\ell \log\left(\frac{1}{\ell} + 1\right) + 3\ell + 4\right)\,, \tag{72}$$

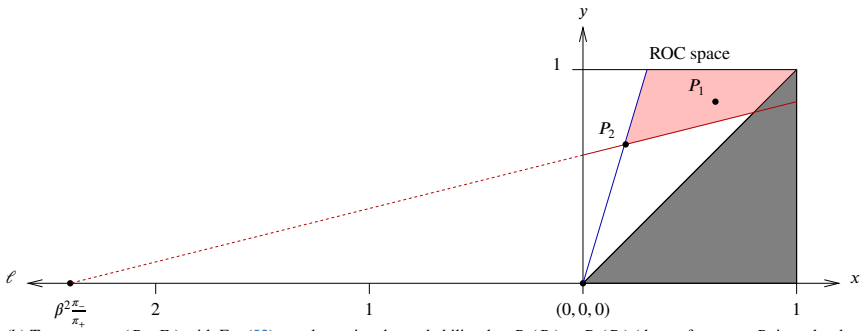
as shown with the code (see Fig. A.3.5c):

```
pLtGt = Integrate[
        Boole[(y1 (x2 + 1) < y2 (x1 + 1)) && (y1 > y2)],
    {x1, 0, 1}, {y1, x1, 1},
    {x2, 0, 1}, {y2, x2, 1},
        Assumptions -> l > 0
] / tot
tau = 1 - 4 * pLtGt
tau = FullSimplify[tau, Assumptions -> l > 0]
```

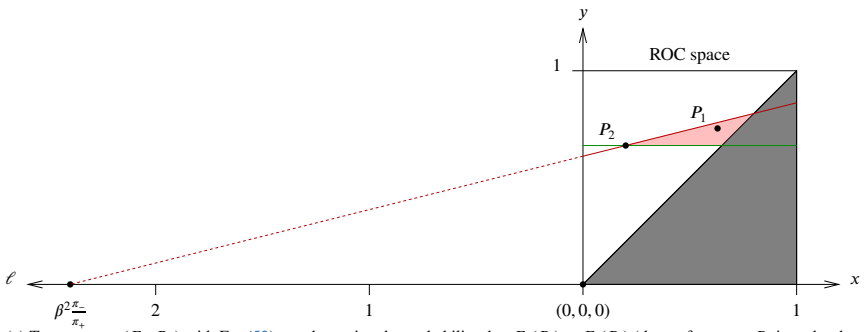
The plots are provided in Fig. A.3.6.



(a) To compute $\tau(Pr; Re)$ with Eq. (52), we determine the probability that $Pr(P_1) < Pr(P_2)$ (the performance P_1 is under the blue line) and $Re(P_1) > Re(P_2)$ (the performance P_1 is above the green line), which is the pink area averaged for all positions of P_2.

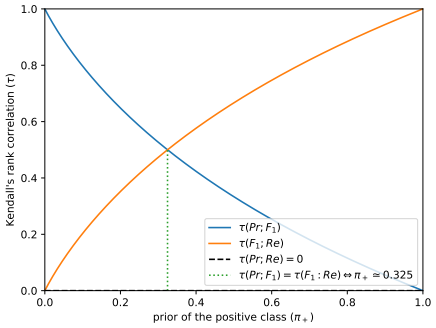


(b) To compute $\tau(Pr; F_\beta)$ with Eq. (52), we determine the probability that $Pr(P_1) < Pr(P_2)$ (the performance P_1 is under the blue line) and $F_\beta(P_1) > F_\beta(P_2)$ (the performance P_1 is above the red line), which is the pink area averaged for all positions of P_2.

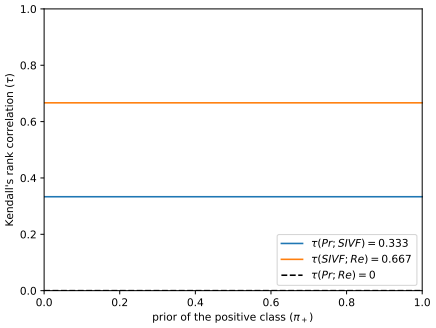


(c) To compute $\tau(F_\beta; Re)$ with Eq. (52), we determine the probability that $F_\beta(P_1) < F_\beta(P_2)$ (the performance P_1 is under the red line) and $Re(P_1) > Re(P_2)$ (the performance P_1 is above the green line), which is the pink area averaged for all positions of P_2.

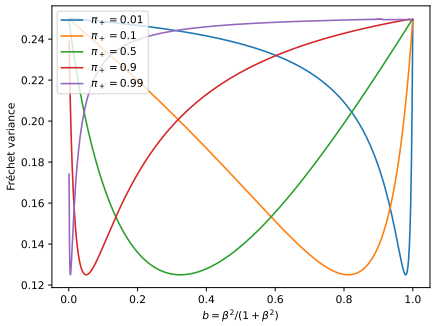
Figure A.3.5. Graphical representation, in and around the ROC space, of the principle that we use to derive the analytical expression for the optimal tradeoff between precision and recall for the purpose of ranking, in the case of uniform distributions with fixed class priors above no-skill. Note that the unshaded area of ROC space is a linear mapping of the triangles depicted in Fig. 3d.



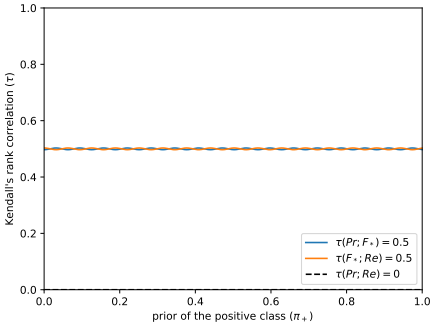
(a) F_1, the traditional (balanced) F-score, is not the optimal tradeoff: $\tau(Pr; F_1) \neq \tau(F_1; Re)$.



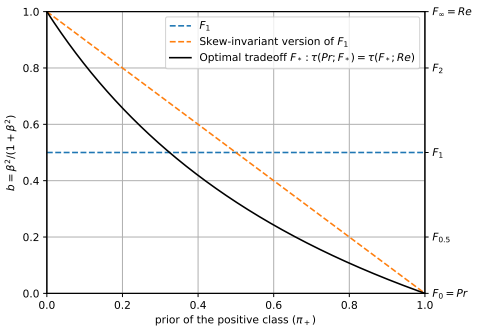
(b) $SIVF$, the skew-insensitive version of F_1 [29], is not the optimal tradeoff: $\tau(Pr; SIVF) \neq \tau(SIVF; Re)$.



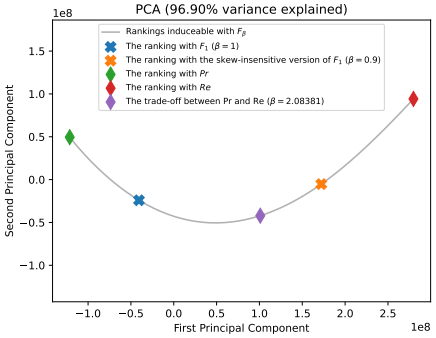
(c) Fréchet variance for various priors. It is defined at Eq. (11) and should be minimized to obtain the optimal tradeoff.



(d) F_* if the optimal tradeoff: $\tau(Pr; F_*) = \tau(F_*; Re)$.



(e) Adaptation of β w.r.t. class priors.



(f) PCA of the manifold (for $\pi_+ = 0.1$).

Figure A.3.6. Results for uniform distributions over the performances with fixed class priors, *i.e.* $\Pi_4(\pi_+)$.

A.3.6. Detailed Results for the Uniform Distributions With Fixed Class Priors, Close to Oracle

To find the probabilities necessary to compute τ, we integrate over a rectangle in the ROC space. Let us denote by $(FPR, TPR) = (x, y)$ the coordinates of a performance in this space. We have

$$Pr(P_1) < Pr(P_2) \Leftrightarrow \frac{y_1}{x_1} < \frac{y_2}{x_2} \tag{73}$$

$$Re(P_1) < Re(P_2) \Leftrightarrow y_1 < y_2 \tag{74}$$

$$F_\beta(P_1) < F_\beta(P_2) \Leftrightarrow \frac{y_1}{x_1 + \ell} < \frac{y_2}{x_2 + \ell} \qquad \text{with} \qquad \ell = \beta^2 \frac{\pi_+}{\pi_-} = \frac{b}{1-b} \frac{\pi_+}{1-\pi_+} \tag{75}$$

Unlike the previous cases, here we can no longer group π_+ and β together to form ℓ, as π_+ appears alone in the integration bounds of x and y. More precisely, we cannot anymore minimize Fréchet variance as a function of just ℓ. The optimal ℓ is not a constant anymore: it is now a function of π_+. This explains why the look of the curve for the adaptation differs from what we had before.

We start with a few initializations.

```
tot = Integrate[
    1,
    {x1, 0, p}, {y1, p, 1},
    {x2, 0, p}, {y2, p, 1},
        Assumptions -> 0 < p < 1
]
```

We have

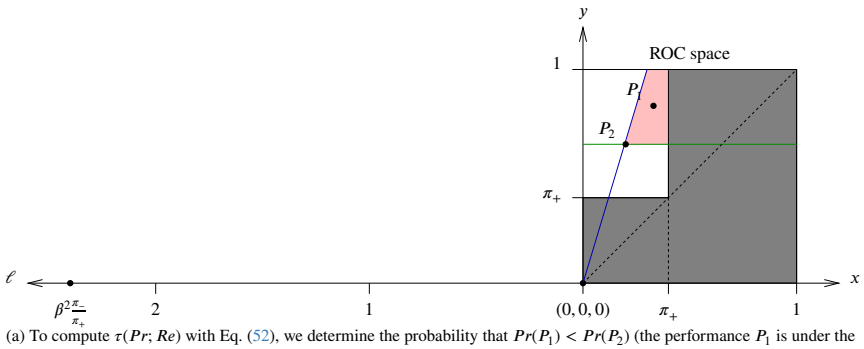
$$\tau(Pr; Re) = 1 - \frac{-\pi_+^4 + 2\pi_+^4 \log(\pi_+) + \pi_+^2}{2(1-\pi_+)^2 \pi_+^2}, \tag{76}$$

as shown with the code (see Fig. A.3.7a):

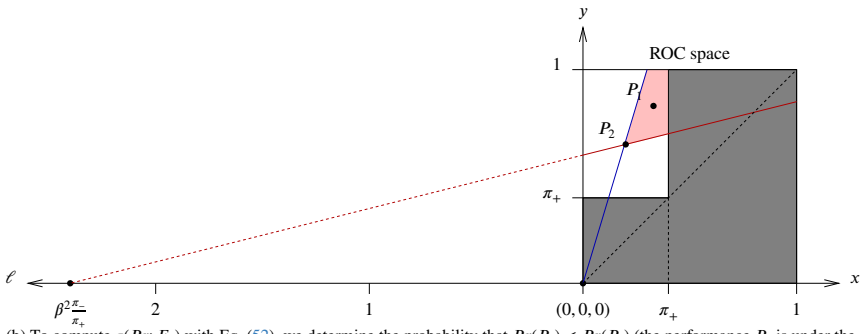
```
pLtGt = Integrate[
        Boole[(y1 x2 < y2 x1) && (y1 > y2)],
    {x1, 0, p}, {y1, p, 1},
    {x2, 0, p}, {y2, p, 1},
        Assumptions -> 0 < p < 1
] / tot
tau = 1 - 4 * pLtGt
```

With Wolfram, we were unable to obtain the analytical expression of $\tau(Pr; F_\beta)$ and $\tau(F_\beta; Re)$ for this family of distributions. Wolfram was unable to provide you with the analytical expression in a reasonable time. We therefore report the results obtained using the Monte-Carlo technique.

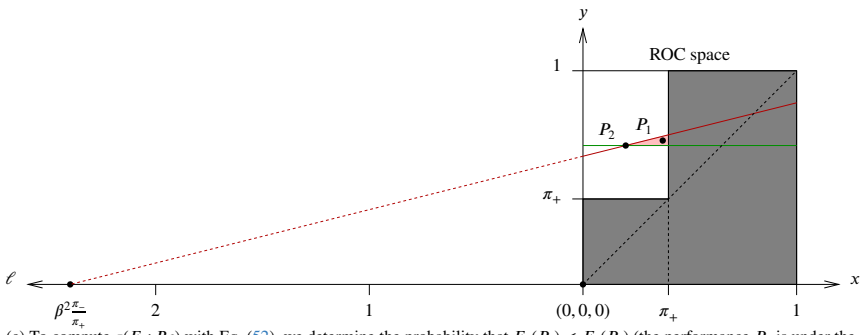
The plots are provided in Fig. A.3.8.



(a) To compute $\tau(Pr; Re)$ with Eq. (52), we determine the probability that $Pr(P_1) < Pr(P_2)$ (the performance P_1 is under the blue line) and $Re(P_1) > Re(P_2)$ (the performance P_1 is above the green line), which is the pink area averaged for all positions of P_2.

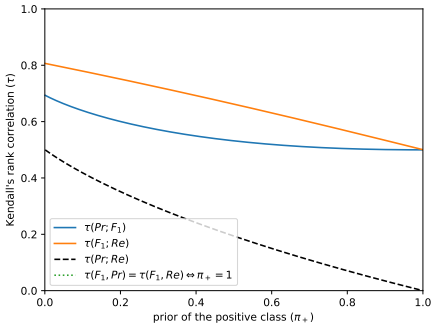


(b) To compute $\tau(Pr; F_\beta)$ with Eq. (52), we determine the probability that $Pr(P_1) < Pr(P_2)$ (the performance P_1 is under the blue line) and $F_\beta(P_1) > F_\beta(P_2)$ (the performance P_1 is above the red line), which is the pink area averaged for all positions of P_2.

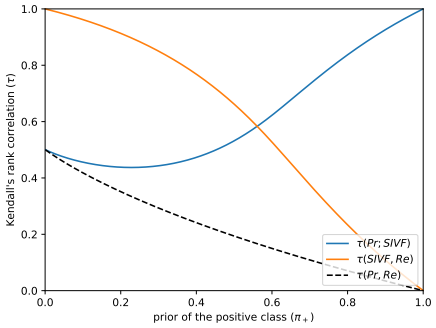


(c) To compute $\tau(F_\beta; Re)$ with Eq. (52), we determine the probability that $F_\beta(P_1) < F_\beta(P_2)$ (the performance P_1 is under the red line) and $Re(P_1) > Re(P_2)$ (the performance P_1 is above the green line), which is the pink area averaged for all positions of P_2.

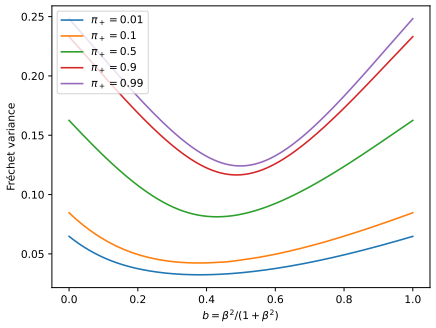
Figure A.3.7. Graphical representation, in and around the ROC space, of the principle that we use to derive the analytical expression for the optimal tradeoff between precision and recall for the purpose of ranking, in the case of uniform distributions with fixed class priors close to oracle. Note that the unshaded area of ROC space is a linear mapping of the squares depicted in Fig. 3e.



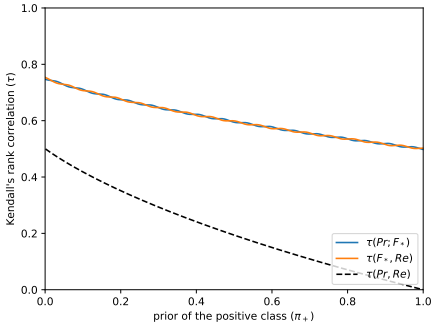
(a) F_1, the traditional (balanced) F-score, is not the optimal tradeoff: $\tau(Pr; F_1) \neq \tau(F_1; Re)$.



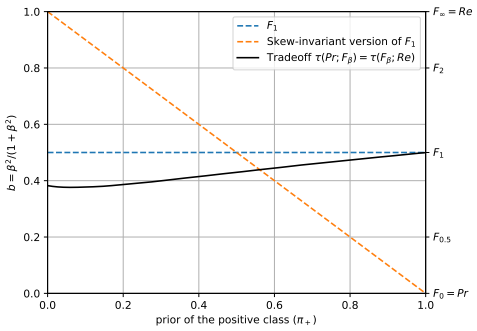
(b) $SIVF$, the skew-insensitive version of F_1 [29], is not the optimal tradeoff: $\tau(Pr; SIVF) \neq \tau(SIVF; Re)$.



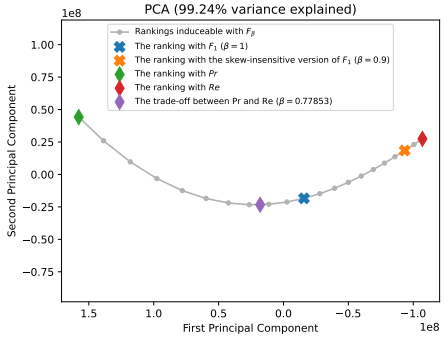
(c) Fréchet variance for various priors. It is defined at Eq. (11) and should be minimized to obtain the optimal tradeoff.



(d) F_* if the optimal tradeoff: $\tau(Pr; F_*) = \tau(F_*; Re)$.



(e) Adaptation of β w.r.t. class priors.



(f) PCA of the manifold (for $\pi_+ = 0.1$).

Figure A.3.8. Results for uniform distributions over the performances with fixed class priors, *i.e.* $\Pi_5(\pi_+)$.

A.3.7. All Results for Some Real Sets of Performances

We provide here the detailed results for each video of *CDnet 2014*.

Category "Baseline"
- The results for video "pedestrians" (ranking of 57 methods; $\pi_+ = 0.0098$) can be found at Fig. A.3.9.
- The results for video "PETS2006" (ranking of 57 methods; $\pi_+ = 0.0130$) can be found at Fig. A.3.10.
- The results for video "office" (ranking of 57 methods; $\pi_+ = 0.0690$) can be found at Fig. A.3.11.
- The results for video "highway" (ranking of 57 methods; $\pi_+ = 0.0593$) can be found at Fig. A.3.12.

Category "Dynamic Background"
- The results for video "overpass" (ranking of 57 methods; $\pi_+ = 0.0134$) can be found at Fig. A.3.13.
- The results for video "canoe" (ranking of 57 methods; $\pi_+ = 0.0354$) can be found at Fig. A.3.14.
- The results for video "fountain01" (ranking of 57 methods; $\pi_+ = 0.0008$) can be found at Fig. A.3.15.
- The results for video "fountain02" (ranking of 57 methods; $\pi_+ = 0.0022$) can be found at Fig. A.3.16.
- The results for video "fall" (ranking of 57 methods; $\pi_+ = 0.0177$) can be found at Fig. A.3.17.
- The results for video "boats" (ranking of 57 methods; $\pi_+ = 0.0063$) can be found at Fig. A.3.18.

Category "Camera Jitter"
- The results for video "boulevard" (ranking of 57 methods; $\pi_+ = 0.0469$) can be found at Fig. A.3.19.
- The results for video "sidewalk" (ranking of 57 methods; $\pi_+ = 0.0261$) can be found at Fig. A.3.20.
- The results for video "badminton" (ranking of 57 methods; $\pi_+ = 0.0343$) can be found at Fig. A.3.21.
- The results for video "traffic" (ranking of 57 methods; $\pi_+ = 0.0623$) can be found at Fig. A.3.22.

Category "Intermittent Object Motion"
- The results for video "abandonedBox" (ranking of 57 methods; $\pi_+ = 0.0481$) can be found at Fig. A.3.23.
- The results for video "winterDriveway" (ranking of 57 methods; $\pi_+ = 0.0075$) can be found at Fig. A.3.24.
- The results for video "sofa" (ranking of 57 methods; $\pi_+ = 0.0437$) can be found at Fig. A.3.25.
- The results for video "tramstop" (ranking of 57 methods; $\pi_+ = 0.1795$) can be found at Fig. A.3.26.
- The results for video "parking" (ranking of 57 methods; $\pi_+ = 0.0773$) can be found at Fig. A.3.27.
- The results for video "streetLight" (ranking of 57 methods; $\pi_+ = 0.0485$) can be found at Fig. A.3.28.

Category "Shadow"
- The results for video "copyMachine" (ranking of 57 methods; $\pi_+ = 0.0693$) can be found at Fig. A.3.29.
- The results for video "bungalows" (ranking of 57 methods; $\pi_+ = 0.0600$) can be found at Fig. A.3.30.
- The results for video "busStation" (ranking of 57 methods; $\pi_+ = 0.0369$) can be found at Fig. A.3.31.
- The results for video "peopleInShade" (ranking of 57 methods; $\pi_+ = 0.0564$) can be found at Fig. A.3.32.
- The results for video "backdoor" (ranking of 57 methods; $\pi_+ = 0.0199$) can be found at Fig. A.3.33.
- The results for video "cubicle" (ranking of 57 methods; $\pi_+ = 0.0196$) can be found at Fig. A.3.34.

Category "Thermal"
- The results for video "lakeSide" (ranking of 57 methods; $\pi_+ = 0.0192$) can be found at Fig. A.3.35.
- The results for video "diningRoom" (ranking of 57 methods; $\pi_+ = 0.0859$) can be found at Fig. A.3.36.
- The results for video "park" (ranking of 57 methods; $\pi_+ = 0.0203$) can be found at Fig. A.3.37.
- The results for video "corridor" (ranking of 57 methods; $\pi_+ = 0.0331$) can be found at Fig. A.3.38.
- The results for video "library" (ranking of 57 methods; $\pi_+ = 0.1928$) can be found at Fig. A.3.39.

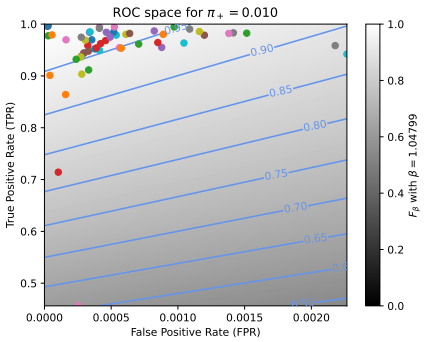
Category "Bad Weather"
- The results for video "skating" (ranking of 57 methods; $\pi_+ = 0.0397$) can be found at Fig. A.3.40.
- The results for video "wetSnow" (ranking of 57 methods; $\pi_+ = 0.0150$) can be found at Fig. A.3.41.
- The results for video "snowFall" (ranking of 57 methods; $\pi_+ = 0.0105$) can be found at Fig. A.3.42.
- The results for video "blizzard" (ranking of 57 methods; $\pi_+ = 0.0115$) can be found at Fig. A.3.43.

Category "Low Framerate"
- The results for video "tunnelExit_0_35fps" (ranking of 57 methods; $\pi_+ = 0.0195$) can be found at Fig. A.3.44.
- The results for video "port_0_17fps" (ranking of 57 methods; $\pi_+ = 0.0002$) can be found at Fig. A.3.45.
- The results for video "tramCrossroad_1fps" (ranking of 57 methods; $\pi_+ = 0.0288$) can be found at Fig. A.3.46.
- The results for video "turnpike_0_5fps" (ranking of 57 methods; $\pi_+ = 0.0581$) can be found at Fig. A.3.47.

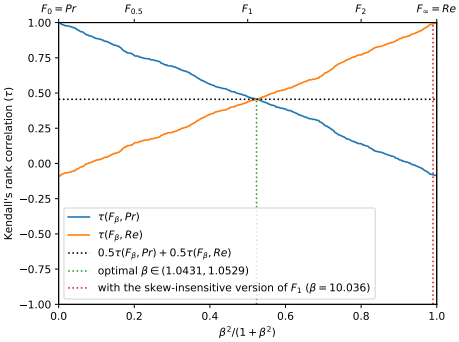
Category "Night Videos"
- The results for video "tramStation" (ranking of 58 methods; $\pi_+ = 0.0339$) can be found at Fig. A.3.48.
- The results for video "busyBoulvard" (ranking of 58 methods; $\pi_+ = 0.0847$) can be found at Fig. A.3.49.
- The results for video "streetCornerAtNight" (ranking of 58 methods; $\pi_+ = 0.0062$) can be found at Fig. A.3.50.
- The results for video "fluidHighway" (ranking of 58 methods; $\pi_+ = 0.0175$) can be found at Fig. A.3.51.
- The results for video "bridgeEntry" (ranking of 58 methods; $\pi_+ = 0.0200$) can be found at Fig. A.3.52.
- The results for video "winterStreet" (ranking of 58 methods; $\pi_+ = 0.0689$) can be found at Fig. A.3.53.

Category "PTZ"
- The results for video "twoPositionPTZCam" (ranking of 56 methods; $\pi_+ = 0.0117$) can be found at Fig. A.3.54.
- The results for video "zoomInZoomOut" (ranking of 56 methods; $\pi_+ = 0.0019$) can be found at Fig. A.3.55.
- The results for video "continuousPan" (ranking of 56 methods; $\pi_+ = 0.0056$) can be found at Fig. A.3.56.
- The results for video "intermittentPan" (ranking of 56 methods; $\pi_+ = 0.0094$) can be found at Fig. A.3.57.

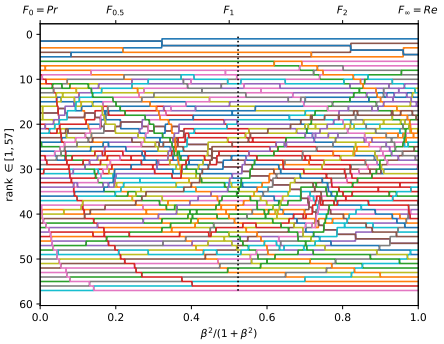
Category "Turbulence"
- The results for video "turbulence2" (ranking of 57 methods; $\pi_+ = 0.0008$) can be found at Fig. A.3.58.
- The results for video "turbulence3" (ranking of 57 methods; $\pi_+ = 0.0121$) can be found at Fig. A.3.59.
- The results for video "turbulence0" (ranking of 57 methods; $\pi_+ = 0.0015$) can be found at Fig. A.3.60.
- The results for video "turbulence1" (ranking of 57 methods; $\pi_+ = 0.0026$) can be found at Fig. A.3.61.



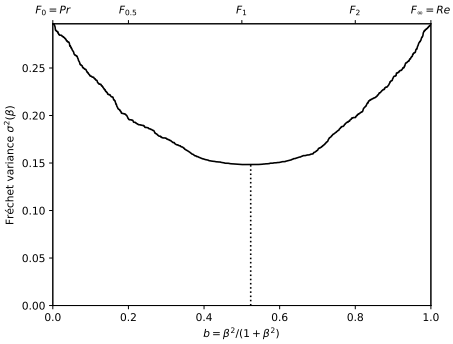
(a) The performances of 57 classifiers (BGS methods) depicted as points in the ROC space, with the isometrics of the optimal tradeoff score, from the ranking point of view, between precision and recall. See Eq. (15).



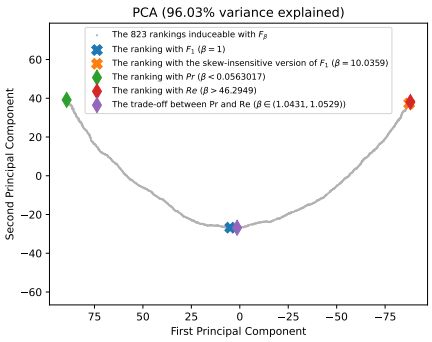
(b) The rank correlations $\tau(Pr; F_\beta)$ and $\tau(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the two curves intersect.



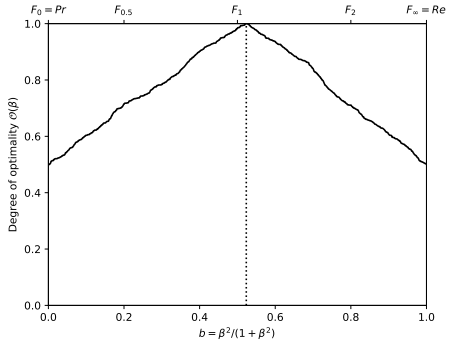
(c) The ranks of each classifier w.r.t. β. The optimal value (or range of optimal values) for β, shown here by the vertical line, is such that the number of swaps on its left is equal to the number of swaps on its right.



(d) The Fréchet variance $\sigma^2(\beta) = d_\tau^2(Pr; F_\beta) + d_\tau^2(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the curve has its minimum.

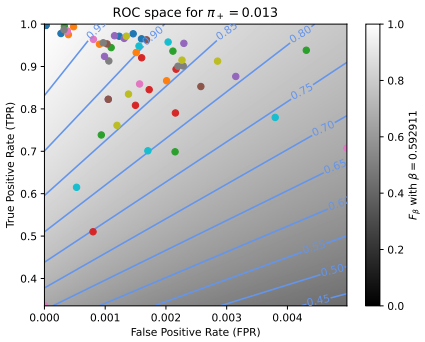


(e) Linear projection (PCA) of the manifold of the rankings induced by the F_β scores. The color points indicate the precision, the recall, F_1, $SIVF$, as well as the optimal tradeoff. The optimal tradeoff is at the same distance of the two extremities when the distance is measured along the manifold, with Kendall's distance d_τ.

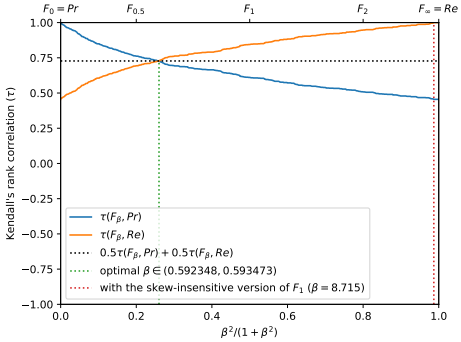


(f) The degree of optimality $\mathcal{O}(\beta)$ w.r.t. β. It is the probability to optimally ordering a pair of classifiers (BGS methods) given that it is not trivial (*i.e.*, that Pr and Re are in contradiction). The optimal value (or range of optimal values) for β is where the curve reaches 100%.

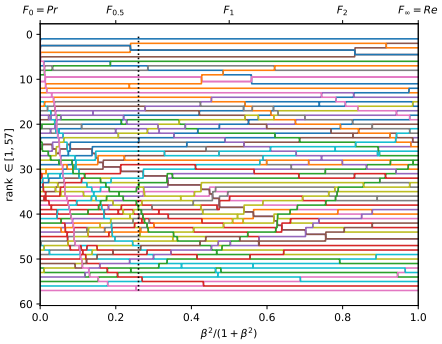
Figure A.3.9. Ranking of 57 BGS methods evaluated on the video "pedestrians" ($\pi_+ = 0.0098$).



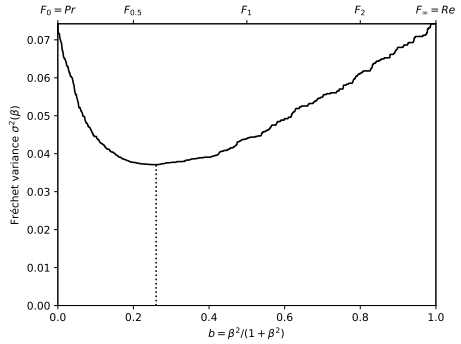
(a) The performances of 57 classifiers (BGS methods) depicted as points in the ROC space, with the isometrics of the optimal tradeoff score, from the ranking point of view, between precision and recall. See Eq. (15).



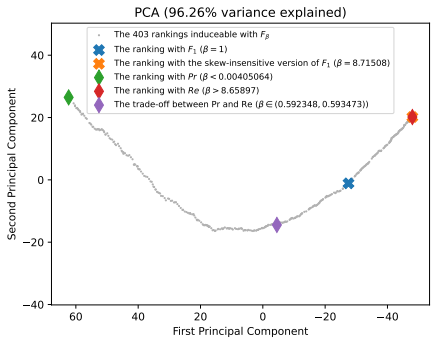
(b) The rank correlations $\tau(Pr; F_\beta)$ and $\tau(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the two curves intersect.



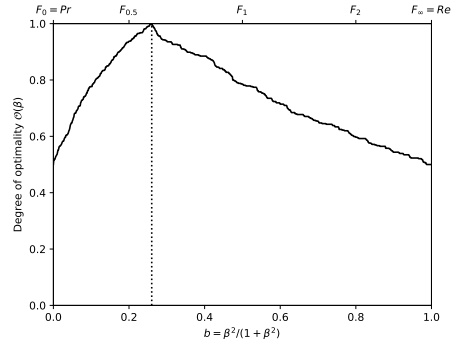
(c) The ranks of each classifier w.r.t. β. The optimal value (or range of optimal values) for β, shown here by the vertical line, is such that the number of swaps on its left is equal to the number of swaps on its right.



(d) The Fréchet variance $\sigma^2(\beta) = d_\tau^2(Pr; F_\beta) + d_\tau^2(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the curve has its minimum.

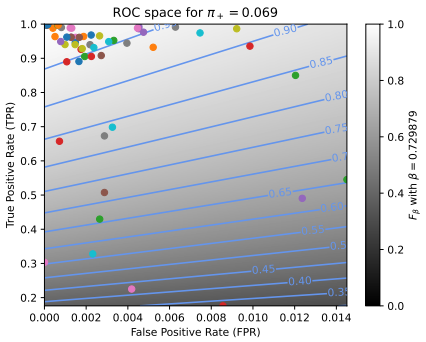


(e) Linear projection (PCA) of the manifold of the rankings induced by the F_β scores. The color points indicate the precision, the recall, F_1, $SIVF$, as well as the optimal tradeoff. The optimal tradeoff is at the same distance of the two extremities when the distance is measured along the manifold, with Kendall's distance d_τ.

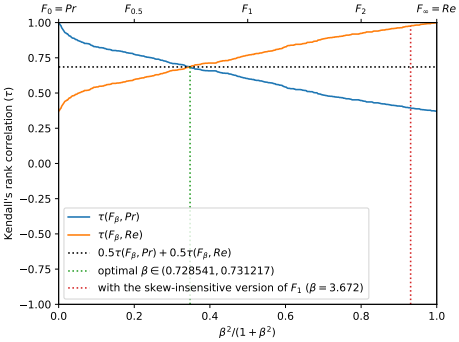


(f) The degree of optimality $\mathcal{O}(\beta)$ w.r.t. β. It is the probability to optimally ordering a pair of classifiers (BGS methods) given that it is not trivial (*i.e.*, that Pr and Re are in contradiction). The optimal value (or range of optimal values) for β is where the curve reaches 100%.

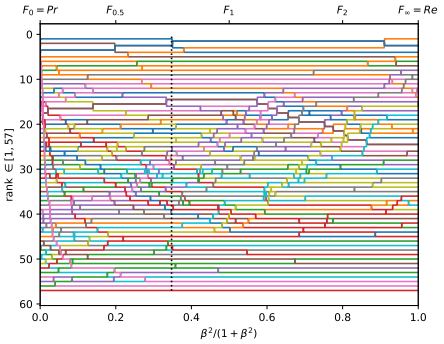
Figure A.3.10. Ranking of 57 BGS methods evaluated on the video "PETS2006" ($\pi_+ = 0.0130$).



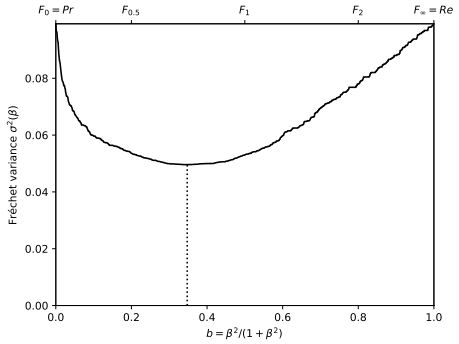
(a) The performances of 57 classifiers (BGS methods) depicted as points in the ROC space, with the isometrics of the optimal tradeoff score, from the ranking point of view, between precision and recall. See Eq. (15).



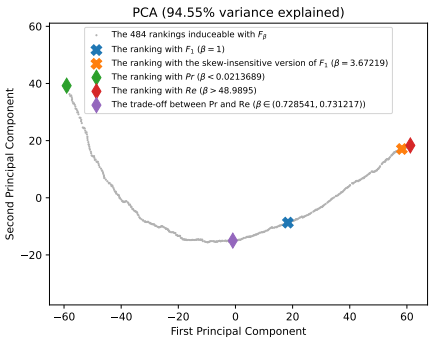
(b) The rank correlations $\tau(Pr; F_\beta)$ and $\tau(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the two curves intersect.



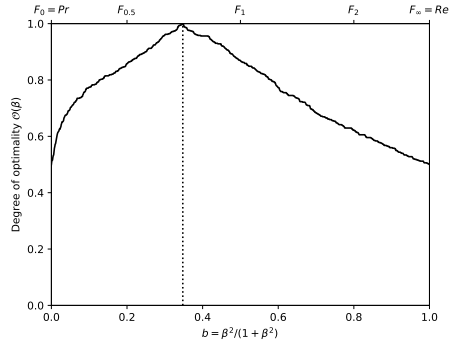
(c) The ranks of each classifier w.r.t. β. The optimal value (or range of optimal values) for β, shown here by the vertical line, is such that the number of swaps on its left is equal to the number of swaps on its right.



(d) The Fréchet variance $\sigma^2(\beta) = d_\tau^2(Pr; F_\beta) + d_\tau^2(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the curve has its minimum.

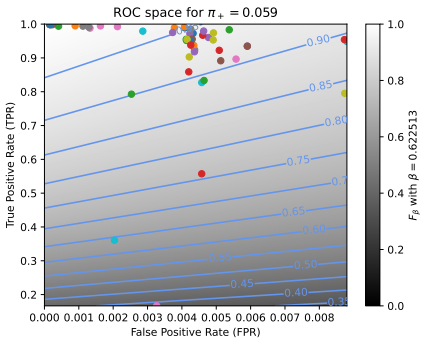


(e) Linear projection (PCA) of the manifold of the rankings induced by the F_β scores. The color points indicate the precision, the recall, F_1, $SIVF$, as well as the optimal tradeoff. The optimal tradeoff is at the same distance of the two extremities when the distance is measured along the manifold, with Kendall's distance d_τ.

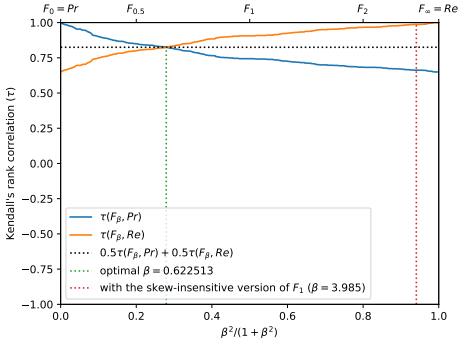


(f) The degree of optimality $\mathcal{O}(\beta)$ w.r.t. β. It is the probability to optimally ordering a pair of classifiers (BGS methods) given that it is not trivial (*i.e.*, that Pr and Re are in contradiction). The optimal value (or range of optimal values) for β is where the curve reaches 100%.

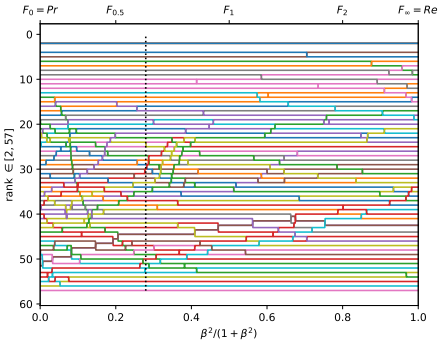
Figure A.3.11. Ranking of 57 BGS methods evaluated on the video "office" ($\pi_+ = 0.0690$).



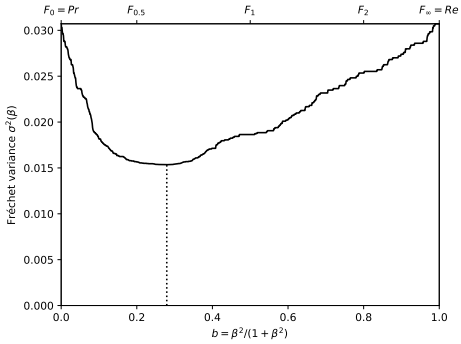
(a) The performances of 57 classifiers (BGS methods) depicted as points in the ROC space, with the isometrics of the optimal tradeoff score, from the ranking point of view, between precision and recall. See Eq. (15).



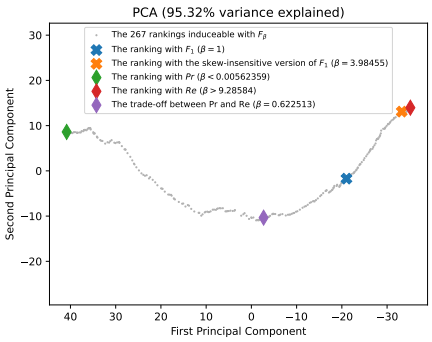
(b) The rank correlations $\tau(Pr; F_\beta)$ and $\tau(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the two curves intersect.



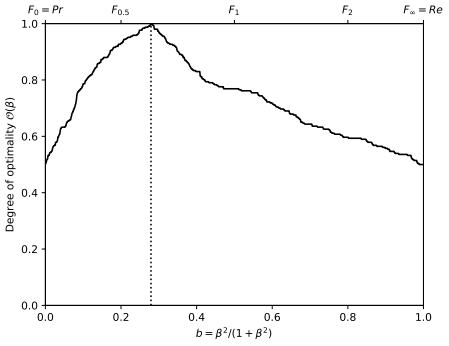
(c) The ranks of each classifier w.r.t. β. The optimal value (or range of optimal values) for β, shown here by the vertical line, is such that the number of swaps on its left is equal to the number of swaps on its right.



(d) The Fréchet variance $\sigma^2(\beta) = d_\tau^2(Pr; F_\beta) + d_\tau^2(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the curve has its minimum.

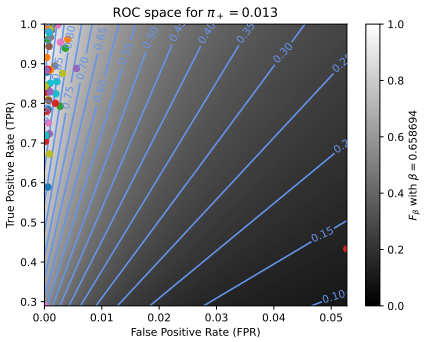


(e) Linear projection (PCA) of the manifold of the rankings induced by the F_β scores. The color points indicate the precision, the recall, F_1, $SIVF$, as well as the optimal tradeoff. The optimal tradeoff is at the same distance of the two extremities when the distance is measured along the manifold, with Kendall's distance d_τ.

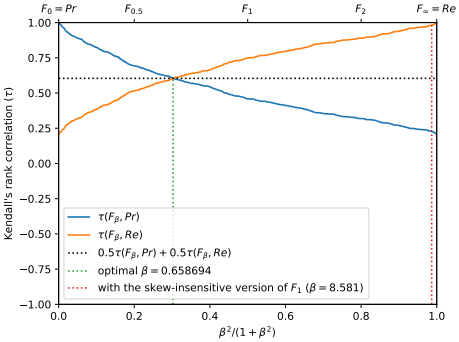


(f) The degree of optimality $\mathcal{O}(\beta)$ w.r.t. β. It is the probability to optimally ordering a pair of classifiers (BGS methods) given that it is not trivial (*i.e.*, that Pr and Re are in contradiction). The optimal value (or range of optimal values) for β is where the curve reaches 100%.

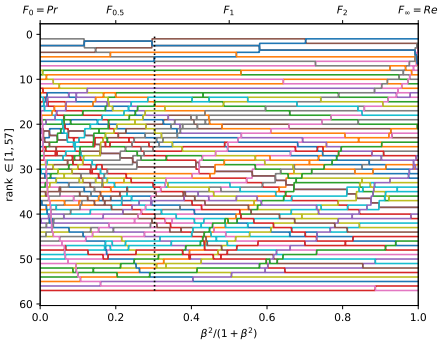
Figure A.3.12. Ranking of 57 BGS methods evaluated on the video "highway" ($\pi_+ = 0.0593$).



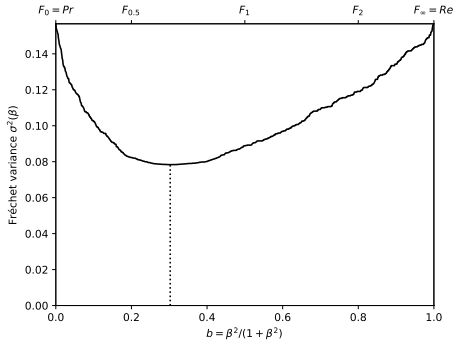
(a) The performances of 57 classifiers (BGS methods) depicted as points in the ROC space, with the isometrics of the optimal tradeoff score, from the ranking point of view, between precision and recall. See Eq. (15).



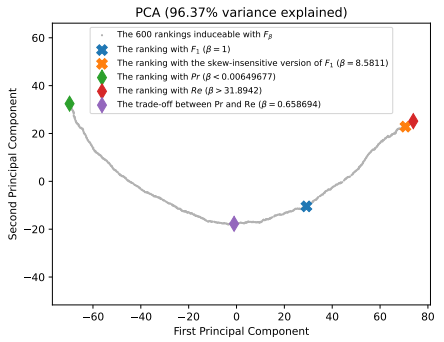
(b) The rank correlations $\tau(Pr; F_\beta)$ and $\tau(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the two curves intersect.



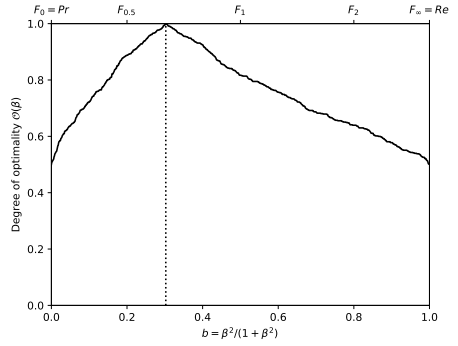
(c) The ranks of each classifier w.r.t. β. The optimal value (or range of optimal values) for β, shown here by the vertical line, is such that the number of swaps on its left is equal to the number of swaps on its right.



(d) The Fréchet variance $\sigma^2(\beta) = d_\tau^2(Pr; F_\beta) + d_\tau^2(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the curve has its minimum.

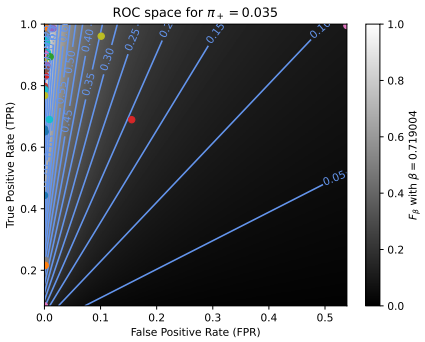


(e) Linear projection (PCA) of the manifold of the rankings induced by the F_β scores. The color points indicate the precision, the recall, F_1, $SIVF$, as well as the optimal tradeoff. The optimal tradeoff is at the same distance of the two extremities when the distance is measured along the manifold, with Kendall's distance d_τ.

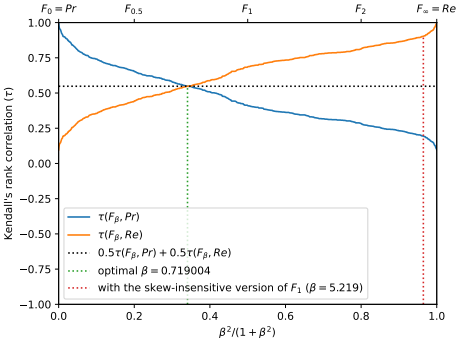


(f) The degree of optimality $\mathcal{O}(\beta)$ w.r.t. β. It is the probability to optimally ordering a pair of classifiers (BGS methods) given that it is not trivial (*i.e.*, that Pr and Re are in contradiction). The optimal value (or range of optimal values) for β is where the curve reaches 100%.

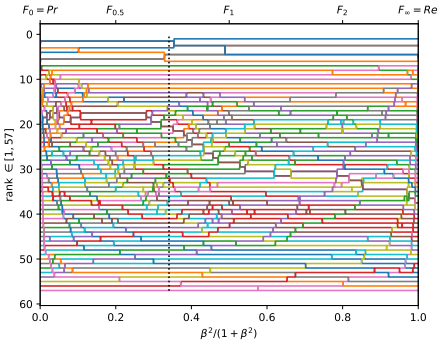
Figure A.3.13. Ranking of 57 BGS methods evaluated on the video "overpass" ($\pi_+ = 0.0134$).



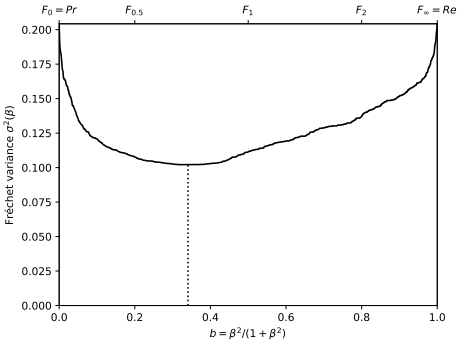
(a) The performances of 57 classifiers (BGS methods) depicted as points in the ROC space, with the isometrics of the optimal tradeoff score, from the ranking point of view, between precision and recall. See Eq. (15).



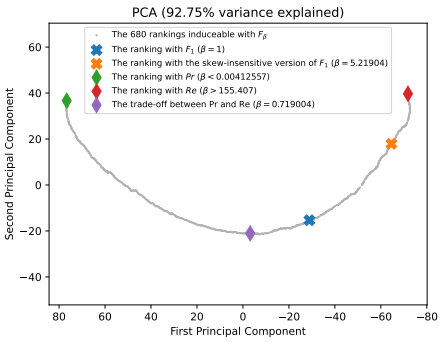
(b) The rank correlations $\tau(Pr; F_\beta)$ and $\tau(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the two curves intersect.



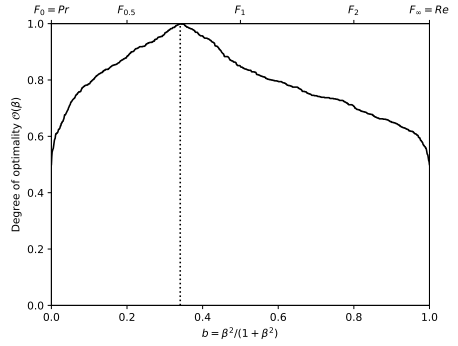
(c) The ranks of each classifier w.r.t. β. The optimal value (or range of optimal values) for β, shown here by the vertical line, is such that the number of swaps on its left is equal to the number of swaps on its right.



(d) The Fréchet variance $\sigma^2(\beta) = d_\tau^2(Pr; F_\beta) + d_\tau^2(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the curve has its minimum.

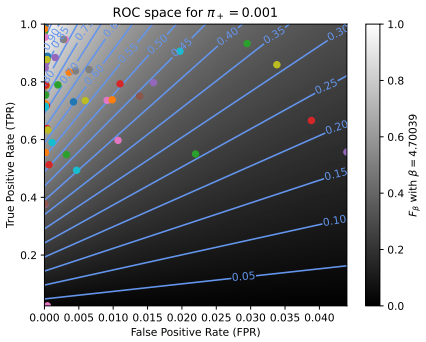


(e) Linear projection (PCA) of the manifold of the rankings induced by the F_β scores. The color points indicate the precision, the recall, F_1, $SIVF$, as well as the optimal tradeoff. The optimal tradeoff is at the same distance of the two extremities when the distance is measured along the manifold, with Kendall's distance d_τ.

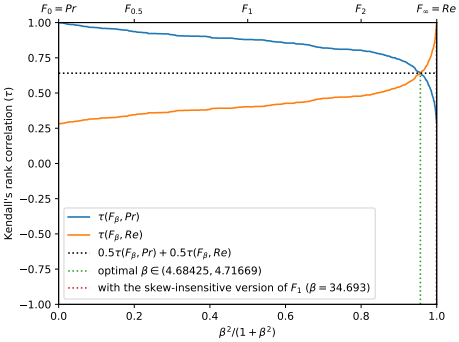


(f) The degree of optimality $\mathcal{O}(\beta)$ w.r.t. β. It is the probability to optimally ordering a pair of classifiers (BGS methods) given that it is not trivial (*i.e.*, that Pr and Re are in contradiction). The optimal value (or range of optimal values) for β is where the curve reaches 100%.

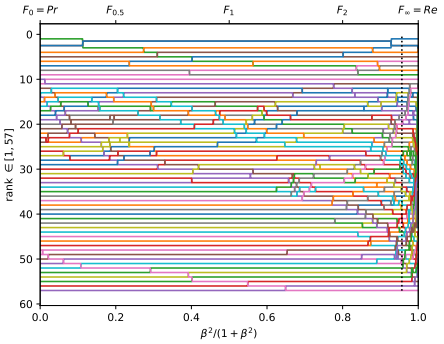
Figure A.3.14. Ranking of 57 BGS methods evaluated on the video "canoe" ($\pi_+ = 0.0354$).



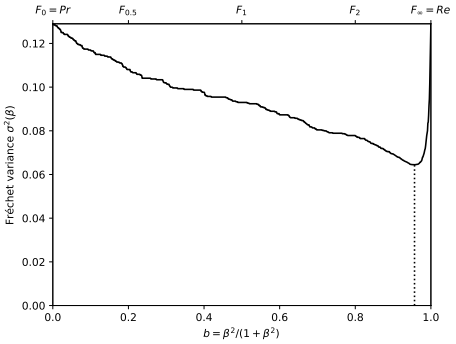
(a) The performances of 57 classifiers (BGS methods) depicted as points in the ROC space, with the isometrics of the optimal tradeoff score, from the ranking point of view, between precision and recall. See Eq. (15).



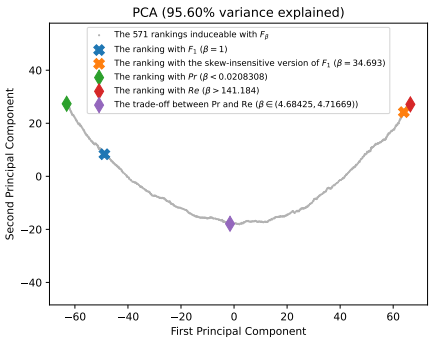
(b) The rank correlations $\tau(Pr; F_\beta)$ and $\tau(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the two curves intersect.



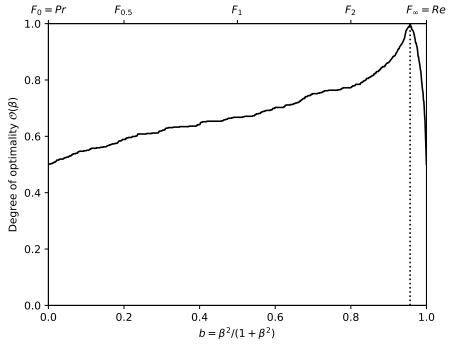
(c) The ranks of each classifier w.r.t. β. The optimal value (or range of optimal values) for β, shown here by the vertical line, is such that the number of swaps on its left is equal to the number of swaps on its right.



(d) The Fréchet variance $\sigma^2(\beta) = d_\tau^2(Pr; F_\beta) + d_\tau^2(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the curve has its minimum.

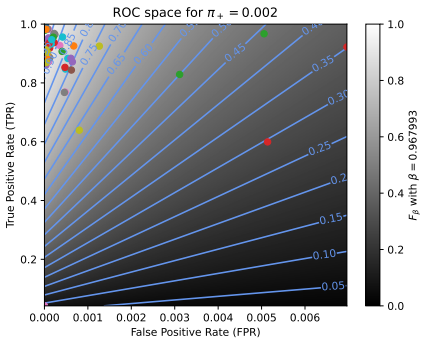


(e) Linear projection (PCA) of the manifold of the rankings induced by the F_β scores. The color points indicate the precision, the recall, F_1, $SIVF$, as well as the optimal tradeoff. The optimal tradeoff is at the same distance of the two extremities when the distance is measured along the manifold, with Kendall's distance d_τ.

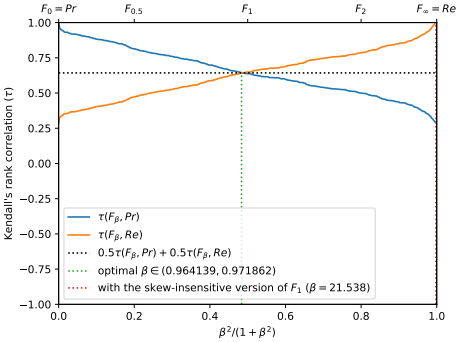


(f) The degree of optimality $\mathcal{O}(\beta)$ w.r.t. β. It is the probability to optimally ordering a pair of classifiers (BGS methods) given that it is not trivial (*i.e.*, that Pr and Re are in contradiction). The optimal value (or range of optimal values) for β is where the curve reaches 100%.

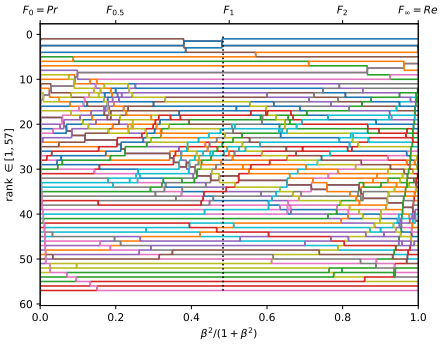
Figure A.3.15. Ranking of 57 BGS methods evaluated on the video "fountain01" ($\pi_+ = 0.0008$).



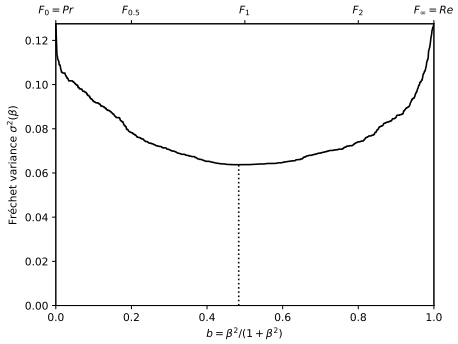
(a) The performances of 57 classifiers (BGS methods) depicted as points in the ROC space, with the isometrics of the optimal tradeoff score, from the ranking point of view, between precision and recall. See Eq. (15).



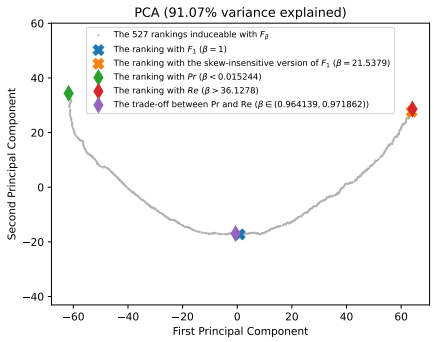
(b) The rank correlations $\tau(Pr; F_\beta)$ and $\tau(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the two curves intersect.



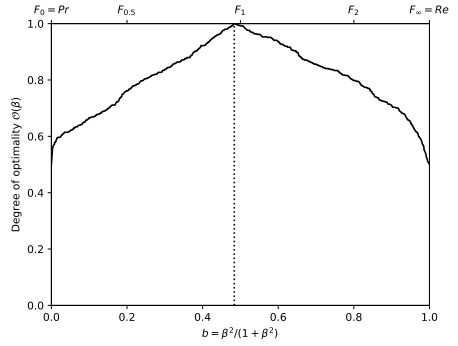
(c) The ranks of each classifier w.r.t. β. The optimal value (or range of optimal values) for β, shown here by the vertical line, is such that the number of swaps on its left is equal to the number of swaps on its right.



(d) The Fréchet variance $\sigma^2(\beta) = d_\tau^2(Pr; F_\beta) + d_\tau^2(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the curve has its minimum.

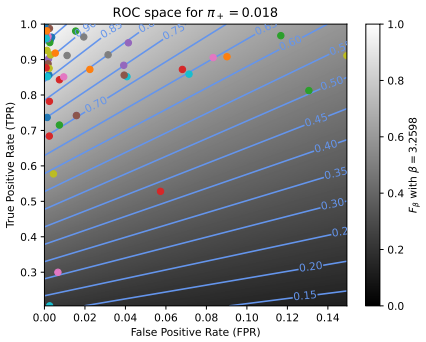


(e) Linear projection (PCA) of the manifold of the rankings induced by the F_β scores. The color points indicate the precision, the recall, F_1, $SIVF$, as well as the optimal tradeoff. The optimal tradeoff is at the same distance of the two extremities when the distance is measured along the manifold, with Kendall's distance d_τ.

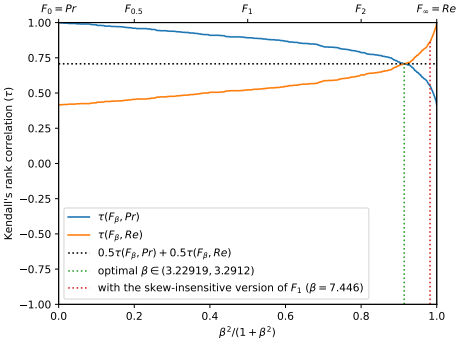


(f) The degree of optimality $\mathcal{O}(\beta)$ w.r.t. β. It is the probability to optimally ordering a pair of classifiers (BGS methods) given that it is not trivial (*i.e.*, that Pr and Re are in contradiction). The optimal value (or range of optimal values) for β is where the curve reaches 100%.

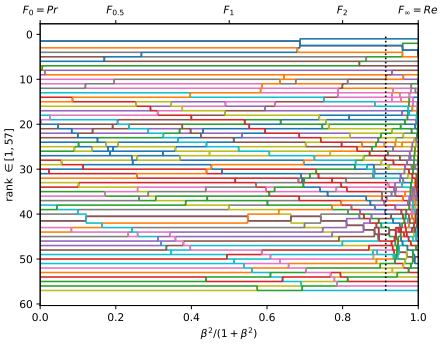
Figure A.3.16. Ranking of 57 BGS methods evaluated on the video "fountain02" ($\pi_+ = 0.0022$).



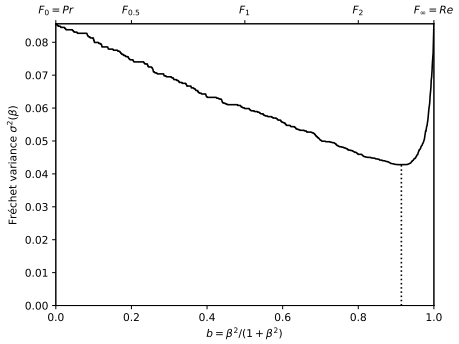
(a) The performances of 57 classifiers (BGS methods) depicted as points in the ROC space, with the isometrics of the optimal tradeoff score, from the ranking point of view, between precision and recall. See Eq. (15).



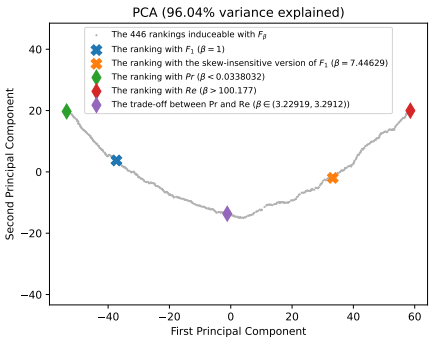
(b) The rank correlations $\tau(Pr; F_\beta)$ and $\tau(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the two curves intersect.



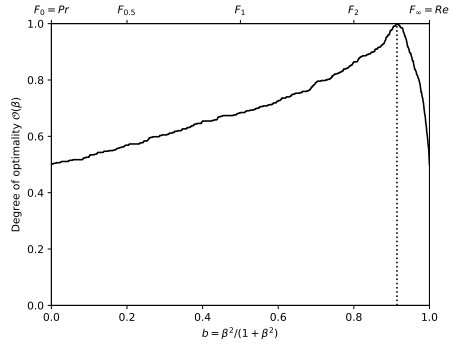
(c) The ranks of each classifier w.r.t. β. The optimal value (or range of optimal values) for β, shown here by the vertical line, is such that the number of swaps on its left is equal to the number of swaps on its right.



(d) The Fréchet variance $\sigma^2(\beta) = d_\tau^2(Pr; F_\beta) + d_\tau^2(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the curve has its minimum.

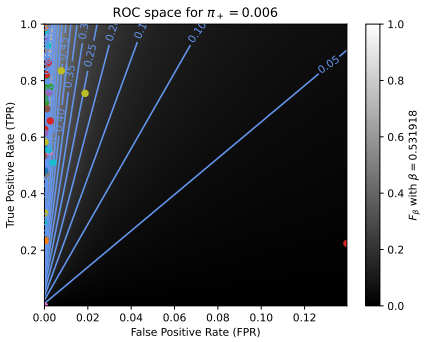


(e) Linear projection (PCA) of the manifold of the rankings induced by the F_β scores. The color points indicate the precision, the recall, F_1, $SIVF$, as well as the optimal tradeoff. The optimal tradeoff is at the same distance of the two extremities when the distance is measured along the manifold, with Kendall's distance d_τ.

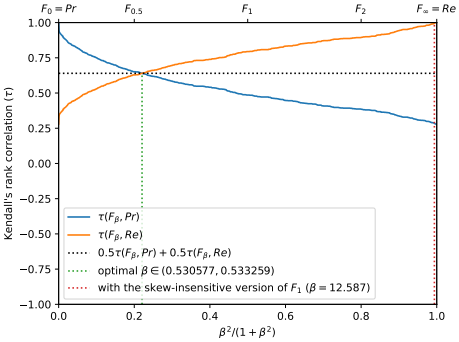


(f) The degree of optimality $\mathcal{O}(\beta)$ w.r.t. β. It is the probability to optimally ordering a pair of classifiers (BGS methods) given that it is not trivial (*i.e.*, that Pr and Re are in contradiction). The optimal value (or range of optimal values) for β is where the curve reaches 100%.

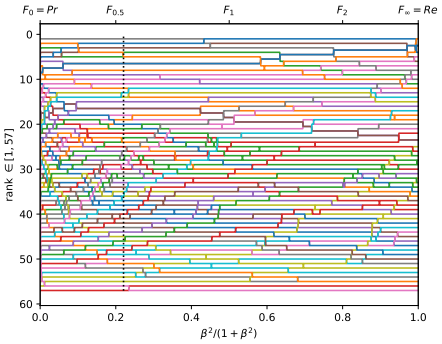
Figure A.3.17. Ranking of 57 BGS methods evaluated on the video "fall" ($\pi_+ = 0.0177$).



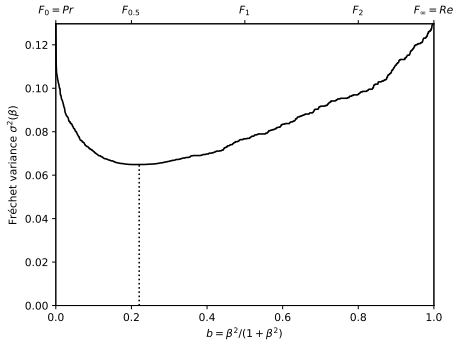
(a) The performances of 57 classifiers (BGS methods) depicted as points in the ROC space, with the isometrics of the optimal tradeoff score, from the ranking point of view, between precision and recall. See Eq. (15).



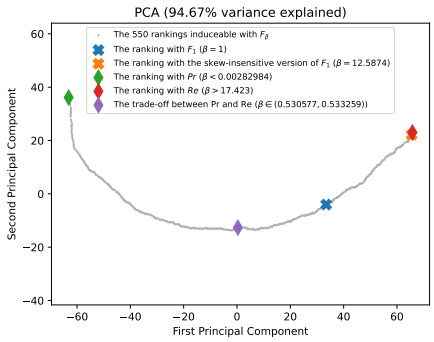
(b) The rank correlations $\tau(Pr; F_\beta)$ and $\tau(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the two curves intersect.



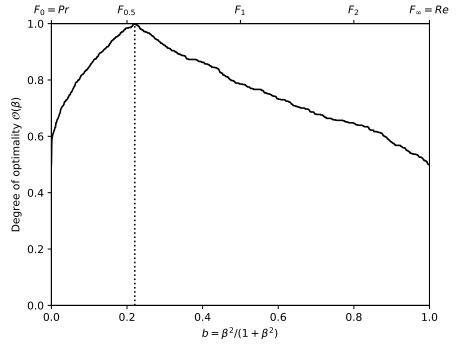
(c) The ranks of each classifier w.r.t. β. The optimal value (or range of optimal values) for β, shown here by the vertical line, is such that the number of swaps on its left is equal to the number of swaps on its right.



(d) The Fréchet variance $\sigma^2(\beta) = d_\tau^2(Pr; F_\beta) + d_\tau^2(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the curve has its minimum.

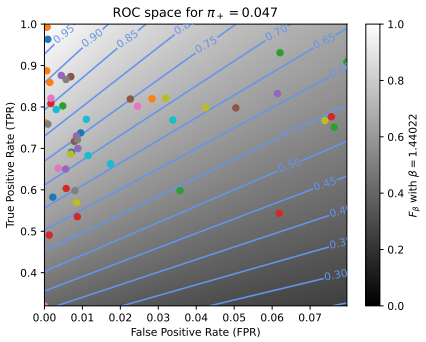


(e) Linear projection (PCA) of the manifold of the rankings induced by the F_β scores. The color points indicate the precision, the recall, F_1, $SIVF$, as well as the optimal tradeoff. The optimal tradeoff is at the same distance of the two extremities when the distance is measured along the manifold, with Kendall's distance d_τ.

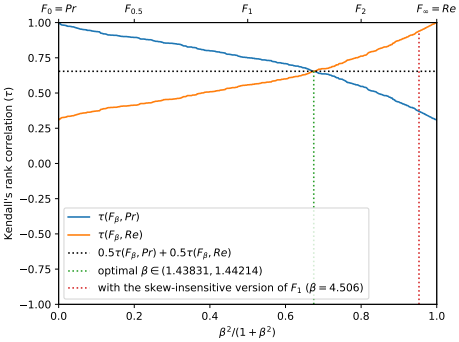


(f) The degree of optimality $\mathcal{O}(\beta)$ w.r.t. β. It is the probability to optimally ordering a pair of classifiers (BGS methods) given that it is not trivial (*i.e.*, that Pr and Re are in contradiction). The optimal value (or range of optimal values) for β is where the curve reaches 100%.

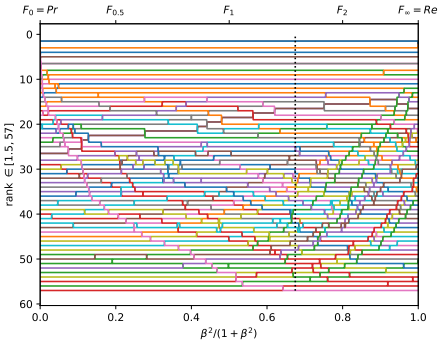
Figure A.3.18. Ranking of 57 BGS methods evaluated on the video "boats" ($\pi_+ = 0.0063$).



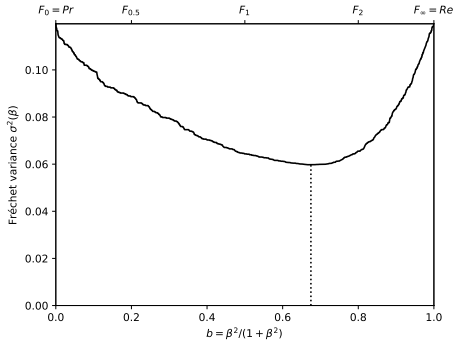
(a) The performances of 57 classifiers (BGS methods) depicted as points in the ROC space, with the isometrics of the optimal tradeoff score, from the ranking point of view, between precision and recall. See Eq. (15).



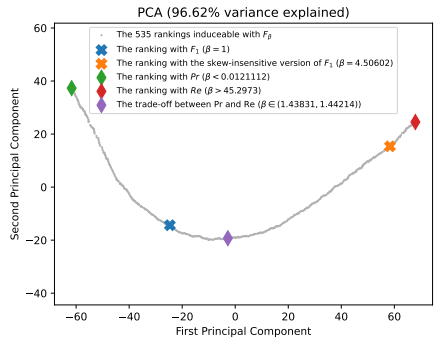
(b) The rank correlations $\tau(Pr; F_\beta)$ and $\tau(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the two curves intersect.



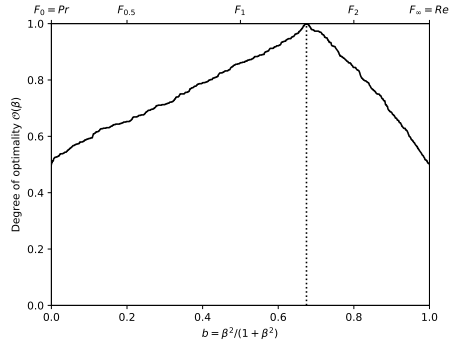
(c) The ranks of each classifier w.r.t. β. The optimal value (or range of optimal values) for β, shown here by the vertical line, is such that the number of swaps on its left is equal to the number of swaps on its right.



(d) The Fréchet variance $\sigma^2(\beta) = d_\tau^2(Pr; F_\beta) + d_\tau^2(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the curve has its minimum.

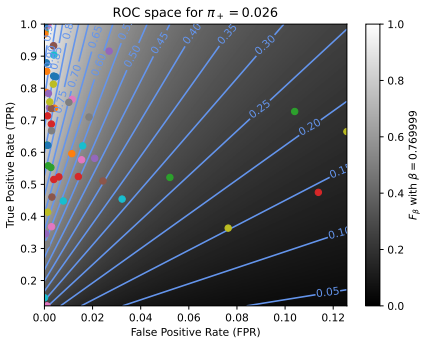


(e) Linear projection (PCA) of the manifold of the rankings induced by the F_β scores. The color points indicate the precision, the recall, F_1, $SIVF$, as well as the optimal tradeoff. The optimal tradeoff is at the same distance of the two extremities when the distance is measured along the manifold, with Kendall's distance d_τ.

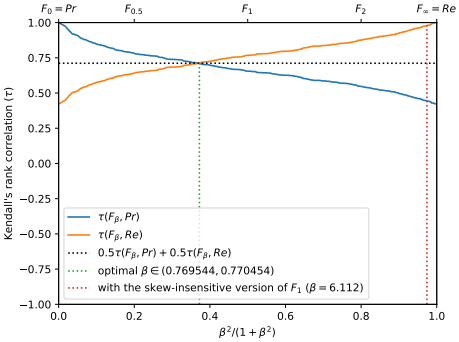


(f) The degree of optimality $\mathcal{O}(\beta)$ w.r.t. β. It is the probability to optimally ordering a pair of classifiers (BGS methods) given that it is not trivial (*i.e.*, that Pr and Re are in contradiction). The optimal value (or range of optimal values) for β is where the curve reaches 100%.

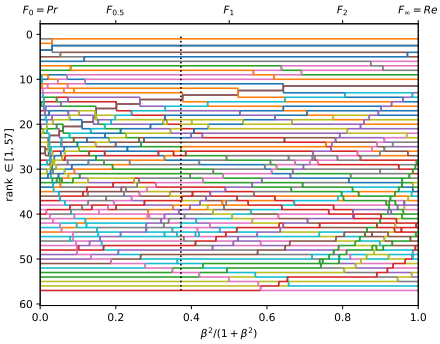
Figure A.3.19. Ranking of 57 BGS methods evaluated on the video "boulevard" ($\pi_+ = 0.0469$).



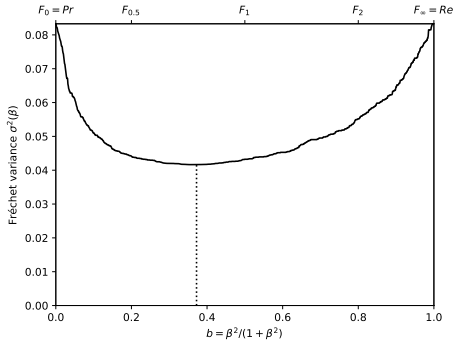
(a) The performances of 57 classifiers (BGS methods) depicted as points in the ROC space, with the isometrics of the optimal tradeoff score, from the ranking point of view, between precision and recall. See Eq. (15).



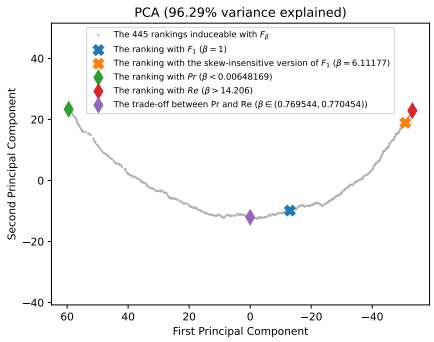
(b) The rank correlations $\tau(Pr; F_\beta)$ and $\tau(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the two curves intersect.



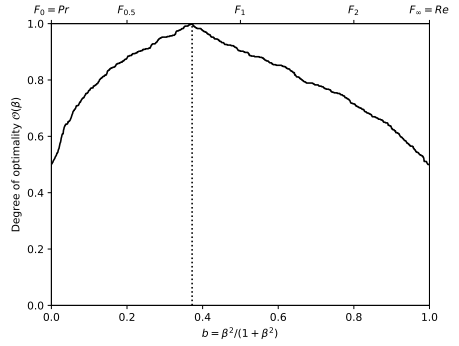
(c) The ranks of each classifier w.r.t. β. The optimal value (or range of optimal values) for β, shown here by the vertical line, is such that the number of swaps on its left is equal to the number of swaps on its right.



(d) The Fréchet variance $\sigma^2(\beta) = d_\tau^2(Pr; F_\beta) + d_\tau^2(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the curve has its minimum.

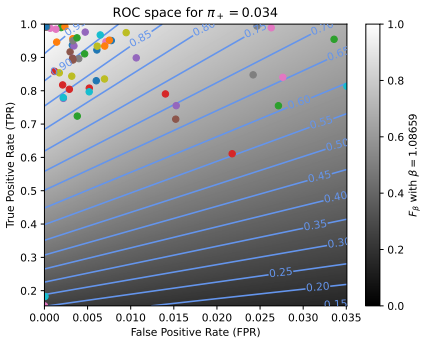


(e) Linear projection (PCA) of the manifold of the rankings induced by the F_β scores. The color points indicate the precision, the recall, F_1, $SIVF$, as well as the optimal tradeoff. The optimal tradeoff is at the same distance of the two extremities when the distance is measured along the manifold, with Kendall's distance d_τ.

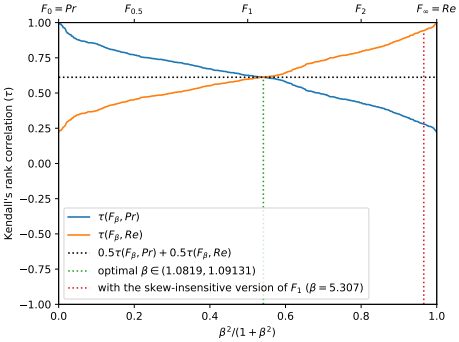


(f) The degree of optimality $\mathcal{O}(\beta)$ w.r.t. β. It is the probability to optimally ordering a pair of classifiers (BGS methods) given that it is not trivial (*i.e.*, that Pr and Re are in contradiction). The optimal value (or range of optimal values) for β is where the curve reaches 100%.

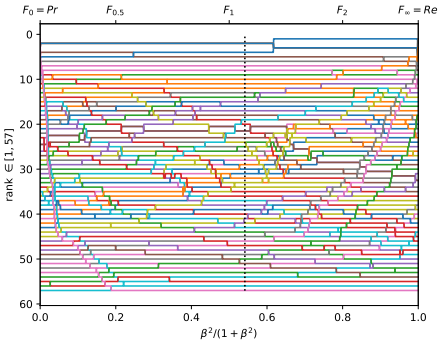
Figure A.3.20. Ranking of 57 BGS methods evaluated on the video "sidewalk" ($\pi_+ = 0.0261$).



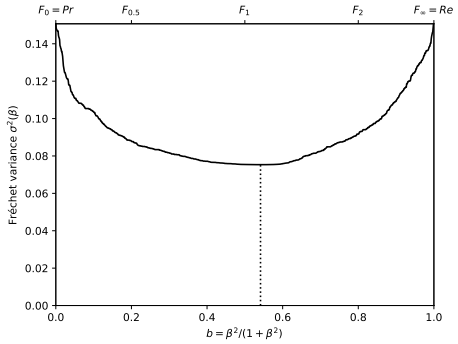
(a) The performances of 57 classifiers (BGS methods) depicted as points in the ROC space, with the isometrics of the optimal tradeoff score, from the ranking point of view, between precision and recall. See Eq. (15).



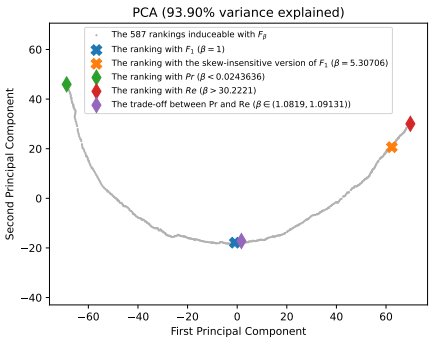
(b) The rank correlations $\tau(Pr; F_\beta)$ and $\tau(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the two curves intersect.



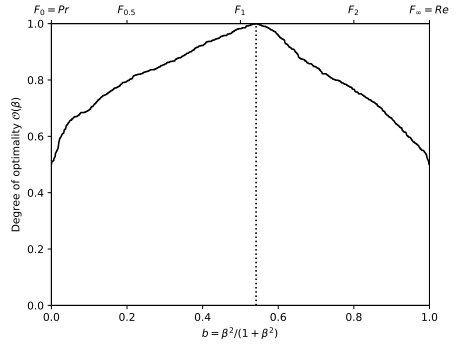
(c) The ranks of each classifier w.r.t. β. The optimal value (or range of optimal values) for β, shown here by the vertical line, is such that the number of swaps on its left is equal to the number of swaps on its right.



(d) The Fréchet variance $\sigma^2(\beta) = d_\tau^2(Pr; F_\beta) + d_\tau^2(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the curve has its minimum.

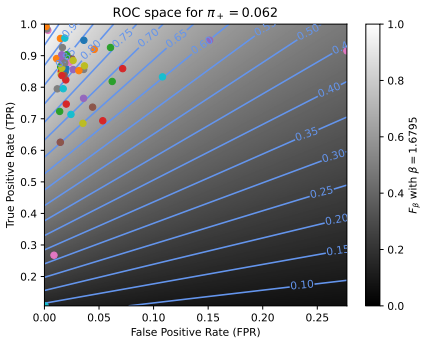


(e) Linear projection (PCA) of the manifold of the rankings induced by the F_β scores. The color points indicate the precision, the recall, F_1, $SIVF$, as well as the optimal tradeoff. The optimal tradeoff is at the same distance of the two extremities when the distance is measured along the manifold, with Kendall's distance d_τ.

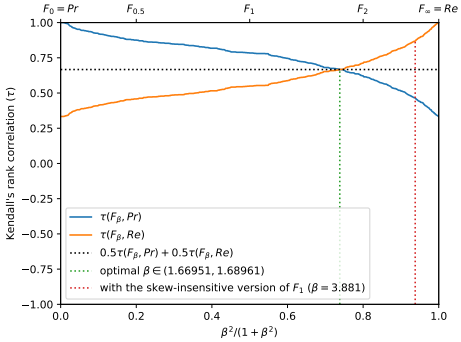


(f) The degree of optimality $\mathcal{O}(\beta)$ w.r.t. β. It is the probability to optimally ordering a pair of classifiers (BGS methods) given that it is not trivial (*i.e.*, that Pr and Re are in contradiction). The optimal value (or range of optimal values) for β is where the curve reaches 100%.

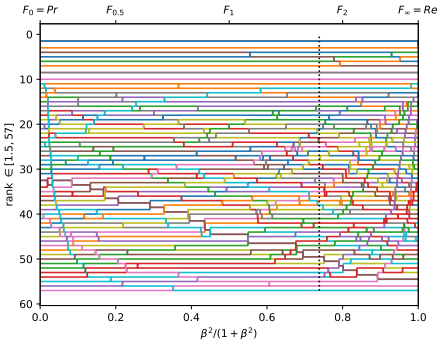
Figure A.3.21. Ranking of 57 BGS methods evaluated on the video "badminton" ($\pi_+ = 0.0343$).



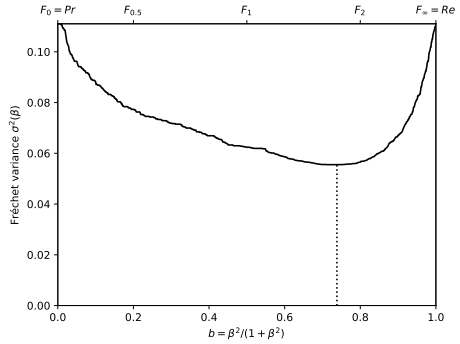
(a) The performances of 57 classifiers (BGS methods) depicted as points in the ROC space, with the isometrics of the optimal tradeoff score, from the ranking point of view, between precision and recall. See Eq. (15).



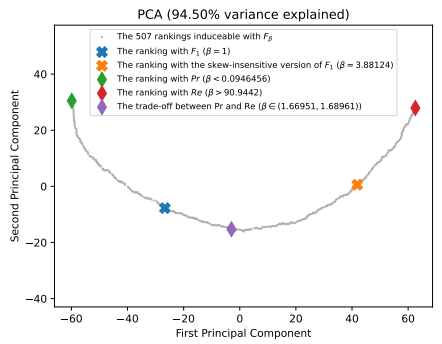
(b) The rank correlations $\tau(Pr; F_\beta)$ and $\tau(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the two curves intersect.



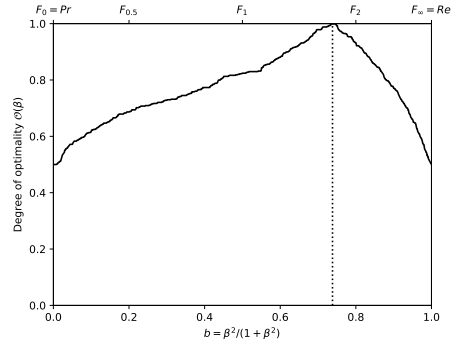
(c) The ranks of each classifier w.r.t. β. The optimal value (or range of optimal values) for β, shown here by the vertical line, is such that the number of swaps on its left is equal to the number of swaps on its right.



(d) The Fréchet variance $\sigma^2(\beta) = d_\tau^2(Pr; F_\beta) + d_\tau^2(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the curve has its minimum.

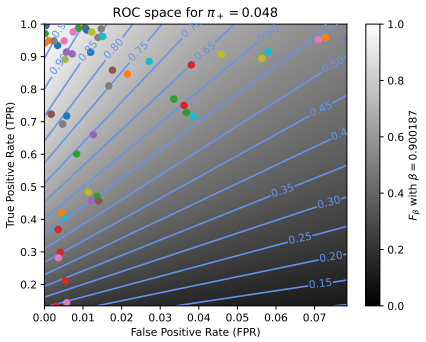


(e) Linear projection (PCA) of the manifold of the rankings induced by the F_β scores. The color points indicate the precision, the recall, F_1, $SIVF$, as well as the optimal tradeoff. The optimal tradeoff is at the same distance of the two extremities when the distance is measured along the manifold, with Kendall's distance d_τ.

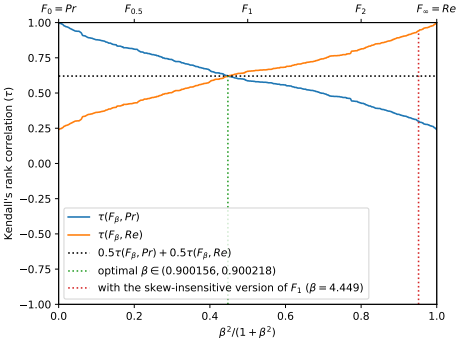


(f) The degree of optimality $\mathcal{O}(\beta)$ w.r.t. β. It is the probability to optimally ordering a pair of classifiers (BGS methods) given that it is not trivial (*i.e.*, that Pr and Re are in contradiction). The optimal value (or range of optimal values) for β is where the curve reaches 100%.

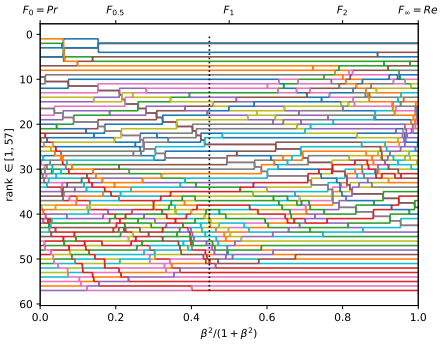
Figure A.3.22. Ranking of 57 BGS methods evaluated on the video "traffic" ($\pi_+ = 0.0623$).



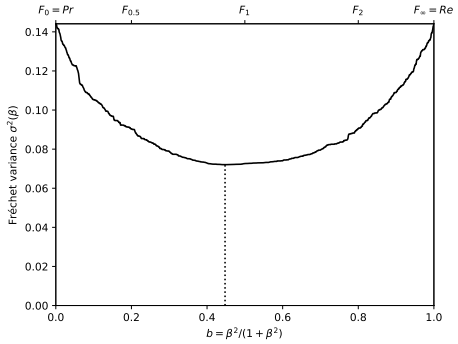
(a) The performances of 57 classifiers (BGS methods) depicted as points in the ROC space, with the isometrics of the optimal tradeoff score, from the ranking point of view, between precision and recall. See Eq. (15).



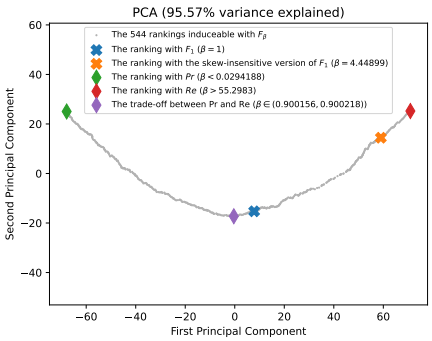
(b) The rank correlations $\tau(Pr; F_\beta)$ and $\tau(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the two curves intersect.



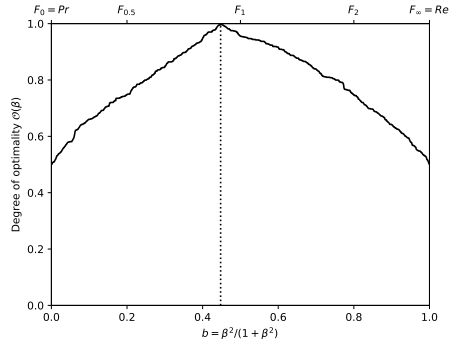
(c) The ranks of each classifier w.r.t. β. The optimal value (or range of optimal values) for β, shown here by the vertical line, is such that the number of swaps on its left is equal to the number of swaps on its right.



(d) The Fréchet variance $\sigma^2(\beta) = d_\tau^2(Pr; F_\beta) + d_\tau^2(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the curve has its minimum.

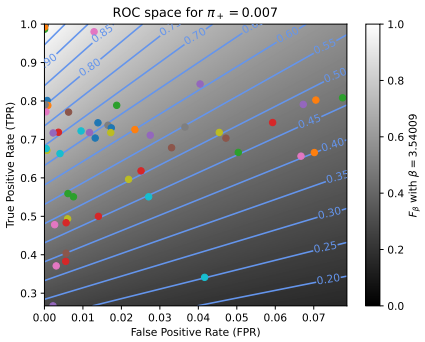


(e) Linear projection (PCA) of the manifold of the rankings induced by the F_β scores. The color points indicate the precision, the recall, F_1, $SIVF$, as well as the optimal tradeoff. The optimal tradeoff is at the same distance of the two extremities when the distance is measured along the manifold, with Kendall's distance d_τ.

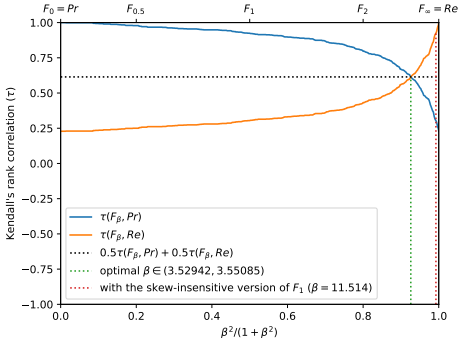


(f) The degree of optimality $\mathcal{O}(\beta)$ w.r.t. β. It is the probability to optimally ordering a pair of classifiers (BGS methods) given that it is not trivial (*i.e.*, that Pr and Re are in contradiction). The optimal value (or range of optimal values) for β is where the curve reaches 100%.

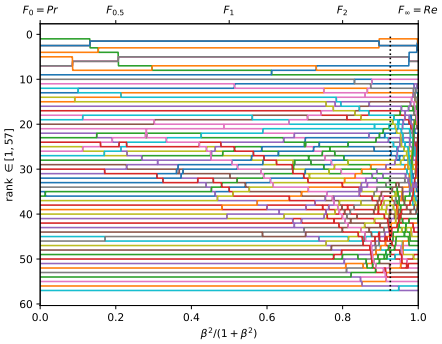
Figure A.3.23. Ranking of 57 BGS methods evaluated on the video "abandonedBox" ($\pi_+ = 0.0481$).



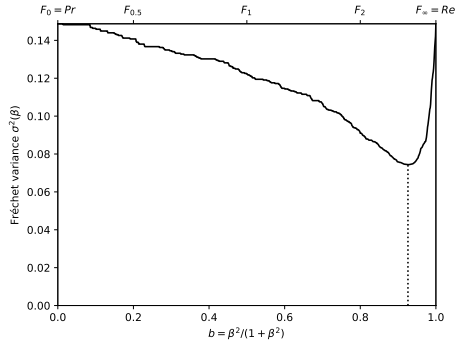
(a) The performances of 57 classifiers (BGS methods) depicted as points in the ROC space, with the isometrics of the optimal tradeoff score, from the ranking point of view, between precision and recall. See Eq. (15).



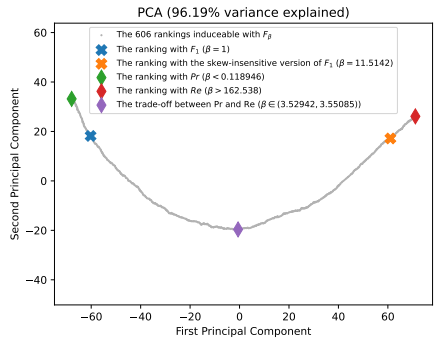
(b) The rank correlations $\tau(Pr; F_\beta)$ and $\tau(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the two curves intersect.



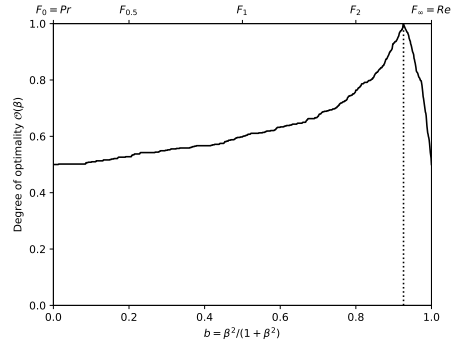
(c) The ranks of each classifier w.r.t. β. The optimal value (or range of optimal values) for β, shown here by the vertical line, is such that the number of swaps on its left is equal to the number of swaps on its right.



(d) The Fréchet variance $\sigma^2(\beta) = d_\tau^2(Pr; F_\beta) + d_\tau^2(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the curve has its minimum.

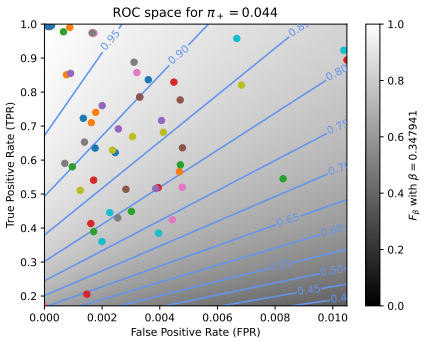


(e) Linear projection (PCA) of the manifold of the rankings induced by the F_β scores. The color points indicate the precision, the recall, F_1, $SIVF$, as well as the optimal tradeoff. The optimal tradeoff is at the same distance of the two extremities when the distance is measured along the manifold, with Kendall's distance d_τ.

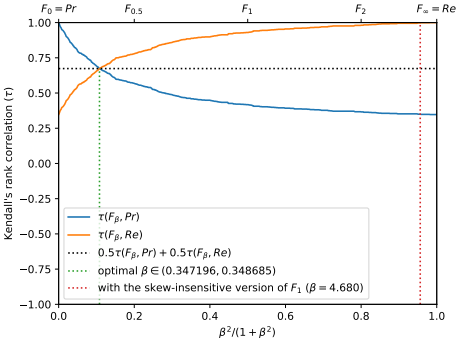


(f) The degree of optimality $\mathcal{O}(\beta)$ w.r.t. β. It is the probability to optimally ordering a pair of classifiers (BGS methods) given that it is not trivial (*i.e.*, that Pr and Re are in contradiction). The optimal value (or range of optimal values) for β is where the curve reaches 100%.

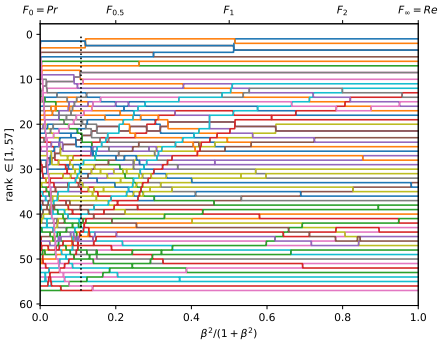
Figure A.3.24. Ranking of 57 BGS methods evaluated on the video "winterDriveway" ($\pi_+ = 0.0075$).



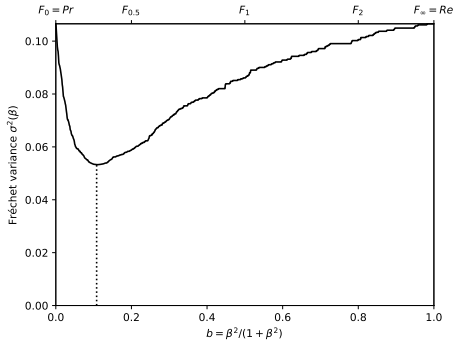
(a) The performances of 57 classifiers (BGS methods) depicted as points in the ROC space, with the isometrics of the optimal tradeoff score, from the ranking point of view, between precision and recall. See Eq. (15).



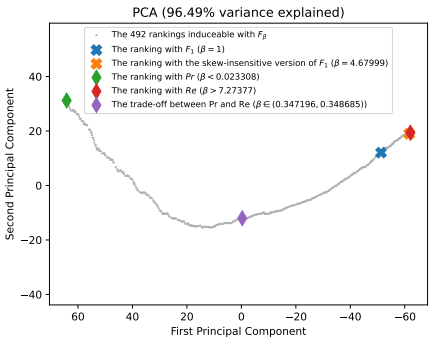
(b) The rank correlations $\tau(Pr; F_\beta)$ and $\tau(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the two curves intersect.



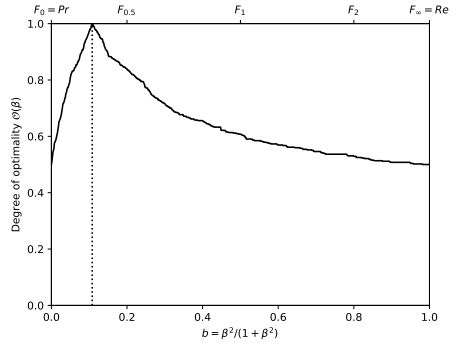
(c) The ranks of each classifier w.r.t. β. The optimal value (or range of optimal values) for β, shown here by the vertical line, is such that the number of swaps on its left is equal to the number of swaps on its right.



(d) The Fréchet variance $\sigma^2(\beta) = d_\tau^2(Pr; F_\beta) + d_\tau^2(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the curve has its minimum.

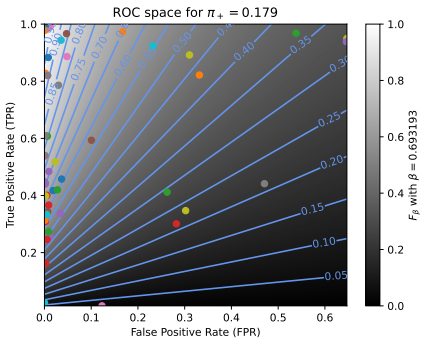


(e) Linear projection (PCA) of the manifold of the rankings induced by the F_β scores. The color points indicate the precision, the recall, F_1, $SIVF$, as well as the optimal tradeoff. The optimal tradeoff is at the same distance of the two extremities when the distance is measured along the manifold, with Kendall's distance d_τ.

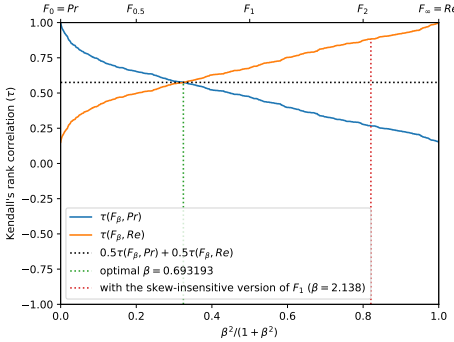


(f) The degree of optimality $\mathcal{O}(\beta)$ w.r.t. β. It is the probability to optimally ordering a pair of classifiers (BGS methods) given that it is not trivial (*i.e.*, that Pr and Re are in contradiction). The optimal value (or range of optimal values) for β is where the curve reaches 100%.

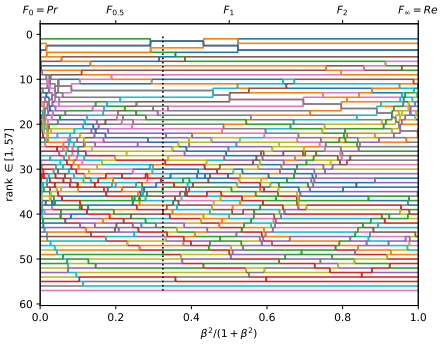
Figure A.3.25. Ranking of 57 BGS methods evaluated on the video "sofa" ($\pi_+ = 0.0437$).



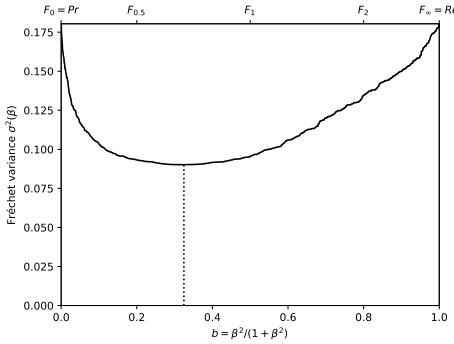
(a) The performances of 57 classifiers (BGS methods) depicted as points in the ROC space, with the isometrics of the optimal tradeoff score, from the ranking point of view, between precision and recall. See Eq. (15).



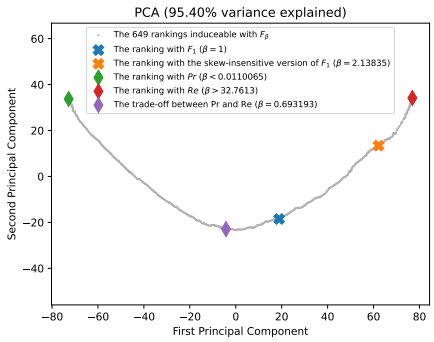
(b) The rank correlations $\tau(Pr; F_\beta)$ and $\tau(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the two curves intersect.



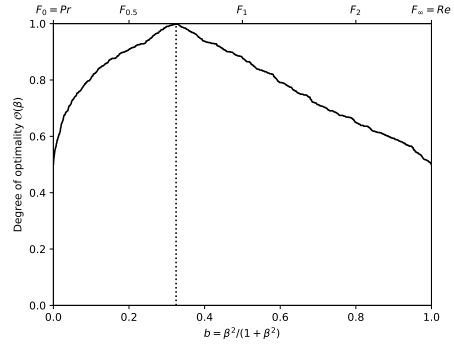
(c) The ranks of each classifier w.r.t. β. The optimal value (or range of optimal values) for β, shown here by the vertical line, is such that the number of swaps on its left is equal to the number of swaps on its right.



(d) The Fréchet variance $\sigma^2(\beta) = d_\tau^2(Pr; F_\beta) + d_\tau^2(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the curve has its minimum.

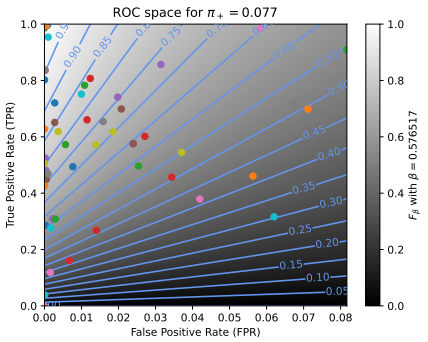


(e) Linear projection (PCA) of the manifold of the rankings induced by the F_β scores. The color points indicate the precision, the recall, F_1, $SIVF$, as well as the optimal tradeoff. The optimal tradeoff is at the same distance of the two extremities when the distance is measured along the manifold, with Kendall's distance d_τ.

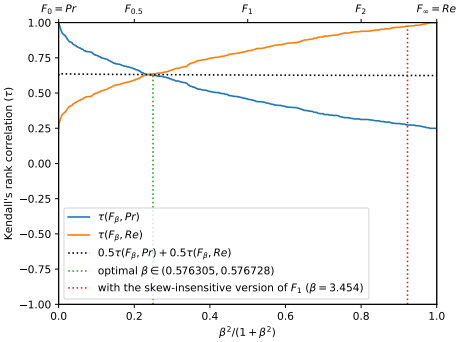


(f) The degree of optimality $\mathcal{O}(\beta)$ w.r.t. β. It is the probability to optimally ordering a pair of classifiers (BGS methods) given that it is not trivial (*i.e.*, that Pr and Re are in contradiction). The optimal value (or range of optimal values) for β is where the curve reaches 100%.

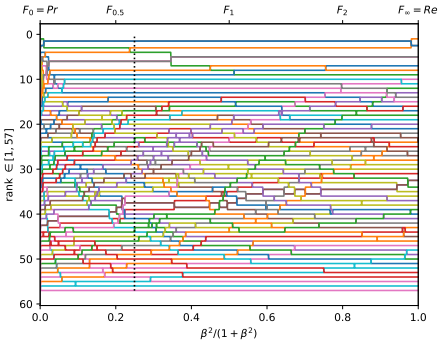
Figure A.3.26. Ranking of 57 BGS methods evaluated on the video "tramstop" ($\pi_+ = 0.1795$).



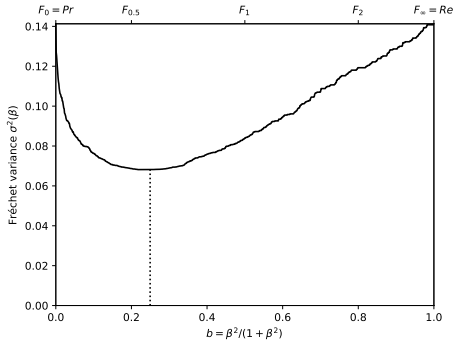
(a) The performances of 57 classifiers (BGS methods) depicted as points in the ROC space, with the isometrics of the optimal tradeoff score, from the ranking point of view, between precision and recall. See Eq. (15).



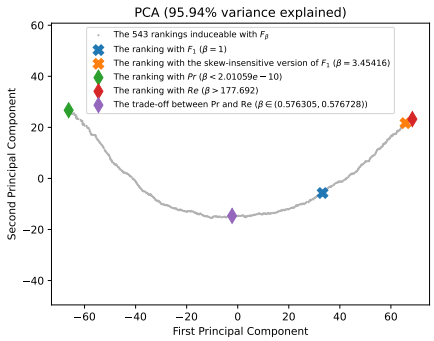
(b) The rank correlations $\tau(Pr; F_\beta)$ and $\tau(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the two curves intersect.



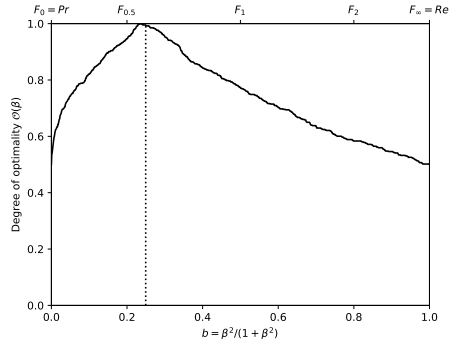
(c) The ranks of each classifier w.r.t. β. The optimal value (or range of optimal values) for β, shown here by the vertical line, is such that the number of swaps on its left is equal to the number of swaps on its right.



(d) The Fréchet variance $\sigma^2(\beta) = d_\tau^2(Pr; F_\beta) + d_\tau^2(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the curve has its minimum.

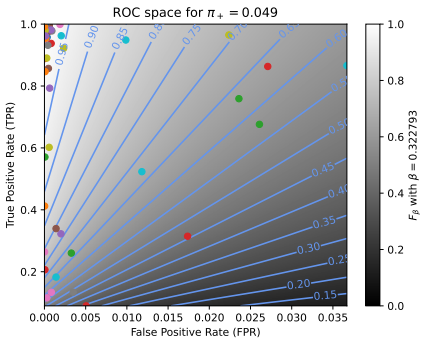


(e) Linear projection (PCA) of the manifold of the rankings induced by the F_β scores. The color points indicate the precision, the recall, F_1, $SIVF$, as well as the optimal tradeoff. The optimal tradeoff is at the same distance of the two extremities when the distance is measured along the manifold, with Kendall's distance d_τ.

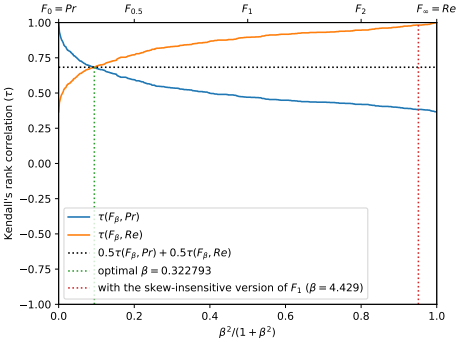


(f) The degree of optimality $\mathcal{O}(\beta)$ w.r.t. β. It is the probability to optimally ordering a pair of classifiers (BGS methods) given that it is not trivial (*i.e.*, that Pr and Re are in contradiction). The optimal value (or range of optimal values) for β is where the curve reaches 100%.

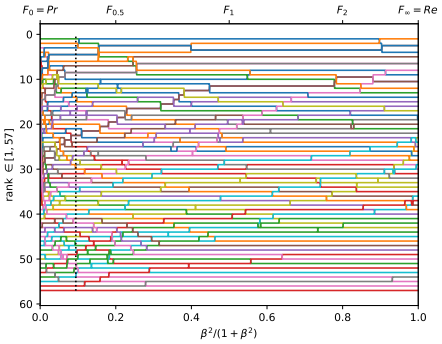
Figure A.3.27. Ranking of 57 BGS methods evaluated on the video "parking" ($\pi_+ = 0.0773$).



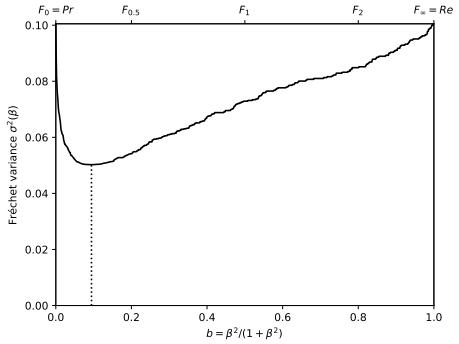
(a) The performances of 57 classifiers (BGS methods) depicted as points in the ROC space, with the isometrics of the optimal tradeoff score, from the ranking point of view, between precision and recall. See Eq. (15).



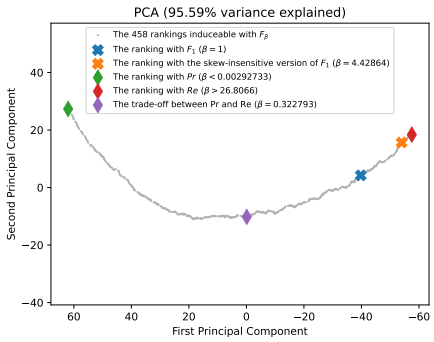
(b) The rank correlations $\tau(Pr; F_\beta)$ and $\tau(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the two curves intersect.



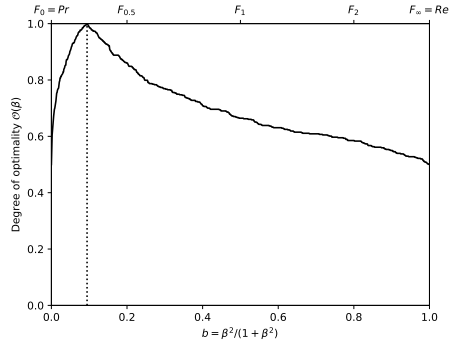
(c) The ranks of each classifier w.r.t. β. The optimal value (or range of optimal values) for β, shown here by the vertical line, is such that the number of swaps on its left is equal to the number of swaps on its right.



(d) The Fréchet variance $\sigma^2(\beta) = d_\tau^2(Pr; F_\beta) + d_\tau^2(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the curve has its minimum.

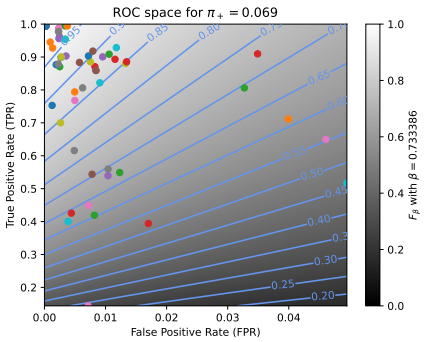


(e) Linear projection (PCA) of the manifold of the rankings induced by the F_β scores. The color points indicate the precision, the recall, F_1, $SIVF$, as well as the optimal tradeoff. The optimal tradeoff is at the same distance of the two extremities when the distance is measured along the manifold, with Kendall's distance d_τ.

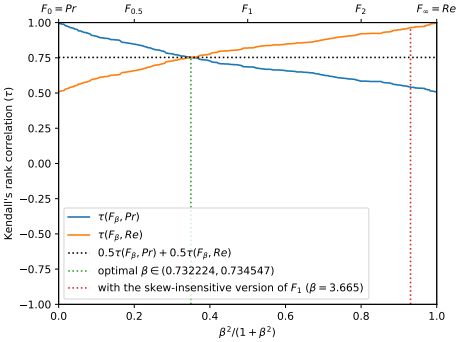


(f) The degree of optimality $\mathcal{O}(\beta)$ w.r.t. β. It is the probability to optimally ordering a pair of classifiers (BGS methods) given that it is not trivial (*i.e.*, that Pr and Re are in contradiction). The optimal value (or range of optimal values) for β is where the curve reaches 100%.

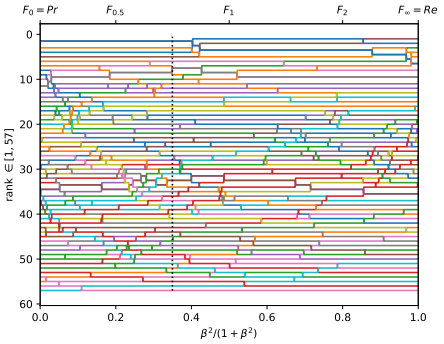
Figure A.3.28. Ranking of 57 BGS methods evaluated on the video "streetLight" ($\pi_+ = 0.0485$).



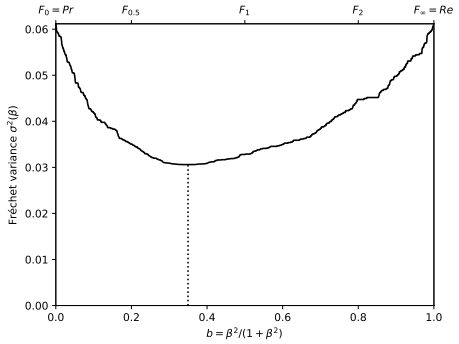
(a) The performances of 57 classifiers (BGS methods) depicted as points in the ROC space, with the isometrics of the optimal tradeoff score, from the ranking point of view, between precision and recall. See Eq. (15).



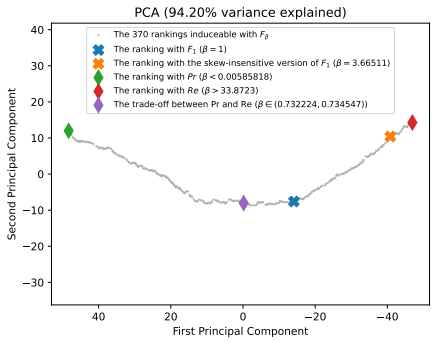
(b) The rank correlations $\tau(Pr; F_\beta)$ and $\tau(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the two curves intersect.



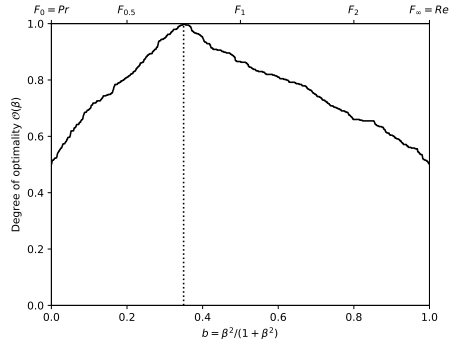
(c) The ranks of each classifier w.r.t. β. The optimal value (or range of optimal values) for β, shown here by the vertical line, is such that the number of swaps on its left is equal to the number of swaps on its right.



(d) The Fréchet variance $\sigma^2(\beta) = d_\tau^2(Pr; F_\beta) + d_\tau^2(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the curve has its minimum.

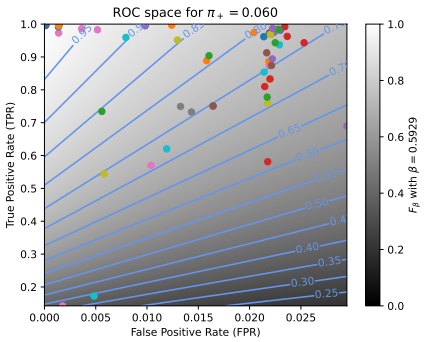


(e) Linear projection (PCA) of the manifold of the rankings induced by the F_β scores. The color points indicate the precision, the recall, F_1, $SIVF$, as well as the optimal tradeoff. The optimal tradeoff is at the same distance of the two extremities when the distance is measured along the manifold, with Kendall's distance d_τ.

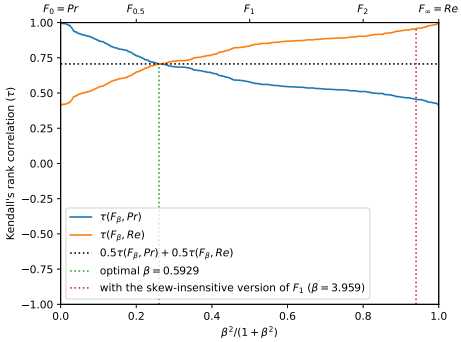


(f) The degree of optimality $\mathcal{O}(\beta)$ w.r.t. β. It is the probability to optimally ordering a pair of classifiers (BGS methods) given that it is not trivial (*i.e.*, that Pr and Re are in contradiction). The optimal value (or range of optimal values) for β is where the curve reaches 100%.

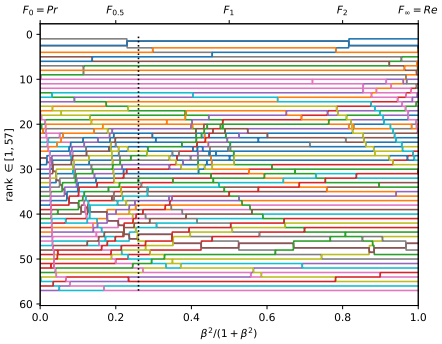
Figure A.3.29. Ranking of 57 BGS methods evaluated on the video "copyMachine" ($\pi_+ = 0.0693$).



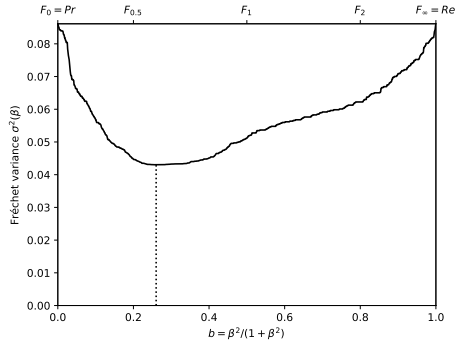
(a) The performances of 57 classifiers (BGS methods) depicted as points in the ROC space, with the isometrics of the optimal tradeoff score, from the ranking point of view, between precision and recall. See Eq. (15).



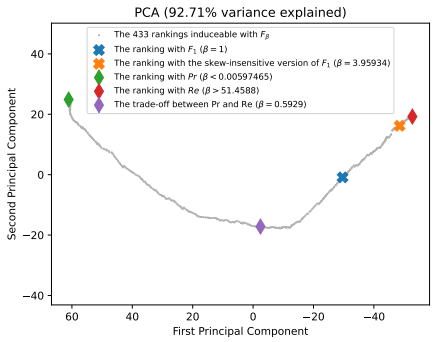
(b) The rank correlations $\tau(Pr; F_\beta)$ and $\tau(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the two curves intersect.



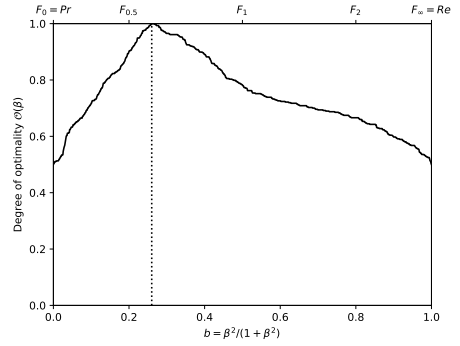
(c) The ranks of each classifier w.r.t. β. The optimal value (or range of optimal values) for β, shown here by the vertical line, is such that the number of swaps on its left is equal to the number of swaps on its right.



(d) The Fréchet variance $\sigma^2(\beta) = d_\tau^2(Pr; F_\beta) + d_\tau^2(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the curve has its minimum.

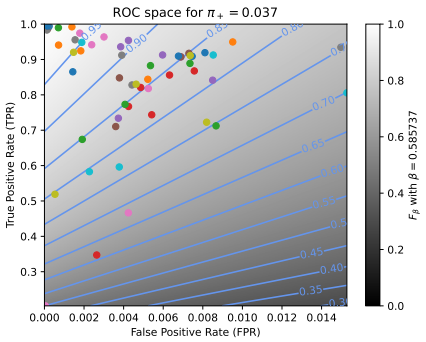


(e) Linear projection (PCA) of the manifold of the rankings induced by the F_β scores. The color points indicate the precision, the recall, F_1, $SIVF$, as well as the optimal tradeoff. The optimal tradeoff is at the same distance of the two extremities when the distance is measured along the manifold, with Kendall's distance d_τ.

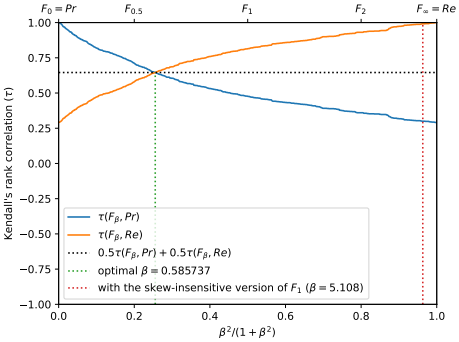


(f) The degree of optimality $\mathcal{O}(\beta)$ w.r.t. β. It is the probability to optimally ordering a pair of classifiers (BGS methods) given that it is not trivial (*i.e.*, that Pr and Re are in contradiction). The optimal value (or range of optimal values) for β is where the curve reaches 100%.

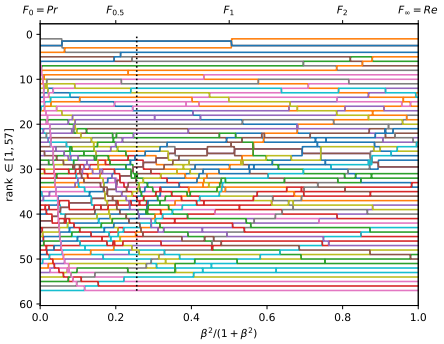
Figure A.3.30. Ranking of 57 BGS methods evaluated on the video "bungalows" ($\pi_+ = 0.0600$).



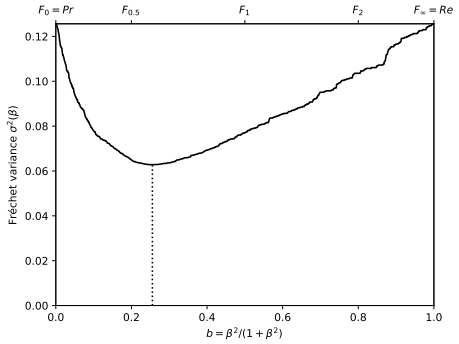
(a) The performances of 57 classifiers (BGS methods) depicted as points in the ROC space, with the isometrics of the optimal tradeoff score, from the ranking point of view, between precision and recall. See Eq. (15).



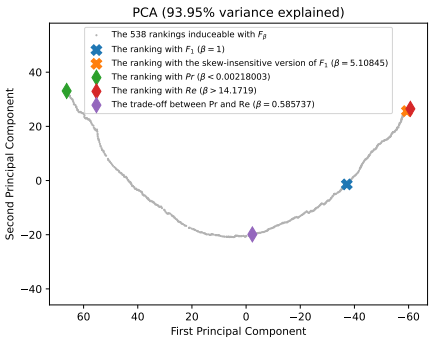
(b) The rank correlations $\tau(Pr; F_\beta)$ and $\tau(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the two curves intersect.



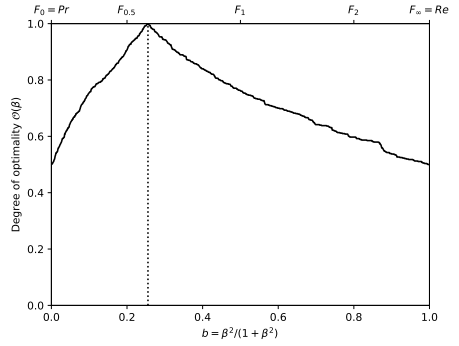
(c) The ranks of each classifier w.r.t. β. The optimal value (or range of optimal values) for β, shown here by the vertical line, is such that the number of swaps on its left is equal to the number of swaps on its right.



(d) The Fréchet variance $\sigma^2(\beta) = d_\tau^2(Pr; F_\beta) + d_\tau^2(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the curve has its minimum.

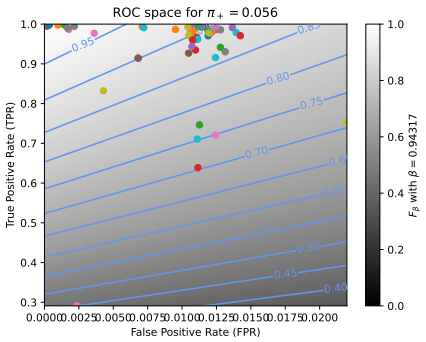


(e) Linear projection (PCA) of the manifold of the rankings induced by the F_β scores. The color points indicate the precision, the recall, F_1, $SIVF$, as well as the optimal tradeoff. The optimal tradeoff is at the same distance of the two extremities when the distance is measured along the manifold, with Kendall's distance d_τ.

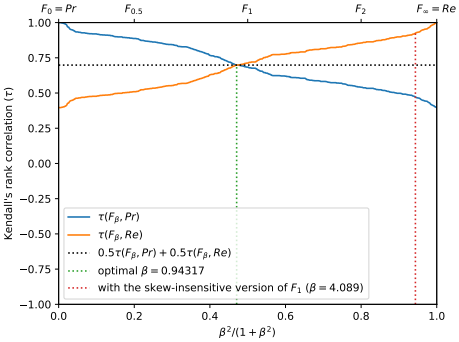


(f) The degree of optimality $\mathcal{O}(\beta)$ w.r.t. β. It is the probability to optimally ordering a pair of classifiers (BGS methods) given that it is not trivial (*i.e.*, that Pr and Re are in contradiction). The optimal value (or range of optimal values) for β is where the curve reaches 100%.

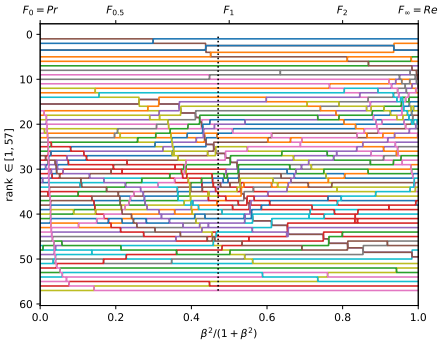
Figure A.3.31. Ranking of 57 BGS methods evaluated on the video "busStation" ($\pi_+ = 0.0369$).



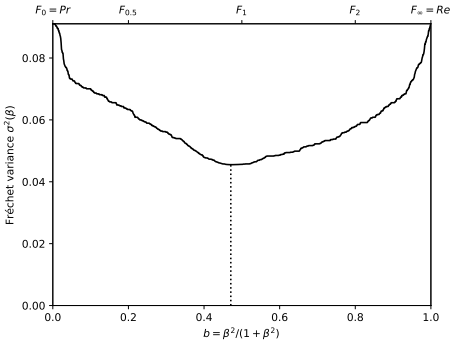
(a) The performances of 57 classifiers (BGS methods) depicted as points in the ROC space, with the isometrics of the optimal tradeoff score, from the ranking point of view, between precision and recall. See Eq. (15).



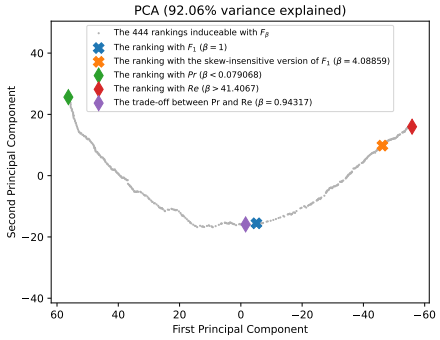
(b) The rank correlations $\tau(Pr; F_\beta)$ and $\tau(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the two curves intersect.



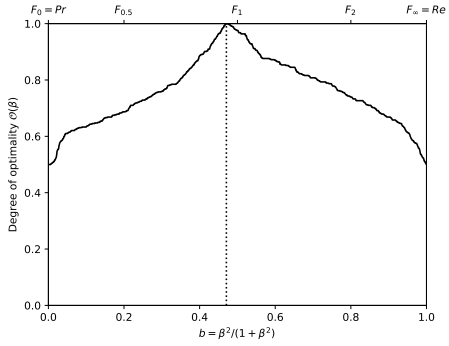
(c) The ranks of each classifier w.r.t. β. The optimal value (or range of optimal values) for β, shown here by the vertical line, is such that the number of swaps on its left is equal to the number of swaps on its right.



(d) The Fréchet variance $\sigma^2(\beta) = d_\tau^2(Pr; F_\beta) + d_\tau^2(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the curve has its minimum.

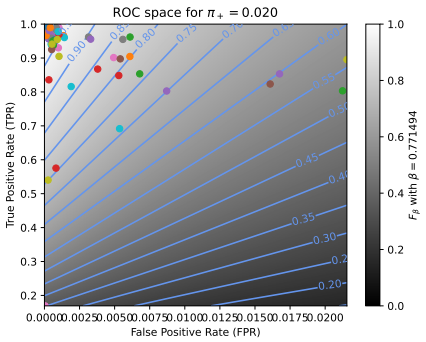


(e) Linear projection (PCA) of the manifold of the rankings induced by the F_β scores. The color points indicate the precision, the recall, F_1, $SIVF$, as well as the optimal tradeoff. The optimal tradeoff is at the same distance of the two extremities when the distance is measured along the manifold, with Kendall's distance d_τ.

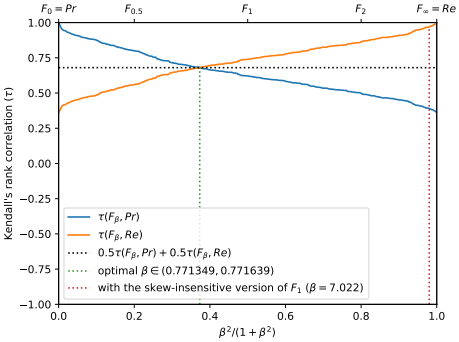


(f) The degree of optimality $\mathcal{O}(\beta)$ w.r.t. β. It is the probability to optimally ordering a pair of classifiers (BGS methods) given that it is not trivial (*i.e.*, that Pr and Re are in contradiction). The optimal value (or range of optimal values) for β is where the curve reaches 100%.

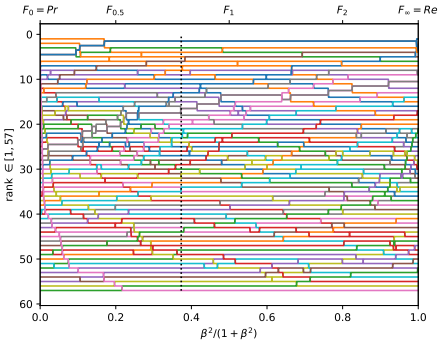
Figure A.3.32. Ranking of 57 BGS methods evaluated on the video "peopleInShade" ($\pi_+ = 0.0564$).



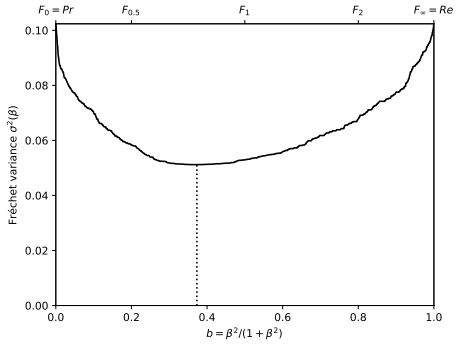
(a) The performances of 57 classifiers (BGS methods) depicted as points in the ROC space, with the isometrics of the optimal tradeoff score, from the ranking point of view, between precision and recall. See Eq. (15).



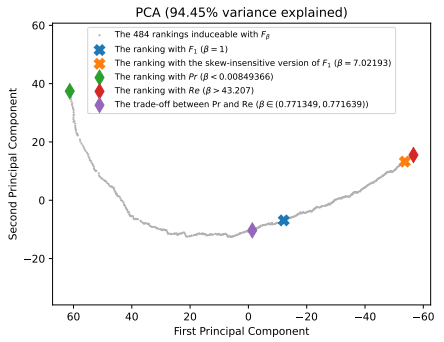
(b) The rank correlations $\tau(Pr; F_\beta)$ and $\tau(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the two curves intersect.



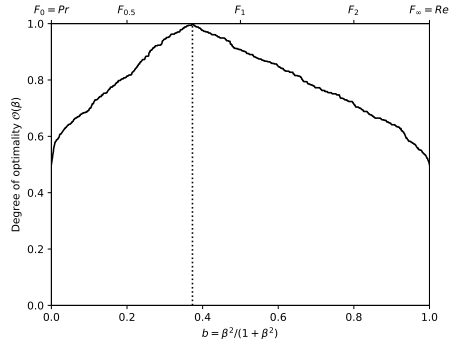
(c) The ranks of each classifier w.r.t. β. The optimal value (or range of optimal values) for β, shown here by the vertical line, is such that the number of swaps on its left is equal to the number of swaps on its right.



(d) The Fréchet variance $\sigma^2(\beta) = d_\tau^2(Pr; F_\beta) + d_\tau^2(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the curve has its minimum.

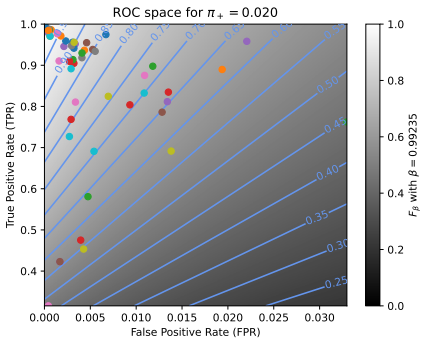


(e) Linear projection (PCA) of the manifold of the rankings induced by the F_β scores. The color points indicate the precision, the recall, F_1, $SIVF$, as well as the optimal tradeoff. The optimal tradeoff is at the same distance of the two extremities when the distance is measured along the manifold, with Kendall's distance d_τ.

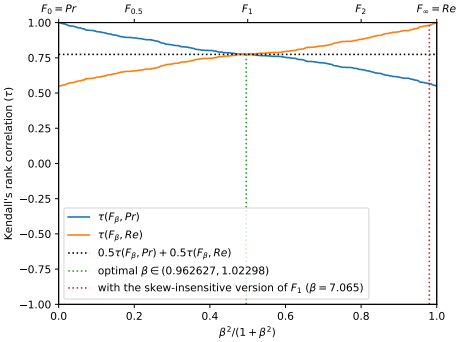


(f) The degree of optimality $\mathcal{O}(\beta)$ w.r.t. β. It is the probability to optimally ordering a pair of classifiers (BGS methods) given that it is not trivial (*i.e.*, that Pr and Re are in contradiction). The optimal value (or range of optimal values) for β is where the curve reaches 100%.

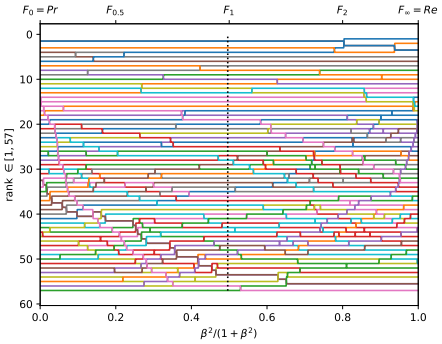
Figure A.3.33. Ranking of 57 BGS methods evaluated on the video "backdoor" ($\pi_+ = 0.0199$).



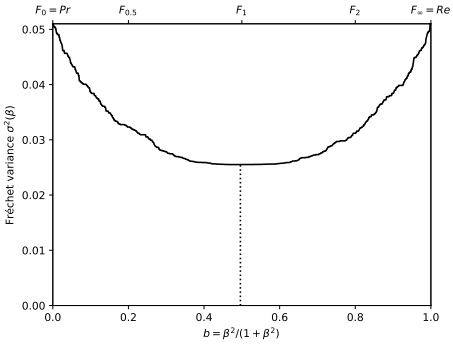
(a) The performances of 57 classifiers (BGS methods) depicted as points in the ROC space, with the isometrics of the optimal tradeoff score, from the ranking point of view, between precision and recall. See Eq. (15).



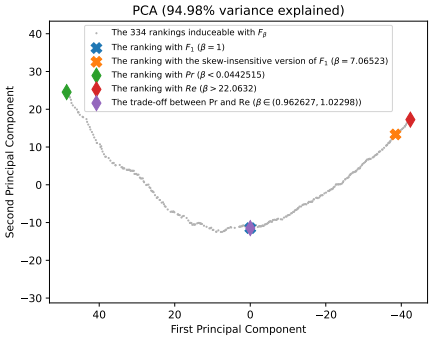
(b) The rank correlations $\tau(Pr; F_\beta)$ and $\tau(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the two curves intersect.



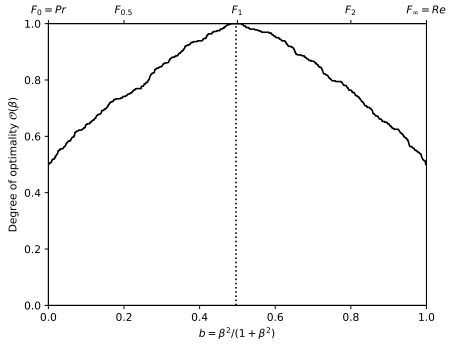
(c) The ranks of each classifier w.r.t. β. The optimal value (or range of optimal values) for β, shown here by the vertical line, is such that the number of swaps on its left is equal to the number of swaps on its right.



(d) The Fréchet variance $\sigma^2(\beta) = d_\tau^2(Pr; F_\beta) + d_\tau^2(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the curve has its minimum.

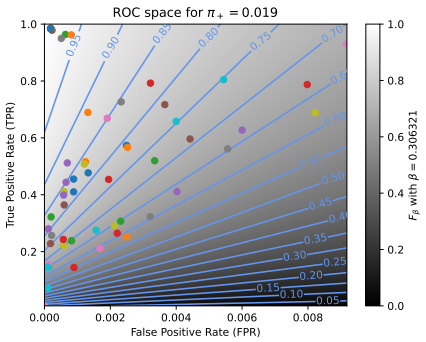


(e) Linear projection (PCA) of the manifold of the rankings induced by the F_β scores. The color points indicate the precision, the recall, F_1, $SIVF$, as well as the optimal tradeoff. The optimal tradeoff is at the same distance of the two extremities when the distance is measured along the manifold, with Kendall's distance d_τ.

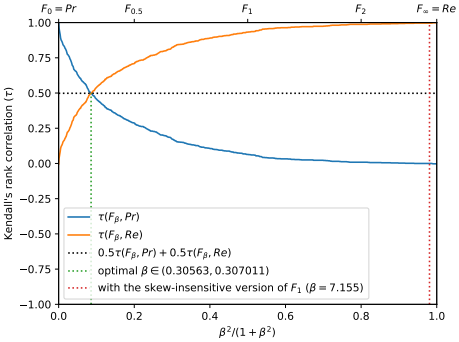


(f) The degree of optimality $\mathcal{O}(\beta)$ w.r.t. β. It is the probability to optimally ordering a pair of classifiers (BGS methods) given that it is not trivial (*i.e.*, that Pr and Re are in contradiction). The optimal value (or range of optimal values) for β is where the curve reaches 100%.

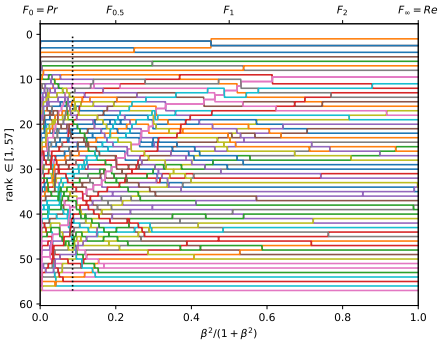
Figure A.3.34. Ranking of 57 BGS methods evaluated on the video "cubicle" ($\pi_+ = 0.0196$).



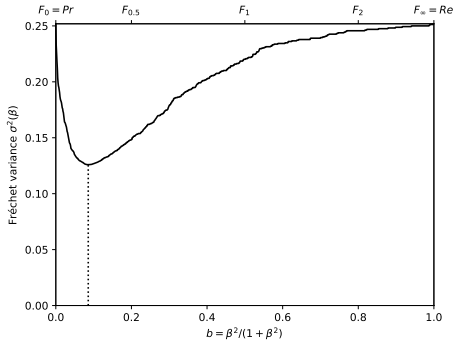
(a) The performances of 57 classifiers (BGS methods) depicted as points in the ROC space, with the isometrics of the optimal tradeoff score, from the ranking point of view, between precision and recall. See Eq. (15).



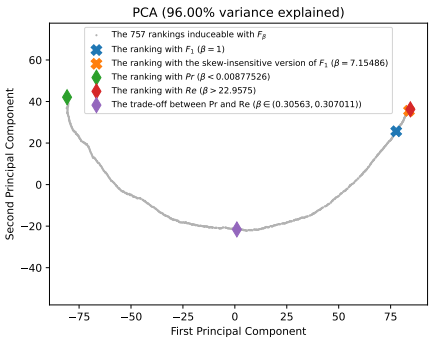
(b) The rank correlations $\tau(Pr; F_\beta)$ and $\tau(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the two curves intersect.



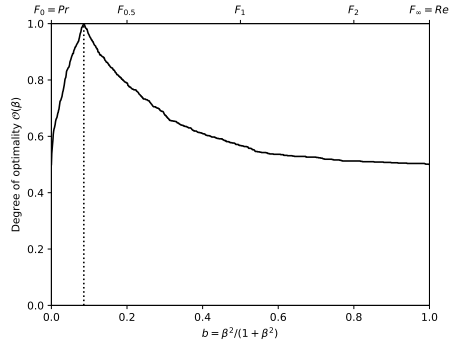
(c) The ranks of each classifier w.r.t. β. The optimal value (or range of optimal values) for β, shown here by the vertical line, is such that the number of swaps on its left is equal to the number of swaps on its right.



(d) The Fréchet variance $\sigma^2(\beta) = d_\tau^2(Pr; F_\beta) + d_\tau^2(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the curve has its minimum.

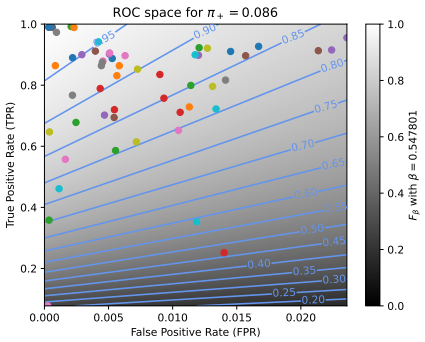


(e) Linear projection (PCA) of the manifold of the rankings induced by the F_β scores. The color points indicate the precision, the recall, F_1, $SIVF$, as well as the optimal tradeoff. The optimal tradeoff is at the same distance of the two extremities when the distance is measured along the manifold, with Kendall's distance d_τ.

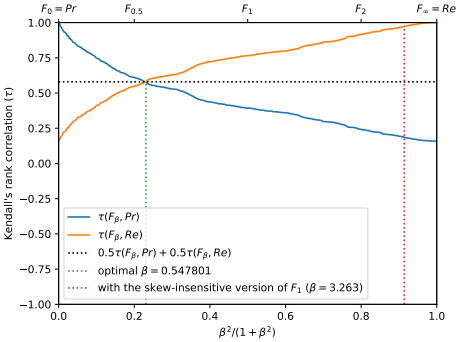


(f) The degree of optimality $\mathcal{O}(\beta)$ w.r.t. β. It is the probability to optimally ordering a pair of classifiers (BGS methods) given that it is not trivial (*i.e.*, that Pr and Re are in contradiction). The optimal value (or range of optimal values) for β is where the curve reaches 100%.

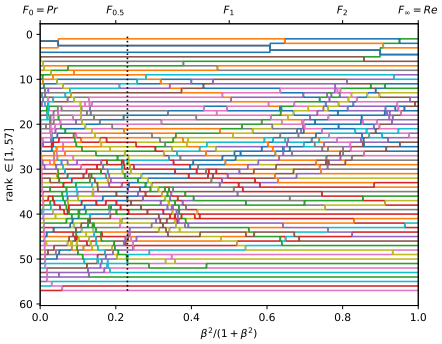
Figure A.3.35. Ranking of 57 BGS methods evaluated on the video "lakeSide" ($\pi_+ = 0.0192$).



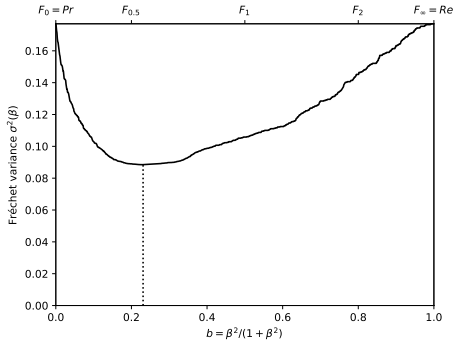
(a) The performances of 57 classifiers (BGS methods) depicted as points in the ROC space, with the isometrics of the optimal tradeoff score, from the ranking point of view, between precision and recall. See Eq. (15).



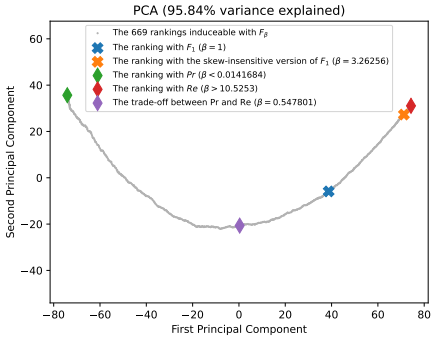
(b) The rank correlations $\tau(Pr; F_\beta)$ and $\tau(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the two curves intersect.



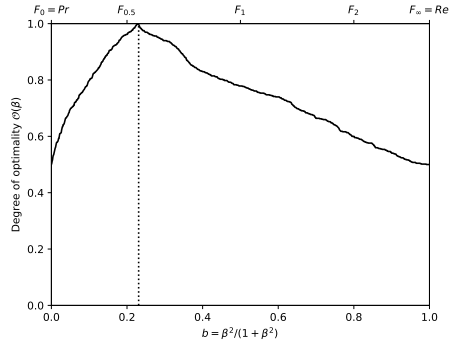
(c) The ranks of each classifier w.r.t. β. The optimal value (or range of optimal values) for β, shown here by the vertical line, is such that the number of swaps on its left is equal to the number of swaps on its right.



(d) The Fréchet variance $\sigma^2(\beta) = d_\tau^2(Pr; F_\beta) + d_\tau^2(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the curve has its minimum.

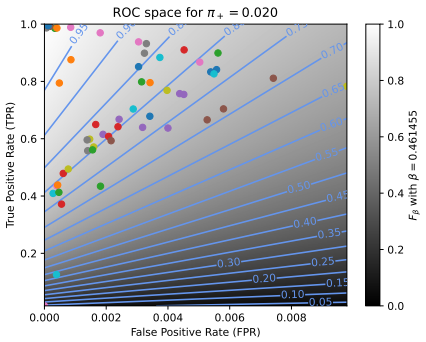


(e) Linear projection (PCA) of the manifold of the rankings induced by the F_β scores. The color points indicate the precision, the recall, F_1, $SIVF$, as well as the optimal tradeoff. The optimal tradeoff is at the same distance of the two extremities when the distance is measured along the manifold, with Kendall's distance d_τ.

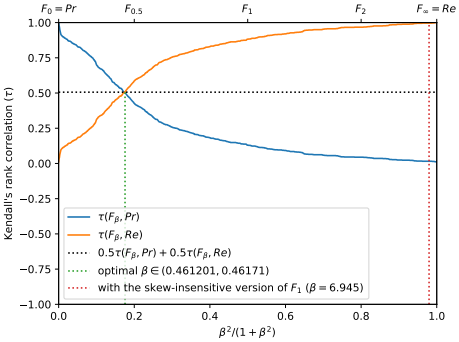


(f) The degree of optimality $\mathcal{O}(\beta)$ w.r.t. β. It is the probability to optimally ordering a pair of classifiers (BGS methods) given that it is not trivial (*i.e.*, that Pr and Re are in contradiction). The optimal value (or range of optimal values) for β is where the curve reaches 100%.

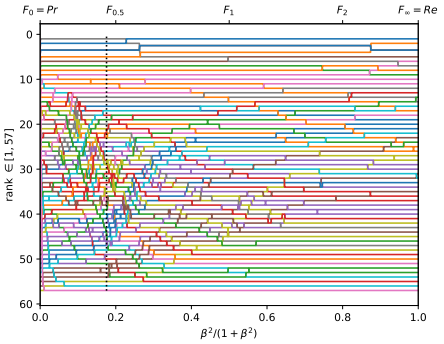
Figure A.3.36. Ranking of 57 BGS methods evaluated on the video "diningRoom" ($\pi_+ = 0.0859$).



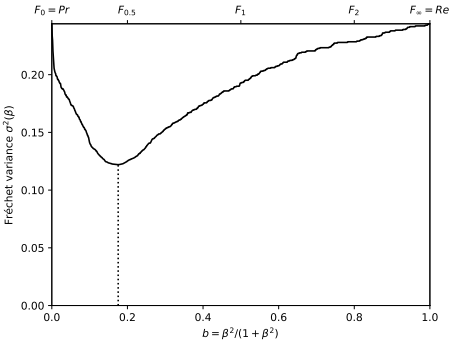
(a) The performances of 57 classifiers (BGS methods) depicted as points in the ROC space, with the isometrics of the optimal tradeoff score, from the ranking point of view, between precision and recall. See Eq. (15).



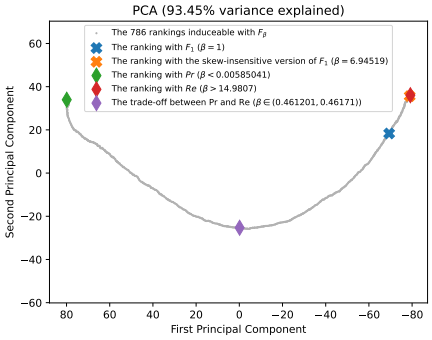
(b) The rank correlations $\tau(Pr; F_\beta)$ and $\tau(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the two curves intersect.



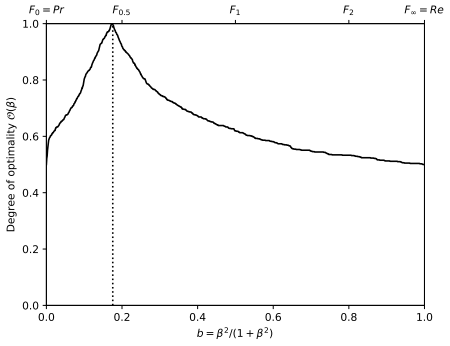
(c) The ranks of each classifier w.r.t. β. The optimal value (or range of optimal values) for β, shown here by the vertical line, is such that the number of swaps on its left is equal to the number of swaps on its right.



(d) The Fréchet variance $\sigma^2(\beta) = d_\tau^2(Pr; F_\beta) + d_\tau^2(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the curve has its minimum.

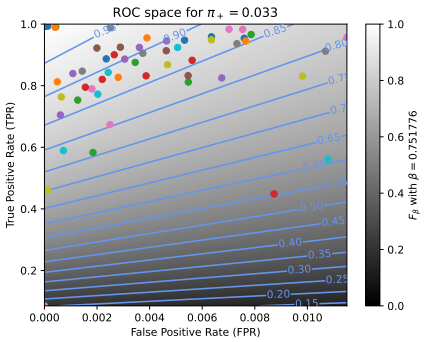


(e) Linear projection (PCA) of the manifold of the rankings induced by the F_β scores. The color points indicate the precision, the recall, F_1, $SIVF$, as well as the optimal tradeoff. The optimal tradeoff is at the same distance of the two extremities when the distance is measured along the manifold, with Kendall's distance d_τ.

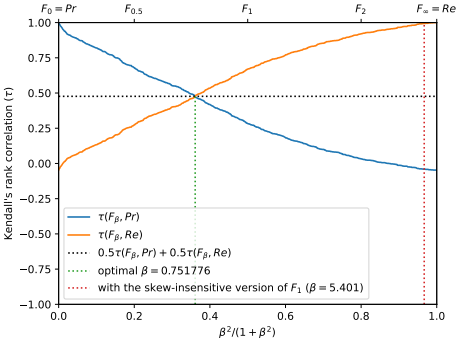


(f) The degree of optimality $\mathcal{O}(\beta)$ w.r.t. β. It is the probability to optimally ordering a pair of classifiers (BGS methods) given that it is not trivial (*i.e.*, that Pr and Re are in contradiction). The optimal value (or range of optimal values) for β is where the curve reaches 100%.

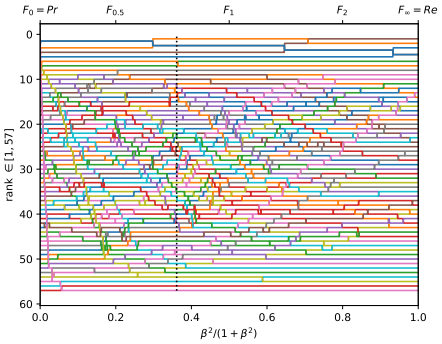
Figure A.3.37. Ranking of 57 BGS methods evaluated on the video "park" ($\pi_+ = 0.0203$).



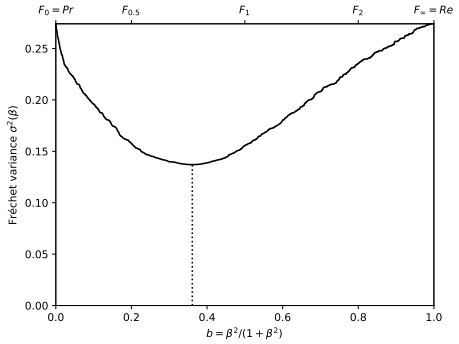
(a) The performances of 57 classifiers (BGS methods) depicted as points in the ROC space, with the isometrics of the optimal tradeoff score, from the ranking point of view, between precision and recall. See Eq. (15).



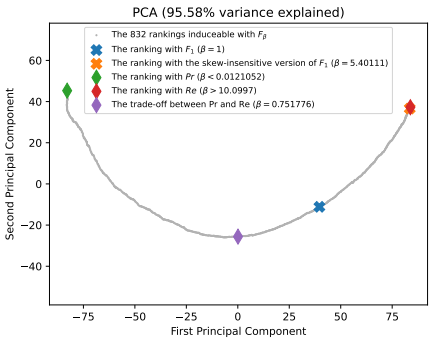
(b) The rank correlations $\tau(Pr; F_\beta)$ and $\tau(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the two curves intersect.



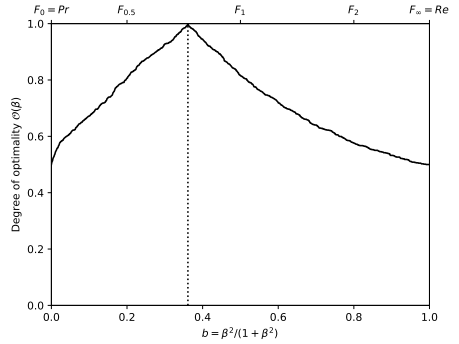
(c) The ranks of each classifier w.r.t. β. The optimal value (or range of optimal values) for β, shown here by the vertical line, is such that the number of swaps on its left is equal to the number of swaps on its right.



(d) The Fréchet variance $\sigma^2(\beta) = d_\tau^2(Pr; F_\beta) + d_\tau^2(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the curve has its minimum.

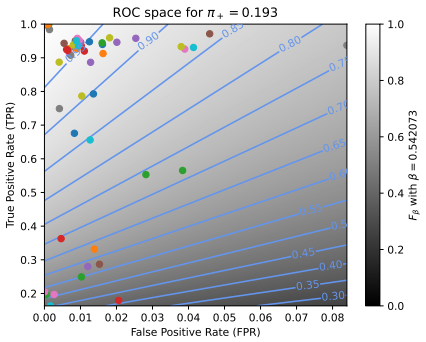


(e) Linear projection (PCA) of the manifold of the rankings induced by the F_β scores. The color points indicate the precision, the recall, F_1, $SIVF$, as well as the optimal tradeoff. The optimal tradeoff is at the same distance of the two extremities when the distance is measured along the manifold, with Kendall's distance d_τ.

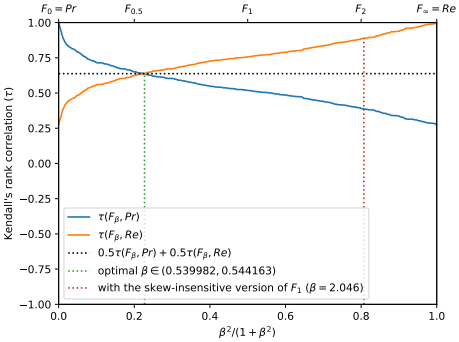


(f) The degree of optimality $\mathcal{O}(\beta)$ w.r.t. β. It is the probability to optimally ordering a pair of classifiers (BGS methods) given that it is not trivial (*i.e.*, that Pr and Re are in contradiction). The optimal value (or range of optimal values) for β is where the curve reaches 100%.

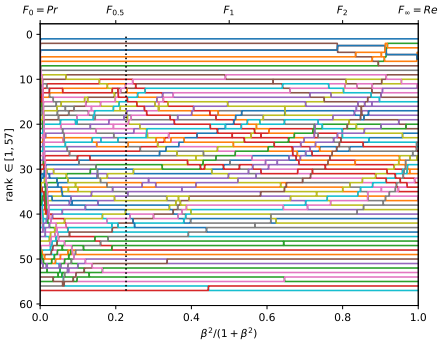
Figure A.3.38. Ranking of 57 BGS methods evaluated on the video "corridor" ($\pi_+ = 0.0331$).



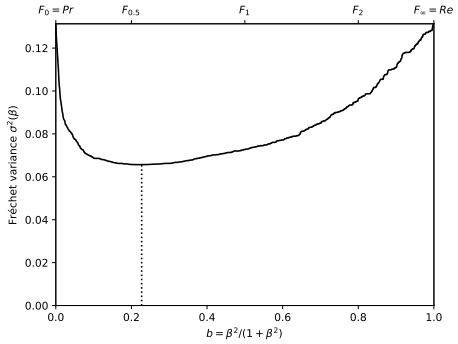
(a) The performances of 57 classifiers (BGS methods) depicted as points in the ROC space, with the isometrics of the optimal tradeoff score, from the ranking point of view, between precision and recall. See Eq. (15).



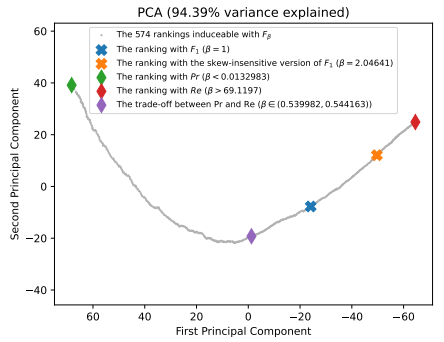
(b) The rank correlations $\tau(Pr; F_\beta)$ and $\tau(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the two curves intersect.



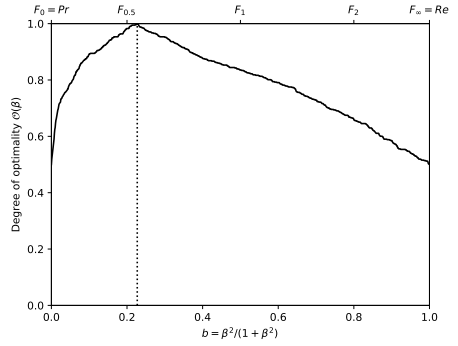
(c) The ranks of each classifier w.r.t. β. The optimal value (or range of optimal values) for β, shown here by the vertical line, is such that the number of swaps on its left is equal to the number of swaps on its right.



(d) The Fréchet variance $\sigma^2(\beta) = d_\tau^2(Pr; F_\beta) + d_\tau^2(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the curve has its minimum.

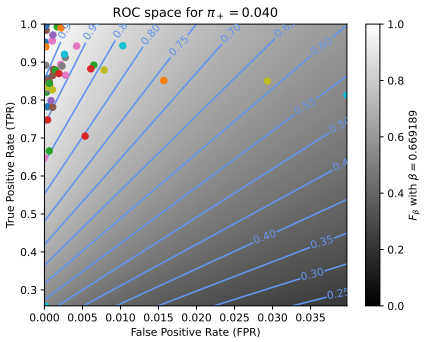


(e) Linear projection (PCA) of the manifold of the rankings induced by the F_β scores. The color points indicate the precision, the recall, F_1, $SIVF$, as well as the optimal tradeoff. The optimal tradeoff is at the same distance of the two extremities when the distance is measured along the manifold, with Kendall's distance d_τ.

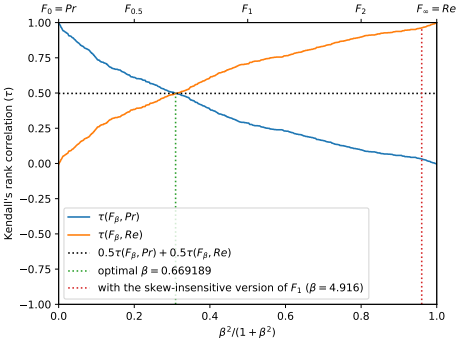


(f) The degree of optimality $\mathcal{O}(\beta)$ w.r.t. β. It is the probability to optimally ordering a pair of classifiers (BGS methods) given that it is not trivial (*i.e.*, that Pr and Re are in contradiction). The optimal value (or range of optimal values) for β is where the curve reaches 100%.

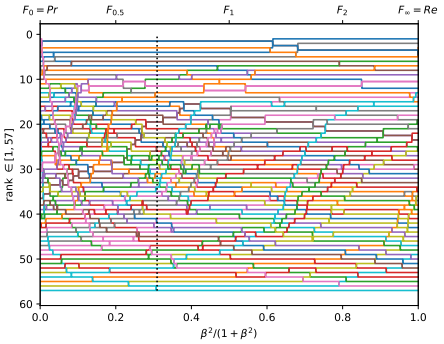
Figure A.3.39. Ranking of 57 BGS methods evaluated on the video "library" ($\pi_+ = 0.1928$).



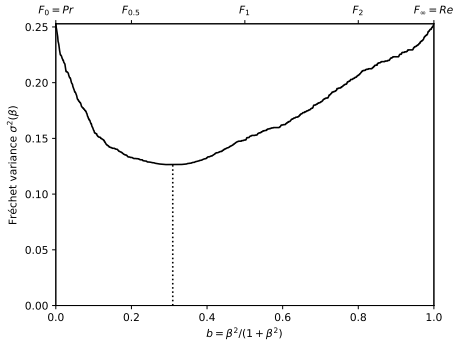
(a) The performances of 57 classifiers (BGS methods) depicted as points in the ROC space, with the isometrics of the optimal tradeoff score, from the ranking point of view, between precision and recall. See Eq. (15).



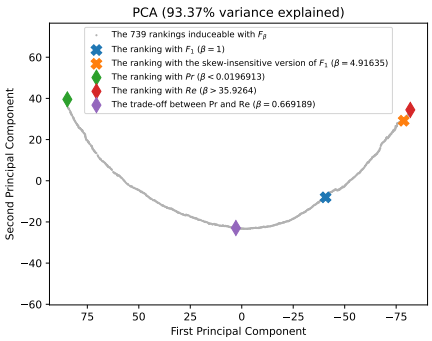
(b) The rank correlations $\tau(Pr; F_\beta)$ and $\tau(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the two curves intersect.



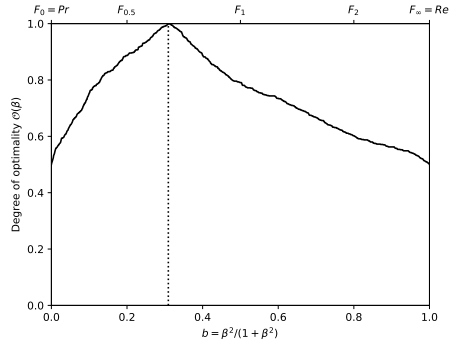
(c) The ranks of each classifier w.r.t. β. The optimal value (or range of optimal values) for β, shown here by the vertical line, is such that the number of swaps on its left is equal to the number of swaps on its right.



(d) The Fréchet variance $\sigma^2(\beta) = d_\tau^2(Pr; F_\beta) + d_\tau^2(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the curve has its minimum.

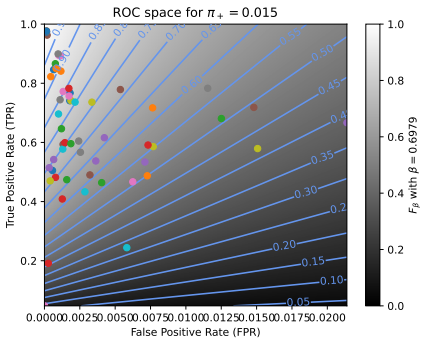


(e) Linear projection (PCA) of the manifold of the rankings induced by the F_β scores. The color points indicate the precision, the recall, F_1, $SIVF$, as well as the optimal tradeoff. The optimal tradeoff is at the same distance of the two extremities when the distance is measured along the manifold, with Kendall's distance d_τ.

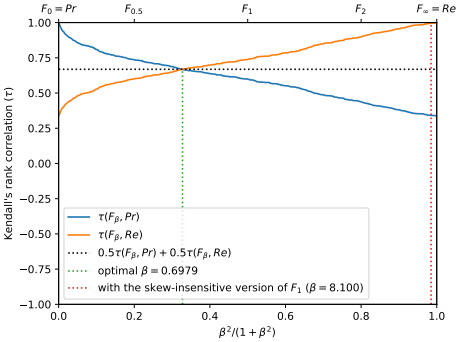


(f) The degree of optimality $\mathcal{O}(\beta)$ w.r.t. β. It is the probability to optimally ordering a pair of classifiers (BGS methods) given that it is not trivial (*i.e.*, that Pr and Re are in contradiction). The optimal value (or range of optimal values) for β is where the curve reaches 100%.

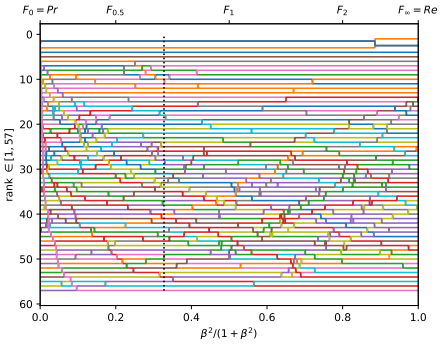
Figure A.3.40. Ranking of 57 BGS methods evaluated on the video "skating" ($\pi_+ = 0.0397$).



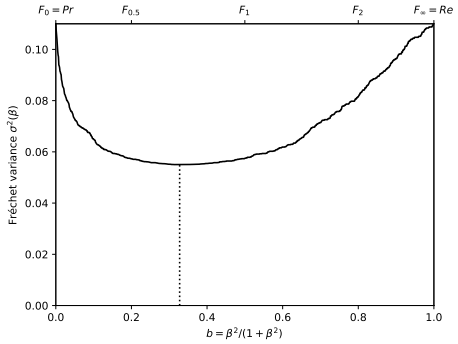
(a) The performances of 57 classifiers (BGS methods) depicted as points in the ROC space, with the isometrics of the optimal tradeoff score, from the ranking point of view, between precision and recall. See Eq. (15).



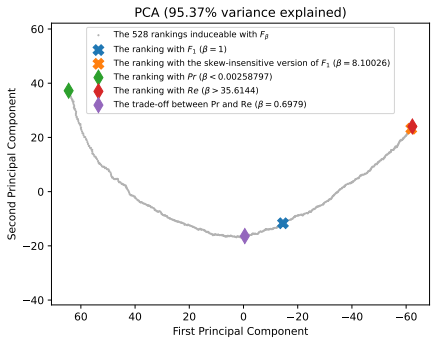
(b) The rank correlations $\tau(Pr; F_\beta)$ and $\tau(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the two curves intersect.



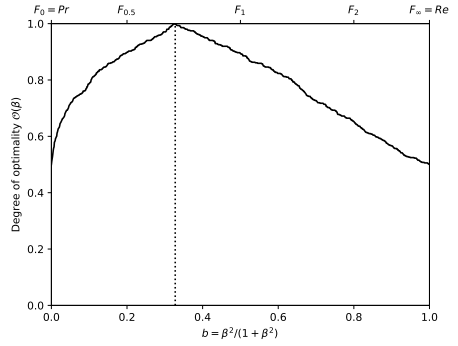
(c) The ranks of each classifier w.r.t. β. The optimal value (or range of optimal values) for β, shown here by the vertical line, is such that the number of swaps on its left is equal to the number of swaps on its right.



(d) The Fréchet variance $\sigma^2(\beta) = d_\tau^2(Pr; F_\beta) + d_\tau^2(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the curve has its minimum.

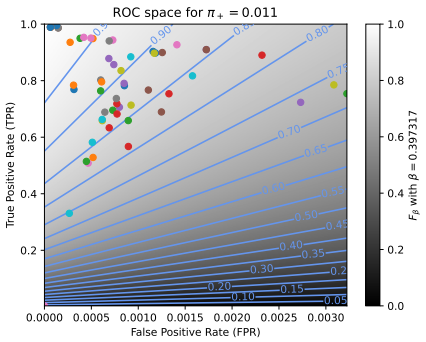


(e) Linear projection (PCA) of the manifold of the rankings induced by the F_β scores. The color points indicate the precision, the recall, F_1, $SIVF$, as well as the optimal tradeoff. The optimal tradeoff is at the same distance of the two extremities when the distance is measured along the manifold, with Kendall's distance d_τ.

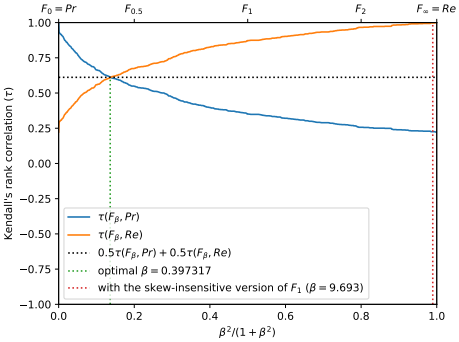


(f) The degree of optimality $\mathcal{O}(\beta)$ w.r.t. β. It is the probability to optimally ordering a pair of classifiers (BGS methods) given that it is not trivial (*i.e.*, that Pr and Re are in contradiction). The optimal value (or range of optimal values) for β is where the curve reaches 100%.

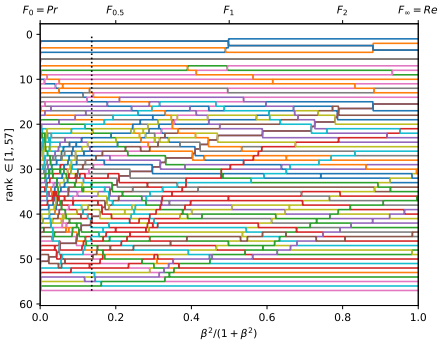
Figure A.3.41. Ranking of 57 BGS methods evaluated on the video "wetSnow" ($\pi_+ = 0.0150$).



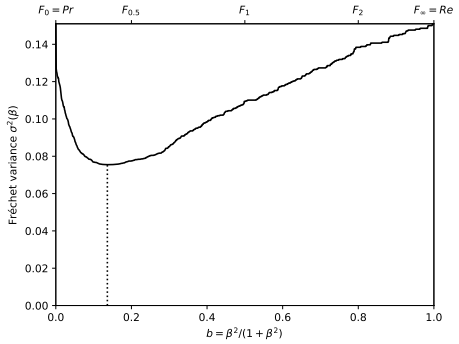
(a) The performances of 57 classifiers (BGS methods) depicted as points in the ROC space, with the isometrics of the optimal tradeoff score, from the ranking point of view, between precision and recall. See Eq. (15).



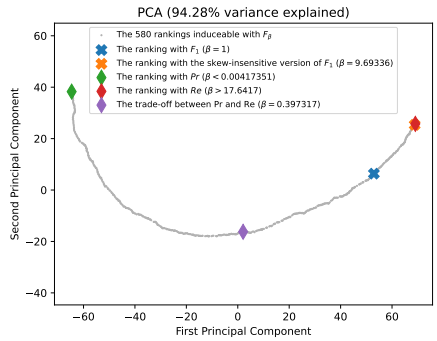
(b) The rank correlations $\tau(Pr; F_\beta)$ and $\tau(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the two curves intersect.



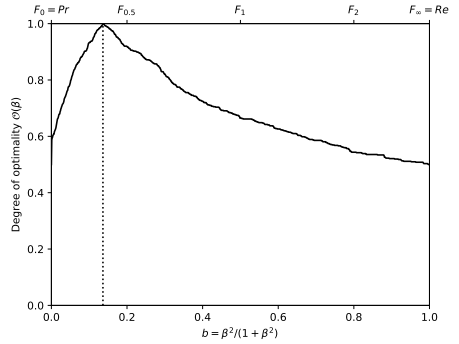
(c) The ranks of each classifier w.r.t. β. The optimal value (or range of optimal values) for β, shown here by the vertical line, is such that the number of swaps on its left is equal to the number of swaps on its right.



(d) The Fréchet variance $\sigma^2(\beta) = d_\tau^2(Pr; F_\beta) + d_\tau^2(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the curve has its minimum.

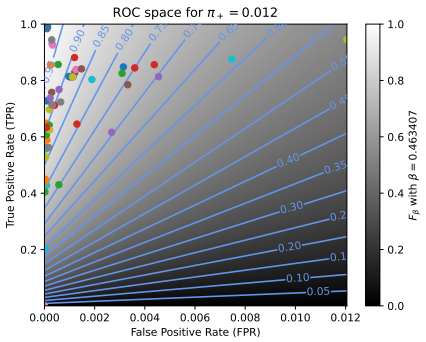


(e) Linear projection (PCA) of the manifold of the rankings induced by the F_β scores. The color points indicate the precision, the recall, F_1, $SIVF$, as well as the optimal tradeoff. The optimal tradeoff is at the same distance of the two extremities when the distance is measured along the manifold, with Kendall's distance d_τ.

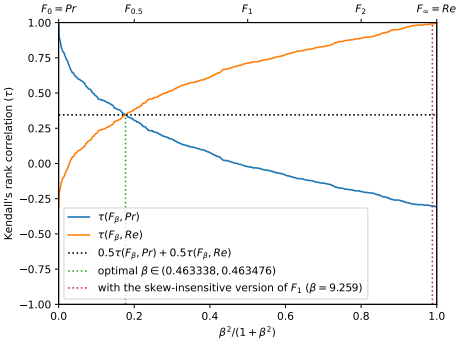


(f) The degree of optimality $\mathcal{O}(\beta)$ w.r.t. β. It is the probability to optimally ordering a pair of classifiers (BGS methods) given that it is not trivial (*i.e.*, that Pr and Re are in contradiction). The optimal value (or range of optimal values) for β is where the curve reaches 100%.

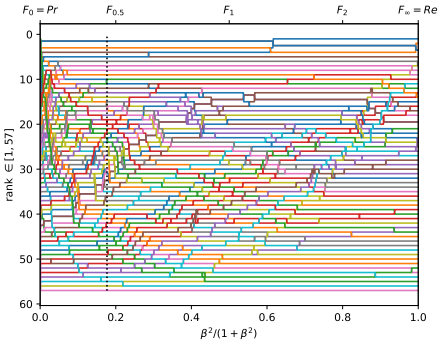
Figure A.3.42. Ranking of 57 BGS methods evaluated on the video "snowFall" ($\pi_+ = 0.0105$).



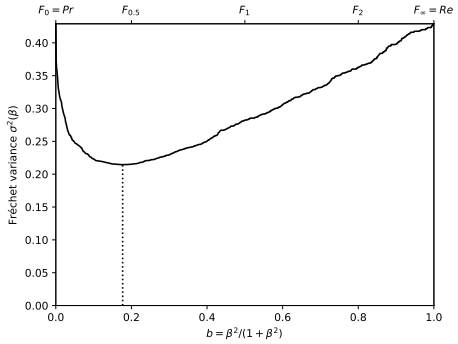
(a) The performances of 57 classifiers (BGS methods) depicted as points in the ROC space, with the isometrics of the optimal tradeoff score, from the ranking point of view, between precision and recall. See Eq. (15).



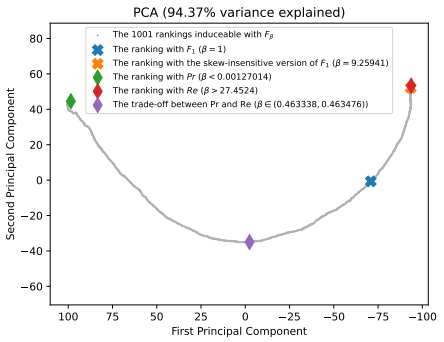
(b) The rank correlations $\tau(Pr; F_\beta)$ and $\tau(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the two curves intersect.



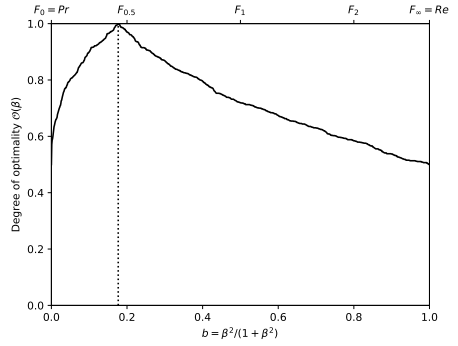
(c) The ranks of each classifier w.r.t. β. The optimal value (or range of optimal values) for β, shown here by the vertical line, is such that the number of swaps on its left is equal to the number of swaps on its right.



(d) The Fréchet variance $\sigma^2(\beta) = d_\tau^2(Pr; F_\beta) + d_\tau^2(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the curve has its minimum.

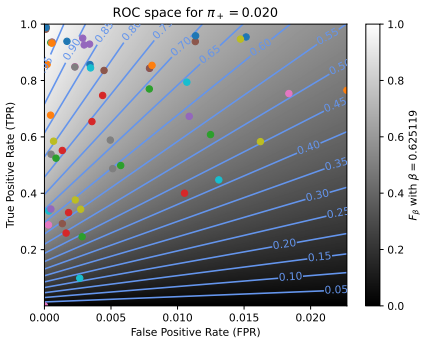


(e) Linear projection (PCA) of the manifold of the rankings induced by the F_β scores. The color points indicate the precision, the recall, F_1, $SIVF$, as well as the optimal tradeoff. The optimal tradeoff is at the same distance of the two extremities when the distance is measured along the manifold, with Kendall's distance d_τ.

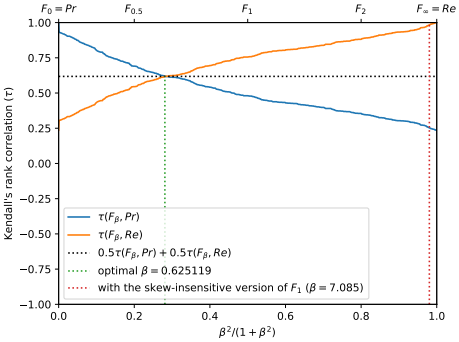


(f) The degree of optimality $\mathcal{O}(\beta)$ w.r.t. β. It is the probability to optimally ordering a pair of classifiers (BGS methods) given that it is not trivial (*i.e.*, that Pr and Re are in contradiction). The optimal value (or range of optimal values) for β is where the curve reaches 100%.

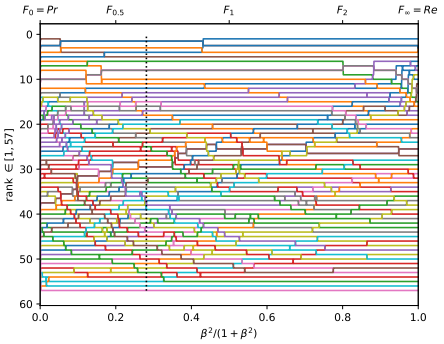
Figure A.3.43. Ranking of 57 BGS methods evaluated on the video "blizzard" ($\pi_+ = 0.0115$).



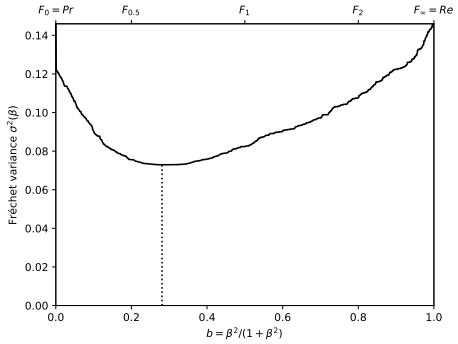
(a) The performances of 57 classifiers (BGS methods) depicted as points in the ROC space, with the isometrics of the optimal tradeoff score, from the ranking point of view, between precision and recall. See Eq. (15).



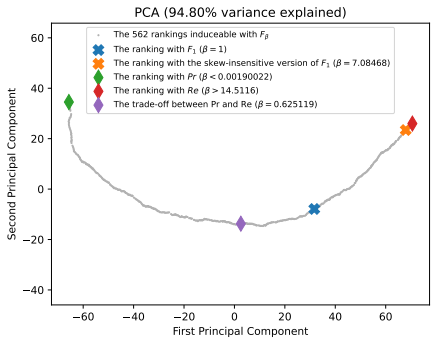
(b) The rank correlations $\tau(Pr; F_\beta)$ and $\tau(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the two curves intersect.



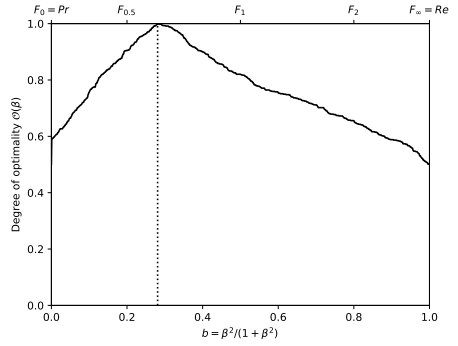
(c) The ranks of each classifier w.r.t. β. The optimal value (or range of optimal values) for β, shown here by the vertical line, is such that the number of swaps on its left is equal to the number of swaps on its right.



(d) The Fréchet variance $\sigma^2(\beta) = d_\tau^2(Pr; F_\beta) + d_\tau^2(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the curve has its minimum.

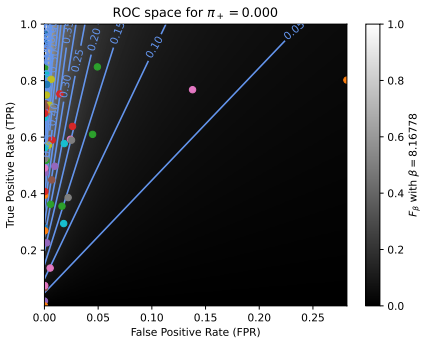


(e) Linear projection (PCA) of the manifold of the rankings induced by the F_β scores. The color points indicate the precision, the recall, F_1, $SIVF$, as well as the optimal tradeoff. The optimal tradeoff is at the same distance of the two extremities when the distance is measured along the manifold, with Kendall's distance d_τ.

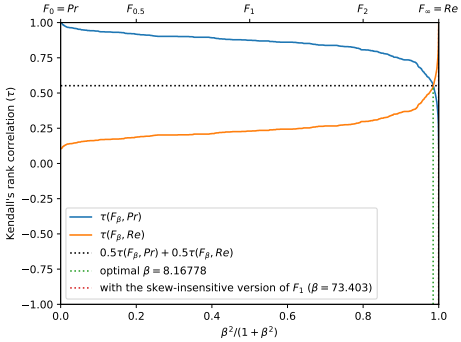


(f) The degree of optimality $\mathcal{O}(\beta)$ w.r.t. β. It is the probability to optimally ordering a pair of classifiers (BGS methods) given that it is not trivial (*i.e.*, that Pr and Re are in contradiction). The optimal value (or range of optimal values) for β is where the curve reaches 100%.

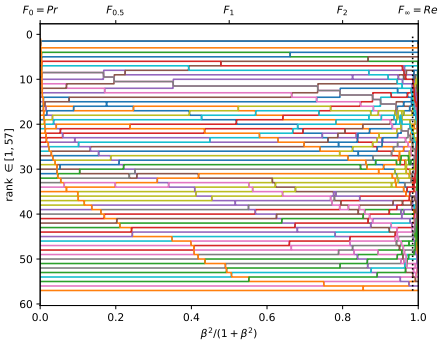
Figure A.3.44. Ranking of 57 BGS methods evaluated on the video "tunnelExit_0_35fps" ($\pi_+ = 0.0195$).



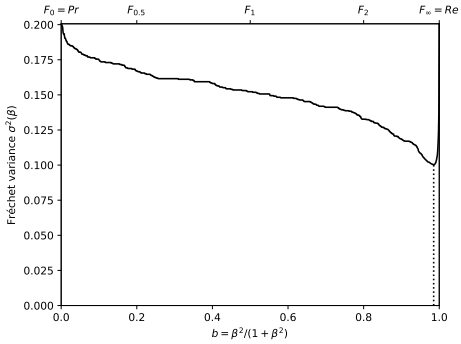
(a) The performances of 57 classifiers (BGS methods) depicted as points in the ROC space, with the isometrics of the optimal tradeoff score, from the ranking point of view, between precision and recall. See Eq. (15).



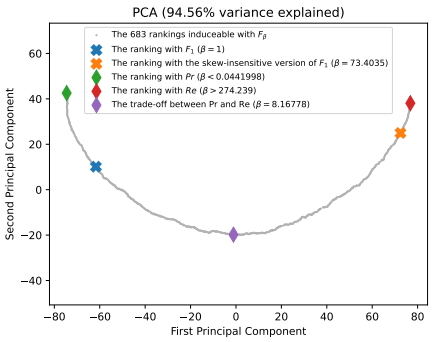
(b) The rank correlations $\tau(Pr; F_\beta)$ and $\tau(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the two curves intersect.



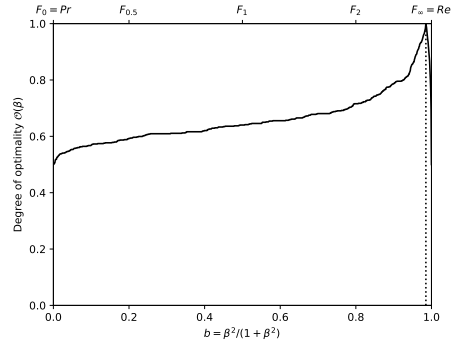
(c) The ranks of each classifier w.r.t. β. The optimal value (or range of optimal values) for β, shown here by the vertical line, is such that the number of swaps on its left is equal to the number of swaps on its right.



(d) The Fréchet variance $\sigma^2(\beta) = d_\tau^2(Pr; F_\beta) + d_\tau^2(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the curve has its minimum.

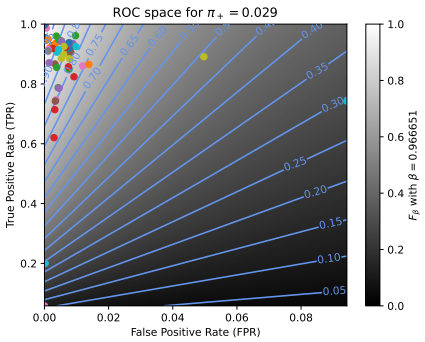


(e) Linear projection (PCA) of the manifold of the rankings induced by the F_β scores. The color points indicate the precision, the recall, F_1, $SIVF$, as well as the optimal tradeoff. The optimal tradeoff is at the same distance of the two extremities when the distance is measured along the manifold, with Kendall's distance d_τ.

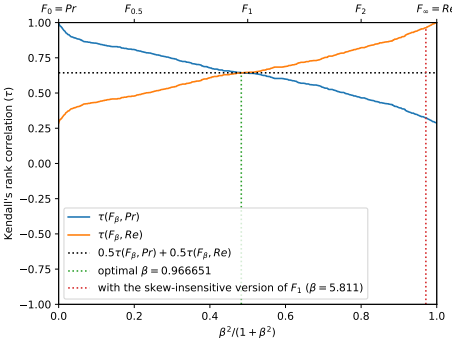


(f) The degree of optimality $\mathcal{O}(\beta)$ w.r.t. β. It is the probability to optimally ordering a pair of classifiers (BGS methods) given that it is not trivial (*i.e.*, that Pr and Re are in contradiction). The optimal value (or range of optimal values) for β is where the curve reaches 100%.

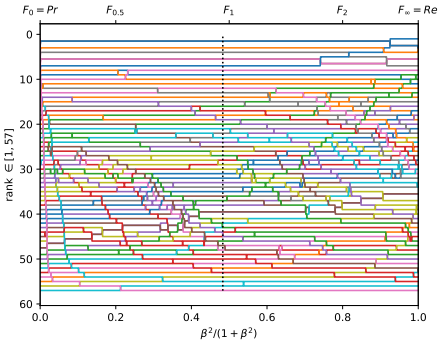
Figure A.3.45. Ranking of 57 BGS methods evaluated on the video "port_0_17fps" ($\pi_+ = 0.0002$).



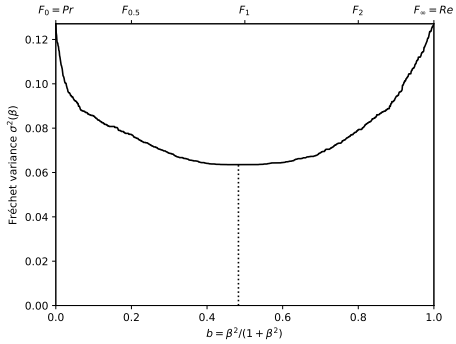
(a) The performances of 57 classifiers (BGS methods) depicted as points in the ROC space, with the isometrics of the optimal tradeoff score, from the ranking point of view, between precision and recall. See Eq. (15).



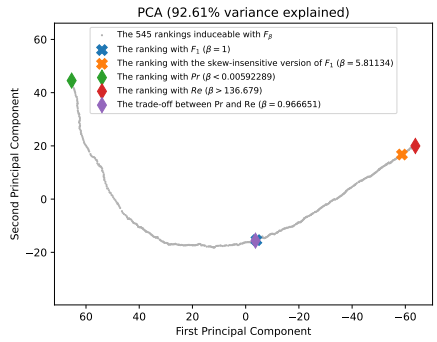
(b) The rank correlations $\tau(Pr; F_\beta)$ and $\tau(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the two curves intersect.



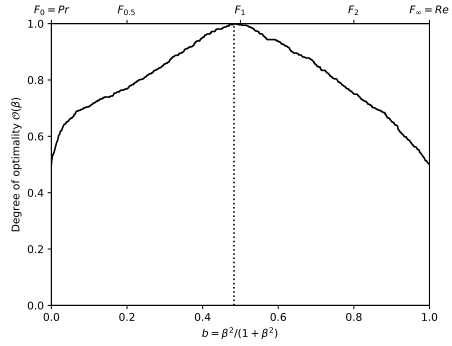
(c) The ranks of each classifier w.r.t. β. The optimal value (or range of optimal values) for β, shown here by the vertical line, is such that the number of swaps on its left is equal to the number of swaps on its right.



(d) The Fréchet variance $\sigma^2(\beta) = d_\tau^2(Pr; F_\beta) + d_\tau^2(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the curve has its minimum.

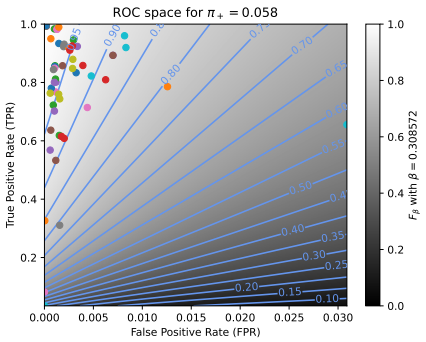


(e) Linear projection (PCA) of the manifold of the rankings induced by the F_β scores. The color points indicate the precision, the recall, F_1, $SIVF$, as well as the optimal tradeoff. The optimal tradeoff is at the same distance of the two extremities when the distance is measured along the manifold, with Kendall's distance d_τ.

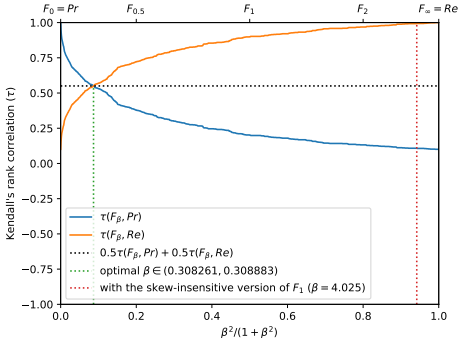


(f) The degree of optimality $\mathcal{O}(\beta)$ w.r.t. β. It is the probability to optimally ordering a pair of classifiers (BGS methods) given that it is not trivial (*i.e.*, that Pr and Re are in contradiction). The optimal value (or range of optimal values) for β is where the curve reaches 100%.

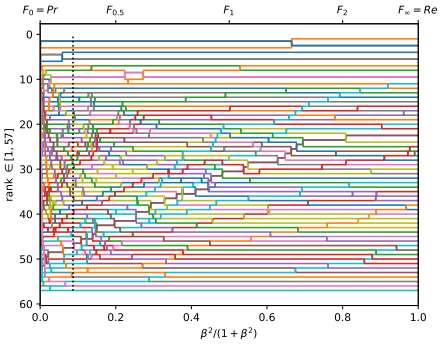
Figure A.3.46. Ranking of 57 BGS methods evaluated on the video "tramCrossroad_1fps" ($\pi_+ = 0.0288$).



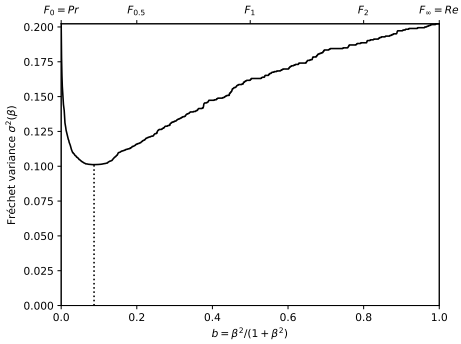
(a) The performances of 57 classifiers (BGS methods) depicted as points in the ROC space, with the isometrics of the optimal tradeoff score, from the ranking point of view, between precision and recall. See Eq. (15).



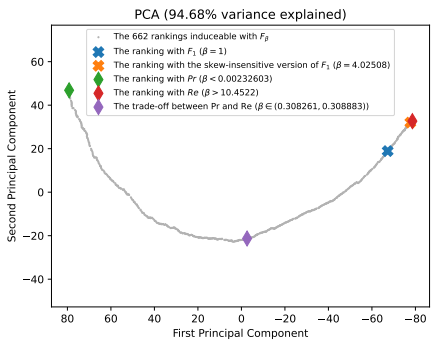
(b) The rank correlations $\tau(Pr; F_\beta)$ and $\tau(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the two curves intersect.



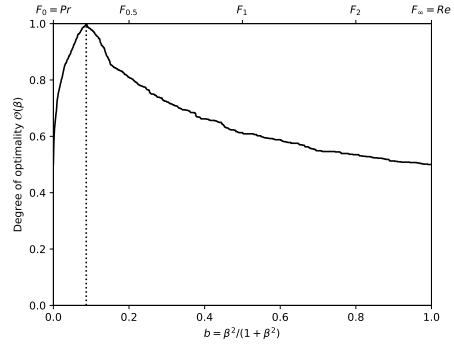
(c) The ranks of each classifier w.r.t. β. The optimal value (or range of optimal values) for β, shown here by the vertical line, is such that the number of swaps on its left is equal to the number of swaps on its right.



(d) The Fréchet variance $\sigma^2(\beta) = d_\tau^2(Pr; F_\beta) + d_\tau^2(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the curve has its minimum.

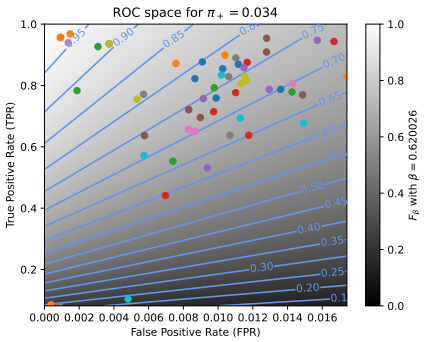


(e) Linear projection (PCA) of the manifold of the rankings induced by the F_β scores. The color points indicate the precision, the recall, F_1, $SIVF$, as well as the optimal tradeoff. The optimal tradeoff is at the same distance of the two extremities when the distance is measured along the manifold, with Kendall's distance d_τ.

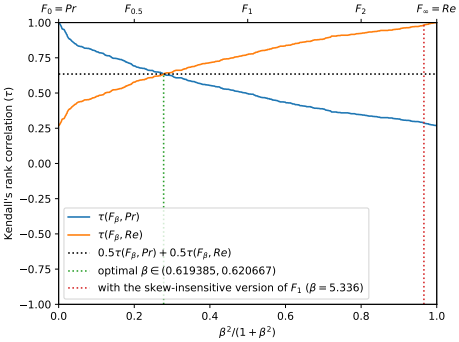


(f) The degree of optimality $\mathcal{O}(\beta)$ w.r.t. β. It is the probability to optimally ordering a pair of classifiers (BGS methods) given that it is not trivial (*i.e.*, that Pr and Re are in contradiction). The optimal value (or range of optimal values) for β is where the curve reaches 100%.

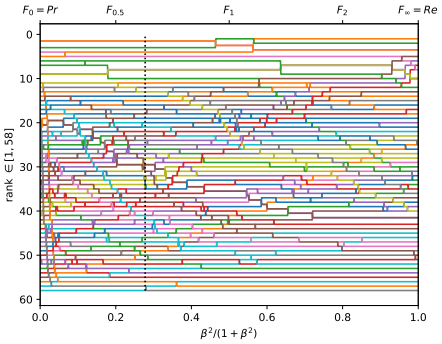
Figure A.3.47. Ranking of 57 BGS methods evaluated on the video "turnpike_0_5fps" ($\pi_+ = 0.0581$).



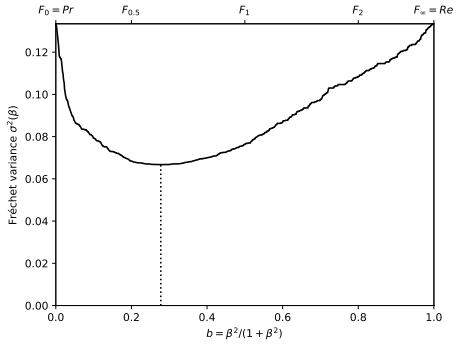
(a) The performances of 58 classifiers (BGS methods) depicted as points in the ROC space, with the isometrics of the optimal tradeoff score, from the ranking point of view, between precision and recall. See Eq. (15).



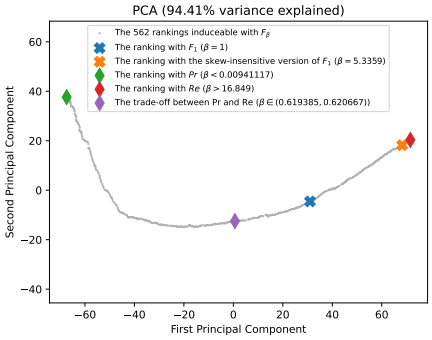
(b) The rank correlations $\tau(Pr; F_\beta)$ and $\tau(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the two curves intersect.



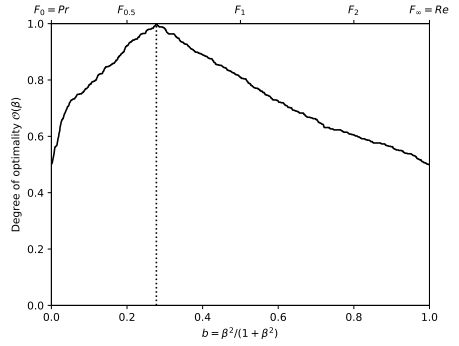
(c) The ranks of each classifier w.r.t. β. The optimal value (or range of optimal values) for β, shown here by the vertical line, is such that the number of swaps on its left is equal to the number of swaps on its right.



(d) The Fréchet variance $\sigma^2(\beta) = d_\tau^2(Pr; F_\beta) + d_\tau^2(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the curve has its minimum.

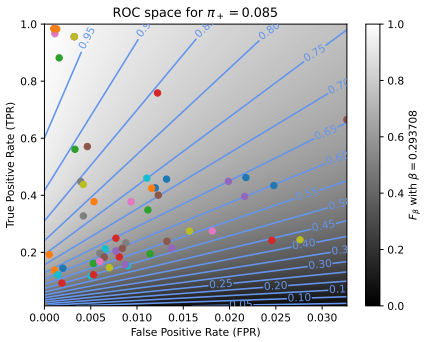


(e) Linear projection (PCA) of the manifold of the rankings induced by the F_β scores. The color points indicate the precision, the recall, F_1, $SIVF$, as well as the optimal tradeoff. The optimal tradeoff is at the same distance of the two extremities when the distance is measured along the manifold, with Kendall's distance d_τ.

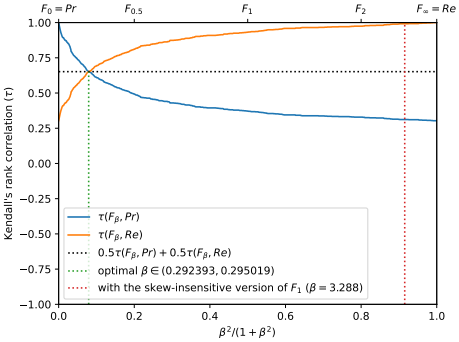


(f) The degree of optimality $\mathcal{O}(\beta)$ w.r.t. β. It is the probability to optimally ordering a pair of classifiers (BGS methods) given that it is not trivial (*i.e.*, that Pr and Re are in contradiction). The optimal value (or range of optimal values) for β is where the curve reaches 100%.

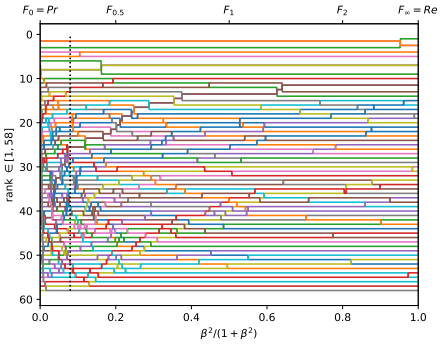
Figure A.3.48. Ranking of 58 BGS methods evaluated on the video "tramStation" ($\pi_+ = 0.0339$).



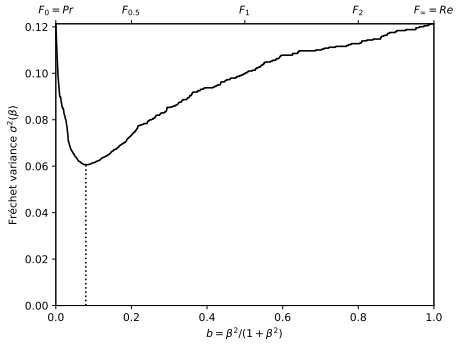
(a) The performances of 58 classifiers (BGS methods) depicted as points in the ROC space, with the isometrics of the optimal tradeoff score, from the ranking point of view, between precision and recall. See Eq. (15).



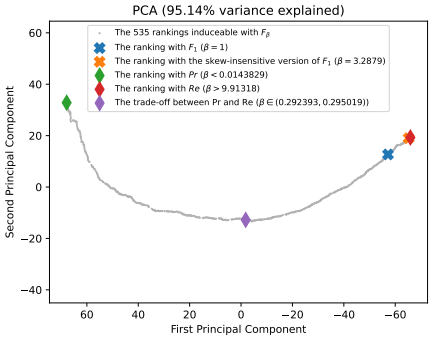
(b) The rank correlations $\tau(Pr; F_\beta)$ and $\tau(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the two curves intersect.



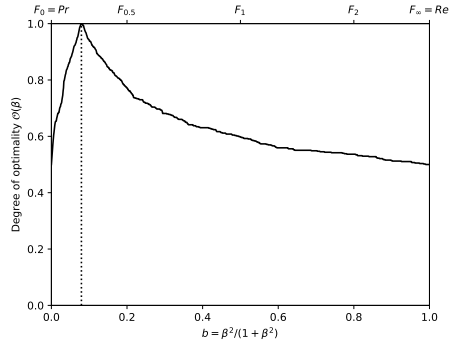
(c) The ranks of each classifier w.r.t. β. The optimal value (or range of optimal values) for β, shown here by the vertical line, is such that the number of swaps on its left is equal to the number of swaps on its right.



(d) The Fréchet variance $\sigma^2(\beta) = d_\tau^2(Pr; F_\beta) + d_\tau^2(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the curve has its minimum.

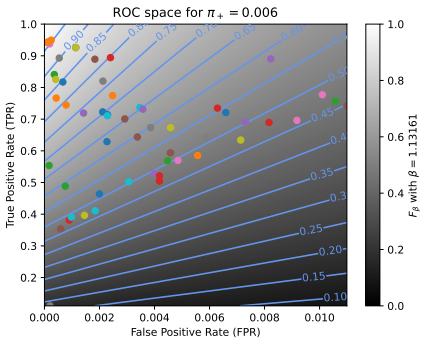


(e) Linear projection (PCA) of the manifold of the rankings induced by the F_β scores. The color points indicate the precision, the recall, F_1, $SIVF$, as well as the optimal tradeoff. The optimal tradeoff is at the same distance of the two extremities when the distance is measured along the manifold, with Kendall's distance d_τ.

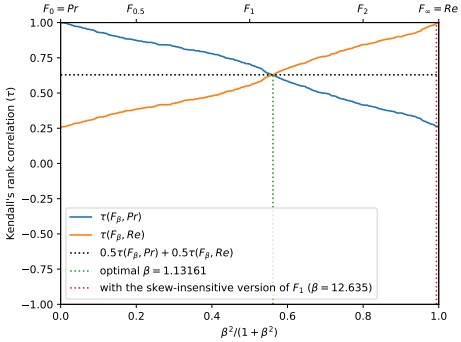


(f) The degree of optimality $\mathcal{O}(\beta)$ w.r.t. β. It is the probability to optimally ordering a pair of classifiers (BGS methods) given that it is not trivial (*i.e.*, that Pr and Re are in contradiction). The optimal value (or range of optimal values) for β is where the curve reaches 100%.

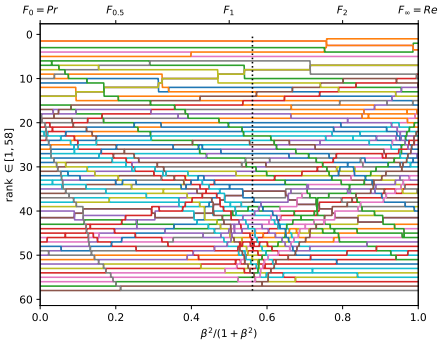
Figure A.3.49. Ranking of 58 BGS methods evaluated on the video "busyBoulvard" ($\pi_+ = 0.0847$).



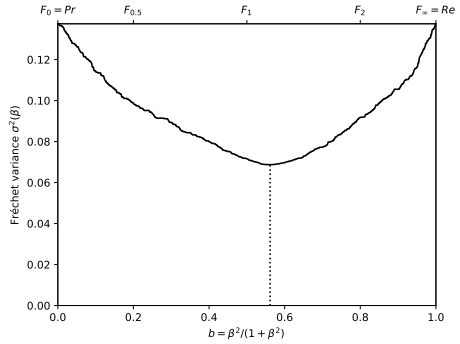
(a) The performances of 58 classifiers (BGS methods) depicted as points in the ROC space, with the isometrics of the optimal tradeoff score, from the ranking point of view, between precision and recall. See Eq. (15).



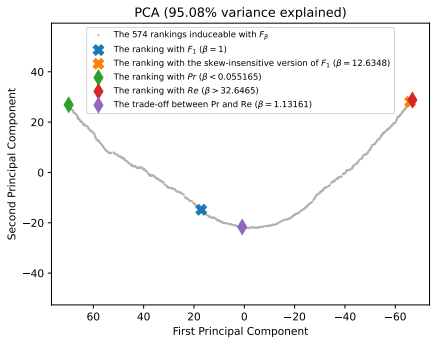
(b) The rank correlations $\tau(Pr; F_\beta)$ and $\tau(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the two curves intersect.



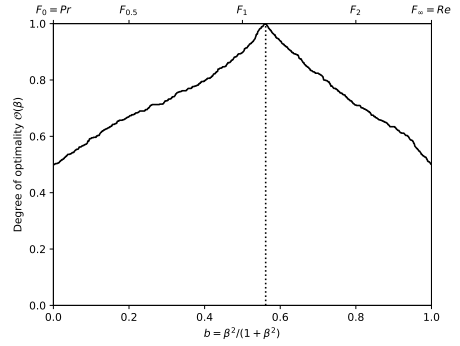
(c) The ranks of each classifier w.r.t. β. The optimal value (or range of optimal values) for β, shown here by the vertical line, is such that the number of swaps on its left is equal to the number of swaps on its right.



(d) The Fréchet variance $\sigma^2(\beta) = d_\tau^2(Pr; F_\beta) + d_\tau^2(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the curve has its minimum.

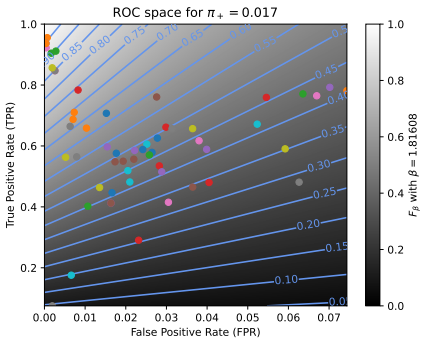


(e) Linear projection (PCA) of the manifold of the rankings induced by the F_β scores. The color points indicate the precision, the recall, F_1, $SIVF$, as well as the optimal tradeoff. The optimal tradeoff is at the same distance of the two extremities when the distance is measured along the manifold, with Kendall's distance d_τ.

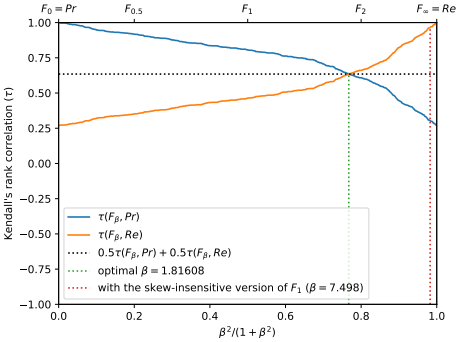


(f) The degree of optimality $\mathcal{O}(\beta)$ w.r.t. β. It is the probability to optimally ordering a pair of classifiers (BGS methods) given that it is not trivial (*i.e.*, that Pr and Re are in contradiction). The optimal value (or range of optimal values) for β is where the curve reaches 100%.

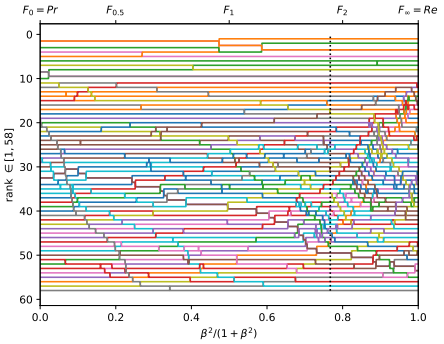
Figure A.3.50. Ranking of 58 BGS methods evaluated on the video "streetCornerAtNight" ($\pi_+ = 0.0062$).



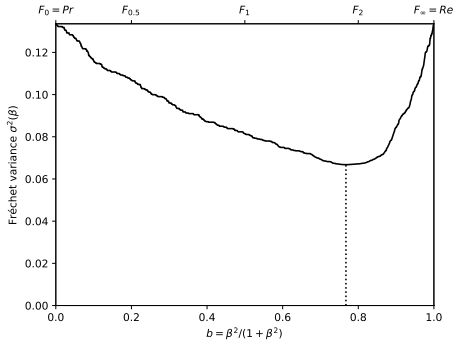
(a) The performances of 58 classifiers (BGS methods) depicted as points in the ROC space, with the isometrics of the optimal tradeoff score, from the ranking point of view, between precision and recall. See Eq. (15).



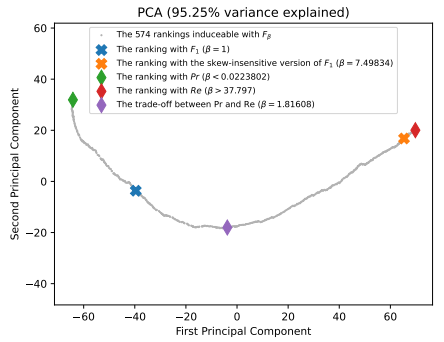
(b) The rank correlations $\tau(Pr; F_\beta)$ and $\tau(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the two curves intersect.



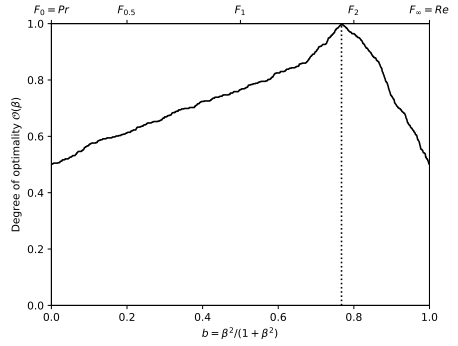
(c) The ranks of each classifier w.r.t. β. The optimal value (or range of optimal values) for β, shown here by the vertical line, is such that the number of swaps on its left is equal to the number of swaps on its right.



(d) The Fréchet variance $\sigma^2(\beta) = d_\tau^2(Pr; F_\beta) + d_\tau^2(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the curve has its minimum.

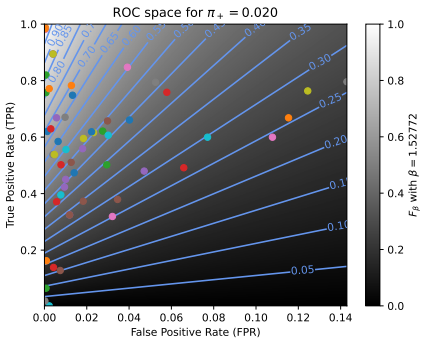


(e) Linear projection (PCA) of the manifold of the rankings induced by the F_β scores. The color points indicate the precision, the recall, F_1, $SIVF$, as well as the optimal tradeoff. The optimal tradeoff is at the same distance of the two extremities when the distance is measured along the manifold, with Kendall's distance d_τ.

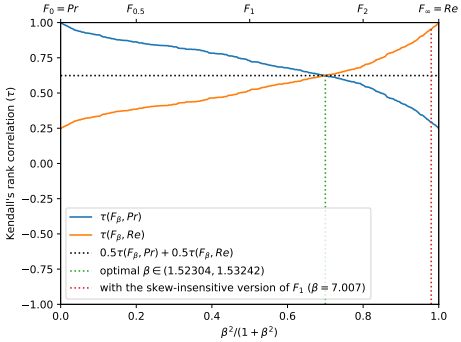


(f) The degree of optimality $\mathcal{O}(\beta)$ w.r.t. β. It is the probability to optimally ordering a pair of classifiers (BGS methods) given that it is not trivial (*i.e.*, that Pr and Re are in contradiction). The optimal value (or range of optimal values) for β is where the curve reaches 100%.

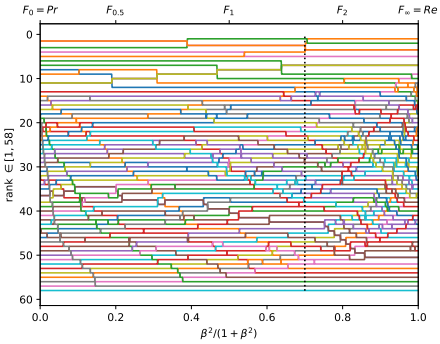
Figure A.3.51. Ranking of 58 BGS methods evaluated on the video "fluidHighway" ($\pi_+ = 0.0175$).



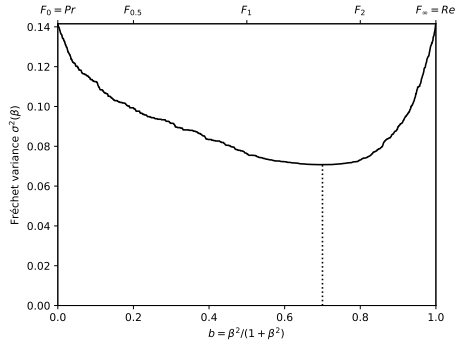
(a) The performances of 58 classifiers (BGS methods) depicted as points in the ROC space, with the isometrics of the optimal tradeoff score, from the ranking point of view, between precision and recall. See Eq. (15).



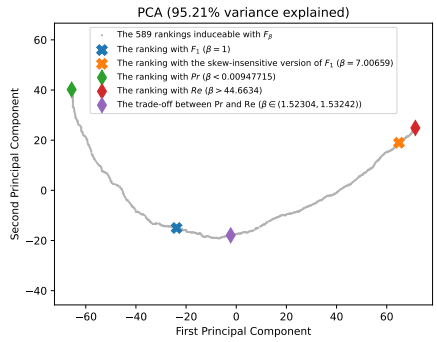
(b) The rank correlations $\tau(Pr; F_\beta)$ and $\tau(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the two curves intersect.



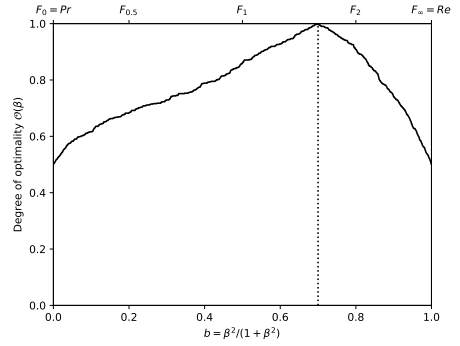
(c) The ranks of each classifier w.r.t. β. The optimal value (or range of optimal values) for β, shown here by the vertical line, is such that the number of swaps on its left is equal to the number of swaps on its right.



(d) The Fréchet variance $\sigma^2(\beta) = d_\tau^2(Pr; F_\beta) + d_\tau^2(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the curve has its minimum.

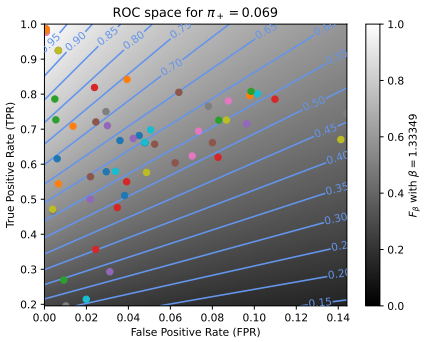


(e) Linear projection (PCA) of the manifold of the rankings induced by the F_β scores. The color points indicate the precision, the recall, F_1, $SIVF$, as well as the optimal tradeoff. The optimal tradeoff is at the same distance of the two extremities when the distance is measured along the manifold, with Kendall's distance d_τ.

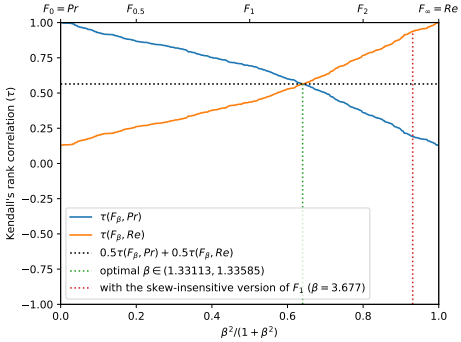


(f) The degree of optimality $\mathcal{O}(\beta)$ w.r.t. β. It is the probability to optimally ordering a pair of classifiers (BGS methods) given that it is not trivial (*i.e.*, that Pr and Re are in contradiction). The optimal value (or range of optimal values) for β is where the curve reaches 100%.

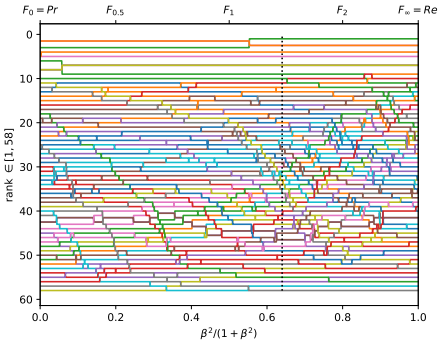
Figure A.3.52. Ranking of 58 BGS methods evaluated on the video "bridgeEntry" ($\pi_+ = 0.0200$).



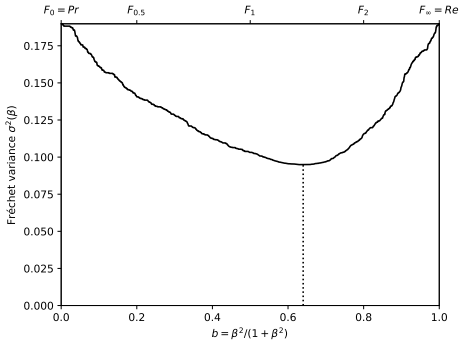
(a) The performances of 58 classifiers (BGS methods) depicted as points in the ROC space, with the isometrics of the optimal tradeoff score, from the ranking point of view, between precision and recall. See Eq. (15).



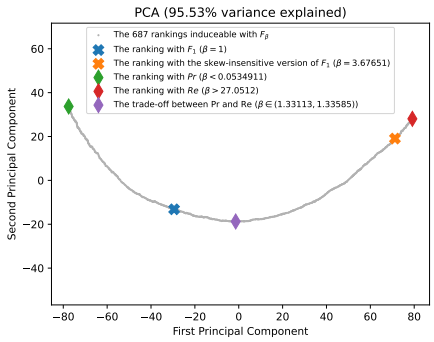
(b) The rank correlations $\tau(Pr; F_\beta)$ and $\tau(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the two curves intersect.



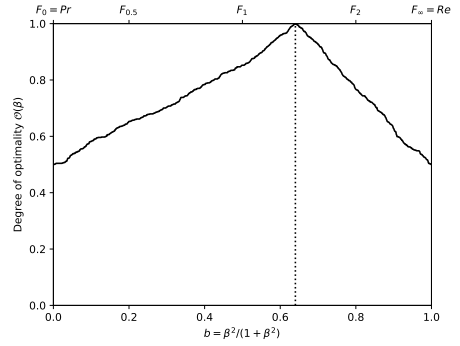
(c) The ranks of each classifier w.r.t. β. The optimal value (or range of optimal values) for β, shown here by the vertical line, is such that the number of swaps on its left is equal to the number of swaps on its right.



(d) The Fréchet variance $\sigma^2(\beta) = d_\tau^2(Pr; F_\beta) + d_\tau^2(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the curve has its minimum.

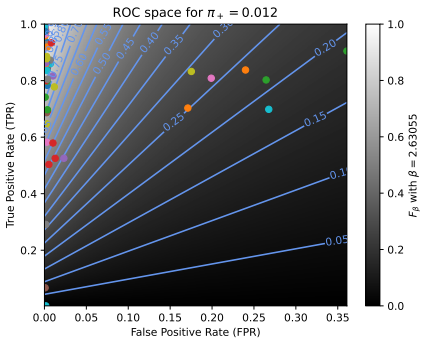


(e) Linear projection (PCA) of the manifold of the rankings induced by the F_β scores. The color points indicate the precision, the recall, F_1, $SIVF$, as well as the optimal tradeoff. The optimal tradeoff is at the same distance of the two extremities when the distance is measured along the manifold, with Kendall's distance d_τ.

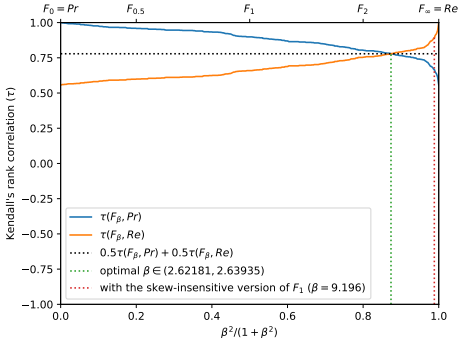


(f) The degree of optimality $\mathcal{O}(\beta)$ w.r.t. β. It is the probability to optimally ordering a pair of classifiers (BGS methods) given that it is not trivial (*i.e.*, that Pr and Re are in contradiction). The optimal value (or range of optimal values) for β is where the curve reaches 100%.

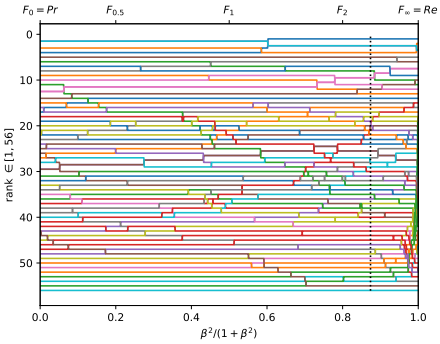
Figure A.3.53. Ranking of 58 BGS methods evaluated on the video "winterStreet" ($\pi_+ = 0.0689$).



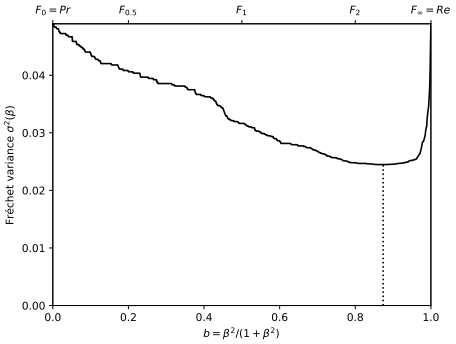
(a) The performances of 56 classifiers (BGS methods) depicted as points in the ROC space, with the isometrics of the optimal tradeoff score, from the ranking point of view, between precision and recall. See Eq. (15).



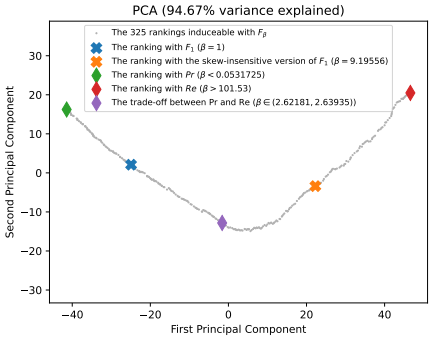
(b) The rank correlations $\tau(Pr; F_\beta)$ and $\tau(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the two curves intersect.



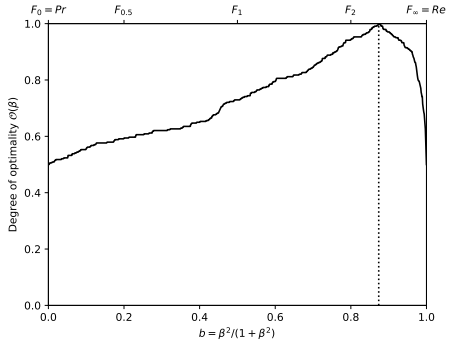
(c) The ranks of each classifier w.r.t. β. The optimal value (or range of optimal values) for β, shown here by the vertical line, is such that the number of swaps on its left is equal to the number of swaps on its right.



(d) The Fréchet variance $\sigma^2(\beta) = d_\tau^2(Pr; F_\beta) + d_\tau^2(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the curve has its minimum.

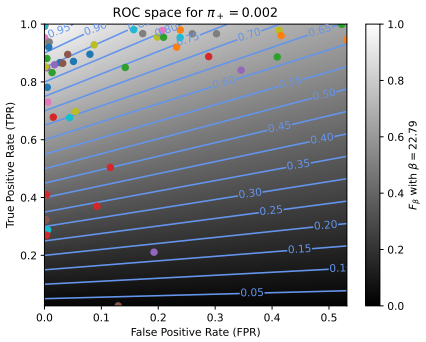


(e) Linear projection (PCA) of the manifold of the rankings induced by the F_β scores. The color points indicate the precision, the recall, F_1, $SIVF$, as well as the optimal tradeoff. The optimal tradeoff is at the same distance of the two extremities when the distance is measured along the manifold, with Kendall's distance d_τ.

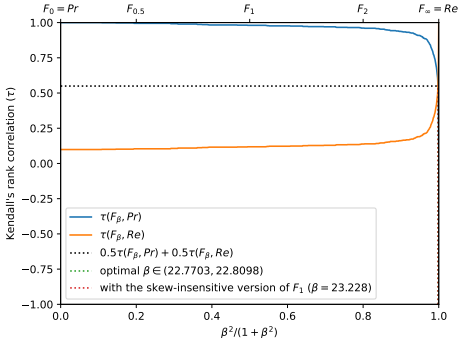


(f) The degree of optimality $\mathcal{O}(\beta)$ w.r.t. β. It is the probability to optimally ordering a pair of classifiers (BGS methods) given that it is not trivial (*i.e.*, that Pr and Re are in contradiction). The optimal value (or range of optimal values) for β is where the curve reaches 100%.

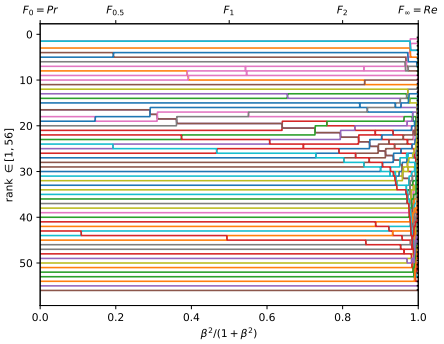
Figure A.3.54. Ranking of 56 BGS methods evaluated on the video "twoPositionPTZCam" ($\pi_+ = 0.0117$).



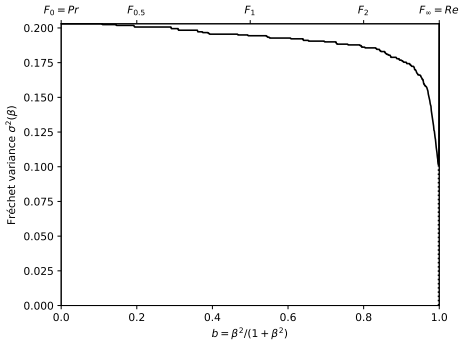
(a) The performances of 56 classifiers (BGS methods) depicted as points in the ROC space, with the isometrics of the optimal tradeoff score, from the ranking point of view, between precision and recall. See Eq. (15).



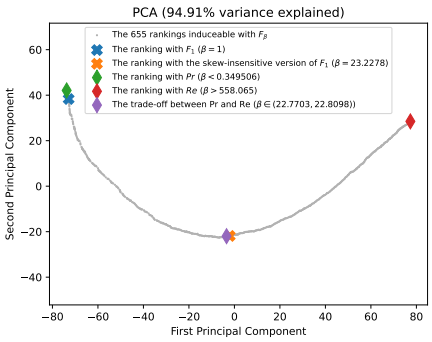
(b) The rank correlations $\tau(Pr; F_\beta)$ and $\tau(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the two curves intersect.



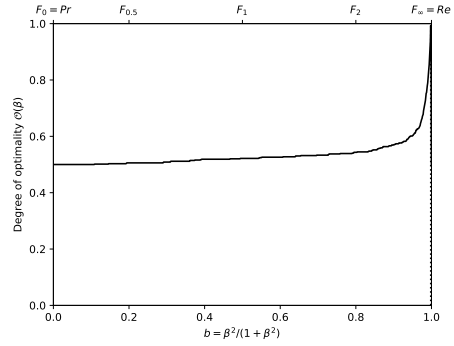
(c) The ranks of each classifier w.r.t. β. The optimal value (or range of optimal values) for β, shown here by the vertical line, is such that the number of swaps on its left is equal to the number of swaps on its right.



(d) The Fréchet variance $\sigma^2(\beta) = d_\tau^2(Pr; F_\beta) + d_\tau^2(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the curve has its minimum.

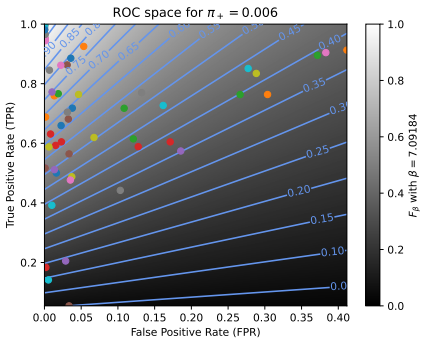


(e) Linear projection (PCA) of the manifold of the rankings induced by the F_β scores. The color points indicate the precision, the recall, F_1, $SIVF$, as well as the optimal tradeoff. The optimal tradeoff is at the same distance of the two extremities when the distance is measured along the manifold, with Kendall's distance d_τ.

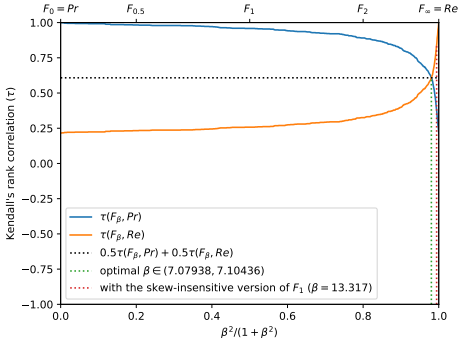


(f) The degree of optimality $\mathcal{O}(\beta)$ w.r.t. β. It is the probability to optimally ordering a pair of classifiers (BGS methods) given that it is not trivial (*i.e.*, that Pr and Re are in contradiction). The optimal value (or range of optimal values) for β is where the curve reaches 100%.

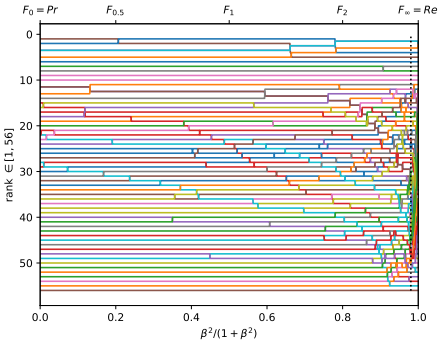
Figure A.3.55. Ranking of 56 BGS methods evaluated on the video "zoomInZoomOut" ($\pi_+ = 0.0019$).



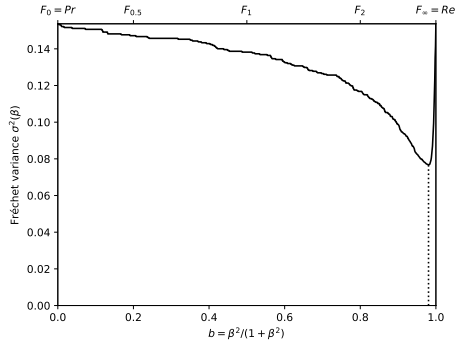
(a) The performances of 56 classifiers (BGS methods) depicted as points in the ROC space, with the isometrics of the optimal tradeoff score, from the ranking point of view, between precision and recall. See Eq. (15).



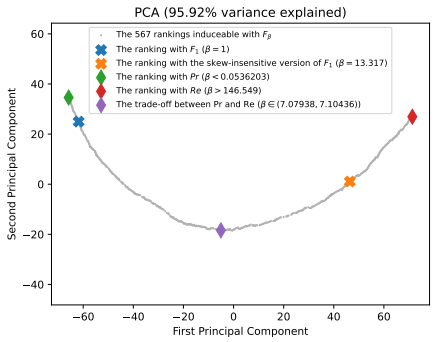
(b) The rank correlations $\tau(Pr; F_\beta)$ and $\tau(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the two curves intersect.



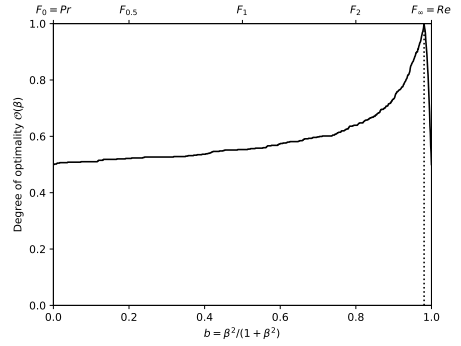
(c) The ranks of each classifier w.r.t. β. The optimal value (or range of optimal values) for β, shown here by the vertical line, is such that the number of swaps on its left is equal to the number of swaps on its right.



(d) The Fréchet variance $\sigma^2(\beta) = d_\tau^2(Pr; F_\beta) + d_\tau^2(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the curve has its minimum.

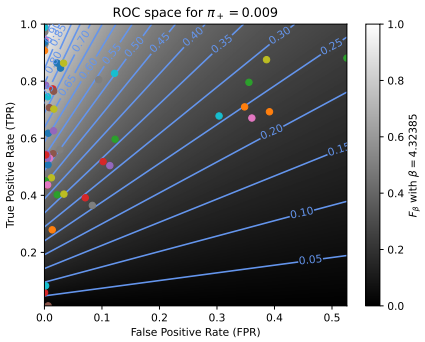


(e) Linear projection (PCA) of the manifold of the rankings induced by the F_β scores. The color points indicate the precision, the recall, F_1, $SIVF$, as well as the optimal tradeoff. The optimal tradeoff is at the same distance of the two extremities when the distance is measured along the manifold, with Kendall's distance d_τ.

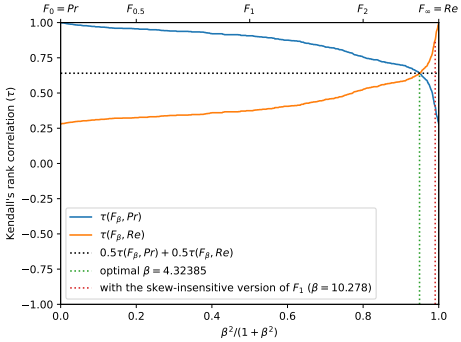


(f) The degree of optimality $\mathcal{O}(\beta)$ w.r.t. β. It is the probability to optimally ordering a pair of classifiers (BGS methods) given that it is not trivial (*i.e.*, that Pr and Re are in contradiction). The optimal value (or range of optimal values) for β is where the curve reaches 100%.

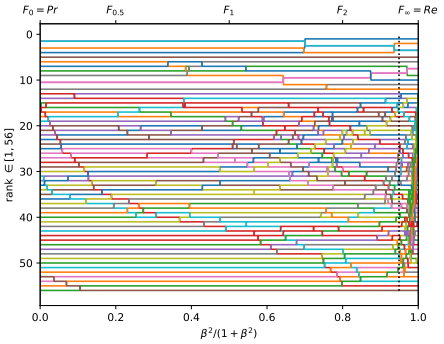
Figure A.3.56. Ranking of 56 BGS methods evaluated on the video "continuousPan" ($\pi_+ = 0.0056$).



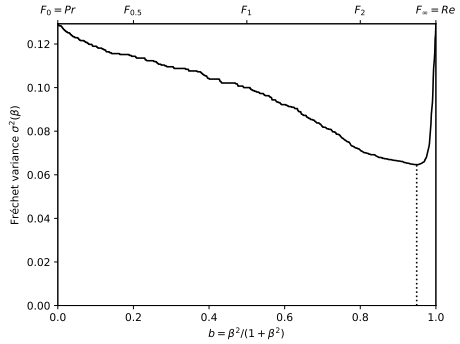
(a) The performances of 56 classifiers (BGS methods) depicted as points in the ROC space, with the isometrics of the optimal tradeoff score, from the ranking point of view, between precision and recall. See Eq. (15).



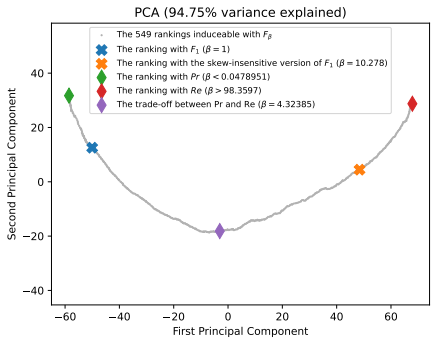
(b) The rank correlations $\tau(Pr; F_\beta)$ and $\tau(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the two curves intersect.



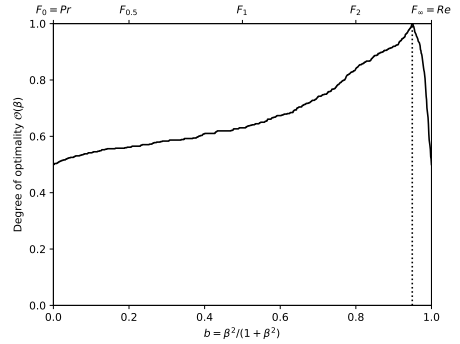
(c) The ranks of each classifier w.r.t. β. The optimal value (or range of optimal values) for β, shown here by the vertical line, is such that the number of swaps on its left is equal to the number of swaps on its right.



(d) The Fréchet variance $\sigma^2(\beta) = d_\tau^2(Pr; F_\beta) + d_\tau^2(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the curve has its minimum.

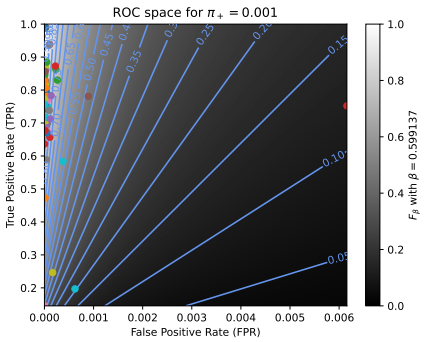


(e) Linear projection (PCA) of the manifold of the rankings induced by the F_β scores. The color points indicate the precision, the recall, F_1, $SIVF$, as well as the optimal tradeoff. The optimal tradeoff is at the same distance of the two extremities when the distance is measured along the manifold, with Kendall's distance d_τ.

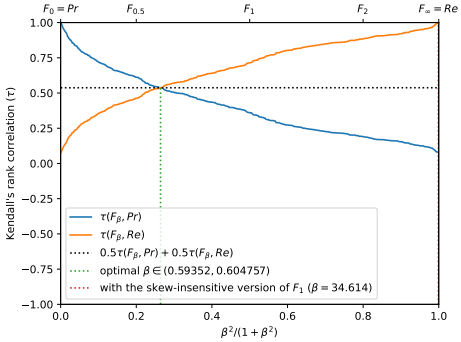


(f) The degree of optimality $\mathcal{O}(\beta)$ w.r.t. β. It is the probability to optimally ordering a pair of classifiers (BGS methods) given that it is not trivial (*i.e.*, that Pr and Re are in contradiction). The optimal value (or range of optimal values) for β is where the curve reaches 100%.

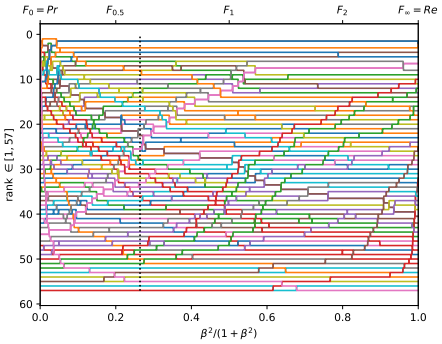
Figure A.3.57. Ranking of 56 BGS methods evaluated on the video "intermittentPan" ($\pi_+ = 0.0094$).



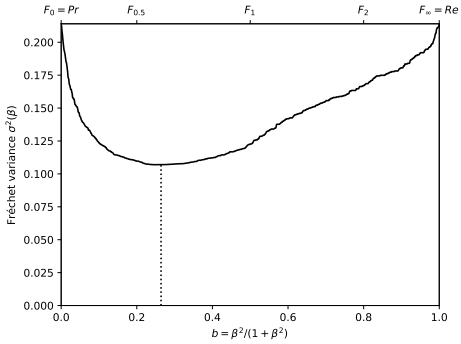
(a) The performances of 57 classifiers (BGS methods) depicted as points in the ROC space, with the isometrics of the optimal tradeoff score, from the ranking point of view, between precision and recall. See Eq. (15).



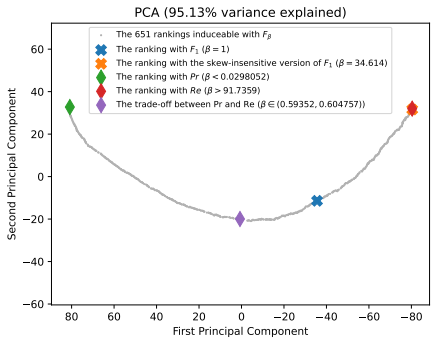
(b) The rank correlations $\tau(Pr; F_\beta)$ and $\tau(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the two curves intersect.



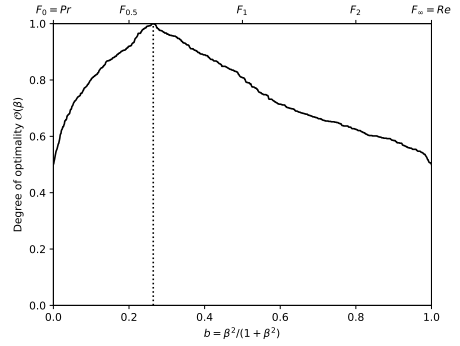
(c) The ranks of each classifier w.r.t. β. The optimal value (or range of optimal values) for β, shown here by the vertical line, is such that the number of swaps on its left is equal to the number of swaps on its right.



(d) The Fréchet variance $\sigma^2(\beta) = d_\tau^2(Pr; F_\beta) + d_\tau^2(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the curve has its minimum.

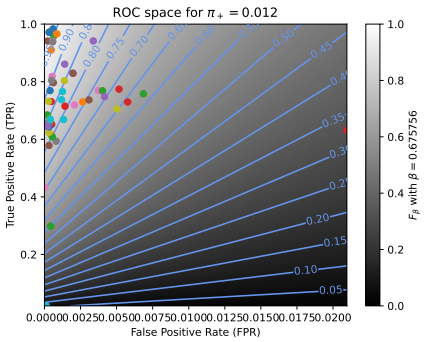


(e) Linear projection (PCA) of the manifold of the rankings induced by the F_β scores. The color points indicate the precision, the recall, F_1, $SIVF$, as well as the optimal tradeoff. The optimal tradeoff is at the same distance of the two extremities when the distance is measured along the manifold, with Kendall's distance d_τ.

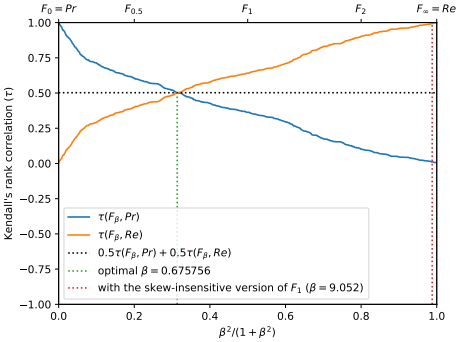


(f) The degree of optimality $\mathcal{O}(\beta)$ w.r.t. β. It is the probability to optimally ordering a pair of classifiers (BGS methods) given that it is not trivial (*i.e.*, that Pr and Re are in contradiction). The optimal value (or range of optimal values) for β is where the curve reaches 100%.

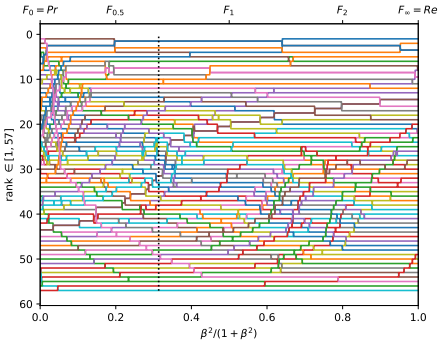
Figure A.3.58. Ranking of 57 BGS methods evaluated on the video "turbulence2" ($\pi_+ = 0.0008$).



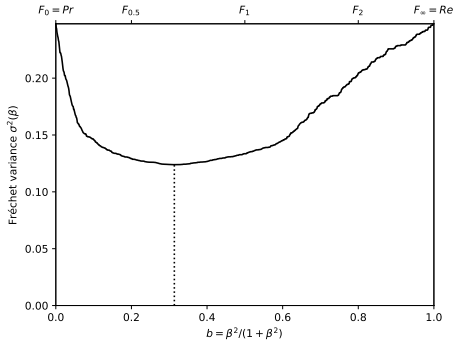
(a) The performances of 57 classifiers (BGS methods) depicted as points in the ROC space, with the isometrics of the optimal tradeoff score, from the ranking point of view, between precision and recall. See Eq. (15).



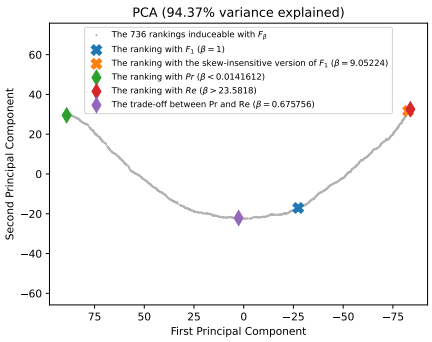
(b) The rank correlations $\tau(Pr; F_\beta)$ and $\tau(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the two curves intersect.



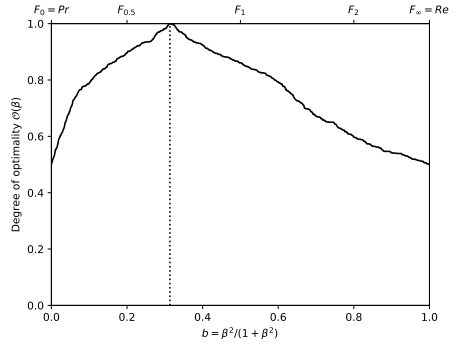
(c) The ranks of each classifier w.r.t. β. The optimal value (or range of optimal values) for β, shown here by the vertical line, is such that the number of swaps on its left is equal to the number of swaps on its right.



(d) The Fréchet variance $\sigma^2(\beta) = d_\tau^2(Pr; F_\beta) + d_\tau^2(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the curve has its minimum.

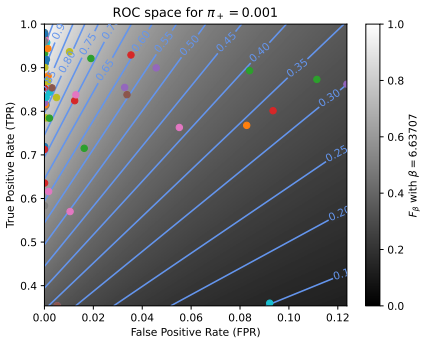


(e) Linear projection (PCA) of the manifold of the rankings induced by the F_β scores. The color points indicate the precision, the recall, F_1, $SIVF$, as well as the optimal tradeoff. The optimal tradeoff is at the same distance of the two extremities when the distance is measured along the manifold, with Kendall's distance d_τ.

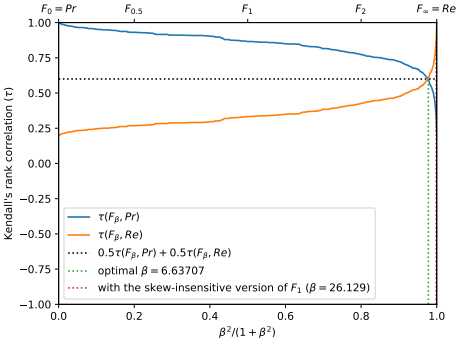


(f) The degree of optimality $\mathcal{O}(\beta)$ w.r.t. β. It is the probability to optimally ordering a pair of classifiers (BGS methods) given that it is not trivial (*i.e.*, that Pr and Re are in contradiction). The optimal value (or range of optimal values) for β is where the curve reaches 100%.

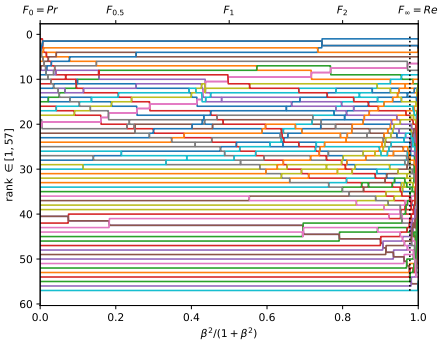
Figure A.3.59. Ranking of 57 BGS methods evaluated on the video "turbulence3" ($\pi_+ = 0.0121$).



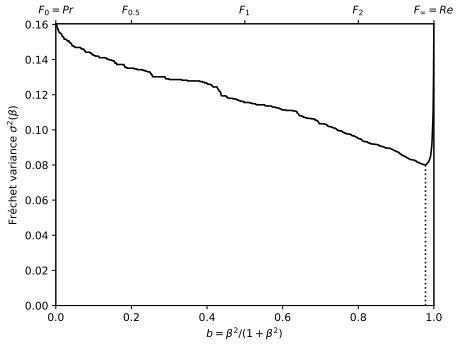
(a) The performances of 57 classifiers (BGS methods) depicted as points in the ROC space, with the isometrics of the optimal tradeoff score, from the ranking point of view, between precision and recall. See Eq. (15).



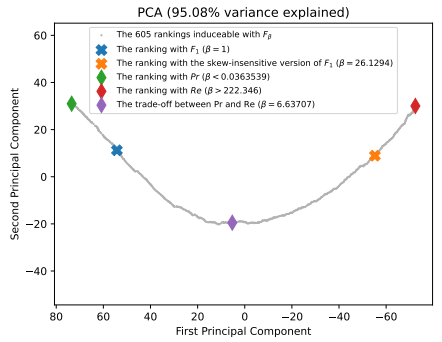
(b) The rank correlations $\tau(Pr; F_\beta)$ and $\tau(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the two curves intersect.



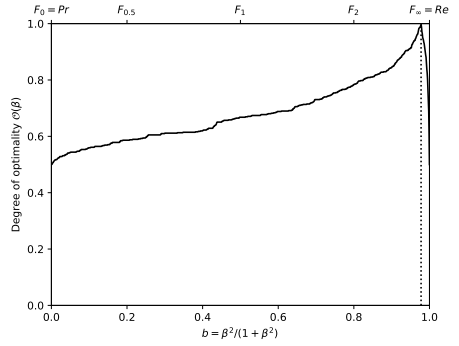
(c) The ranks of each classifier w.r.t. β. The optimal value (or range of optimal values) for β, shown here by the vertical line, is such that the number of swaps on its left is equal to the number of swaps on its right.



(d) The Fréchet variance $\sigma^2(\beta) = d_\tau^2(Pr; F_\beta) + d_\tau^2(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the curve has its minimum.

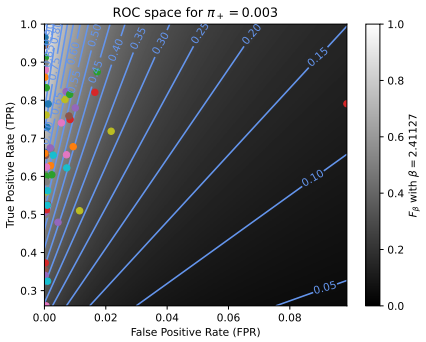


(e) Linear projection (PCA) of the manifold of the rankings induced by the F_β scores. The color points indicate the precision, the recall, F_1, $SIVF$, as well as the optimal tradeoff. The optimal tradeoff is at the same distance of the two extremities when the distance is measured along the manifold, with Kendall's distance d_τ.

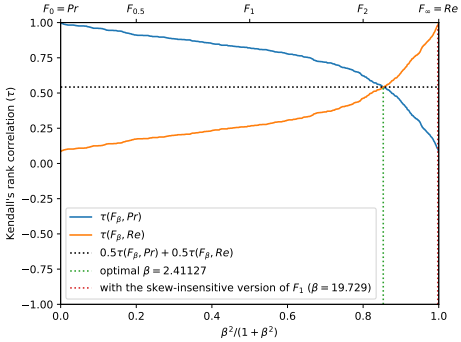


(f) The degree of optimality $\mathcal{O}(\beta)$ w.r.t. β. It is the probability to optimally ordering a pair of classifiers (BGS methods) given that it is not trivial (*i.e.*, that Pr and Re are in contradiction). The optimal value (or range of optimal values) for β is where the curve reaches 100%.

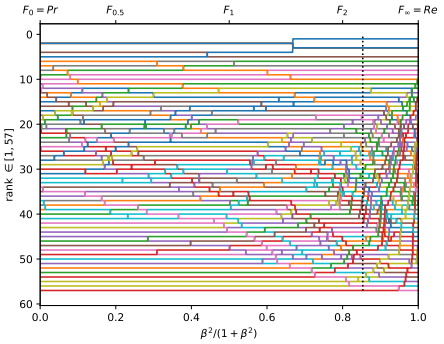
Figure A.3.60. Ranking of 57 BGS methods evaluated on the video "turbulence0" ($\pi_+ = 0.0015$).



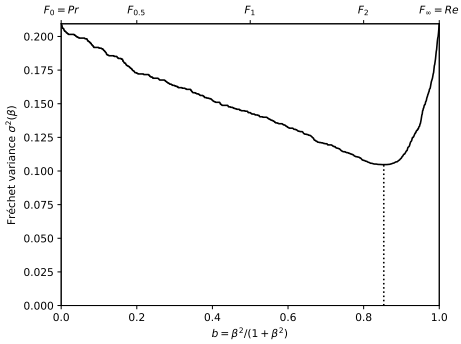
(a) The performances of 57 classifiers (BGS methods) depicted as points in the ROC space, with the isometrics of the optimal tradeoff score, from the ranking point of view, between precision and recall. See Eq. (15).



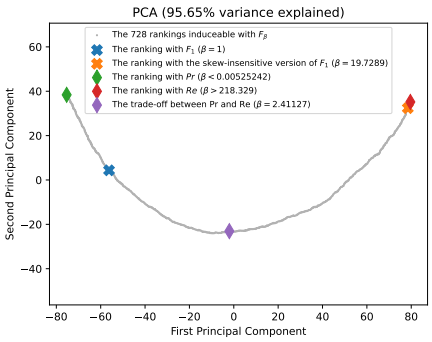
(b) The rank correlations $\tau(Pr; F_\beta)$ and $\tau(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the two curves intersect.



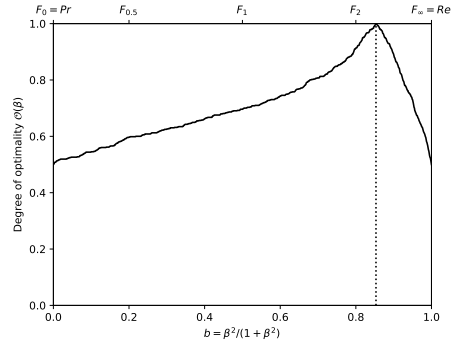
(c) The ranks of each classifier w.r.t. β. The optimal value (or range of optimal values) for β, shown here by the vertical line, is such that the number of swaps on its left is equal to the number of swaps on its right.



(d) The Fréchet variance $\sigma^2(\beta) = d_\tau^2(Pr; F_\beta) + d_\tau^2(F_\beta; Re)$ w.r.t. β. The optimal value (or range of optimal values) for β is where the curve has its minimum.



(e) Linear projection (PCA) of the manifold of the rankings induced by the F_β scores. The color points indicate the precision, the recall, F_1, $SIVF$, as well as the optimal tradeoff. The optimal tradeoff is at the same distance of the two extremities when the distance is measured along the manifold, with Kendall's distance d_τ.



(f) The degree of optimality $\mathcal{O}(\beta)$ w.r.t. β. It is the probability to optimally ordering a pair of classifiers (BGS methods) given that it is not trivial (*i.e.*, that Pr and Re are in contradiction). The optimal value (or range of optimal values) for β is where the curve reaches 100%.

Figure A.3.61. Ranking of 57 BGS methods evaluated on the video "turbulence1" ($\pi_+ = 0.0026$).

C. Supplementary material for the 4th article

This appendix contains the supplementary material provided with my 4th selected article [4], presented in Chapter 4.

A. Supplementary Material

Contents

A.1. List of symbols . . . 249
A.1.1. Mathematical Symbols . . . 249
A.1.2. Symbols related to our mathematical framework of [2] . . . 249
A.1.3. Symbols used for operations on performances . . . 250
A.1.4. Symbols used in the performance ordering and performance-based ranking theory . . . 250
A.1.5. Symbols used for the particular case of two-class crisp classifications . . . 250
A.2. Supplementary material about Sec. 4.1 . . . 252
A.2.1. Operations on performances . . . 252
A.2.2. When the priors are fixed: superimposing a grid on the Tile . . . 254
A.3. Supplementary material about Sec. 4.2 . . . 255
A.3.1. Placing performance orderings induced by probabilistic scores on the Tile . . . 255
A.3.2. Placing performance orderings induced by the scores F_β on the Tile . . . 256
A.3.3. Placing performance orderings induced by the score κ on the Tile . . . 256
A.3.4. Placing performance orderings induced by the score WA on the Tile . . . 259
A.3.5. Placing performance orderings induced by some other non-probabilistic scores on the Tile . . . 260
A.4. Supplementary material about Sec. 4.3 . . . 263
A.4.1. Algorithmic contributions . . . 263
A.5. Supplementary material about Sec. 4.4 . . . 264
A.5.1. What is achievable by no-skilled classifiers? . . . 264
A.6. Averaging all canonical ranking scores: the *Volume Under Tile* . . . 265
A.6.1. Definition . . . 265
A.6.2. Closed-form expression . . . 265
A.6.3. Discussion: can we use it to rank? . . . 266

A.1. List of symbols

A.1.1. Mathematical Symbols

- $\mathbf{1}_U$: the 0-1 indicator function of subset U
- $\mathbb{R}$: the real numbers
- $\mathcal{R}$: a relation
- conv: the set of convex combinations
- $\vee$: the *inclusive disjunction* (*i.e.*, logical or)
- $\wedge$: the *conjunction* (*i.e.*, logical and)
- $\circ$: the composition of functions, *i.e.* $(g \circ f)(x) = g(f(x))$
- $\mathbf{E}$: the mathematical expectation

A.1.2. Symbols related to our mathematical framework of [2]

We organize these symbols according to the 6 pillars depicted in the graphical abstract of [2].

Symbols related to the 1st pillar

- Ω: the sample space (universe)
- ω: a sample (*i.e.*, an element of Ω)
- Σ: the event space (a σ-algebra on Ω, *e.g.* 2^Ω)
- E: an event (*i.e.*, an element of Σ)
- (Ω, Σ): the measurable space
- $\mathbb{P}_{(\Omega,\Sigma)}$: all performances on (Ω, Σ)
- Π: a set of performances ($\Pi \subseteq \mathbb{P}_{(\Omega,\Sigma)}$)
- P: a performance (*i.e.*, an element of $\mathbb{P}_{(\Omega,\Sigma)}$)

Symbols related to the 2nd pillar

- $\lesssim$: binary relation *worse or equivalent* on $\mathbb{P}_{(\Omega,\Sigma)}$
- $\gtrsim$: binary relation *better or equivalent* on $\mathbb{P}_{(\Omega,\Sigma)}$
- $\sim$: binary relation *equivalent* on $\mathbb{P}_{(\Omega,\Sigma)}$
- $>$: binary relation *better* on $\mathbb{P}_{(\Omega,\Sigma)}$
- $<$: binary relation *worse* on $\mathbb{P}_{(\Omega,\Sigma)}$
- $\lesseqgtr$: binary relation *incomparable* on $\mathbb{P}_{(\Omega,\Sigma)}$

Symbols related to the 3rd pillar

- S: the random variable *Satisfaction*

Symbols related to the 4th pillar

- $\mathbb{E}$: the set of entities to rank
- ϵ: an entity, *i.e.* an element of $\mathbb{E}$
- eval: the performance *evaluation* function
- Φ: some performances that are for sure achievable

Symbols related to the 5th pillar

- $\mathbb{X}_{(\Omega,\Sigma)}$: all scores on (Ω, Σ)
- X: a score
- $\mathrm{dom}(X)$: the domain of the score X
- X_E^U: the *unconditional probabilistic score* parameterized by the event E
- $X_{E_1|E_2}^C$: the *conditional probabilistic score* parameterized by the events E_1 and E_2

Symbols related to the 6th pillar

- I: the random variable *Importance*

A.1.3. Symbols used for operations on performances

- filter_I: the *filtering* operation, parameterized by a random variable I, as defined in [2]
- no–skill: the operation that transforms a performance P into P' such that $P'(Y, \hat{Y}) = P(Y)P(\hat{Y})$
- $\mathrm{shift}_{\pi \to \pi'}$: the *prior/target shift* operation [94]
- $\mathrm{change}_{\hat{Y}}$: the operation that changes the predicted class $\hat{Y}$
- change_{Y}: the operation that changes the ground-truth class Y
- $\mathrm{swap}_{Y \leftrightarrow \hat{Y}}$: the operation that swaps the predicted ($\hat{Y}$) and ground-truth (Y) classes
- $\mathrm{swap}_{c_- \leftrightarrow c_+}$: the operation that swaps the classes c_- and c_+.

A.1.4. Symbols used in the performance ordering and performance-based ranking theory

- $\mathrm{rank}_{\mathbb{E}}$: the *ranking* function, w.r.t. the set of entities $\mathbb{E}$
- $\lesssim_X$: the ordering induced by the score X
- $\gtrsim_X$: the dual (inverted) ordering induced by the score X
- R_I: the *ranking score* parameterized by the importance I
- $R_{I_{a,b}}$: the *canonical ranking score* parameterized by the parameters a and b
- a: the parameter specifying the relative importance given to the incorrect outcomes ($S = 0$), it corresponds to the horizontal axis of the Tile
- b: the parameter specifying the relative importance given to the correct outcomes ($S = 1$), it corresponds to the vertical axis of the Tile
- τ: the rank correlation coefficient of Kendall

A.1.5. Symbols used for the particular case of two-class crisp classifications

Particularization of the mathematical framework

- tn: the sample *true negative*
- fp: the sample *false positive*, *a.k.a.* type I error
- fn: the sample *false negative*, *a.k.a.* type II error
- tp: the sample *true positive*

Extensions to the mathematical framework

- ROC: the *Receiver Operating Characteristic* space, *i.e.* $FPR \times TPR$
- PR: the *Precision-Recall* space, *i.e.* $TPR \times PPV$
- Y: the random variable for the ground truth
- $\hat{Y}$: the random variable for the prediction
- $\mathbb{C}$: the set of classes
- c: a class (*i.e.*, an element of $\mathbb{C}$)
- c_-: the negative class
- c_+: the positive class
- γ_π: the locus, on the Tile, of all the canonical ranking scores that put the no-skill performances on the same footing, for fixed class priors
- γ_τ: the locus, on the Tile, of all the canonical ranking scores that put the no-skill performances on the same footing, for fixed rates of predictions

Some unconditional probabilistic scores

- PTN: the probability of the elementary event *true negative*, *a.k.a. rejection rate*
- PFP: the probability of the elementary event *false positive*
- PFN: the probability of the elementary event *false negative*
- PTP: the probability of the elementary event *true positive*, *a.k.a. detection rate*
- π_-: the *prior of the negative class*
- π_+: the *prior of the positive class*, *a.k.a. prevalence*
- τ_-: the *rate of negative predictions*
- τ_+: the *rate of positive predictions*
- A: the *accuracy*, *a.k.a. matching coefficient*

Some conditional probabilistic scores

- TNR: the *True Negative Rate*, *a.k.a. specificity*, *selectivity*, *inverse recall*
- FPR: the *False Positive Rate*
- TPR: the *True Positive Rate*, *a.k.a. sensitivity*, *recall*
- FNR: the *False Negative Rate*
- NPV: the *Negative Predictive Value*, *a.k.a. inverse precision*
- FOR: the *False Omission Rate*
- NPV: the *Positive Predictive Value*, *a.k.a. precision*
- FDR: the *False Discovery Rate*
- J_-: Jaccard index for the negative class
- J_+: Jaccard index for the positive class, *a.k.a. Tanimoto coefficient*, *similarity*, *intersection over union*, *critical success index* [68], *G-measure* [29]

Some other scores

- S_B: Bennett, Alpert and Goldstein's S
- F_β: the F-scores
- F_1: the F-one score, *a.k.a. Dice-Sørensen coefficient*
- $SNPV$: the score *Standardized Negative Predictive Value* [70]
- $SPPV$: the score *Standardized Positive Predictive Value* [70]
- NLR: the score *Negative Likelihood Ratio*
- PLR: the score *Positive Likelihood Ratio*
- κ: Cohen's kappa statistic, *a.k.a. Heidke Skill Score* [76]
- π: Scott's pi statistic
- κ: Fleiss's kappa statistic
- BI: the *Bias Index*, as defined in [77]
- WA: the *Weighted Accuracy*
- BA: the *Balanced Accuracy*
- J_Y: Youden's index [82], *a.k.a.* Youden's J statistic, *informedness*, *Peirce Skill Score* [76]
- $|C|$: the determinant of the (normalized) confusion matrix or contingency matrix
- ACP: the *Average Conditional Probability*, *i.e.* the arithmetic mean of the Tile's four corners [86]
- P_4: the harmonic mean of the Tile's four corners [87]
- VUT: the score *Volume Under Tile*, *i.e.* the arithmetic mean of all canonical scores (see Sec. A.6)

Table 1. Summary of the effects on the Tile of 5 operations on performances.

operation on performances		the ordering that was at $(a_{\text{origin}}, b_{\text{origin}})$ is moved at $(a_{\text{adapted}}, b_{\text{adapted}})$		note
$\text{change}_{\hat{Y}}$	(see Lemma 1)	$a_{\text{adapted}} = b_{\text{origin}}$	$b_{\text{adapted}} = a_{\text{origin}}$	the preorder is inverted (dual)
change_{Y}	(see Lemma 2)	$a_{\text{adapted}} = 1 - b_{\text{origin}}$	$b_{\text{adapted}} = 1 - a_{\text{origin}}$	the preorder is inverted (dual)
$\text{swap}_{Y \leftrightarrow \hat{Y}}$	(see Lemma 3)	$a_{\text{adapted}} = a_{\text{origin}}$	$b_{\text{adapted}} = 1 - b_{\text{origin}}$	the preorder is unchanged
$\text{swap}_{c_- \leftrightarrow c_+}$	(see Lemma 4)	$a_{\text{adapted}} = 1 - a_{\text{origin}}$	$b_{\text{adapted}} = 1 - b_{\text{origin}}$	the preorder is unchanged
$\text{shift}_{\pi \to \pi'}$	(see Lemma 5)	$a_{\text{adapted}} = f^{-1}(a_{\text{origin}})$	$b_{\text{adapted}} = f^{-1}(b_{\text{origin}})$	the preorder is unchanged

A.2. Supplementary material about Sec. 4.1

A.2.1. Operations on performances

We present here the proofs for the geometric properties of the Tile that are related to some operations that can be applied to performances. A summary is provided in Tab. 1.

Lemma 1. *Let* $\text{change}_{\hat{Y}} : \mathbb{P}_{(\Omega,\Sigma)} \to \mathbb{P}_{(\Omega,\Sigma)}$ *be the operation that changes the predicted class* $\hat{Y}$*. We have* $R_{I_{a,b}} \circ \text{change}_{\hat{Y}} = 1 - R_{I_{b,a}}$.

Proof. Let P be a two-class classification performance and $P' = \text{change}_{\hat{Y}}(P)$. If $P' \in \text{dom}(R_{I_{a,b}})$, then $P \in \text{dom}(R_{I_{b,a}})$ and

$$\begin{aligned} R_{I_{a,b}}(P') &= \frac{(1-a)P'(\{tn\}) + aP'(\{tp\})}{(1-a)P'(\{tn\}) + (1-b)P'(\{fp\}) + bP'(\{fn\}) + aP'(\{tp\})} \\ &= \frac{(1-a)P(\{fp\}) + aP(\{fn\})}{(1-a)P(\{fp\}) + (1-b)P(\{tn\}) + bP(\{tp\}) + aP(\{fn\})} \\ &= 1 - \frac{(1-b)P(\{tn\}) + bP(\{tp\})}{(1-b)P(\{tn\}) + (1-a)P(\{fp\}) + aP(\{fn\}) + bP(\{tp\})} \\ &= 1 - R_{I_{b,a}}(P) \end{aligned}$$

□

Lemma 2. *Let* $\text{change}_{Y} : \mathbb{P}_{(\Omega,\Sigma)} \to \mathbb{P}_{(\Omega,\Sigma)}$ *be the operation that changes the ground-truth class* Y*. We have* $R_{I_{a,b}} \circ \text{change}_{Y} = 1 - R_{I_{1-b,1-a}}$.

Proof. Let P be a two-class classification performance and $P' = \text{change}_{Y}(P)$. If $P' \in \text{dom}(R_{I_{a,b}})$, then $P \in \text{dom}(R_{I_{1-b,1-a}})$ and

$$\begin{aligned} R_{I_{a,b}}(P') &= \frac{(1-a)P'(\{tn\}) + aP'(\{tp\})}{(1-a)P'(\{tn\}) + (1-b)P'(\{fp\}) + bP'(\{fn\}) + aP'(\{tp\})} \\ &= \frac{(1-a)P(\{fn\}) + aP(\{fp\})}{(1-a)P(\{fn\}) + (1-b)P(\{tp\}) + bP(\{tn\}) + aP(\{fp\})} \\ &= 1 - \frac{bP(\{tn\}) + (1-b)P(\{tp\})}{bP(\{tn\}) + aP(\{fp\}) + (1-a)P(\{fn\}) + (1-b)P(\{tp\})} \\ &= 1 - R_{I_{1-b,1-a}}(P) \end{aligned}$$

□

Lemma 3. *Let* $\text{swap}_{Y \leftrightarrow \hat{Y}} : \mathbb{P}_{(\Omega,\Sigma)} \to \mathbb{P}_{(\Omega,\Sigma)}$ *be the operation that swaps the predicted* ($\hat{Y}$) *and ground-truth* (Y) *classes. We have* $R_{I_{a,b}} \circ \text{swap}_{Y \leftrightarrow \hat{Y}} = R_{I_{a,1-b}}$.

Proof. Let P be a two-class classification performance and $P' = \text{swap}_{Y \leftrightarrow \hat{Y}}(P)$. If $P' \in \text{dom}(R_{I_{a,b}})$, then $P \in \text{dom}(R_{I_{a,1-b}})$ and

$$\begin{aligned}
R_{I_{a,b}}(P') &= \frac{(1-a)P'(\{tn\}) + aP'(\{tp\})}{(1-a)P'(\{tn\}) + (1-b)P'(\{fp\}) + bP'(\{fn\}) + aP'(\{tp\})} \\
&= \frac{(1-a)P(\{tn\}) + aP(\{tp\})}{(1-a)P(\{tn\}) + (1-b)P(\{fn\}) + bP(\{fp\}) + aP(\{tp\})} \\
&= \frac{(1-a)P(\{tn\}) + aP(\{tp\})}{(1-a)P(\{tn\}) + bP(\{fp\}) + (1-b)P(\{fn\}) + aP(\{tp\})} \\
&= R_{I_{a,1-b}}(P)
\end{aligned}$$

□

Lemma 4. *Let* $\text{swap}_{c_- \leftrightarrow c_+} : \mathbb{P}_{(\Omega,\Sigma)} \to \mathbb{P}_{(\Omega,\Sigma)}$ *be the operation that swaps the classes* c_- *and* c_+*. We have* $R_{I_{a,b}} \circ \text{swap}_{c_- \leftrightarrow c_+} = R_{I_{1-a,1-b}}$.

Proof. Let P be a two-class classification performance and $P' = \text{swap}_{c_- \leftrightarrow c_+}(P)$. If $P' \in \text{dom}(R_{I_{a,b}})$, then $P \in \text{dom}(R_{I_{1-a,1-b}})$ and

$$\begin{aligned}
R_{I_{a,b}}(P') &= \frac{(1-a)P'(\{tn\}) + aP'(\{tp\})}{(1-a)P'(\{tn\}) + (1-b)P'(\{fp\}) + bP'(\{fn\}) + aP'(\{tp\})} \\
&= \frac{(1-a)P(\{tp\}) + aP(\{tn\})}{(1-a)P(\{tp\}) + (1-b)P(\{fn\}) + bP(\{fp\}) + aP(\{tn\})} \\
&= \frac{aP(\{tn\}) + (1-a)P(\{tp\})}{aP(\{tn\}) + bP(\{fp\}) + (1-b)P(\{fn\}) + (1-a)P(\{tp\})} \\
&= R_{I_{1-a,1-b}}(P)
\end{aligned}$$

□

Lemma 5. *Let* $\text{shift}_{\pi \to \pi'}$ *be the operation that applies a prior/target shift on the distribution* $P(Y)$*, transforming the priors* (π_-, π_+) *into* (π'_-, π'_+)*. The ordering induced by* $R_{I_{a,b}} \circ \text{shift}_{\pi \to \pi'}$ *is the same as the ordering induced by* $R_{I_{f(a),f(b)}}$*, where* f *is the function*

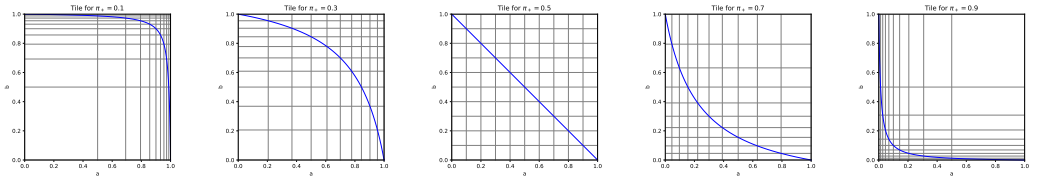
$$f : x \mapsto f(x) = \frac{x\frac{\pi'_+}{\pi_+}}{(1-x)\frac{\pi'_-}{\pi_-} + x\frac{\pi'_+}{\pi_+}}\,.$$

Proof. Let P be a two-class classification performance and $P' = \text{shift}_{\pi \to \pi'}(P)$. We have:

$$\begin{aligned}
R_{I_{a,b}}(P') &= \frac{(1-a)P'(\{tn\}) + aP'(\{tp\})}{(1-a)P'(\{tn\}) + (1-b)P'(\{fp\}) + bP'(\{fn\}) + aP'(\{tp\})} \\
&= \frac{(1-a)P(\{tn\})\frac{\pi'_-}{\pi_-} + aP(\{tp\})\frac{\pi'_+}{\pi_+}}{(1-a)P(\{tn\})\frac{\pi'_-}{\pi_-} + (1-b)P(\{fp\})\frac{\pi'_-}{\pi_-} + bP(\{fn\})\frac{\pi'_+}{\pi_+} + aP(\{tp\})\frac{\pi'_+}{\pi_+}} \\
&= \frac{I'(tn)P(\{tn\}) + I'(tp)P(\{tp\})}{I'(tn)P(\{tn\}) + I'(fp)P(\{fp\}) + I'(fn)P(\{fn\}) + I'(tp)P(\{tp\})} \\
&= R_{I'}(P)
\end{aligned}$$

with

$$\begin{cases}
I'(tn) = g(a)\,(1 - f(a)) \\
I'(fp) = g(b)\,(1 - f(b)) \\
I'(fn) = g(b)f(b) \\
I'(tp) = g(a)f(a)
\end{cases}$$



(a) Shift from $(\pi_-, \pi_+) = (0.5, 0.5)$ to $(0.9, 0.1)$.

(b) Shift from $(\pi_-, \pi_+) = (0.5, 0.5)$ to $(0.7, 0.3)$.

(c) Reference.

(d) Shift from $(\pi_-, \pi_+) = (0.5, 0.5)$ to $(0.3, 0.7)$.

(e) Shift from $(\pi_-, \pi_+) = (0.5, 0.5)$ to $(0.1, 0.9)$.

Figure 8. Visualization of how the Tile deforms with a shift for the distribution $P(Y)$. The central figure shows an initial Tile, in which we have arbitrarily drawn a regular grid. The other figures show the result of Lemma 5, representing the way the Tile deforms. We've also drawn the descending diagonal on our reference Tile. Its deformation gives birth to the family of curves denoted by γ_π in our paper.

and

$$g : x \mapsto g(x) = (1-x)\frac{\pi'_-}{\pi_-} + x\frac{\pi'_+}{\pi_+}\,.$$

The random variable I' can be rewritten as $I' = (\mathbf{1}_{S=0}g(b) + \mathbf{1}_{S=1}g(a))I_{f(a),f(b)}$. Using Property 2, we conclude that

$$\begin{aligned}
&\frac{\partial R_{I'}}{\partial R_{I_{f(a),f(b)}}} > 0 \\
\Rightarrow &\frac{\partial\left(R_{I_{a,b}} \circ \text{shift}_{\pi\to\pi'}\right)}{\partial R_{I_{f(a),f(b)}}} > 0 \\
\Rightarrow &\lesssim_{R_{I_{a,b}} \circ \text{shift}_{\pi\to\pi'}} = \lesssim_{R_{I_{f(a),f(b)}}}\,.
\end{aligned}$$

□

A.2.2. When the priors are fixed: superimposing a grid on the Tile

We found it very useful, when priors are fixed, to represent the displacement that would have occurred if we had started with uniform priors and applied the target/prior shift. This can be done thanks to Lemma 5. In practice, we do this by superimposing a grid on the Tile, as done in Fig. 5 and Fig. 7. We provide more information on the subject in Fig. 8. Vertical lines are drawn at

$$a = f^{-1}(x) = \frac{x\frac{0.5}{\pi_+}}{(1-x)\frac{0.5}{\pi_-} + x\frac{0.5}{\pi_+}} \text{ for } x = 0.0, 0.1, 0.2, \ldots, 0.9, 1.0\,, \tag{4}$$

and horizontal lines are drawn at

$$b = f^{-1}(y) = \frac{y\frac{0.5}{\pi_+}}{(1-y)\frac{0.5}{\pi_-} + y\frac{0.5}{\pi_+}} \text{ for } y = 0.0, 0.1, 0.2, \ldots, 0.9, 1.0\,. \tag{5}$$

Table 2. The nine ranking scores for two-class classifications that belong to the family of probabilistic scores.

Ranking score R_I	Probabilistic writing	$I(tn)$	$I(fp)$	$I(fn)$	$I(tp)$	Canonical?	location of $\lesssim_{R_I}$ on the Tile
Negative Predictive Value NPV	$P(S=1 \mid \hat{Y}=c_-)$	1	0	1	0	yes	$(a,b)=(0,1)$
$X^C_{\{tn,tp\}\mid\{tn,fn,tp\}}$	$P(S=1 \mid Y=c_+ \vee \hat{Y}=c_-)$	1	0	1	1	no	$(a,b)=(\frac{1}{2},1)$
True Positive Rate TPR	$P(S=1 \mid Y=c_+)$	0	0	1	1	yes	$(a,b)=(1,1)$
Jaccard's index for the negative class J_-	$P(S=1 \mid Y=c_- \vee \hat{Y}=c_-)$	1	1	1	0	no	$(a,b)=(0,\frac{1}{2})$
Accuracy A	$P(S=1)$	$1/2$	$1/2$	$1/2$	$1/2$	yes	$(a,b)=(\frac{1}{2},\frac{1}{2})$
Jaccard's index for the positive class J_+	$P(S=1 \mid Y=c_+ \vee \hat{Y}=c_+)$	0	1	1	1	no	$(a,b)=(1,\frac{1}{2})$
True Negative Rate TNR	$P(S=1 \mid Y=c_-)$	1	1	0	0	yes	$(a,b)=(0,0)$
$X^C_{\{tn,tp\}\mid\{tn,fp,tp\}}$	$P(S=1 \mid Y=c_- \vee \hat{Y}=c_+)$	1	1	0	1	no	$(a,b)=(\frac{1}{2},0)$
Positive Predictive Value PPV	$P(S=1 \mid \hat{Y}=c_+)$	0	1	0	1	yes	$(a,b)=(1,0)$

A.3. Supplementary material about Sec. 4.2

A.3.1. Placing performance orderings induced by probabilistic scores on the Tile

The following lemma explains that some probabilistic scores are ranking scores, and specifies the importance values for them. Based on this, placing the performance orderings induced by probabilistic scores is straightforward, as the performance ordering induced by the ranking score R_I is located at $(a,b)=(\frac{I(tp)}{I(tn)+I(tp)},\frac{I(fn)}{I(fp)+I(fn)})$.

Lemma 6. *Let S be a binary-valued satisfaction. A probabilistic score $X^C_{E_1|E_2}$, with $\emptyset \subsetneq E_1 \subsetneq E_2 \subseteq \Omega$, is a ranking score R_I if $I = k\mathbf{1}_{E_2}$, $k>0$, and $E_1 = E_2 \cap E_S$, with $E_S = \{\omega \in \Omega : S(\omega)=1\}$.*

Proof. Let us first check the equality of the domains. We have

$$\mathrm{dom}(R_I) = \left\{P \in \mathbb{P}_{(\Omega,\Sigma)} : \mathbf{E}_P[I] \neq 0\right\}$$

and

$$\mathrm{dom}(X^C_{E_1|E_2}) = \left\{P \in \mathbb{P}_{(\Omega,\Sigma)} : P(E_2) \neq 0\right\}.$$

Using the "fundamental bridge", we have

$$P(E_2) \neq 0 \Leftrightarrow \mathbf{E}_P[\mathbf{1}_{E_2}] \neq 0 \Leftrightarrow \mathbf{E}_P[I] \neq 0\,,$$

and thus

$$\mathrm{dom}(R_I) = \mathrm{dom}(X^C_{E_1|E_2})\,.$$

We now check the equality of the values taken by both scores. We have, for all $P \in \mathrm{dom}(X^C_{E_1|E_2})$,

$$X^C_{E_1|E_2}(P) = P(E_1|E_2) = \frac{P(E_1 \cap E_2)}{P(E_2)} = \frac{P(E_1)}{P(E_2)}\,.$$

Using again the "fundamental bridge",

$$X^C_{E_1|E_2}(P) = \frac{\mathbf{E}_P[\mathbf{1}_{E_1}]}{\mathbf{E}_P[\mathbf{1}_{E_2}]} = \frac{\mathbf{E}_P[\mathbf{1}_{E_2\cap E_S}]}{\mathbf{E}_P[\mathbf{1}_{E_2}]} = \frac{\mathbf{E}_P[\mathbf{1}_{E_2}S]}{\mathbf{E}_P[\mathbf{1}_{E_2}]} = \frac{\mathbf{E}_P[\frac{I}{k}S]}{\mathbf{E}_P[\frac{I}{k}]} = \frac{\mathbf{E}_P[IS]}{\mathbf{E}_P[I]} = R_I(P)\,.$$

□

Thanks to Lemma 6, we can place on the Tile the probabilistic scores that are ranking scores. In the particular case of two-class classification, there exist 9 pairs (E_1,E_2) such that $\emptyset \subsetneq E_1 \subsetneq E_2 \subseteq \Omega$ and $E_1 = E_2 \cap E_S$. There are thus 9 ranking scores that are also probabilistic scores. Among them, 5 are canonical and can be placed directly on the Tile. For the 4 others, the induced performance ordering can be placed on the Tile. A summary is provided in Table 2.

A.3.2. Placing performance orderings induced by the scores F_β on the Tile

Lemma 7 (F-scores)**.** *In two-class classification, all F-scores are canonical ranking scores: $F_\beta = R_I$ with $I(tn) = 0$, $I(fp) = \frac{1}{1+\beta^2}$, $I(fn) = \frac{\beta^2}{1+\beta^2}$, and $I(tp) = 1$.*

Proof. For the sake of concision, let us pose $TN = P(\{tn\})$, $FP = P(\{fp\})$, $FN = P(\{fn\})$, and $TP = P(\{tp\})$.

We first check that F_β and R_I have the same domain.

$$\begin{aligned}\mathrm{dom}(F_\beta) &= \left\{P \in \mathbb{P}_{(\Omega,\Sigma)} : (1+\beta^2)TP + \beta^2 FN + FP \neq 0\right\}\\ &= \left\{P \in \mathbb{P}_{(\Omega,\Sigma)} : (1+\beta^2)\mathbf{E}_P[I] \neq 0\right\}\\ &= \left\{P \in \mathbb{P}_{(\Omega,\Sigma)} : \mathbf{E}_P[I] \neq 0\right\}\\ &= \mathrm{dom}(R_I)\end{aligned}$$

Then, we check the equality of the values taken by both scores. By definition of F_β, we have, for all $P \in \mathrm{dom}(F_\beta)$,

$$\begin{aligned}F_\beta(P) &= \frac{(1+\beta^2)TP}{(1+\beta^2)TP + \beta^2 FN + FP}\\ &= \frac{0TN + 1TP}{0TN + \frac{1}{1+\beta^2}FP + \frac{\beta^2}{1+\beta^2}FN + 1TP}\\ &= \frac{I(tn)TN + I(tp)TP}{I(tn)TN + I(fp)FP + I(fn)FN + I(tp)TP}\\ &= R_I(P)\,.\end{aligned}$$

In conclusion, $F_\beta = R_I$. As Moreover, as $I(tn) + I(tp) = I(fp) + I(fn)$, R_I is canonical. □

A.3.3. Placing performance orderings induced by the score κ on the Tile

Lemma 8 (Cohen's κ statistic)**.** *Cohen's κ statistic increases linearly with a canonical ranking score when the priors are fixed. In two-class classification, let the priors of the negative and positive classes be fixed and denoted, respectively, by π_- and π_+. In this case, $(\pi_-^2 + \pi_+^2)\kappa + 2\pi_-\pi_+ = R_I$ with $I(tn) = \frac{\pi_+^2}{\pi_-^2+\pi_+^2}$, $I(fp) = \frac{1}{2}$, $I(fn) = \frac{1}{2}$, and $I(tp) = \frac{\pi_-^2}{\pi_-^2+\pi_+^2}$.*

Proof. Let us begin by introducing the set of performances with fixed and strictly positive priors:

$$\mathbb{P}^* = \left\{P \in \mathbb{P}_{(\Omega,\Sigma)} : P(\{fn, tp\}) = \pi_+\right\} \qquad \text{with } \pi_+ \in (0,1)\ .$$

For the sake of concision, let us pose $TN = P(\{tn\})$, $FP = P(\{fp\})$, $FN = P(\{fn\})$, $TP = P(\{tp\})$, $\tau_- = P(\{fn, tn\})$, and $\tau_+ = P(\{fp, tp\})$.

Let us first observe that $\mathrm{dom}(\kappa) \cap \mathbb{P}^* = \mathrm{dom}(R_I) \cap \mathbb{P}^*$. By definition of R_I,

$$\mathrm{dom}(R_I) = \left\{P \in \mathbb{P}_{(\Omega,\Sigma)} : \mathbf{E}_P[I] \neq 0\right\}\ ,$$

and by definition of κ,

$$\mathrm{dom}(\kappa) = \left\{P \in \mathbb{P}_{(\Omega,\Sigma)} : \pi_-\tau_- + \pi_+\tau_+ \neq 1\right\}\ .$$

We discuss three cases.

1. When $\pi_+ = 0$, which implies that $FN = 0 \wedge TP = 0$, we have

$$\begin{aligned}\mathrm{dom}(\kappa) \cap \mathbb{P}^* &= \left\{P \in \mathbb{P}_{(\Omega,\Sigma)} : \pi_-\tau_- + \pi_+\tau_+ \neq 1\right\} \cap \mathbb{P}^*\\ &= \left\{P \in \mathbb{P}_{(\Omega,\Sigma)} : \tau_- \neq 1\right\} \cap \mathbb{P}^*\\ &= \left\{P \in \mathbb{P}_{(\Omega,\Sigma)} : FN + TN \neq 1\right\} \cap \mathbb{P}^*\\ &= \left\{P \in \mathbb{P}_{(\Omega,\Sigma)} : TN \neq 1\right\} \cap \mathbb{P}^*\\ &= \left\{P \in \mathbb{P}_{(\Omega,\Sigma)} : \mathbf{E}_P[I] \neq 0\right\} \cap \mathbb{P}^*\\ &= \mathrm{dom}(R_I) \cap \mathbb{P}^*\,.\end{aligned}$$

2. When $\pi_+ \in]0,1[$, we have

$$\begin{aligned}\mathrm{dom}(\kappa) \cap \mathbb{P}^* &= \left\{P \in \mathbb{P}_{(\Omega,\Sigma)} : \pi_-\tau_- + \pi_+\tau_+ \neq 1\right\} \cap \mathbb{P}^* \\ &= \left\{P \in \mathbb{P}_{(\Omega,\Sigma)}\right\} \cap \mathbb{P}^* \\ &= \mathrm{dom}(R_I) \cap \mathbb{P}^* \,.\end{aligned}$$

3. When $\pi_+ = 1$, which implies that $TN = 0 \wedge FP = 0$, we have

$$\begin{aligned}\mathrm{dom}(\kappa) \cap \mathbb{P}^* &= \left\{P \in \mathbb{P}_{(\Omega,\Sigma)} : \pi_-\tau_- + \pi_+\tau_+ \neq 1\right\} \cap \mathbb{P}^* \\ &= \left\{P \in \mathbb{P}_{(\Omega,\Sigma)} : \tau_+ \neq 1\right\} \cap \mathbb{P}^* \\ &= \left\{P \in \mathbb{P}_{(\Omega,\Sigma)} : FP + TP \neq 1\right\} \cap \mathbb{P}^* \\ &= \left\{P \in \mathbb{P}_{(\Omega,\Sigma)} : TP \neq 1\right\} \cap \mathbb{P}^* \\ &= \left\{P \in \mathbb{P}_{(\Omega,\Sigma)} : \mathbf{E}_P[I] \neq 0\right\} \cap \mathbb{P}^* \\ &= \mathrm{dom}(R_I) \cap \mathbb{P}^* \,.\end{aligned}$$

The restricted domains are thus equal in all cases:

$$\mathrm{dom}(\kappa) \cap \mathbb{P}^* = \mathrm{dom}(R_I) \cap \mathbb{P}^* \,.$$

We now check the equality of the values taken by both scores. By definition, for all $P \in \mathrm{dom}(\kappa)$,

$$\begin{aligned}\kappa(P) &= \frac{A - EA}{1 - EA} \text{ avec } \begin{cases} A = P(\{tn, tp\}) & \text{(the accuracy)} \\ EA = \pi_-\tau_- + \pi_+\tau_+ & \text{(the expected accuracy)} \end{cases} \\ &= \frac{(TN + TP) - (\pi_-TN + \pi_+FP + \pi_-FN + \pi_+TP)}{(TN + FP + FN + TP) - (\pi_-TN + \pi_+FP + \pi_-FN + \pi_+TP)} \\ &= \frac{(1-\pi_-)TN - \pi_+FP - \pi_-FN + (1-\pi_+)TP}{(1-\pi_-)TN + (1-\pi_+)FP + (1-\pi_-)FN + (1-\pi_+)TP} \\ &= \frac{\pi_+TN - \pi_+FP - \pi_-FN + \pi_-TP}{\pi_+TN + \pi_-FP + \pi_+FN + \pi_-TP}\end{aligned}$$

Thus,

$$(\pi_-^2 + \pi_+^2)\kappa(P) + 2\pi_-\pi_+ = \frac{\lambda_{tn}TN + \lambda_{fp}FP + \lambda_{fn}FN + \lambda_{tp}TP}{\pi_+TN + \pi_-FP + \pi_+FN + \pi_-TP}$$

with

$$\begin{aligned}\lambda_{tn} &= \pi_+(\pi_-^2 + \pi_+^2) + \pi_+(2\pi_-\pi_+) \\ &= \pi_+(\pi_-^2 + 2\pi_-\pi_+ + \pi_+^2) \\ &= \pi_+(\pi_- + \pi_+)^2 \\ &= \pi_+\end{aligned}$$

$$\begin{aligned}\lambda_{fp} &= -\pi_+(\pi_-^2 + \pi_+^2) + \pi_-(2\pi_-\pi_+) \\ &= \pi_+(\pi_-^2 - \pi_+^2) \\ &= \pi_+(\pi_- - \pi_+)(\pi_- + \pi_+) \\ &= \pi_+(\pi_- - \pi_+)\end{aligned}$$

$$\begin{aligned}\lambda_{fn} &= -\pi_-(\pi_-^2 + \pi_+^2) + \pi_+(2\pi_-\pi_+) \\ &= \pi_-(\pi_+^2 - \pi_-^2) \\ &= \pi_-(\pi_+ - \pi_-)(\pi_+ + \pi_-) \\ &= \pi_-(\pi_+ - \pi_-)\end{aligned}$$

$$\begin{aligned}\lambda_{tp} &= \pi_-(\pi_-^2 + \pi_+^2) + \pi_-(2\pi_-\pi_+)\\ &= \pi_-(\pi_-^2 + 2\pi_-\pi_+ + \pi_+^2)\\ &= \pi_-(\pi_- + \pi_+)^2\\ &= \pi_-\end{aligned}$$

We continue by eliminating FP and FN from the equation. For all $P \in \mathrm{dom}(\kappa) \cap \mathbb{P}^*$,

$$\begin{aligned}(\pi_-^2 + \pi_+^2)\kappa(P) + 2\pi_-\pi_+ &= \frac{\lambda_{tn}TN + \lambda_{fp}(\pi_- - TN) + \lambda_{fn}(\pi_+ - TP) + \lambda_{tp}TP}{\pi_+TN + \pi_-(\pi_- - TN) + \pi_+(\pi_+ - TP) + \pi_-TP}\\ &= \frac{(\lambda_{tn} - \lambda_{fp})TN + (\lambda_{fp}\pi_- + \lambda_{fn}\pi_+) + (\lambda_{tp} - \lambda_{fn})TP}{(\pi_+ - \pi_-)TN + (\pi_-^2 + \pi_+^2) + (\pi_- - \pi_+)TP}\end{aligned}$$

We have

$$\begin{aligned}\lambda_{tn} - \lambda_{fp} &= \pi_+ - \pi_+(\pi_- - \pi_+)\\ &= \pi_+(\pi_- + \pi_+) - \pi_+(\pi_- - \pi_+)\\ &= \pi_+\pi_- + \pi_+^2 - \pi_+\pi_- + \pi_+^2\\ &= 2\pi_+^2\end{aligned}$$

$$\begin{aligned}\lambda_{fp}\pi_- + \lambda_{fn}\pi_+ &= \pi_+(\pi_- - \pi_+)\pi_- + \pi_-(\pi_+ - \pi_-)\pi_+\\ &= \pi_-\pi_+(\pi_- - \pi_+ + \pi_+ - \pi_-)\\ &= 0\end{aligned}$$

$$\begin{aligned}\lambda_{tp} - \lambda_{fn} &= \pi_- - \pi_-(\pi_+ - \pi_-)\\ &= \pi_-(\pi_+ + \pi_-) - \pi_-(\pi_+ - \pi_-)\\ &= \pi_-\pi_+ + \pi_-^2 - \pi_-\pi_+ + \pi_-^2\\ &= 2\pi_-^2\end{aligned}$$

So, for all $P \in \mathrm{dom}(\kappa) \cap \mathbb{P}^*$,

$$(\pi_-^2 + \pi_+^2)\kappa(P) + 2\pi_-\pi_+ = \frac{2\pi_+^2TN + 2\pi_-^2TP}{(\pi_+ - \pi_-)TN + (\pi_-^2 + \pi_+^2) + (\pi_- - \pi_+)TP}$$

Let us now rework the denominator.

$$\begin{aligned}&(\pi_+ - \pi_-)TN + (\pi_-^2 + \pi_+^2) + (\pi_- - \pi_+)TP\\ =&(\pi_+ - \pi_-)(\pi_+ + \pi_-)TN + (\pi_-^2 + \pi_+^2) + (\pi_- - \pi_+)(\pi_- + \pi_+)TP\\ =&(\pi_+^2 - \pi_-^2)TN + (\pi_-^2 + \pi_+^2) + (\pi_-^2 - \pi_+^2)TP\\ =&(\pi_+^2 - \pi_-^2)TN + (\pi_-^2 + \pi_+^2)(\pi_- + \pi_+) + (\pi_-^2 - \pi_+^2)TP\\ =&(\pi_+^2 - \pi_-^2)TN + (\pi_-^2 + \pi_+^2)\pi_- + (\pi_-^2 + \pi_+^2)\pi_+ + (\pi_-^2 - \pi_+^2)TP\\ =&2\pi_+^2TN + (\pi_-^2 + \pi_+^2)(\pi_- - TN) + (\pi_-^2 + \pi_+^2)(\pi_+ - TP) + 2\pi_-^2TP\\ =&2\pi_+^2TN + (\pi_-^2 + \pi_+^2)FP + (\pi_-^2 + \pi_+^2)FN + 2\pi_-^2TP\end{aligned}$$

And thus, for all $P \in \mathrm{dom}(\kappa) \cap \mathbb{P}^*$,

$$\begin{aligned}(\pi_-^2+\pi_+^2)\kappa(P) + 2\pi_-\pi_+ =& \frac{2\pi_+^2 TN + 2\pi_-^2 TP}{2\pi_+^2 TN + (\pi_-^2+\pi_+^2)FP + (\pi_-^2+\pi_+^2)FN + 2\pi_-^2 TP} \\ =& \frac{\frac{\pi_+^2}{\pi_-^2+\pi_+^2}TN + \frac{\pi_-^2}{\pi_-^2+\pi_+^2}TP}{\frac{\pi_+^2}{\pi_-^2+\pi_+^2}TN + \frac{1}{2}FP + \frac{1}{2}FN + \frac{\pi_-^2}{\pi_-^2+\pi_+^2}TP} \\ =& \frac{I(tn)TN + I(tp)TP}{I(tn)TN + I(fp)FP + I(fn)FN + I(tp)TP} \\ =& R_I(P)\end{aligned}$$

In conclusion, $(\pi_-^2+\pi_+^2)\kappa + 2\pi_-\pi_+ = R_I$ on $\mathrm{dom}(\kappa)\cap\mathbb{P}^* = \mathrm{dom}(R_I)\cap\mathbb{P}^*$. Moreover, as $I(tn)+I(tp) = I(fp)+I(fn)$, R_I is canonical. □

A.3.4. Placing performance orderings induced by the score WA on the Tile

Lemma 9 (The weighted accuracies)**.** *In two-class classification, let the priors of the negative and positive classes be fixed, strictly positive, and denoted, respectively, by π_- and π_+. In this case, all weighted accuracies $WA = (1-\alpha)TNR + \alpha TPR$, $\alpha \in [0,1]$, are canonical ranking scores: $WA = R_I$ with*

$$\begin{cases} I(tn) = I(fp) = \frac{\frac{1-\alpha}{\pi_-}}{\frac{1-\alpha}{\pi_-}+\frac{\alpha}{\pi_+}} \\ I(fn) = I(tp) = \frac{\frac{\alpha}{\pi_+}}{\frac{1-\alpha}{\pi_-}+\frac{\alpha}{\pi_+}} \end{cases} .$$

Proof. Let us begin by introducing the set of performances with fixed and strictly positive priors:

$$\mathbb{P}^* = \left\{P \in \mathbb{P}_{(\Omega,\Sigma)} : P(\{fn, tp\}) = \pi_+\right\} \qquad \text{with } \pi_+ \in (0,1) .$$

For the sake of concision, we pose $TN = P(\{tn\})$, $FP = P(\{fp\})$, $FN = P(\{fn\})$, and $TP = P(\{tp\})$.

We first check the equality of the restricted domains. The domain of WA is $\mathrm{dom}(TNR) \cap \mathrm{dom}(TPR)$, that is

$$\mathrm{dom}(WA) = \left\{P \in \mathbb{P}_{(\Omega,\Sigma)} : P(\{fn, tp\}) \notin \{0,1\}\right\} ,$$

and thus the restricted domain of the weighted accuracy WA is

$$\mathrm{dom}(WA) \cap \mathbb{P}^* = \mathbb{P}^* .$$

For the domain of R_I, we have to take into account the fact that I should be well defined ($\pi_- \neq 0$ and $\pi_+ \neq 0$) and that the mathematical expectation of the importance should be non-zero, which is always the case:

$$\begin{aligned}\mathbf{E}_P[I] \neq 0 &\Leftrightarrow \frac{\frac{1-\alpha}{\pi_-}}{\frac{1-\alpha}{\pi_-}+\frac{\alpha}{\pi_+}}(TN+FP) + \frac{\frac{\alpha}{\pi_+}}{\frac{1-\alpha}{\pi_-}+\frac{\alpha}{\pi_+}}(FN+TP) \neq 0 \\ &\Leftrightarrow \frac{1-\alpha}{\pi_-}(TN+FP) + \frac{\alpha}{\pi_+}(FN+TP) \neq 0 \\ &\Leftrightarrow \frac{1-\alpha}{\pi_-}\pi_- + \frac{\alpha}{\pi_+}\pi_+ \neq 0 \\ &\Leftrightarrow 1 \neq 0\end{aligned}$$

Thus, the restricted domain of the ranking score R_I is

$$\mathrm{dom}(R_I) \cap \mathbb{P}^* = \mathbb{P}^* .$$

$$\mathrm{dom}(R_I) = \left\{P \in \mathbb{P}_{(\Omega,\Sigma)} : \mathbf{E}_P[I] \neq 0\right\}$$

The restricted domains are thus equal:

$$\mathrm{dom}(WA) \cap \mathbb{P}^* = \mathrm{dom}(R_I) \cap \mathbb{P}^* \,.$$

We now check the equality of the values taken by both scores. For all $P \in \mathrm{dom}(WA) \cap \mathbb{P}^*$, we have:

$$\begin{aligned}
WA(P) &= (1-\alpha)TNR(P) + \alpha TPR(P) \\
&= (1-\alpha)\frac{TN}{\pi_-} + \alpha\frac{TP}{\pi_+} \\
&= \frac{1-\alpha}{\pi_-}TN + \frac{\alpha}{\pi_+}TP \\
&= \frac{\frac{1-\alpha}{\pi_-}TN + \frac{\alpha}{\pi_+}TP}{\frac{1-\alpha}{\pi_-}\pi_- + \frac{\alpha}{\pi_+}\pi_+} \\
&= \frac{\frac{1-\alpha}{\pi_-}TN + \frac{\alpha}{\pi_+}TP}{\frac{1-\alpha}{\pi_-}(TN+FP) + \frac{\alpha}{\pi_+}(FN+TP)} \\
&= \frac{\frac{1-\alpha}{\pi_-}TN + \frac{\alpha}{\pi_+}TP}{\frac{1-\alpha}{\pi_-}TN + \frac{1-\alpha}{\pi_-}FP + \frac{\alpha}{\pi_+}FN + \frac{\alpha}{\pi_+}TP} \\
&= \frac{I(tn)TN + I(tp)TP}{I(tn)TN + I(fp)FP + I(fn)FN + I(tp)TP} \\
&= R_I(P)
\end{aligned}$$

In conclusion, $WA = R_I$ on $\mathrm{dom}(WA) \cap \mathbb{P}^* = \mathrm{dom}(R_I) \cap \mathbb{P}^*$. Moreover, as $I(tn) + I(tp) = I(fp) + I(fn)$, R_I is canonical. □

The previous lemma can be particularized for the balanced accuracy and for the accuracy. Taking $\alpha = \frac{1}{2}$, we obtain $BA = WA = R_I$ with $I(tn) = I(fp) = \pi_+$ and $I(fn) = I(tp) = \pi_-$. Taking $\alpha = \pi_+$, we obtain $A = WA = R_I$ with $I(tn) = I(fp) = 1/2$ and $I(fn) = I(tp) = 1/2$.

A.3.5. Placing performance orderings induced by some other non-probabilistic scores on the Tile

Lemma 10 (Other particular scores)**.** *Let us consider performance orderings induced by scores by the mechanism described in the 1st theorem of [2]. In two-class classification, Youden's index J_Y leads to the same performance ordering as the balanced accuracy BA, and Jaccard's coefficient for the positive class J_+ leads to the same performance ordering as F_1. Moreover, when the class priors are fixed and non-zero, the standardized negative predictive value $SNPV$ leads to the same performance ordering as the negative predictive value NPV when $P(\{tn, fn\}) \neq 0$, the negative likelihood ratio leads to the dual performance ordering of NPV, the standardized positive predictive value $SPPV$ leads to the same performance ordering as the positive predictive value PPV when $P(\{tp, fp\}) \neq 0$, and the positive likelihood ratio PLR leads also to the same performance ordering as PPV.*

Proof. For the sake of concision, let us pose $TN = P(\{tn\})$, $FP = P(\{fp\})$, $FN = P(\{fn\})$, and $TP = P(\{tp\})$.

- Youden's index is defined as $J_Y = TNR + TPR - 1$, and the balanced accuracy as $BA = 1/2 TNR + 1/2 TPR$. They have the same domain: $\mathrm{dom}(J_Y) = \mathrm{dom}(BA) = \mathrm{dom}(TNR) \cap \mathrm{dom}(TPR)$. Trivially,

$$J_Y = 2BA - 1 \,.$$

  Thus,

$$\frac{\partial J_Y}{\partial BA} = 2 > 0 \,.$$

  As there is a strictly increasing relationship between J_Y and BA, these two scores lead to same performance ordering.
- Jaccard's coefficient for the positive class is defined as $J_+ = X_{\{tp\}|\{fp,fn,tp\}}$ and the F-one score as $F_1 : P \mapsto \frac{2TP}{FP+FN+2TP}$. Their respective domains

$$\begin{aligned}
\mathrm{dom}(J_+) &= \left\{P \in \mathbb{P}_{(\Omega,\Sigma)} : P(\{fp, fn, tp\}) \neq 0\right\} \\
&= \left\{P \in \mathbb{P}_{(\Omega,\Sigma)} : FP \neq 0 \vee FN \neq 0 \vee TP \neq 0\right\}
\end{aligned}$$

and

$$\begin{aligned}\mathrm{dom}(F_1) &= \left\{P \in \mathbb{P}_{(\Omega,\Sigma)} : FP + FN + 2TP \neq 0\right\} \\ &= \left\{P \in \mathbb{P}_{(\Omega,\Sigma)} : FP \neq 0 \vee FN \neq 0 \vee TP \neq 0\right\}\end{aligned}$$

are equal. Trivially,

$$F_1 = \frac{2J_+}{1+J_+}\,.$$

Thus,

$$\frac{\partial F_1}{\partial J_+} = \frac{2}{(1+J_+)^2} > 0\,.$$

As there is a strictly increasing relationship between F_1 and J_+, these two scores lead to same performance ordering.

- The standardized negative predictive value is defined as $SNPV = \frac{TNR}{TNR+(1-TPR)}$ and the negative predictive value as $NPV = X_{\{tn\}|\{tn,fn\}}$. Let us consider the performances such that $TN + FN \neq 0$, $TN + FP = \pi_- \neq 0$, and $FN + TP = \pi_+ \neq 0$. We have:

$$\begin{aligned}SNPV = \frac{TNR}{TNR+(1-TPR)} &= \frac{\frac{TN}{\pi_-}}{\frac{TN}{\pi_-} + \frac{FN}{\pi_+}} \\ &= \frac{\frac{TN}{TN+FN}\frac{1}{\pi_-}}{\frac{TN}{TN+FN}\frac{1}{\pi_-} + \frac{FN}{TN+FN}\frac{1}{\pi_+}} = \frac{NPV\frac{1}{\pi_-}}{NPV\frac{1}{\pi_-} + (1-NPV)\frac{1}{\pi_+}}\,.\end{aligned}$$

Thus,

$$\frac{\partial SNPV}{\partial NPV} = \frac{\frac{1}{\pi_-}\frac{1}{\pi_+}}{\left(NPV\frac{1}{\pi_-} + (1-NPV)\frac{1}{\pi_+}\right)^2} > 0\,.$$

As there is a strictly increasing relationship between $SNPV$ and NPV, these two scores lead to same performance ordering.

- The negative likelihood ratio is defined as $NLR = \frac{1-TPR}{TNR}$ and the negative predictive value as $NPV = X_{\{tn\}|\{tn,fn\}}$. Let us consider the performances such that $TN \neq 0$, $TN + FP = \pi_-$, and $FN + TP = \pi_+ \neq 0$. We have:

$$NLR = \frac{1-TPR}{TNR} = \frac{\frac{FN}{\pi_+}}{\frac{TN}{\pi_-}} = \frac{FN}{TN}\frac{\pi_-}{\pi_+} = \frac{1-NPV}{NPV}\frac{\pi_-}{\pi_+}\,.$$

Thus,

$$\frac{\partial NLR}{\partial NPV} = -\frac{1}{NPV^2}\frac{\pi_-}{\pi_+} < 0\,.$$

As there is a strictly decreasing relationship between NLR and NPV, these two scores lead to dual performance orderings.

- The standardized positive predictive value is defined as $SPPV = \frac{TPR}{(1-TNR)+TPR}$ and the positive predictive value as $PPV = X_{\{tp\}|\{tp,fp\}}$. Let us consider the performances such that $TP + FP \neq 0$, $TN + FP = \pi_- \neq 0$, and $FN + TP = \pi_+ \neq 0$. We have:

$$\begin{aligned}SPPV = \frac{TPR}{(1-TNR)+TPR} &= \frac{\frac{TP}{\pi_+}}{\frac{FP}{\pi_-} + \frac{TP}{\pi_+}} \\ &= \frac{\frac{TP}{TP+FP}\frac{1}{\pi_+}}{\frac{FP}{TP+FP}\frac{1}{\pi_-} + \frac{TP}{TP+FP}\frac{1}{\pi_+}} = \frac{PPV\frac{1}{\pi_+}}{(1-PPV)\frac{1}{\pi_-} + PPV\frac{1}{\pi_+}}\end{aligned}$$

Thus,

$$\frac{\partial SPPV}{\partial PPV} = \frac{\frac{1}{\pi_-}\frac{1}{\pi_+}}{\left((1-PPV)\frac{1}{\pi_-} + PPV\frac{1}{\pi_+}\right)^2} > 0\,.$$

As there is a strictly increasing relationship between $SPPV$ and PPV, these two scores lead to same performance ordering.

- The positive likelihood ratio is defined as $PLR = \frac{TPR}{1-TNR}$ and the positive predictive value as $PPV = X_{\{tp\}|\{tp,fp\}}$. Let us consider the performances such that $FP \neq 0$, $TN + FP = \pi_-$, and $FN + TP = \pi_+ \neq 0$. We have:

$$PLR = \frac{TPR}{1-TNR} = \frac{\frac{TP}{\pi_+}}{\frac{FP}{\pi_-}} = \frac{TP}{FP}\frac{\pi_-}{\pi_+} = \frac{PPV}{1-PPV}\frac{\pi_-}{\pi_+}\,.$$

Thus,

$$\frac{\partial PLR}{\partial PPV} = \frac{1}{(1-PPV)^2}\frac{\pi_-}{\pi_+} > 0\,.$$

As there is a strictly increasing relationship between PLR and PPV, these two scores lead to same performance ordering. □

A.4. Supplementary material about Sec. 4.3

A.4.1. Algorithmic contributions

It is easy to calculate, for a given importance I, which entity ϵ is the best among a set $\mathbb{E}$, based on their performances: it is the entity that has the highest score value. By performing this calculation at numerous points on the Tile, it is possible to get a good idea of the set of importances for which a given entity is the best, occupies the r-th place, or is ranked last. However, it is very interesting to be able to give an explicit representation of this set, in an exact way, by computing it efficiently. We now present some algorithmic contributions that make this possible. We present them for the particular case in which all compared performances have the same class priors.

When $\pi_- = \pi_+$. Let us first consider the case in which two performances with balanced priors, P_1, P_2, are compared. In our toy example (Fig. 5), we made the observation that the regions where the different classifiers are ranked first are convex polygons when $\pi_- = \pi_+$. In fact, one can show that the canonical importances $I_{a,b}$ for which P_1 is better or equivalent than P_2 are given by a linear inequality in the variables a and b:

$$P_1 \gtrsim P_2 \Leftrightarrow \lambda_a(P_1, P_2)a + \lambda_b(P_1, P_2)b + \lambda_0(P_1, P_2) \geq 0 \tag{6}$$

with

$$\begin{cases} \lambda_a(P_1, P_2) = FPR(P_1)FNR(P_2) - FNR(P_1)FPR(P_2) \\ \lambda_b(P_1, P_2) = TPR(P_1)TNR(P_2) - TNR(P_1)TPR(P_2) \\ \lambda_0(P_1, P_2) = TNR(P_1) - TNR(P_2) \end{cases} . \tag{7}$$

Given that the Tile is bounded by four other linear inequalities in the variables a and b,

$$\begin{cases} (1)a + (0)b + (0) \geq 0\,. \\ (-1)a + (0)b + (1) \geq 0\,. \\ (0)a + (1)b + (0) \geq 0\,. \\ (0)a + (-1)b + (1) \geq 0\,. \end{cases} \tag{8}$$

the zone in the Tile where P_1 is better or equivalent than P_2 is the intersection between 5 half-planes, thus either an empty set or a convex polygon. This can be trivially generalized to the computation of the zone in which all performances in a set Π_1 are better or equivalent to all performances in a set Π_2. In this case, we obtain either an empty set or a convex polygon that is the intersection of $|\Pi_1||\Pi_2| + 4$ half planes. If needed, the representation of this polygon as a set of linear inequalities can be converted in an equivalent representation based on its vertices and edges.

When $\pi_- \neq \pi_+$. In this case, we proceed in three steps: (1) we apply a tartget/prior shift to the performances to fall back in the case of uniform priors, then (2) we calculate the polygon as described above, and finally (3) we cancel the effect of a tartget/prior shift by applying a deformation to this polygon according to what was described in Sec. A.2.1. This transformation is continuous and invertible, so that the border of the computed zone corresponds to the contour of the polygon. For that reason, it is convenient to represent the polygons based on their vertices and edges. Applying the transformation to them can be done by simply discretizing the edges and applying the transformation to the points. In this way, the resulting zone is approximated as a polygon.

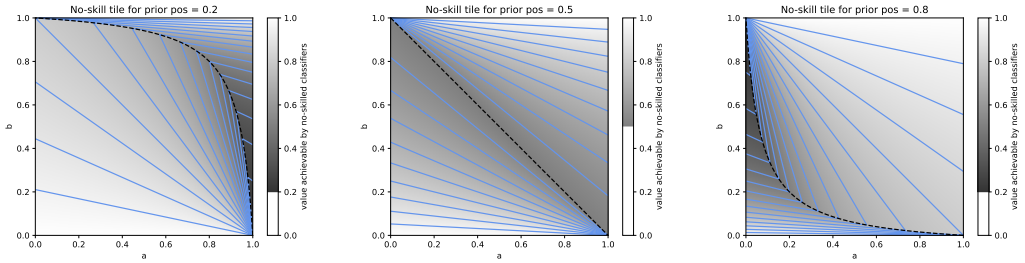


Figure 9. Tiles showing the values of canonical ranking scores achievable by no-skill classifiers when the prior of the positive class is 0.2 (left), 0.5 (center), and 0.8 (right). In all cases, the Tile is divided into two parts. On the bottom left part, the best no-skill classifier is the one predicting always the negative class. In the upper-right part, the best no-skill classifier is the one predicting always the positive class. Note that the limit between the two parts is the curve γ_π given in Fig. 6.

A.5. Supplementary material about Sec. 4.4

A.5.1. What is achievable by no-skilled classifiers?

Figure 9 uses the Tile to depict, for all canonical ranking scores, the maximum value achievable by the no-skill performances.

A.6. Averaging all canonical ranking scores: the *Volume Under Tile*

A.6.1. Definition

Let us study the score VUT (*Volume Under Tile*):

$$VUT : \mathbb{P}_{(\Omega,\Sigma)} \to [0,1] : P \mapsto VUT(P) = \int_{a=0}^{1} \int_{b=0}^{1} R_{I_{a,b}}(P)\, \mathrm{d}b\, \mathrm{d}a\,. \tag{9}$$

A.6.2. Closed-form expression

It can be showed that:

1. When $P(\{tp\}) = P(\{tn\})$ and $P(\{fn\}) = P(\{fp\})$:

$$\begin{aligned} VUT(P) &= R_{I_{a,b}}(P)\, \forall (a,b) \in [0,1] && (10)\\ &= A(P) = TNR(P) = TPR(P) && (11)\\ &= NPV(P) = PPV(P) = F_\beta(P) = \dots && (12) \end{aligned}$$

2. When $P(\{tp\}) = P(\{tn\})$ and $P(\{fn\}) \neq P(\{fp\})$:

$$VUT(P) = \frac{P(\{tn\})}{P(\{fn\}) - P(\{fp\})} \left(\ln\left(P(\{tn, fn\})\right) - \ln\left(P(\{tn, fp\})\right)\right) \tag{13}$$

3. When $P(\{tp\}) \neq P(\{tn\})$ and $P(\{fn\}) = P(\{fp\})$:

$$VUT(P) = 1 - \frac{P(\{fn\})}{P(\{tp\}) - P(\{tn\})} \left(\ln\left(P(\{tp, fn\})\right) - \ln\left(P(\{tn, fn\})\right)\right) \tag{14}$$

4. When $P(\{tp\}) \neq P(\{tn\})$ and $P(\{fn\}) \neq P(\{fp\})$:

$$\begin{aligned} VUT(P) = \frac{1}{2} - \frac{1}{2} \frac{\begin{array}{rl} & \left(P(\{tn\})^2 - P(\{fn\})^2\right) \ln\left(P(\{tn, fn\})\right)\\ + & \left(P(\{tp\})^2 - P(\{fp\})^2\right) \ln\left(P(\{tp, fp\})\right)\\ + & \left(P(\{fp\})^2 - P(\{tn\})^2\right) \ln\left(P(\{fp, tn\})\right)\\ + & \left(P(\{fn\})^2 - P(\{tp\})^2\right) \ln\left(P(\{fn, tp\})\right) \end{array}}{\left(P(\{tp\}) - P(\{tn\})\right)\left(P(\{fn\}) - P(\{fp\})\right)} \end{aligned} \tag{15}$$

Proof of Eq. (15). For the sake of concision, let us rewrite under volume under the Tile as the following double integral, where a, b, c, d are positive numbers:

$\int_{x=0}^{1} \int_{y=0}^{1} \frac{(1-x)a+xd}{(1-x)a+(1-y)b+yc+xd} dxdy$

$= \int_{x=0}^{1} \int_{y=0}^{1} \frac{(d-a)x+a}{(d-a)x+a+b+(c-b)y} dxdy$

Let's substitute: $p = (d-a), q = a, r = (d-a) = p, s = a+b+(c-b)y$

Knowing that $\int \frac{px+q}{rx+s} dx = \int f(x)dx = \frac{px}{r} + \frac{1}{r}\left(q - \frac{ps}{r}\right) \ln|rx+s| + C$, we have

$\int_0^1 f(x)dx = \frac{p}{r} + \frac{1}{r}\left(q - \frac{ps}{r}\right) \ln|r+s| - \frac{1}{r}\left(q - \frac{ps}{r}\right) \ln|s|$

As $p = r$, we can simplify the equation as:

$\int_0^1 f(x)dx = 1 + \left(\frac{q}{r} - \frac{s}{r}\right) \ln|r+s| - \left(\frac{q}{r} - \frac{s}{r}\right) \ln|s|$

$= 1 + \frac{a-a-b+(b-c)y}{d-a} \ln|d-a+a+b+(c-b)y| - \left(\frac{-b+(b-c)y}{d-a}\right) \ln|a+b+(c-b)y|$

$= 1 + \left[\frac{(b-c)}{d-a}y - \frac{b}{d-a}\right] \ln|b+d+(c-b)y| - \left[\frac{(b-c)}{d-a}y - \frac{b}{d-a}\right] \ln|a+b+(c-b)y|$

Let's now substitute: $\alpha = \frac{b-c}{d-a}, \beta = -\frac{b}{d-a}, \gamma = b+d, \delta = (c-b), \varepsilon = a+b$

$\to \int_0^1 1dy + \int_0^1 (\alpha y + \beta) \ln \underbrace{|\delta y + \gamma|}_{\geqslant 0} dy - \int_0^1 (\alpha y + \beta) \ln \underbrace{|\delta y + \varepsilon|}_{\geqslant 0} dy$

Knowing that $\int (ax+b)\ln(cx+d)dx = \int g(x)dx$

$= \frac{1}{4c^2}[2(cx+d)\ln(cx+d)(acx - ad + 2bc) - cx(acx - 2ad + 4bc)] + C\,,$

we have

$$\begin{aligned}\int_0^1 g(x)dx =& \frac{1}{4c^2}[2(c+d)\ln(c+d)(ac-ad+2bc)-c(ac-2ad+4bc)]\\ &-\frac{1}{4c^2}[2d\ln(d)(2bc-ac)]\end{aligned}$$

$$\begin{aligned}
&= 1+\frac{1}{4\delta^2}\left[2(\delta+\gamma)\ln(\delta+\gamma)(\alpha\delta-\alpha\gamma+2\beta\delta)\right.\\
&\quad\left.-\delta\left(\alpha\delta-2\alpha\gamma+4\beta\delta\right)-2\gamma\ln(2\beta\delta-\alpha\delta)\right]\\
&\quad-\frac{1}{4\delta^2}\left[2(\delta+\varepsilon)\ln(\delta+\varepsilon)(\alpha\delta-\alpha\varepsilon+2\beta\delta)\right.\\
&\quad\left.-\delta\left(\alpha\delta-2\alpha\varepsilon+4\beta\delta\right)-2\varepsilon\ln(2\beta\delta-\alpha\delta)\right]\\
&= 1+\frac{1}{4(c-b)^2(d-a)}\left[2(c+d)\ln(c+d)\left[-(b-c)^2-(b-c)(b+d)+2b(b-c)\right]\right.\\
&\left.-2(c+a)\ln(c+a)\left[-(b-c)^2-(b-c)(b+a)+2b(b-c)\right]\right.\\
&\left.-2(b+d)\ln(b+d)[2(c-b)(-b)+(c-b)(b+d)]\right.\\
&\left.+2(a+b)\ln(a+b)[2(c-b)(-b)+(c-b)(b+a)]\right.\\
&\left.-(c-b)\left[-(c-b)^2-2(b-c)(b+d)-4b(c-b)\right]\right.\\
&\left.+(c-b)\left[-(c-b)^2-2(b-c)(b+a)-4b(c-b)\right]\right]\\
&= 1+\frac{1}{4(c-b)(d-a)}[2(c+d)\ln(c+d)[(b-c)+(b+d)-2b]\\
&-2(c+a)\ln(c+a)[(b-c)+(b+a)-2b]\\
&-2(b+d)\ln(b+d)[d-b]\\
&+2(a+b)\ln(a+b)[a-b]\\
&+\left[(c-b)^2+2(b-c)(b+d)+4b(c-b)\right]\\
&-\left[(c-b)^2+2(b-c)(b+a)+4b(c-b)\right]
\end{aligned}$$

The volume under the Tile is thus analytically expressed as:

$$\begin{aligned}VUT = \frac{1}{2}&-\frac{1}{2(c-b)(d-a)}\left[\left(c^2-d^2\right)\ln(c+d)\right.\\
&+\left(a^2-c^2\right)\ln(a+c)\\
&+\left(d^2-b^2\right)\ln(b+d)\\
&\left.+\left(b^2-a^2\right)\ln(a+b)\right]\end{aligned}$$

□ □

A.6.3. Discussion: can we use it to rank?

The ordering induced by VUT does not satisfy Axiom 3. Nevertheless, it is interesting to note that it has a very high rank correlation with the accuracy A (Spearman's ρ is about 0.996 for a uniform performance distribution, *i.e.* a Dirichlet distribution with all concentration parameters set to $\alpha = 1$).

D. Supplementary material for the 5th article

This appendix contains the supplementary material provided with my 5th selected article [5], presented in Chapter 5.

A. Supplementary Material

We provide in this supplementary material (1) the list of symbols used in this paper (Section A.1), (2) a description of the software in the Python Jupyter Notebook (Sec. A.2), (3) a definition of the Relative-Skill Tile (Sec. A.3), (4) more details about the illustration used in the paper (Sec. A.4), (5) an analysis of the influence of a learning set (Sec. A.5), (6) an analysis of the behavior of scores, using the Correlation Tile, explaining the patterns observed in the various tiles presented in the paper (Sec. A.6), and finally (7) the comprehensive report generated by the Python Jupyter Notebook for our illustration (Sec. A.7).

A.1. List of Symbols

Symbols used for the probability theory

- Ω: the sample space (universe)
- Σ: the event space (a σ-algebra on Ω, *e.g.* 2^{Ω})
- (Ω, Σ): the measurable space

Symbols used for our probabilistic framework for performances

- P: a performance, *i.e.*, a probability measure
- P_ϵ: the performance of an entity ϵ
- $\mathbb{P}_{(\Omega,\Sigma)}$: all performances on (Ω, Σ)
- X: a score
- $\mathbb{X}_{(\Omega,\Sigma)}$: all scores on (Ω, Σ)
- $\mathrm{dom}(X)$: the domain of the score X

Symbols used for two-class classifications

- Y: the random variable for the ground truth
- $\hat{Y}$: the random variable for the prediction
- c_-: the negative class
- c_+: the positive class
- tn: the sample *true negative*
- fp: the sample *false positive*
- fn: the sample *false negative*
- tp: the sample *true positive*
- A: the score *accuracy*
- TNR: the score *true negative rate*
- TPR: the score *true positive rate*
- FPR: the score *false positive rate*
- NPV: the score *negative predictive value*
- PPV: the score *positive predictive value*
- F_β: the F-scores
- π_+: the score *prior of the positive class*
- π_-: the score *prior of the negative class*
- ROC: Receiver Operating Characteristic
- PR: Precision-Recall
- AUC-ROC: Area Under The ROC curve
- AUC-PR: Area Under The PR curve

Symbols used for the performance-based ranking of entities

- $\mathrm{rank}_{\mathbb{E}}$: the *ranking* function, relative to the set of entities $\mathbb{E}$
- $\mathbb{E}$: the set of entities to rank
- ϵ: an entity (*i.e.*, an element of $\mathbb{E}$)
- I: the random variable *Importance*
- R_I: the *ranking score* parameterized by the importance I
- $\lesssim_{R_I}$: the ordering induced by the ranking score R_I
- a: the parameter specifying the relative importance given to the incorrect outcomes (*i.e.*, tp and tn), it corresponds to the horizontal axis of the Tile
- b: the parameter specifying the relative importance given to the correct outcomes (*i.e.*, fn and fp), it corresponds to the vertical axis of the Tile

Symbols used for our illustration

- ♠: the learning set *Cityscapes*
- ♥: the learning set *ADE20K*
- ♦: the learning set *Pascal VOC 2012*
- ♣: the learning set *COCO-Stuff 164k*

Other symbols

- $\mathbb{R}$: the real numbers
- mIoU: mean intersection over union
- r: the linear correlation coefficient of Pearson
- ρ: the rank correlation coefficient of Spearman
- τ: the rank correlation coefficient of Kendall

A.2. Software Description

The *Python Jupyter Notebook* generates a comprehensive report related to the illustration of this paper, featuring all the tiles presented in the paper and more. It takes as input a list of two-class classification performances (*i.e.*, the values of $P(\{\omega\})$, for all $\omega \in \{tn, fp, fn, tp\}$). The various tiles are obtained by integrating two types of information. The first involves point-specific data, such as the value of a ranking score, a rank, an entity, or a correlation value, as detailed in this paper. This is achieved by discretizing the Tile using a grid size parameter along the two axis, a and b, and mapping the respective data, accordingly to the ranking scores in the Tile, to a point (a, b). The second type, relevant when priors are fixed (*i.e.*, when all compared performances have the same class priors), pertains to area-based information. The algorithm used to identify these areas of interest, described in [4], determines the areas (and their respective boundaries) within the Tile where an entity holds a specific rank (typically, the first or the last) without requiring a discretization of the Tile. The hatched areas on the various tiles provided in the paper (*e.g.*, the Value Tile or Ranking Tile) are drawn using this algorithm, as well as the white boundary lines on the Baseline Value Tile and the State-of-the-Art Value Tile.

A.2.1. Three Methods to Compute Value Tiles

In practice, several methods can be used to compute the *Value Tile*, depending on the availability of data.

1. *Direct computation:* If the values of $P(\{fp\})$, $P(\{fn\})$, $P(\{tp\})$, and $P(\{tn\})$ are known, the first method consists in computing the values of the ranking scores in each point of the Tile using Eq. (3).
2. *Interpolation:* Knowing the values taken by the scores TPR, TNR, NPV, and PPV at the four corners of the Tile, the values taken by the remaining scores can be determined by averaging, using f-means, vertically with $f : x \mapsto x^{-1}$ (*i.e.* the harmonic mean) and horizontally with $f : x \mapsto (1-x)^{-1}$ [4]. As one can see, the non-linearities differ between the vertical and horizontal axes.
3. *Equation system resolution:* The third method involves first, determining the values of $P(\{fp\})$, $P(\{fn\})$, $P(\{tp\})$, and $P(\{tn\})$ by solving a system of 4 equations with 4 unknowns, then, computing the values of the ranking scores using Eq. (3). Specifically, if the values of three ranking scores from the Value Tile are known, or if two ranking scores and the priors (for fixed priors) with $\pi_- = P(\{tn\}) + P(\{fp\})$ and $\pi_+ = P(\{fn\}) + P(\{tp\})$, then, knowing that $P(\{fp\}) + P(\{fn\}) + P(\{tp\}) + P(\{tn\}) = 1$ makes the system complete.

These three methods require discretizing the Tile. This can be implemented with a grid size parameter, which defines linearly spaced values for the two axes, a and b. The provided code implements the direct computation method.

A.3. Relative-Skill Tile

In this section, we define two new tiles, namely the No-Skill Tile and the Relative-Skill Tile. For this purpose, we first define two sets of performances.

We say that a performance P is *no-skilled* when the ground truth Y and the prediction $\hat{Y}$ are independent. We denote the set of all non-skilled performances on (Ω, Σ) by $\mathbb{P}^{Y \perp\!\!\!\perp \hat{Y}}_{(\Omega,\Sigma)}$:

$$\mathbb{P}^{Y \perp\!\!\!\perp \hat{Y}}_{(\Omega,\Sigma)} = \left\{ P \in \mathbb{P}_{(\Omega,\Sigma)} : P(Y,\hat{Y}) = P(Y)P(\hat{Y}) \right\} . \quad (10)$$

Therefore, the classifier always predicting the positive class c_+ and the classifier always predicting the negative class c_- have both no-skilled performances.

When the priors are the same for all compared performances, we can define the set of all performances at a given positive prior, π_+:

$$\mathbb{P}^{\pi_+}_{(\Omega,\Sigma)} = \left\{ P \in \mathbb{P}_{(\Omega,\Sigma)} : P(Y = c_+) = \pi_+ \right\} . \quad (11)$$

Definition 8. *The* No-Skill Tile *is the mapping*

$$noskill_I : [0,1]^2 \to [0,1] : (a,b) \mapsto \max_{P \in \mathbb{P}^{Y \perp\!\!\!\perp \hat{Y}}_{(\Omega,\Sigma)} \cap \mathbb{P}^{\pi_+}_{(\Omega,\Sigma)}} R_I(P) .$$

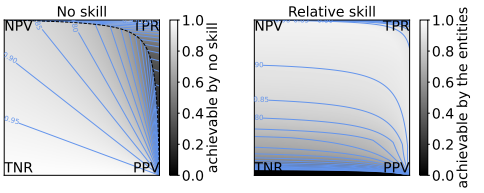


Figure 9. **No-Skill Tile (left) and Relative-Skill Tile (right).** These tiles show, on the left, the score values of the best no-skilled performances and, on the right, the relative performance of the state of the art compared to no-skilled performances for the 74 semantic segmentation models. The blue lines are iso-value lines, *i.e.*, lines on which all scores have the same value. The dashed line indicates boundaries between different entities.

The No-Skill Tile thus displays the value of the ranking scores in each point of the tile for the best-performing non-skilled performance. Note that, *e.g.*, the hatched areas on Fig. 5 highlight regions where the No-Skill Tile exhibits a higher value than the Value Tile.

Definition 9. *The* Relative-Skill Tile *is the mapping*

$$skill_I : [0,1]^2 \to [0,1] : (a,b) \mapsto \frac{SOTA_I - noskill_I}{1 - noskill_I} ,$$

where $SOTA_I$ corresponds to the values of the State-of-the-Art Value Tile.

Figure 9 illustrates the No-Skill Tile and the Relative-Skill Tile. Note that the code to obtain these tiles is also available in the Python Jupyter Notebook.

A.4. More Details on the Illustration

Our illustration is for an arbitrarily chosen pixel-based two-class classification task derived from a pixel-based semantic segmentation task. The performances are those of 74 models evaluated on 8,000 images of the BDD100K dataset [148]: this is our testing set. These models have been trained on the ♠ Cityscapes [109], ♥ ADE20K [110], ♦ Pascal VOC 2012 [111], or ♣ COCO-Stuff 164k [112] datasets: these are the learning sets. The set of semantic labels in BDD100K are identical to those in ♠ Cityscapes, but differ from the ones in ♥ ADE20K, ♦ Pascal VOC 2012, and ♣ COCO-Stuff 164k.

Defining the two classes for the testing set. We arbitrarily took the union of the first 11 semantic labels of BDD100K as the negative class, corresponding to background objects (*e.g.*, road, sidewalk, or building), and the union of the 8 remaining ones as the positive class (person, rider, car, truck, bus, train, motorcycle, and bicycle).

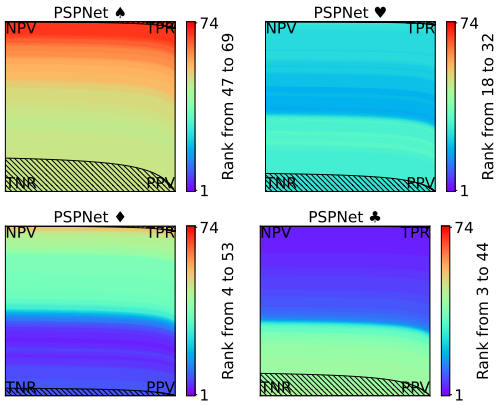


Figure 10. **Ranking Tiles showing the rank of PSPNet trained on different learning datasets (from left to right: ♠ Cityscapes, ♥ ADE20K, ♦ Pascal VOC 2012, and ♣ COCO-Stuff 164k).** We observe that the learning dataset strongly influences the ranking behavior.

Defining the two classes for the 4 learning sets. The semantic labels predicted by the models are those from the corresponding learning set. We therefore also had to map all these labels to our negative and positive classes. To this end, for each learning set, taking into account all the models learned on this learning set, we computed the proportion of pixels that are positive for each semantic label. The semantic label has been attributed to the positive class when this proportion is greater than the positive prior π_+ of BDD100K (about 0.1242).

All the results presented in our paper, and also hereafter, are specific to these arbitrary choices. In particular, we noticed that thresholding the posteriors at 0.5 instead of π_+ leads to a significantly different two-class classification problem, for which the ranking of the models is different.

A.5. Influence of a Learning Set

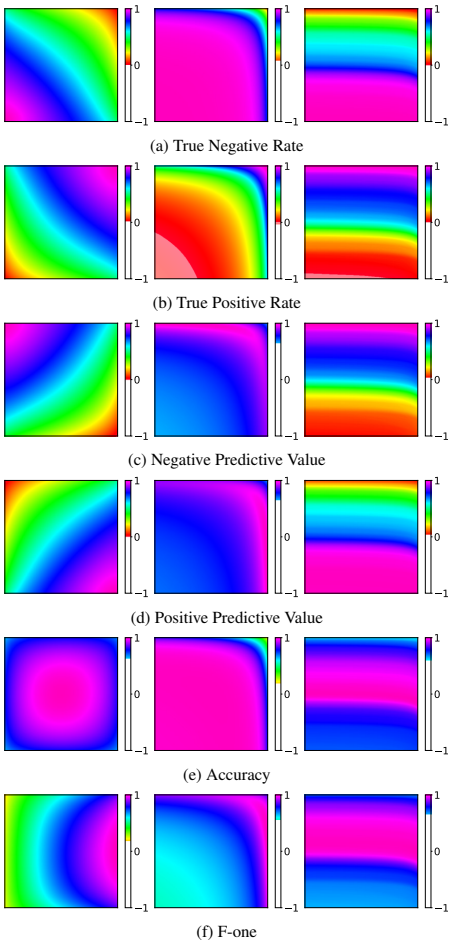
The influence of a learning set on the ranking of a semantic segmentation model can be seen by placing the Ranking Tiles corresponding to that model trained on the different learning sets side by side. As shown in Fig. 10, training PSPNet on ♦ Pascal VOC 2012 and ♣ COCO-Stuff 164k performs better than training it on ♠ Cityscapes. But the best stability over the tile is achieved by training PSPNet on ♥ ADE20K.

A.6. Behavior of Scores

The Correlation Tile allows to depict the behavior of any score, showing the rank correlations between that score and all canonical ranking scores, for a given performance distribution. On Fig. 11, we analyze the rank correlation, using Spearman correlation coefficient (Spearman's ρ), for 6 scores belonging to the ranking scores (namely TNR, TPR, NPV, PPV, A, and F_1) and 3 distributions of performances (all 3 distributions are uniform but are on different sets of performances). We analyze these scores (1) for a uniform distribution over all performances, (2) for a uniform distribution of performances having fixed priors corresponding to those of BDD100K, and (3) for a uniform distribution of performances, where the set of performances is the one of our illustration. Comparing the figures on the left with those on the center, we see the effect of having fixed priors with $\pi_+ = 0.1242$. The figures on the right show the effect of having fixed priors and an arbitrary choice of performances (the ones of the 74 state-of-the-art semantic segmentation models). For the explanation of the quasi-horizontal bands, see Tab. 2.

	Set 1	Set 2	Set 3
Num. perfs. (size)	1000	1000	74
$\min(\theta)$	0.000628	0.0000634	0.0971
$\max(\theta)$	1.57	1.56	0.170
$\mu(\theta)$	0.806	0.242	0.125
$\sigma(\theta)$	0.496	0.281	0.00997

Table 2. **Comparison of the three sets of two-class classification performances used in Fig. 11:** (1) a uniform distribution over all possible performances, (2) over all performances with a prior of the positive class equal to the one in our illustration ($\pi_+ = 0.124227$), and (3) over the 74 performances compared in our illustration. The score θ is given by $\theta = \arctan(\frac{P(\{tp\})}{P(\{tn\})})$. The set for which the standard deviation $\sigma(\theta)$ is minimal is the one for which $P(\{tn\})$ is the closer to be proportional to $P(\{tp\})$. It is the third one. This is why one can observe quasi horizontal bands in the Correlation Tiles on the right hand side of Fig. 11.



(a) True Negative Rate

(b) True Positive Rate

(c) Negative Predictive Value

(d) Positive Predictive Value

(e) Accuracy

(f) F-one

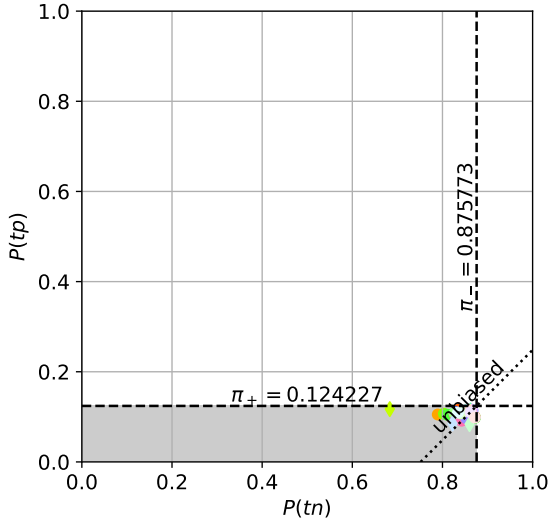
Figure 11. **Behavior of** 6 **scores for** 3 **sets of performances.** The Correlation Tile shows the estimated rank correlations (with the Spearman's ρ) between 6 scores and all canonical ranking scores, for a uniform distribution over all possible performances (left), over all performances with a prior of the positive class equal to the one in our illustration ($\pi_+ = 0.124227$) (center), and over the 74 performances compared in our illustration (right).

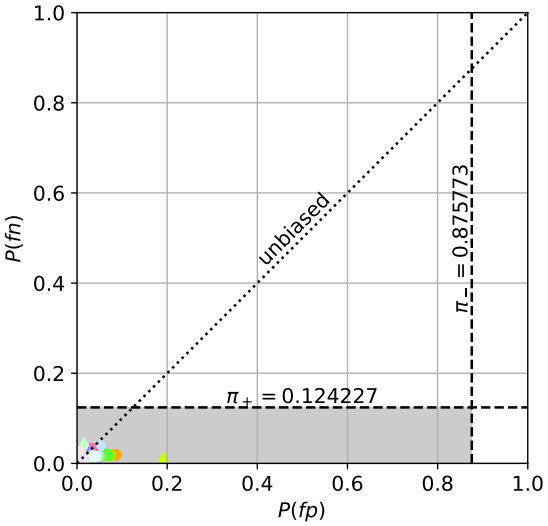
A.7. Report Generated by the Python Jupyter Notebook for our Illustration

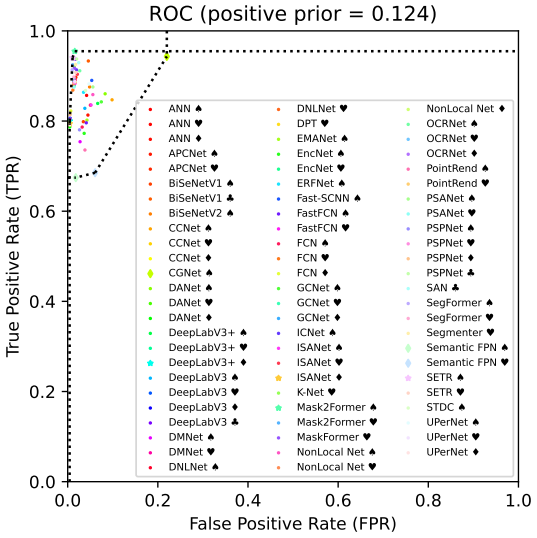
The performances of the two-class classification entities

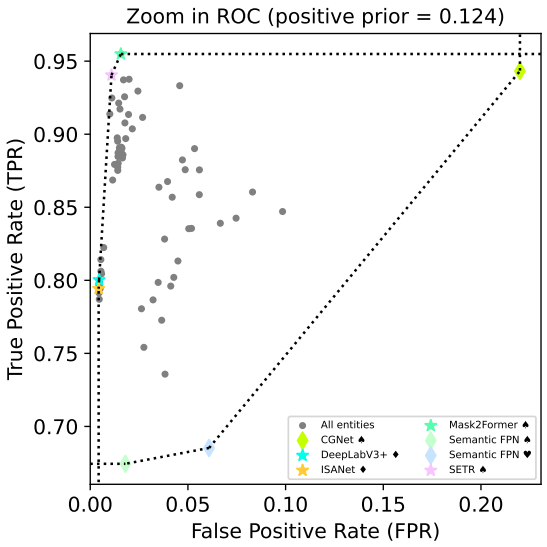
entity	$P(\{tn\})$	$P(\{fp\})$	$P(\{fn\})$	$P(\{tp\})$
ANN ♠	0.8390	0.0368	0.0178	0.1064
ANN ♥	0.8615	0.0142	0.0136	0.1106
ANN ♦	0.8713	0.0045	0.0255	0.0987
APCNet ♠	0.8365	0.0393	0.0232	0.1010
APCNet ♥	0.8599	0.0159	0.0128	0.1114
BiSeNetV1 ♠	0.8453	0.0304	0.0250	0.0992
BiSeNetV1 ♣	0.8357	0.0401	0.0083	0.1159
BiSeNetV2 ♠	0.8332	0.0426	0.0154	0.1088
CCNet ♠	0.7896	0.0862	0.0190	0.1052
CCNet ♥	0.8621	0.0136	0.0142	0.1101
CCNet ♦	0.8719	0.0039	0.0253	0.0990
CGNet ♠	0.6831	0.1926	0.0070	0.1172
DANet ♠	0.8030	0.0728	0.0173	0.1069
DANet ♥	0.8626	0.0131	0.0136	0.1107
DANet ♦	0.8717	0.0040	0.0265	0.0978
DeepLabV3+ ♠	0.8175	0.0583	0.0200	0.1042
DeepLabV3+ ♥	0.8627	0.0131	0.0136	0.1106
DeepLabV3+ ♦	0.8718	0.0040	0.0248	0.0994
DeepLabV3 ♠	0.8424	0.0333	0.0213	0.1029
DeepLabV3 ♥	0.8633	0.0125	0.0140	0.1102
DeepLabV3 ♦	0.8710	0.0048	0.0241	0.1002
DeepLabV3 ♣	0.8609	0.0149	0.0078	0.1164
DMNet ♠	0.8517	0.0240	0.0305	0.0937
DMNet ♥	0.8618	0.0140	0.0135	0.1107
DNLNet ♠	0.8316	0.0441	0.0205	0.1038
DNLNet ♥	0.8633	0.0124	0.0143	0.1099
DPT ♥	0.8635	0.0123	0.0150	0.1092
EMANet ♠	0.8383	0.0375	0.0246	0.0996
EncNet ♠	0.8437	0.0320	0.0282	0.0960
EncNet ♥	0.8633	0.0125	0.0149	0.1093
ERFNet ♠	0.8450	0.0308	0.0169	0.1073
Fast-SCNN ♠	0.8290	0.0468	0.0136	0.1106
FastFCN ♠	0.8397	0.0361	0.0253	0.0989
FastFCN ♥	0.8613	0.0144	0.0144	0.1098
FCN ♠	0.8476	0.0282	0.0265	0.0977
FCN ♥	0.8657	0.0100	0.0163	0.1079
FCN ♦	0.8697	0.0061	0.0221	0.1022
GCNet ♠	0.8104	0.0654	0.0196	0.1047
GCNet ♥	0.8635	0.0123	0.0152	0.1090
GCNet ♦	0.8715	0.0043	0.0253	0.0989
ICNet ♠	0.8586	0.0172	0.0107	0.1135
ISANet ♠	0.8304	0.0454	0.0204	0.1038
ISANet ♥	0.8569	0.0189	0.0120	0.1123
ISANet ♦	0.8720	0.0038	0.0256	0.0987
K-Net ♥	0.8630	0.0127	0.0098	0.1144
Mask2Former ♠	0.8621	0.0137	0.0056	0.1186
Mask2Former ♥	0.8661	0.0097	0.0093	0.1149
MaskFormer ♥	0.8624	0.0133	0.0103	0.1139

NonLocal Net ♠	0.8269	0.0489	0.0176	0.1067
NonLocal Net ♥	0.8649	0.0109	0.0150	0.1092
NonLocal Net ♦	0.8716	0.0042	0.0250	0.0992
OCRNet ♠	0.8523	0.0235	0.0110	0.1132
OCRNet ♥	0.8635	0.0123	0.0130	0.1112
OCRNet ♦	0.8707	0.0051	0.0243	0.0999
PointRend ♠	0.8422	0.0335	0.0328	0.0914
PointRend ♥	0.8611	0.0147	0.0142	0.1101
PSANet ♠	0.8268	0.0489	0.0155	0.1088
PSANet ♥	0.8634	0.0123	0.0155	0.1087
PSPNet ♠	0.8528	0.0230	0.0273	0.0970
PSPNet ♥	0.8629	0.0129	0.0142	0.1100
PSPNet ♦	0.8710	0.0048	0.0231	0.1011
PSPNet ♣	0.8584	0.0174	0.0078	0.1165
SAN ♣	0.8544	0.0214	0.0088	0.1155
SegFormer ♠	0.8670	0.0088	0.0107	0.1135
SegFormer ♥	0.8637	0.0121	0.0127	0.1115
Segmenter ♥	0.8604	0.0154	0.0092	0.1150
Semantic FPN ♠	0.8600	0.0158	0.0404	0.0838
Semantic FPN ♥	0.8225	0.0533	0.0391	0.0851
SETR ♠	0.8663	0.0095	0.0074	0.1168
SETR ♥	0.8602	0.0155	0.0115	0.1128
STDC ♠	0.8411	0.0347	0.0164	0.1078
UPerNet ♠	0.8344	0.0413	0.0146	0.1096
UPerNet ♥	0.8619	0.0138	0.0139	0.1103
UPerNet ♦	0.8718	0.0039	0.0259	0.0983

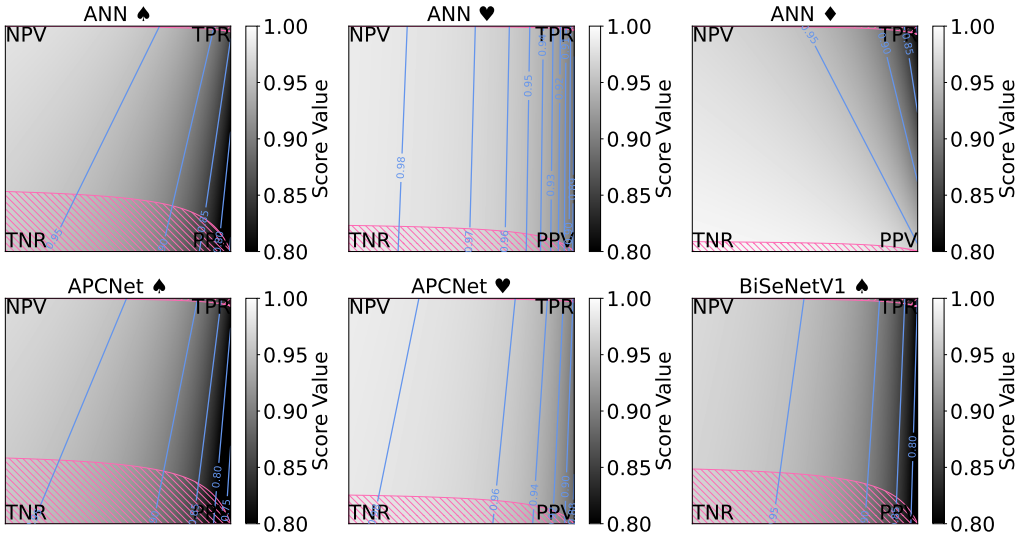


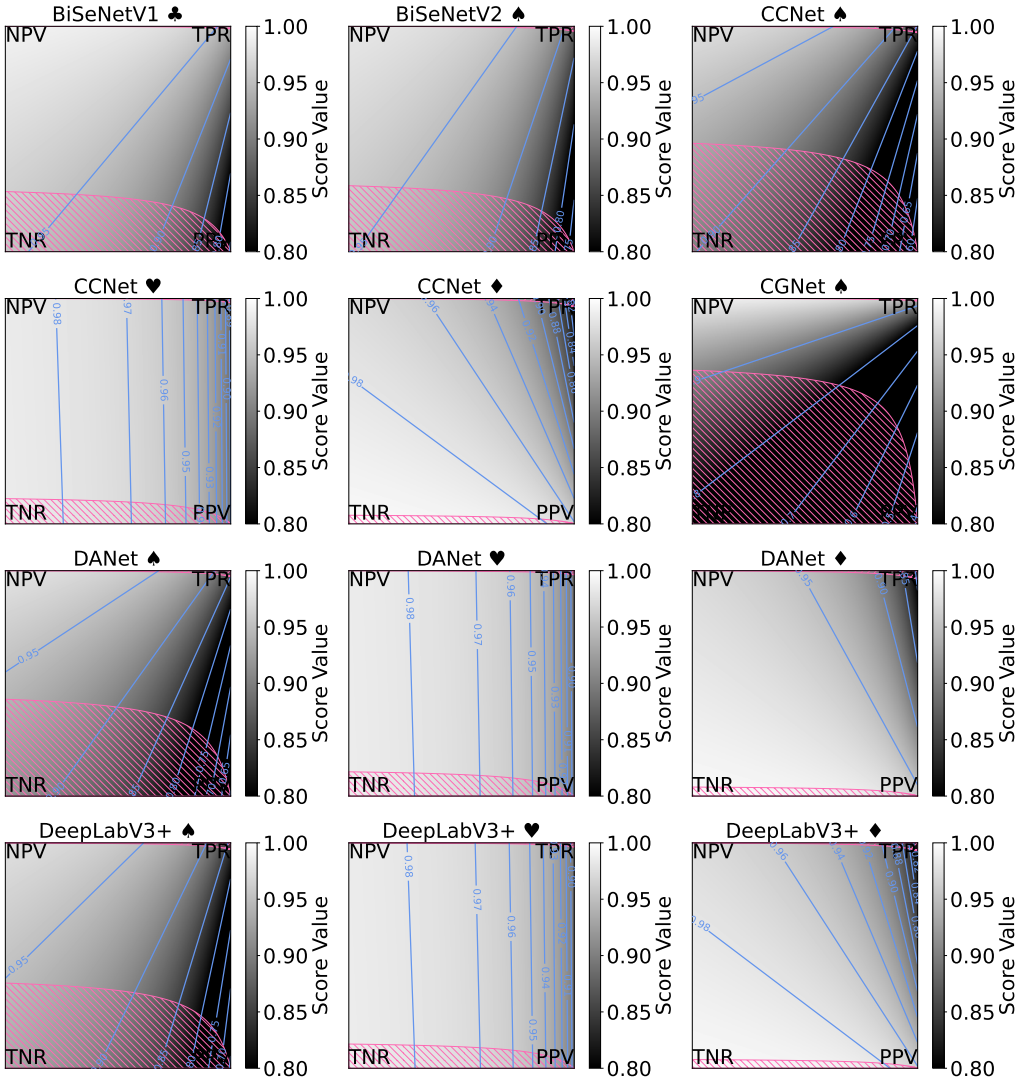

1.0
0.8
0.6
0.4
0.2
0.0
P(fn)
unbiased
$\pi_- = 0.875773$
$\pi_+ = 0.124227$
0.0
0.2
0.4
0.6
0.8
1.0
P(fp)

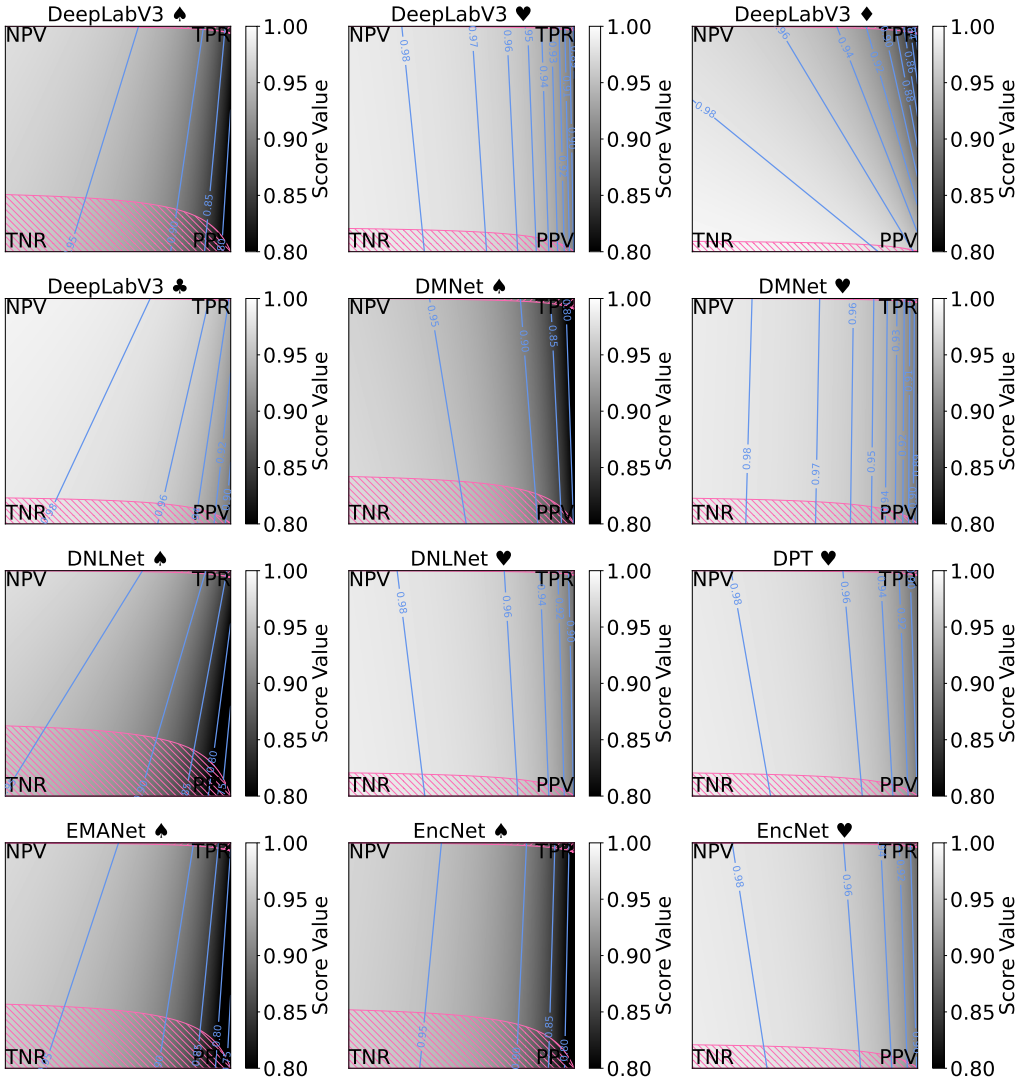

ROC (positive prior = 0.124)
True Positive Rate (TPR)
False Positive Rate (FPR)
ANN ♠
ANN ♥
ANN ♦
APCNet ♠
APCNet ♥
BiSeNetV1 ♠
BiSeNetV1 ♣
BiSeNetV2 ♠
CCNet ♠
CCNet ♥
CCNet ♦
CGNet ♠
DANet ♠
DANet ♥
DANet ♦
DeepLabV3+ ♠
DeepLabV3+ ♥
DeepLabV3+ ♦
DeepLabV3 ♠
DeepLabV3 ♥
DeepLabV3 ♦
DeepLabV3 ♣
DMNet ♠
DMNet ♥
DNLNet ♠
DNLNet ♥
DPT ♥
EMANet ♠
EncNet ♠
EncNet ♥
ERFNet ♠
Fast-SCNN ♠
FastFCN ♠
FastFCN ♥
FCN ♠
FCN ♥
FCN ♦
GCNet ♠
GCNet ♥
GCNet ♦
ICNet ♠
ISANet ♠
ISANet ♥
ISANet ♦
K-Net ♥
Mask2Former ♠
Mask2Former ♥
MaskFormer ♥
NonLocal Net ♠
NonLocal Net ♥
NonLocal Net ♦
OCRNet ♠
OCRNet ♥
OCRNet ♦
PointRend ♠
PointRend ♥
PSANet ♠
PSANet ♥
PSPNet ♠
PSPNet ♥
PSPNet ♦
PSPNet ♣
SAN ♣
SegFormer ♠
SegFormer ♥
Segmenter ♥
Semantic FPN ♠
Semantic FPN ♥
SETR ♠
SETR ♥
STDC ♠
UPerNet ♠
UPerNet ♥
UPerNet ♦

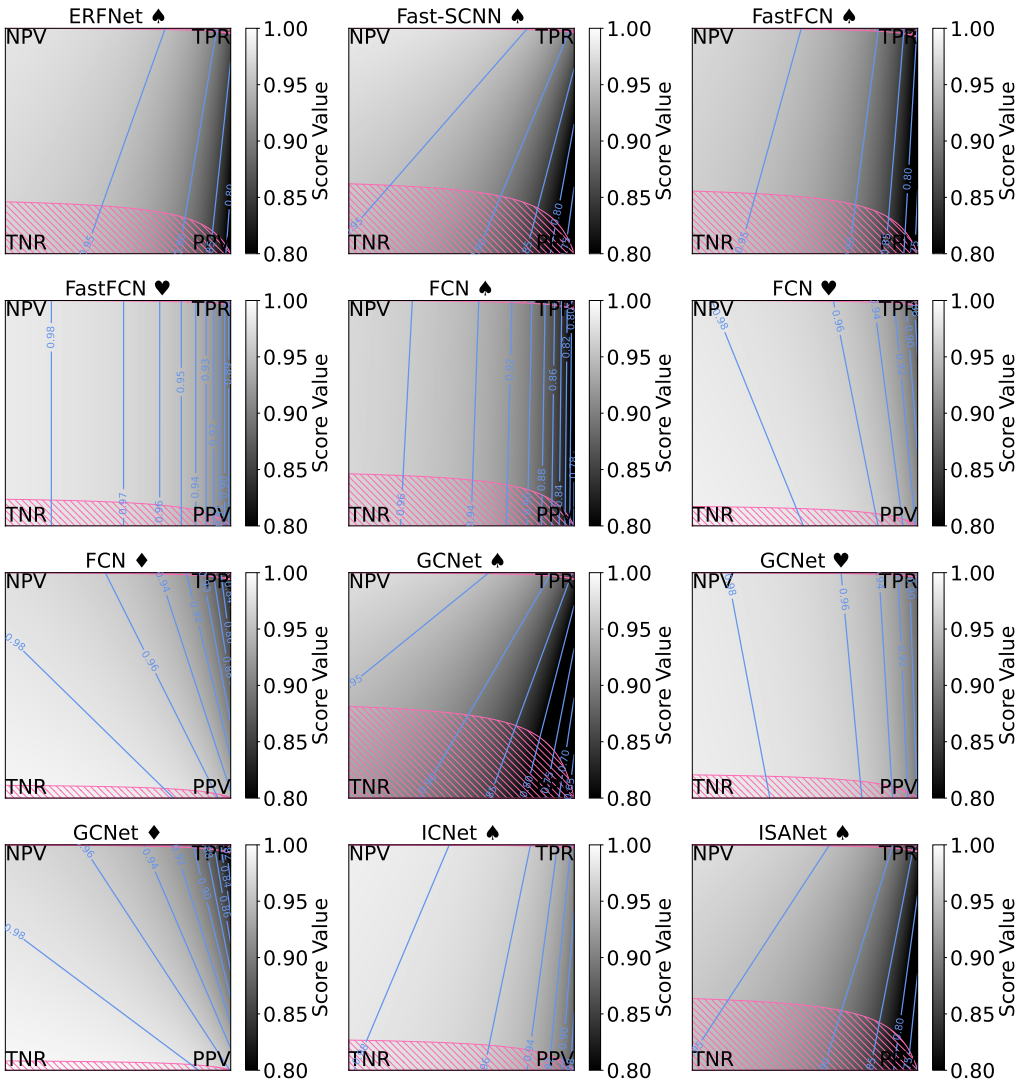


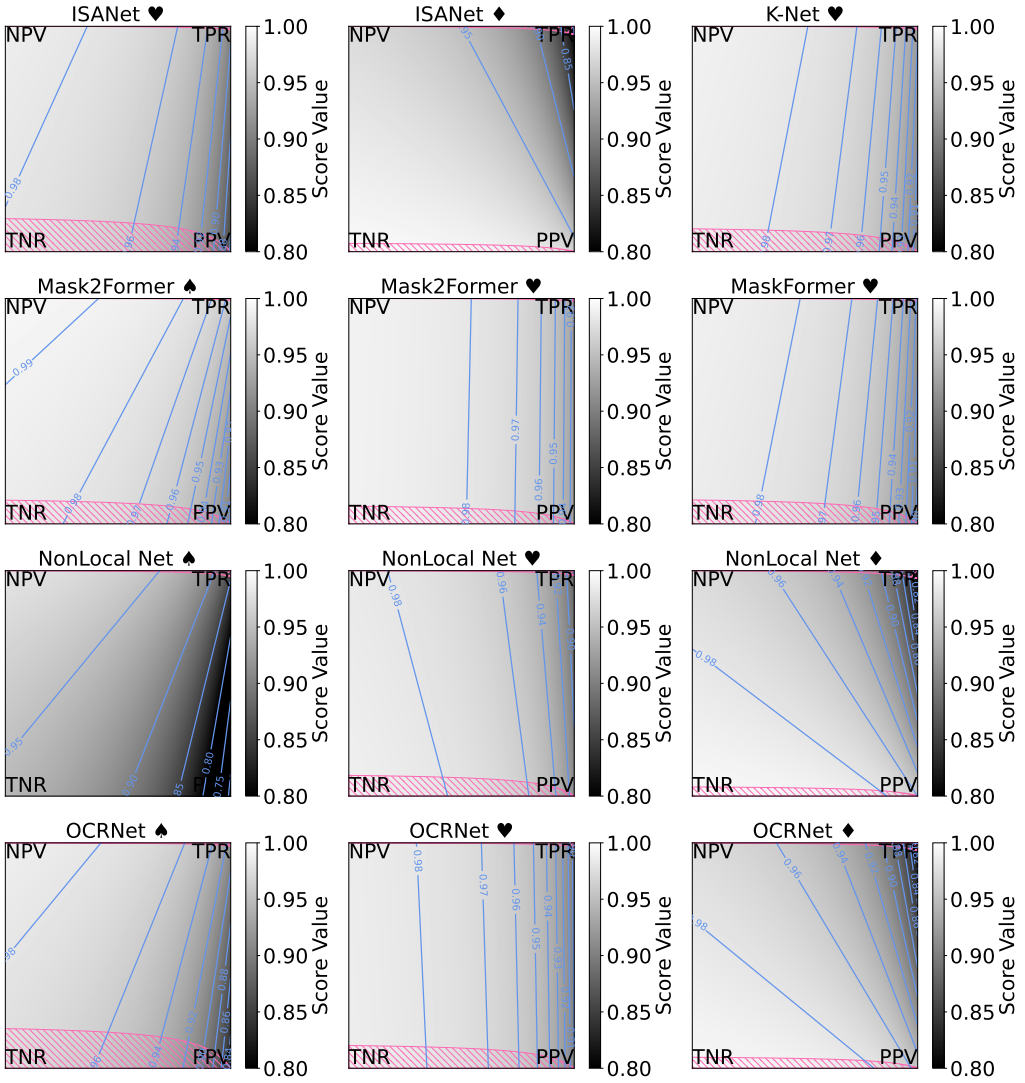
Value Tile: Using the tile to show the canonical ranking scores values for each entity

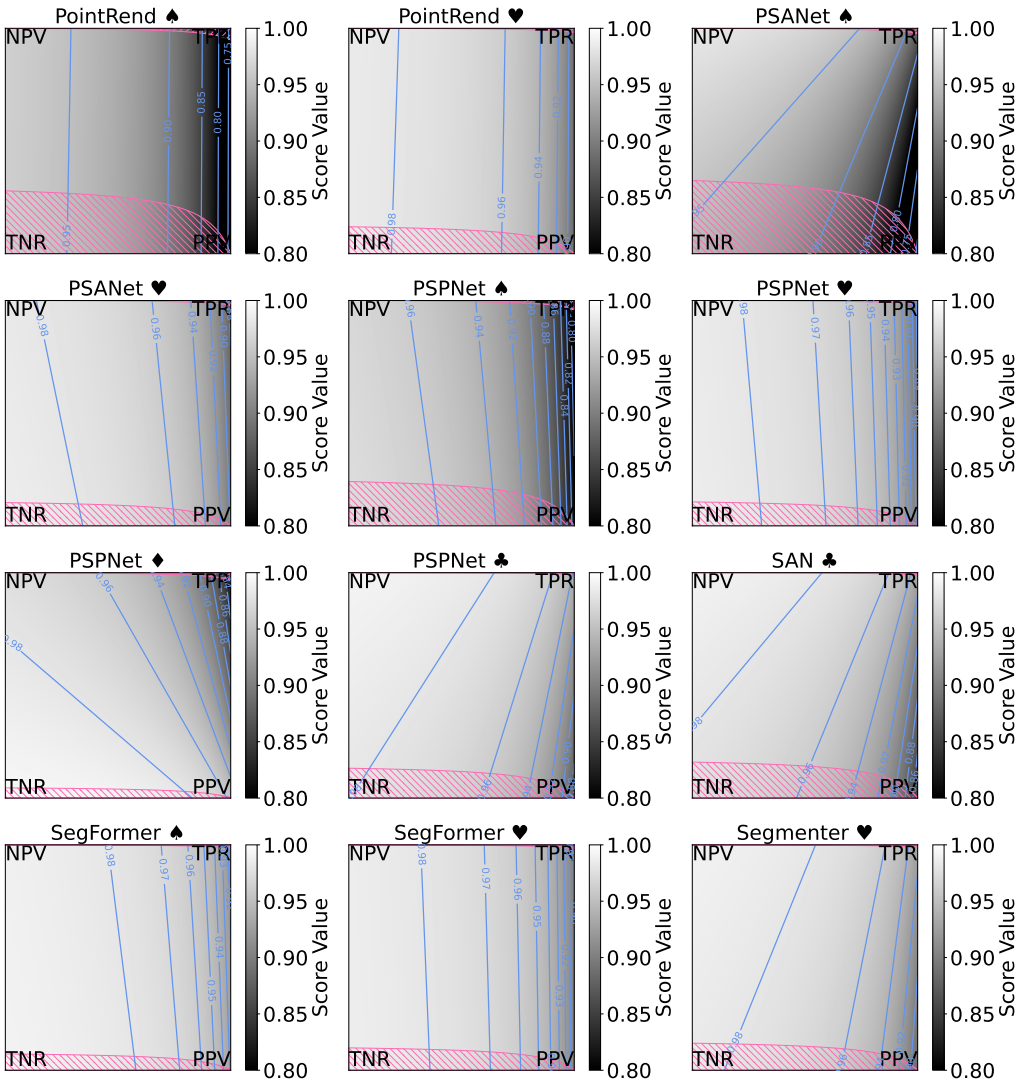


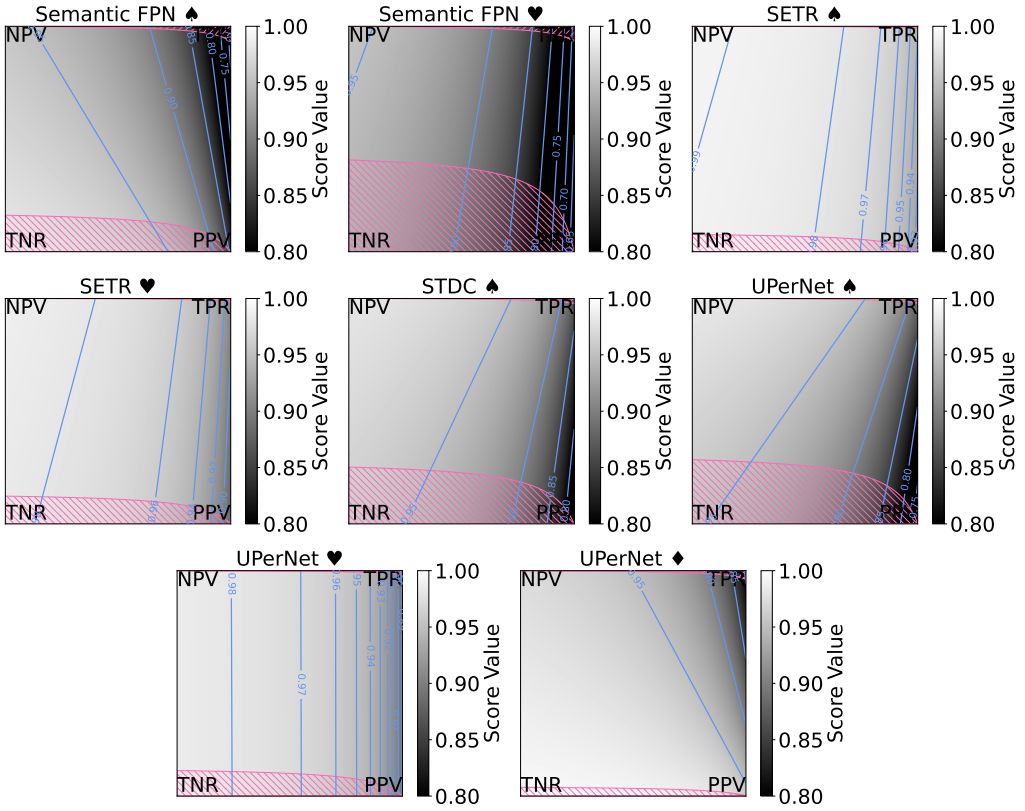

BiSeNetV1 ♣
BiSeNetV2 ♠
CCNet ♠
CCNet ♥
CCNet ♦
CGNet ♠
DANet ♠
DANet ♥
DANet ♦
DeepLabV3+ ♠
DeepLabV3+ ♥
DeepLabV3+ ♦
NPV
TPR
TNR
PPV
Score Value
1.00
0.95
0.90
0.85
0.80


DeepLabV3 ♠
DeepLabV3 ♥
DeepLabV3 ♦
DeepLabV3 ♣
DMNet ♠
DMNet ♥
DNLNet ♠
DNLNet ♥
DPT ♥
EMANet ♠
EncNet ♠
EncNet ♥
NPV
TPR
TNR
PPV
Score Value
1.00
0.95
0.90
0.85
0.80

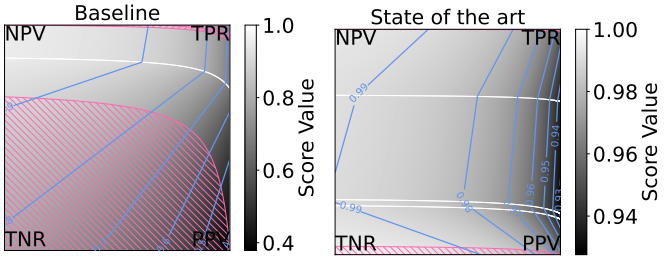

ERFNet ♠
Fast-SCNN ♠
FastFCN ♠
FastFCN ♥
FCN ♠
FCN ♥
FCN ♦
GCNet ♠
GCNet ♥
GCNet ♦
ICNet ♠
ISANet ♠
NPV
TPR
TNR
PPV
Score Value
1.00
0.95
0.90
0.85
0.80


ISANet ♥
ISANet ♦
K-Net ♥
Mask2Former ♠
Mask2Former ♥
MaskFormer ♥
NonLocal Net ♠
NonLocal Net ♥
NonLocal Net ♦
OCRNet ♠
OCRNet ♥
OCRNet ♦
NPV
TPR
TNR
PPV
Score Value
1.00
0.95
0.90
0.85
0.80

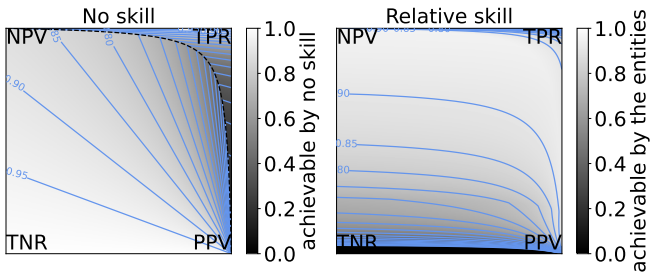

PointRend ♠
PointRend ♥
PSANet ♠
PSANet ♥
PSPNet ♠
PSPNet ♥
PSPNet ♦
PSPNet ♣
SAN ♣
SegFormer ♠
SegFormer ♥
Segmenter ♥
NPV
TPR
TNR
PPV
Score Value
1.00
0.95
0.90
0.85
0.80



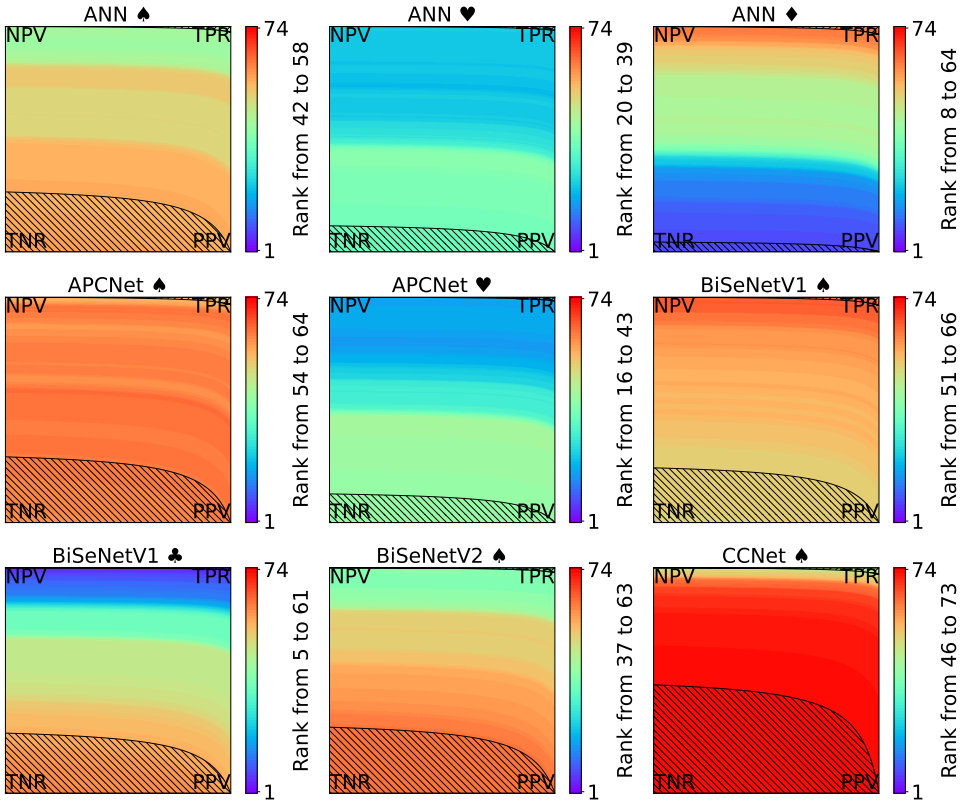
Baseline Value Tile and State-of-the-Art Value Tile: Using the tile to show the 'Baseline' and 'State of the Art'

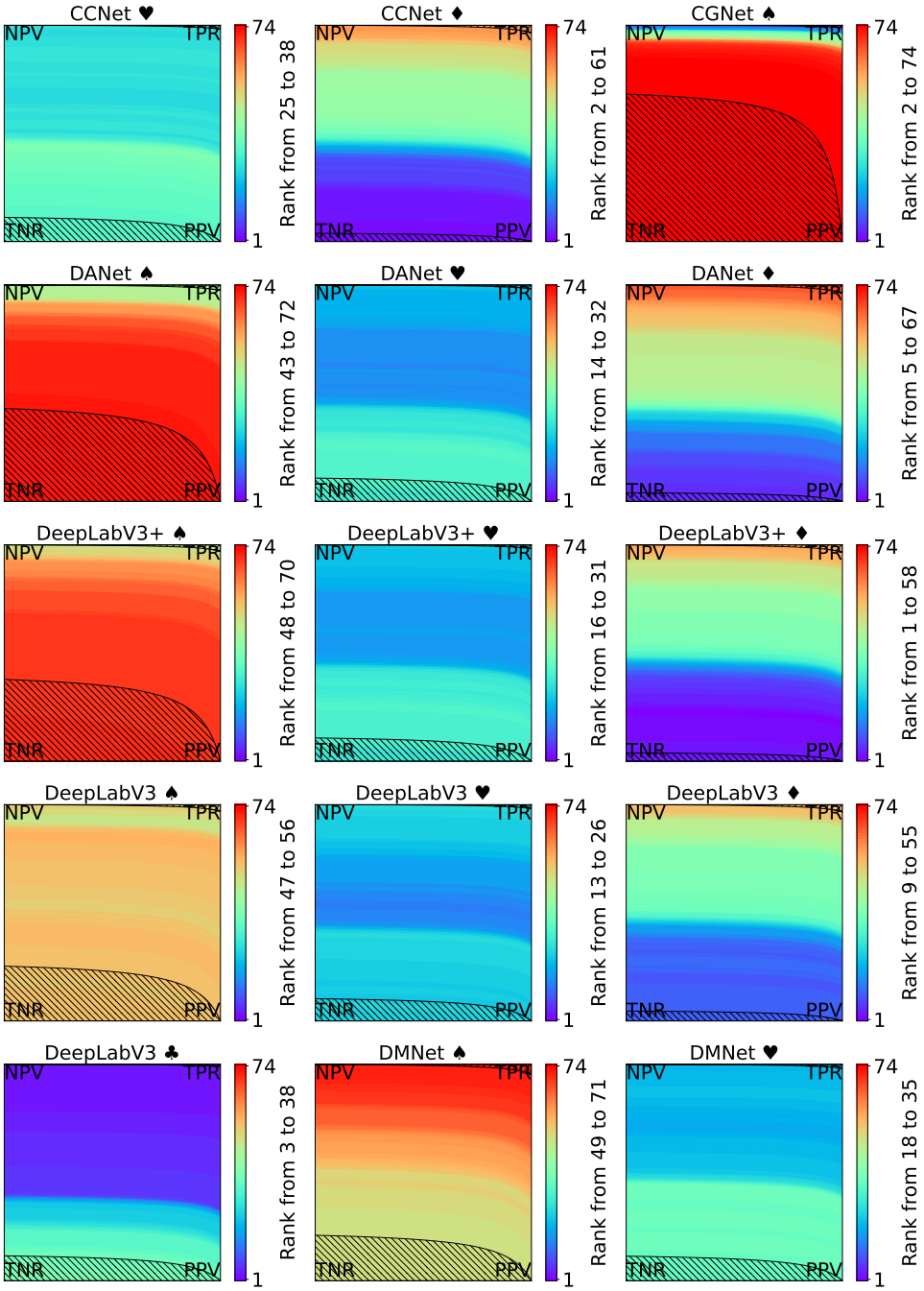


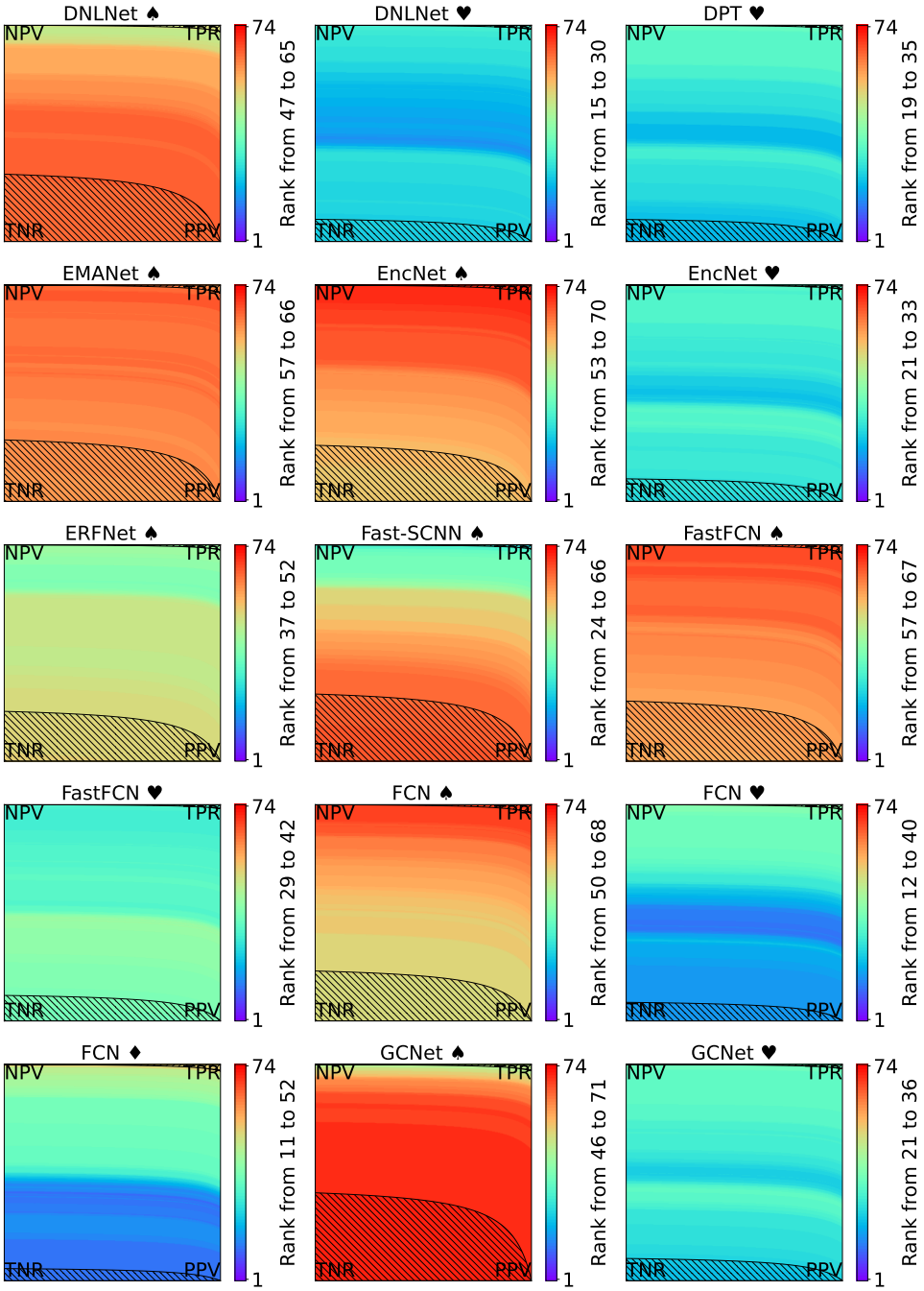
No-Skill Tile and Relative-Skill Tile: Using the tile to show the no-skill values and the relative skill values

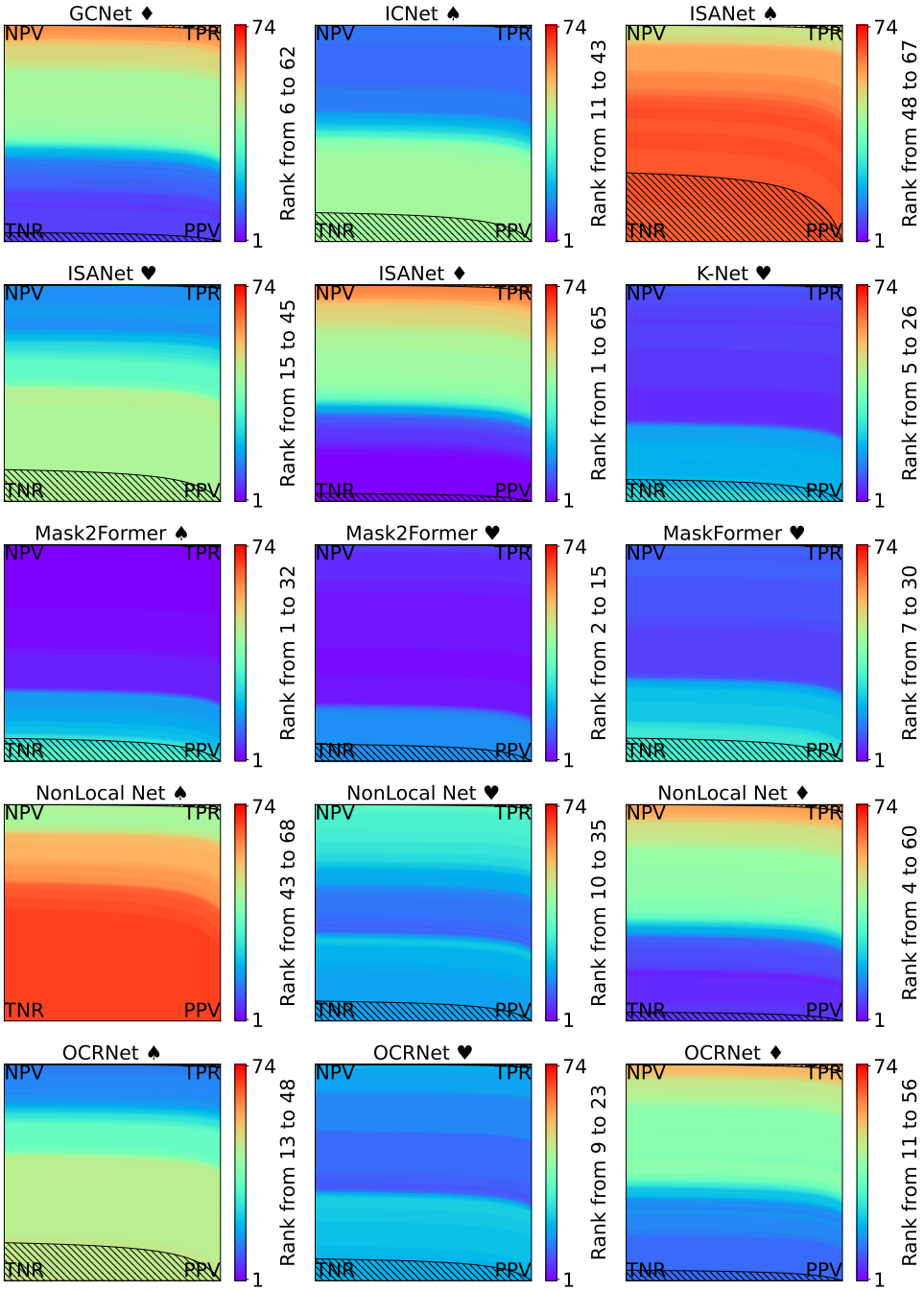


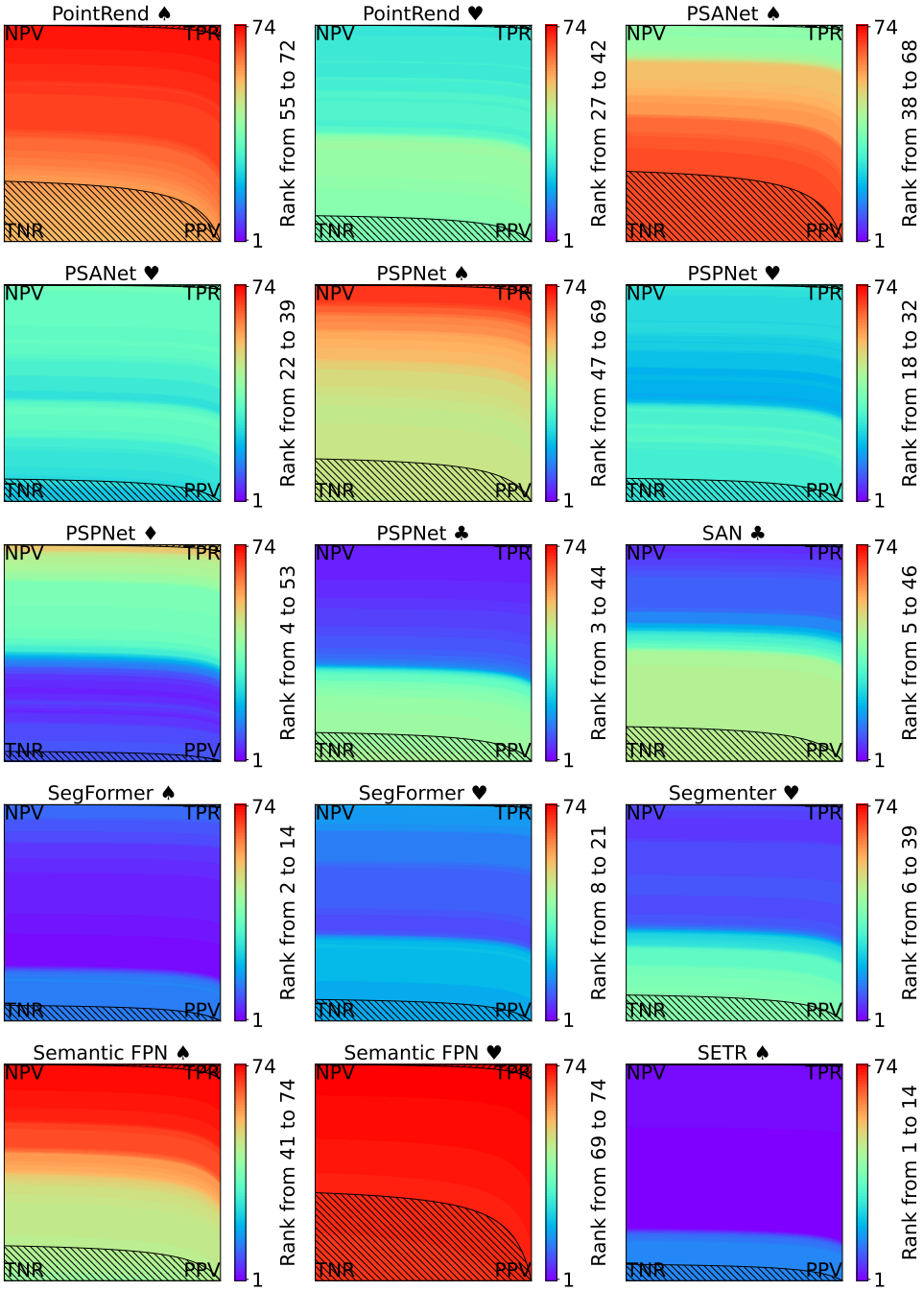
Ranking Tile: Using the tile to show the ranks for each entity

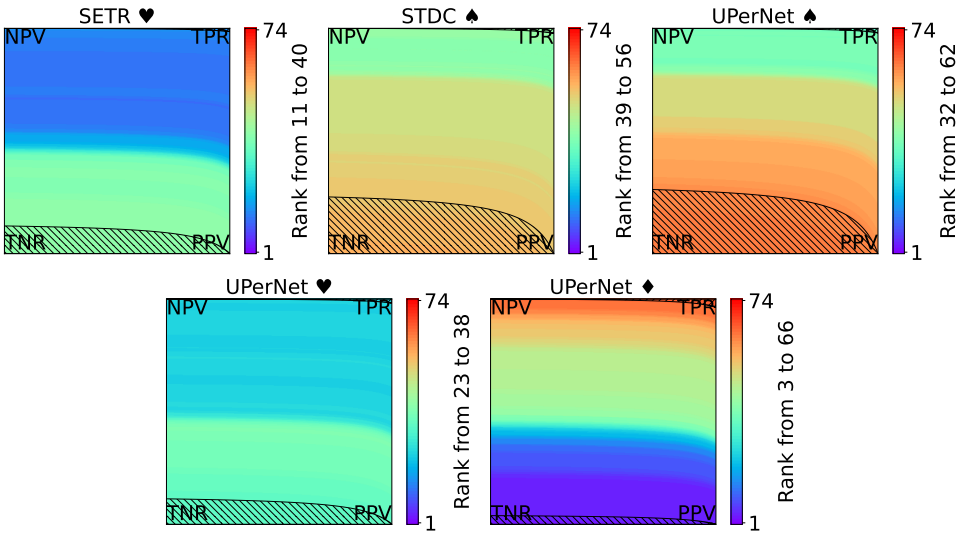



CCNet ♥
NPV
TPR
TNR
PPV
74
1
Rank from 25 to 38
CCNet ♦
NPV
TPR
TNR
PPV
74
1
Rank from 2 to 61
CGNet ♠
NPV
TPR
TNR
PPV
74
1
Rank from 2 to 74
DANet ♠
NPV
TPR
TNR
PPV
74
1
Rank from 43 to 72
DANet ♥
NPV
TPR
TNR
PPV
74
1
Rank from 14 to 32
DANet ♦
NPV
TPR
TNR
PPV
74
1
Rank from 5 to 67
DeepLabV3+ ♠
NPV
TPR
TNR
PPV
74
1
Rank from 48 to 70
DeepLabV3+ ♥
NPV
TPR
TNR
PPV
74
1
Rank from 16 to 31
DeepLabV3+ ♦
NPV
TPR
TNR
PPV
74
1
Rank from 1 to 58
DeepLabV3 ♠
NPV
TPR
TNR
PPV
74
1
Rank from 47 to 56
DeepLabV3 ♥
NPV
TPR
TNR
PPV
74
1
Rank from 13 to 26
DeepLabV3 ♦
NPV
TPR
TNR
PPV
74
1
Rank from 9 to 55
DeepLabV3 ♣
NPV
TPR
TNR
PPV
74
1
Rank from 3 to 38
DMNet ♠
NPV
TPR
TNR
PPV
74
1
Rank from 49 to 71
DMNet ♥
NPV
TPR
TNR
PPV
74
1
Rank from 18 to 35


DNLNet ♠
NPV
TPR
TNR
PPV
74
1
Rank from 47 to 65
DNLNet ♥
NPV
TPR
TNR
PPV
74
1
Rank from 15 to 30
DPT ♥
NPV
TPR
TNR
PPV
74
1
Rank from 19 to 35
EMANet ♠
NPV
TPR
TNR
PPV
74
1
Rank from 57 to 66
EncNet ♠
NPV
TPR
TNR
PPV
74
1
Rank from 53 to 70
EncNet ♥
NPV
TPR
TNR
PPV
74
1
Rank from 21 to 33
ERFNet ♠
NPV
TPR
TNR
PPV
74
1
Rank from 37 to 52
Fast-SCNN ♠
NPV
TPR
TNR
PPV
74
1
Rank from 24 to 66
FastFCN ♠
NPV
TPR
TNR
PPV
74
1
Rank from 57 to 67
FastFCN ♥
NPV
TPR
TNR
PPV
74
1
Rank from 29 to 42
FCN ♠
NPV
TPR
TNR
PPV
74
1
Rank from 50 to 68
FCN ♥
NPV
TPR
TNR
PPV
74
1
Rank from 12 to 40
FCN ♦
NPV
TPR
TNR
PPV
74
1
Rank from 11 to 52
GCNet ♠
NPV
TPR
TNR
PPV
74
1
Rank from 46 to 71
GCNet ♥
NPV
TPR
TNR
PPV
74
1
Rank from 21 to 36


GCNet ♦
NPV
TPR
TNR
PPV
74
1
Rank from 6 to 62
ICNet ♠
NPV
TPR
TNR
PPV
74
1
Rank from 11 to 43
ISANet ♠
NPV
TPR
TNR
PPV
74
1
Rank from 48 to 67
ISANet ♥
NPV
TPR
TNR
PPV
74
1
Rank from 15 to 45
ISANet ♦
NPV
TPR
TNR
PPV
74
1
Rank from 1 to 65
K-Net ♥
NPV
TPR
TNR
PPV
74
1
Rank from 5 to 26
Mask2Former ♠
NPV
TPR
TNR
PPV
74
1
Rank from 1 to 32
Mask2Former ♥
NPV
TPR
TNR
PPV
74
1
Rank from 2 to 15
MaskFormer ♥
NPV
TPR
TNR
PPV
74
1
Rank from 7 to 30
NonLocal Net ♠
NPV
TPR
TNR
PPV
74
1
Rank from 43 to 68
NonLocal Net ♥
NPV
TPR
TNR
PPV
74
1
Rank from 10 to 35
NonLocal Net ♦
NPV
TPR
TNR
PPV
74
1
Rank from 4 to 60
OCRNet ♠
NPV
TPR
TNR
PPV
74
1
Rank from 13 to 48
OCRNet ♥
NPV
TPR
TNR
PPV
74
1
Rank from 9 to 23
OCRNet ♦
NPV
TPR
TNR
PPV
74
1
Rank from 11 to 56


PointRend ♠
NPV
TPR
TNR
PPV
74
1
Rank from 55 to 72
PointRend ♥
Rank from 27 to 42
PSANet ♠
Rank from 38 to 68
PSANet ♥
Rank from 22 to 39
PSPNet ♠
Rank from 47 to 69
PSPNet ♥
Rank from 18 to 32
PSPNet ♦
Rank from 4 to 53
PSPNet ♣
Rank from 3 to 44
SAN ♣
Rank from 5 to 46
SegFormer ♠
Rank from 2 to 14
SegFormer ♥
Rank from 8 to 21
Segmenter ♥
Rank from 6 to 39
Semantic FPN ♠
Rank from 41 to 74
Semantic FPN ♥
Rank from 69 to 74
SETR ♠
Rank from 1 to 14



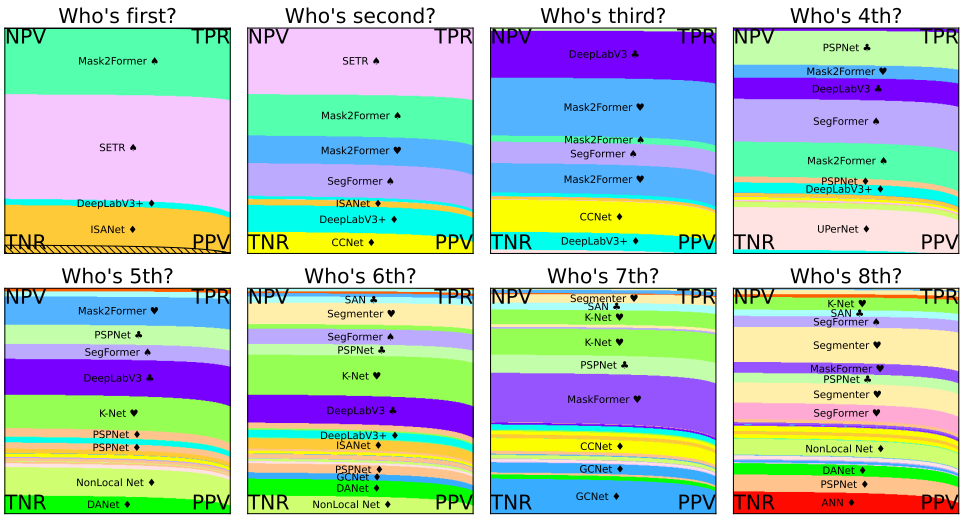
Analysis. The following entities mimimize the maximum rank:

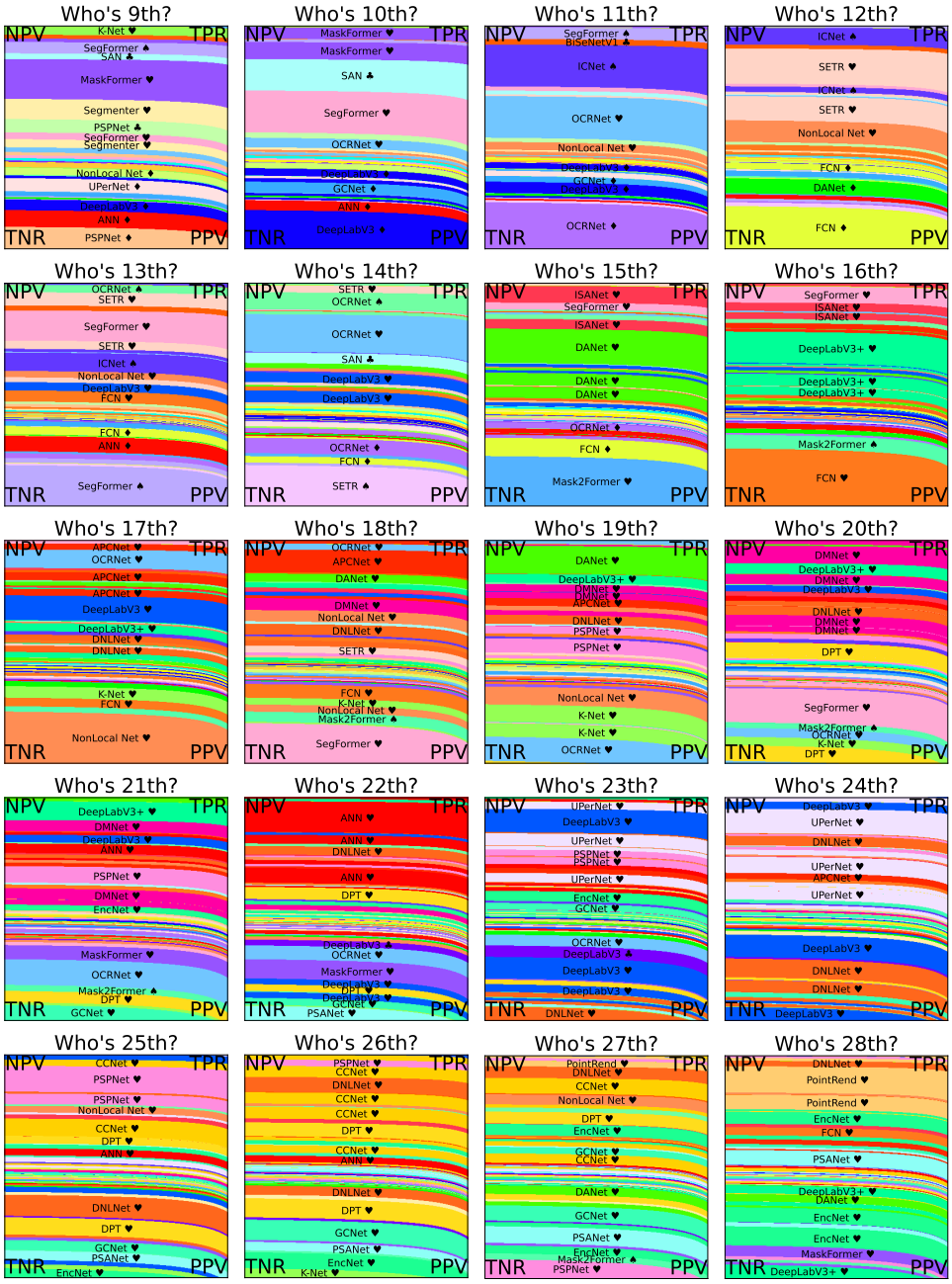
- SegFormer ♠ (max rank: 14, mean rank: 6.85568)
- SETR ♠ (max rank: 14, mean rank: 4.04938)

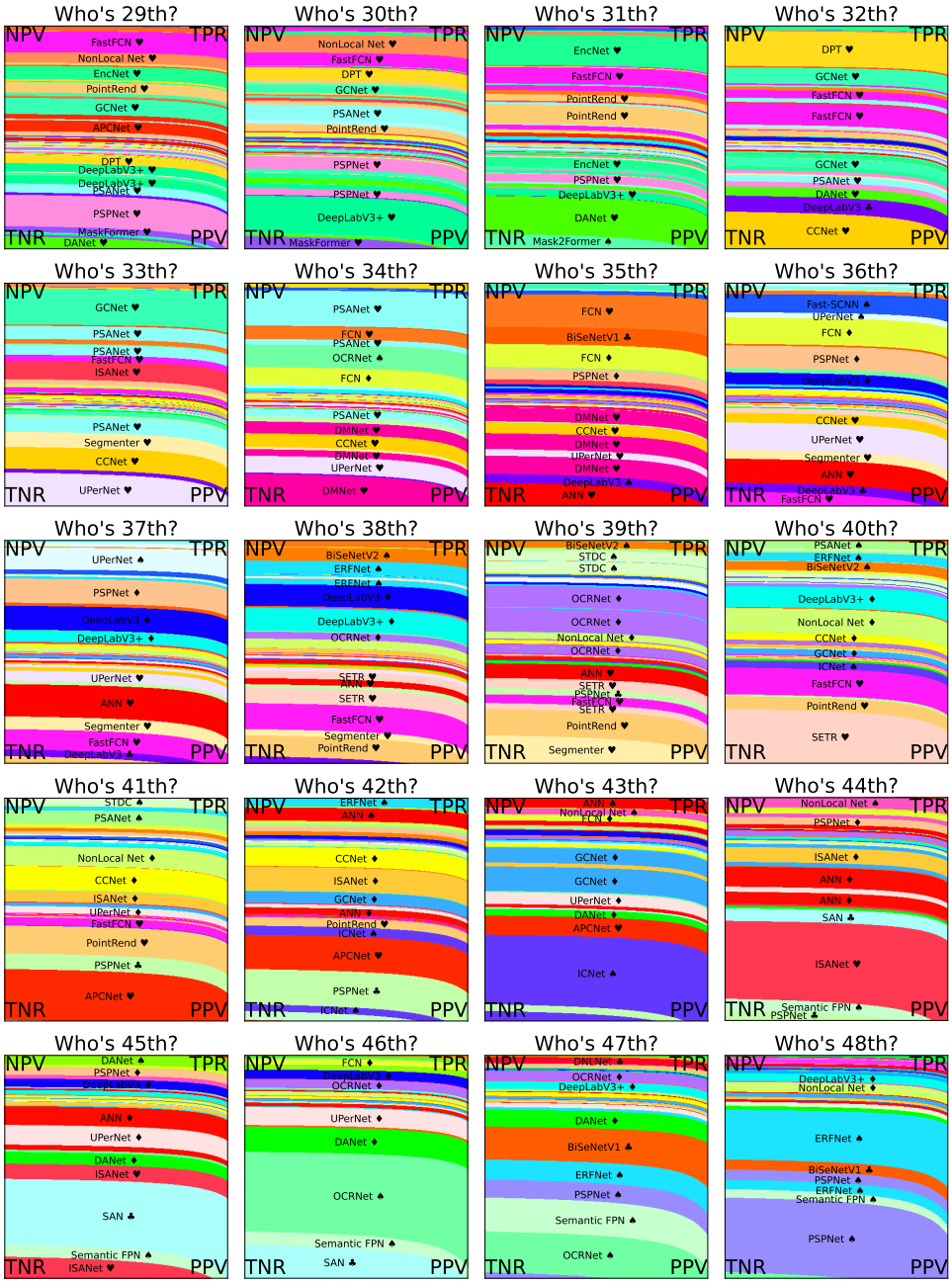
Among them, the following entities have the lowest mean rank:

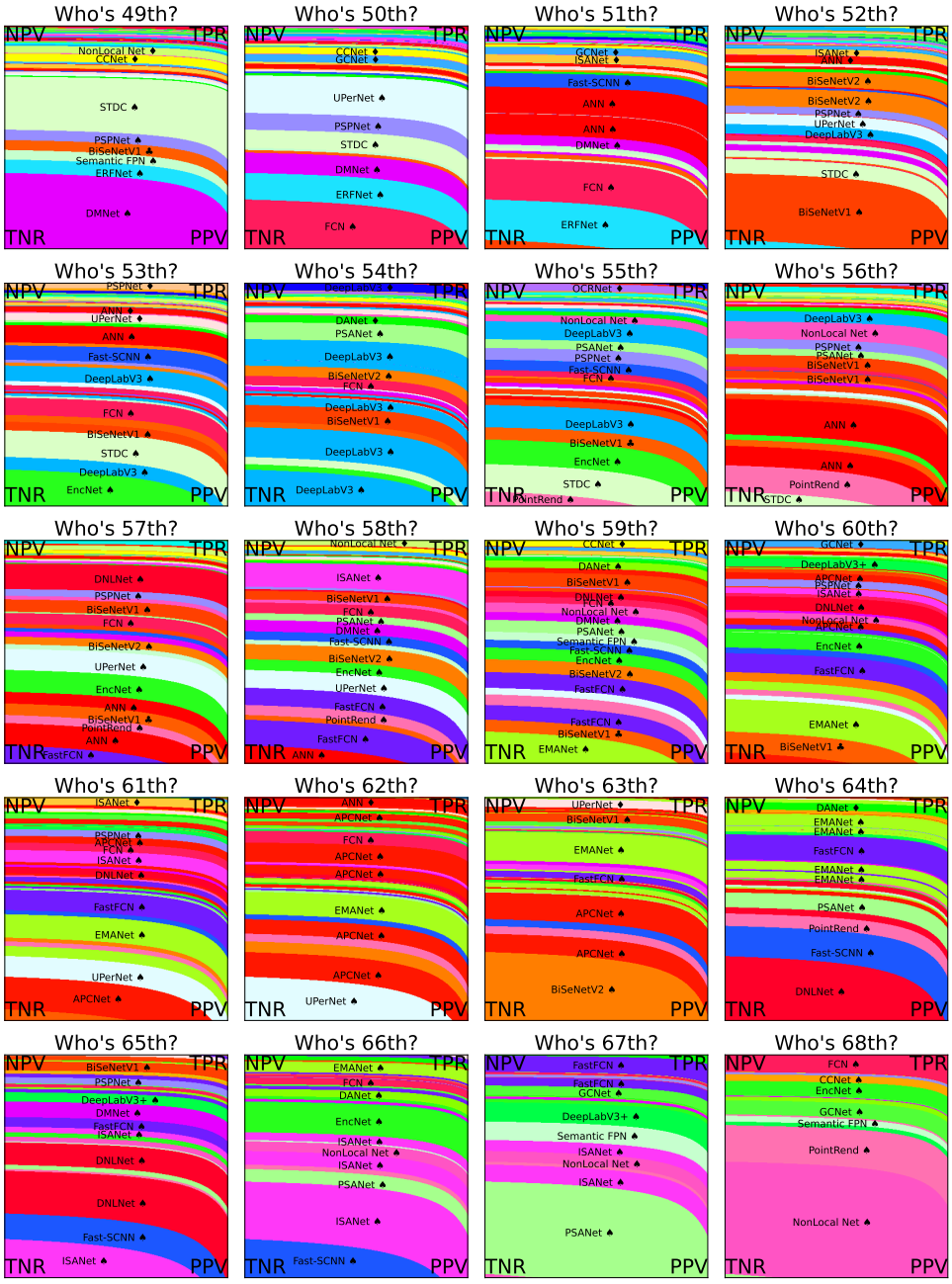
- SETR ♠

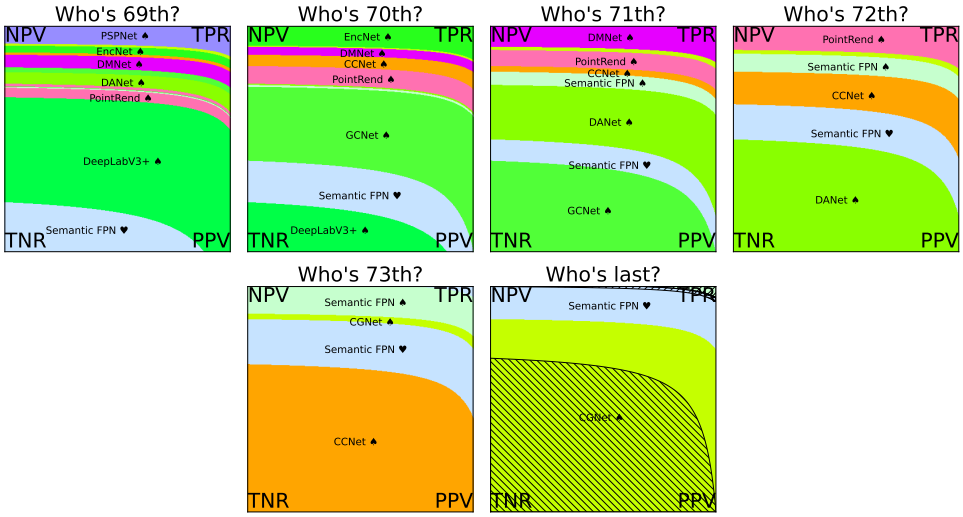
Entity Tile: Using the tile to show the entities for each rank




Who's 9th?
Who's 10th?
Who's 11th?
Who's 12th?
Who's 13th?
Who's 14th?
Who's 15th?
Who's 16th?
Who's 17th?
Who's 18th?
Who's 19th?
Who's 20th?
Who's 21th?
Who's 22th?
Who's 23th?
Who's 24th?
Who's 25th?
Who's 26th?
Who's 27th?
Who's 28th?
NPV
TPR
TNR
PPV

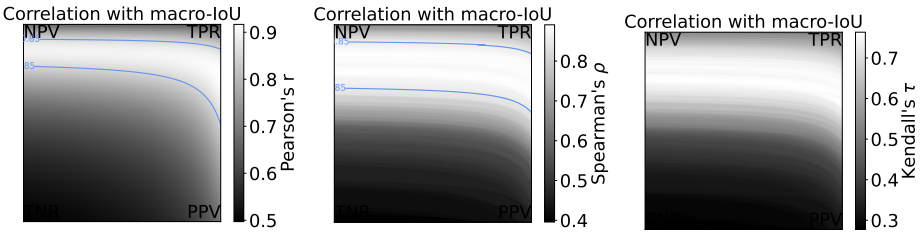

Who's 29th?
NPV TPR
FastFCN ♥
NonLocal Net ♥
EncNet ♥
PointRend ♥
GCNet ♥
APCNet ♥
DPT ♥
DeepLabV3+ ♥
DeepLabV3+ ♥
PSANet ♥
PSPNet ♥
MaskFormer ♥
TNR DANet ♥ PPV
Who's 30th?
NPV TPR
NonLocal Net ♥
FastFCN ♥
DPT ♥
GCNet ♥
PSANet ♥
PointRend ♥
PSPNet ♥
PSPNet ♥
DeepLabV3+ ♥
TNR MaskFormer ♥ PPV
Who's 31th?
NPV TPR
EncNet ♥
FastFCN ♥
PointRend ♥
PointRend ♥
EncNet ♥
PSPNet ♥
DeepLabV3+ ♥
DANet ♥
TNR Mask2Former ♣ PPV
Who's 32th?
NPV TPR
DPT ♥
GCNet ♥
FastFCN ♥
FastFCN ♥
GCNet ♥
PSANet ♥
DANet ♥
DeepLabV3 ♣
TNR CCNet ♥ PPV
Who's 33th?
NPV TPR
GCNet ♥
PSANet ♥
PSANet ♥
FastFCN ♥
ISANet ♥
PSANet ♥
Segmenter ♥
CCNet ♥
TNR UPerNet ♥ PPV
Who's 34th?
NPV TPR
PSANet ♥
FCN ♥
PSANet ♥
OCRNet ♠
FCN ♦
PSANet ♥
DMNet ♥
CCNet ♥
DMNet ♥
UPerNet ♥
TNR DMNet ♥ PPV
Who's 35th?
NPV TPR
FCN ♥
BiSeNetV1 ♣
FCN ♦
PSPNet ♦
DMNet ♥
CCNet ♥
DMNet ♥
UPerNet ♥
DMNet ♥
DeepLabV3 ♣
TNR ANN ♥ PPV
Who's 36th?
NPV TPR
Fast-SCNN ♣
UPerNet ♣
FCN ♦
PSPNet ♦
DeepLabV3 ♦
CCNet ♥
UPerNet ♥
Segmenter ♥
ANN ♥
DeepLabV3 ♣
TNR FastFCN ♥ PPV
Who's 37th?
NPV TPR
UPerNet ♠
PSPNet ♦
DeepLabV3 ♦
DeepLabV3+ ♦
UPerNet ♥
ANN ♥
Segmenter ♥
FastFCN ♥
TNR DeepLabV3 ♣ PPV
Who's 38th?
NPV TPR
BiSeNetV2 ♠
ERFNet ♠
ERFNet ♠
DeepLabV3 ♦
DeepLabV3+ ♦
OCRNet ♦
SETR ♥
ANN ♥
SETR ♥
FastFCN ♥
Segmenter ♥
TNR PointRend ♥ PPV
Who's 39th?
NPV TPR
BiSeNetV2 ♠
STDC ♠
STDC ♠
OCRNet ♦
OCRNet ♦
NonLocal Net ♦
OCRNet ♦
ANN ♥
SETR ♥
PSPNet ♣
FastFCN ♥
SETR ♥
PointRend ♥
TNR Segmenter ♥ PPV
Who's 40th?
NPV TPR
PSANet ♠
ERFNet ♠
BiSeNetV2 ♠
DeepLabV3+ ♦
NonLocal Net ♦
CCNet ♦
GCNet ♦
ICNet ♣
FastFCN ♥
PointRend ♥
TNR SETR ♥ PPV
Who's 41th?
NPV TPR
STDC ♠
PSANet ♠
NonLocal Net ♦
CCNet ♦
ISANet ♦
UPerNet ♦
FastFCN ♥
PointRend ♥
PSPNet ♣
TNR APCNet ♥ PPV
Who's 42th?
NPV TPR
ERFNet ♠
ANN ♠
CCNet ♦
ISANet ♦
GCNet ♦
ANN ♦
PointRend ♥
ICNet ♣
APCNet ♥
PSPNet ♣
TNR ICNet ♣ PPV
Who's 43th?
NPV TPR
ANN ♠
NonLocal Net ♠
FCN ♠
GCNet ♦
GCNet ♦
UPerNet ♦
DANet ♦
APCNet ♥
TNR ICNet ♣ PPV
Who's 44th?
NPV TPR
NonLocal Net ♠
PSPNet ♦
ISANet ♦
ANN ♦
ANN ♦
SAN ♣
ISANet ♥
TNR Semantic FPN ♣ PPV
PSPNet ♣
Who's 45th?
NPV TPR
DANet ♠
PSPNet ♦
DeepLabV3 ♦
ANN ♦
UPerNet ♦
DANet ♦
ISANet ♥
SAN ♣
TNR Semantic FPN ♠ PPV
ISANet ♥
Who's 46th?
NPV TPR
FCN ♦
DeepLabV3 ♦
OCRNet ♦
UPerNet ♦
DANet ♦
OCRNet ♠
Semantic FPN ♠
TNR SAN ♣ PPV
Who's 47th?
NPV TPR
DNLNet ♠
OCRNet ♦
DeepLabV3+ ♦
DANet ♦
BiSeNetV1 ♣
ERFNet ♠
PSPNet ♠
Semantic FPN ♠
TNR OCRNet ♠ PPV
Who's 48th?
NPV TPR
DeepLabV3+ ♦
NonLocal Net ♦
ERFNet ♠
BiSeNetV1 ♣
PSPNet ♠
ERFNet ♠
Semantic FPN ♠
PSPNet ♠
TNR PPV


Who's 49th?
NPV TPR
NonLocal Net ♦
CCNet ♦
STDC ♠
PSPNet ♠
BiSeNetV1 ♣
Semantic FPN ♠
ERFNet ♠
DMNet ♠
TNR PPV
Who's 50th?
NPV TPR
CCNet ♦
GCNet ♦
UPerNet ♠
PSPNet ♠
STDC ♠
DMNet ♠
ERFNet ♠
FCN ♠
TNR PPV
Who's 51th?
NPV TPR
GCNet ♦
ISANet ♦
Fast-SCNN ♠
ANN ♠
ANN ♠
DMNet ♠
FCN ♠
ERFNet ♠
TNR PPV
Who's 52th?
NPV TPR
ISANet ♦
ANN ♦
BiSeNetV2 ♠
BiSeNetV2 ♠
PSPNet ♠
UPerNet ♠
DeepLabV3 ♠
STDC ♠
BiSeNetV1 ♠
TNR PPV
Who's 53th?
NPV TPR
PSPNet ♦
ANN ♦
UPerNet ♦
ANN ♠
Fast-SCNN ♠
DeepLabV3 ♠
FCN ♠
BiSeNetV1 ♠
STDC ♠
DeepLabV3 ♠
EncNet ♠
TNR PPV
Who's 54th?
NPV TPR
DeepLabV3 ♦
DANet ♦
PSANet ♠
DeepLabV3 ♠
BiSeNetV2 ♠
FCN ♠
DeepLabV3 ♠
BiSeNetV1 ♠
DeepLabV3 ♠
TNR DeepLabV3 ♠ PPV
Who's 55th?
NPV TPR
OCRNet ♦
NonLocal Net ♠
DeepLabV3 ♠
PSANet ♠
PSPNet ♠
Fast-SCNN ♠
FCN ♠
DeepLabV3 ♠
BiSeNetV1 ♣
EncNet ♠
STDC ♠
TNR PPV
Who's 56th?
NPV TPR
DeepLabV3 ♠
NonLocal Net ♠
PSPNet ♠
PSANet ♠
BiSeNetV1 ♠
BiSeNetV1 ♠
ANN ♠
ANN ♠
PointRend ♠
TNR STDC ♠ PPV
Who's 57th?
NPV TPR
DNLNet ♠
PSPNet ♠
BiSeNetV1 ♠
FCN ♠
BiSeNetV2 ♠
UPerNet ♠
EncNet ♠
ANN ♠
BiSeNetV1 ♠
PointRend ♠
ANN ♠
TNR FastFCN ♠ PPV
Who's 58th?
NPV TPR
NonLocal Net ♦
ISANet ♠
BiSeNetV1 ♠
FCN ♠
PSANet ♠
DMNet ♠
Fast-SCNN ♠
BiSeNetV2 ♠
EncNet ♠
UPerNet ♠
FastFCN ♠
PointRend ♠
TNR ANN ♠ PPV
Who's 59th?
NPV TPR
CCNet ♦
DANet ♠
BiSeNetV1 ♠
DNLNet ♠
FCN ♠
NonLocal Net ♠
DMNet ♠
PSANet ♠
Semantic FPN ♠
Fast-SCNN ♠
EncNet ♠
BiSeNetV2 ♠
FastFCN ♠
FastFCN ♠
BiSeNetV1 ♣
TNR EMANet ♠ PPV
Who's 60th?
NPV TPR
GCNet ♦
DeepLabV3+ ♠
APCNet ♠
PSPNet ♠
ISANet ♠
DNLNet ♠
NonLocal Net ♠
APCNet ♠
EncNet ♠
FastFCN ♠
EMANet ♠
TNR BiSeNetV1 ♣ PPV
Who's 61th?
NPV TPR
ISANet ♦
PSPNet ♠
APCNet ♠
FCN ♠
ISANet ♠
DNLNet ♠
FastFCN ♠
EMANet ♠
UPerNet ♠
TNR APCNet ♠ PPV
Who's 62th?
NPV TPR
ANN ♠
APCNet ♠
FCN ♠
APCNet ♠
APCNet ♠
EMANet ♠
APCNet ♠
APCNet ♠
TNR UPerNet ♠ PPV
Who's 63th?
NPV TPR
UPerNet ♠
BiSeNetV1 ♠
EMANet ♠
FastFCN ♠
APCNet ♠
APCNet ♠
BiSeNetV2 ♠
TNR PPV
Who's 64th?
NPV TPR
DANet ♠
EMANet ♠
EMANet ♠
FastFCN ♠
EMANet ♠
EMANet ♠
PSANet ♠
PointRend ♠
Fast-SCNN ♠
DNLNet ♠
TNR PPV
Who's 65th?
NPV TPR
BiSeNetV1 ♠
PSPNet ♠
DeepLabV3+ ♠
DMNet ♠
FastFCN ♠
ISANet ♠
DNLNet ♠
DNLNet ♠
Fast-SCNN ♠
TNR ISANet ♠ PPV
Who's 66th?
NPV TPR
EMANet ♠
FCN ♠
DANet ♠
EncNet ♠
ISANet ♠
NonLocal Net ♠
ISANet ♠
PSANet ♠
ISANet ♠
TNR Fast-SCNN ♠ PPV
Who's 67th?
NPV TPR
FastFCN ♠
FastFCN ♠
GCNet ♠
DeepLabV3+ ♠
Semantic FPN ♠
ISANet ♠
NonLocal Net ♠
ISANet ♠
PSANet ♠
TNR PPV
Who's 68th?
NPV TPR
FCN ♠
CCNet ♠
EncNet ♠
GCNet ♠
Semantic FPN ♠
PointRend ♠
NonLocal Net ♠
TNR PPV



Analysis. The following entities are ranked first:
- DeepLabV3+ ♦ (2.66% of the tile)
- ISANet ♦ (20.52% of the tile)
- Mask2Former ♠ (29.85% of the tile)
- SETR ♠ (46.97% of the tile)

Correlation Tile: Using the tile to show the rank and linear correlations with the mean-IoU



Analysis for the linear correlation with Pearson's r.
- In 93.9% of the zone where $r \geq 0.85$ in the tile, the best is Mask2Former ♠
- In 6.1% of the zone where $r \geq 0.85$ in the tile, the best is SETR ♠

Analysis for the rank correlation with Spearman's ρ.
- In 81.5% of the zone where $\rho \geq 0.85$ in the tile, the best is Mask2Former ♠
- In 18.5% of the zone where $\rho \geq 0.85$ in the tile, the best is SETR ♠

Analysis for the rank correlation with Kendall's τ.
- WARNING: There is no zone where $\tau \geq 0.85$ in the tile! Maybe do you want to change the threshold?

E. Supplementary material for the 7th article

This appendix contains the supplementary material provided with my 7th selected article [7], presented in Chapter 7.

APPENDIX

A. Useful resources

Useful resources can be found at:

https://github.com/pierard/performance

B. Enlarged version of Fig. 3

Fig. 9 is an enlarged version of Fig. 3 on which it is easier to observe the 40 layers of the mille-feuilles.

C. BGS methods considered in Sec. IV

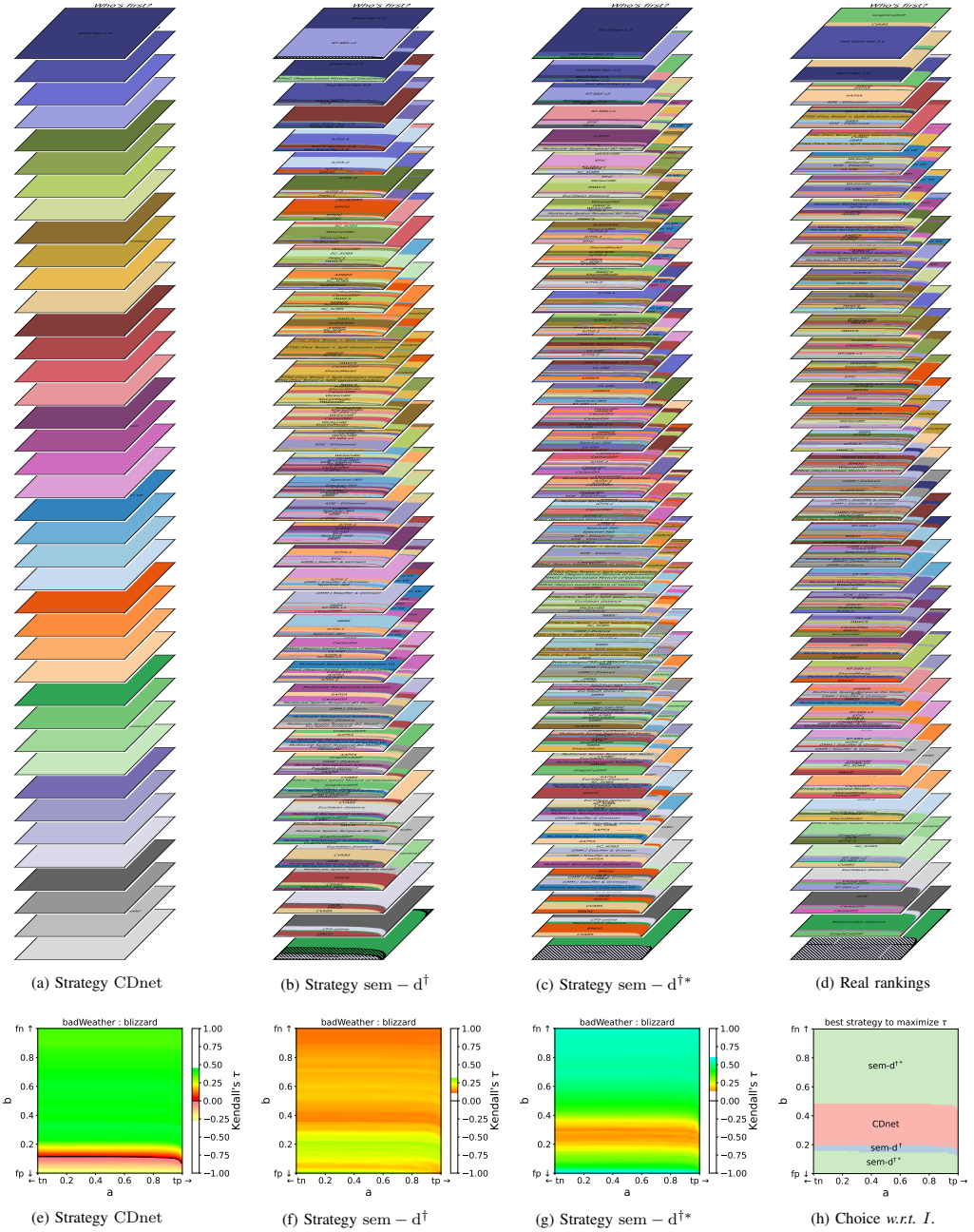
We considered 40 unsupervised BGS methods. More precisely, we selected those for which the output masks were available on changedetection.net, for all videos of *CDnet 2014*, on March 11th 2025, and excluded those tagged as "supervised" on the platform.

1) BSUV-Net 2.0
2) Fast BSUV-Net 2.0
3) IUTIS-5
4) RT-SBS-v2
5) IUTIS-3
6) WisenetMD
7) PAWCS
8) WeSamBE
9) SuBSENSE
10) FTSG (Flux Tensor with Split Gaussian mdoels)
11) SharedModel
12) CVABS
13) M4CD Version 2.0
14) SWCD
15) CwisarDRP
16) RT-SBS-v1
17) C-EFIC
18) Multimode Background Subtraction
19) CwisarDH
20) EFIC
21) Multimode Background Subtraction Version 0 (MBS V0)
22) Spectral-360
23) Sample based background subtractor (SBBS)
24) IUTIS-2
25) BMOG
26) AMBER
27) IUTIS-1
28) AAPSA
29) Mahalanobis distance
30) GraphCutDiff
31) RMoG (Region-based Mixture of Gaussians)
32) SC_SOBS
33) CL-VID
34) KDE - ElGammal
35) GMM | Stauffer & Grimson
36) CP3-online
37) DCB
38) GMM | Zivkovic
39) Multiscale Spatio-Temporal BG Model
40) Euclidean distance

D. Detailed results for the first experiment

For the 11 strategies involved in our 1st experiment, we provide the correlation Tiles for each domain (the 53 videos of *CDnet 2014*), as well as the average (mean correlation Tile) and worst-case (minimum correlation Tile) over all domains. These are the intermediate results that are summarized in Fig. 6.

- Fig. 10: detailed results for CDnet
- Fig. 11: detailed results for CDnet*
- Fig. 12: detailed results for mean-P
- Fig. 13: detailed results for mean-P*
- Fig. 14: detailed results for sem-P†
- Fig. 15: detailed results for sem-P$^{\dagger *}$
- Fig. 16: detailed results for $\mathrm{sem}-\mathrm{d}^\dagger$
- Fig. 17: detailed results for $\mathrm{sem}-\mathrm{d}^{\dagger *}$
- Fig. 18: detailed results for avg†
- Fig. 19: detailed results for avg$^{\dagger *}$
- Fig. 20: detailed results for all†



(a) Strategy CDnet (b) Strategy sem − d† (c) Strategy sem − d†* (d) Real rankings

(e) Strategy CDnet (f) Strategy sem − d† (g) Strategy sem − d†* (h) Choice *w.r.t. I*.

Fig. 9: Enlarged version of Fig. 3.

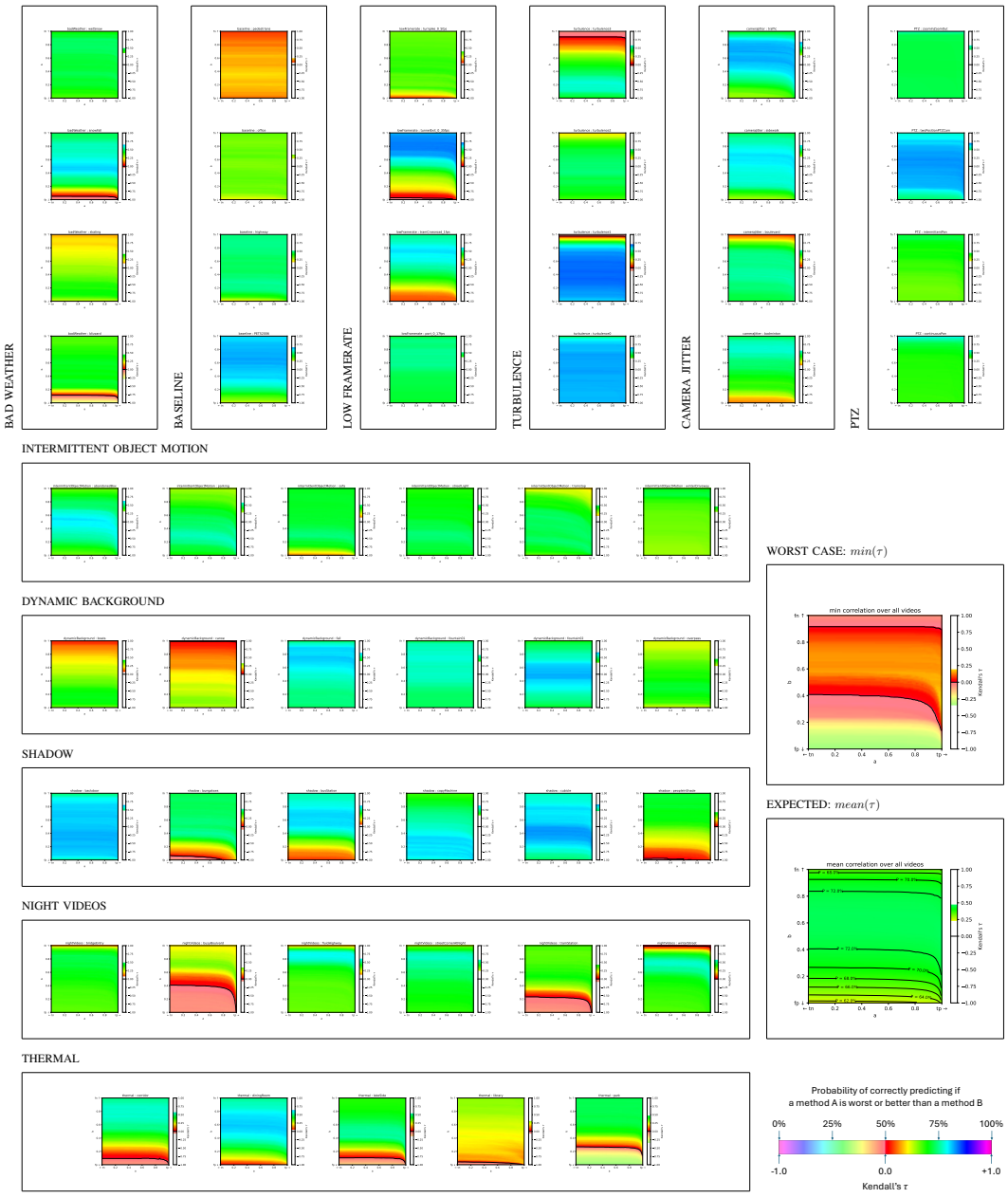


Fig. 10: Tiles showing the correlations (Kendall's τ), *w.r.t.* the application-specific preferences I, between the true ranking and the one predicted with the strategy "CDnet", for 53 videos, as well as in the behavior in the worst case and in average. These rankings involve 40 BGS methods.

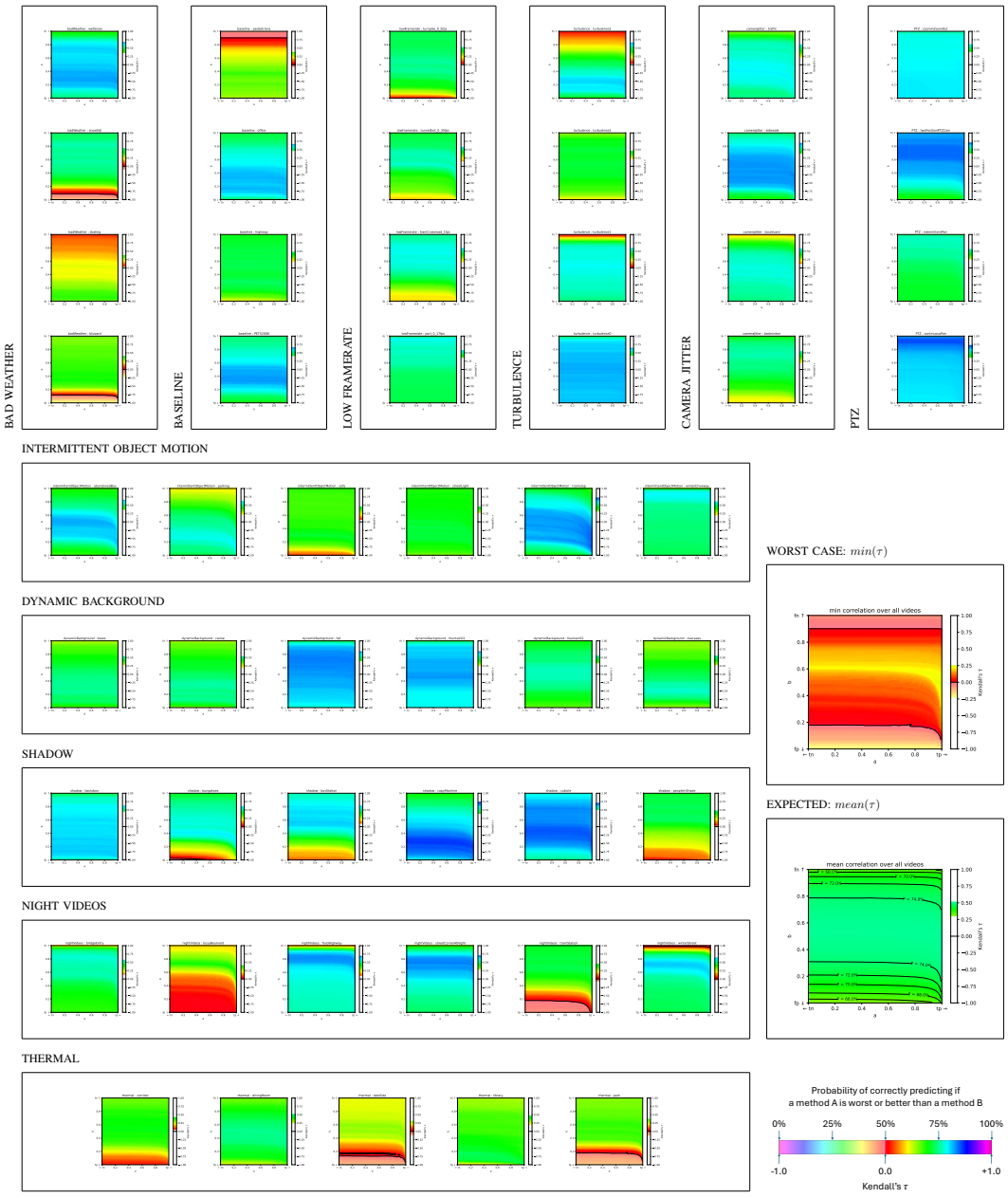


Fig. 11: Tiles showing the correlations (Kendall's τ), *w.r.t.* the application-specific preferences I, between the true ranking and the one predicted with the strategy "CDnet*", for 53 videos, as well as in the behavior in the worst case and in average. These rankings involve 40 BGS methods.

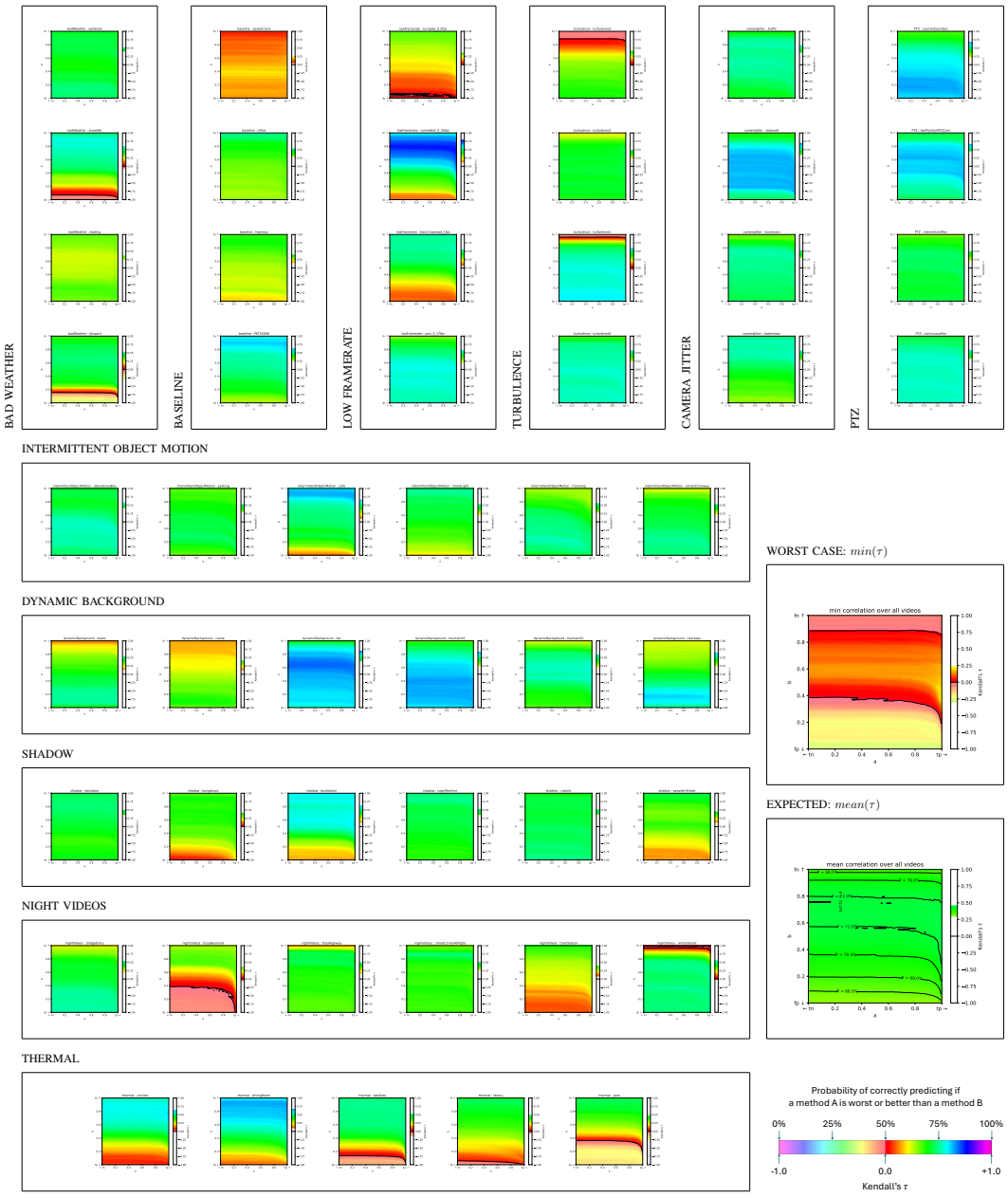


Fig. 12: Tiles showing the correlations (Kendall's τ), *w.r.t.* the application-specific preferences I, between the true ranking and the one predicted with the strategy "mean-P", for 53 videos, as well as in the behavior in the worst case and in average. These rankings involve 40 BGS methods.

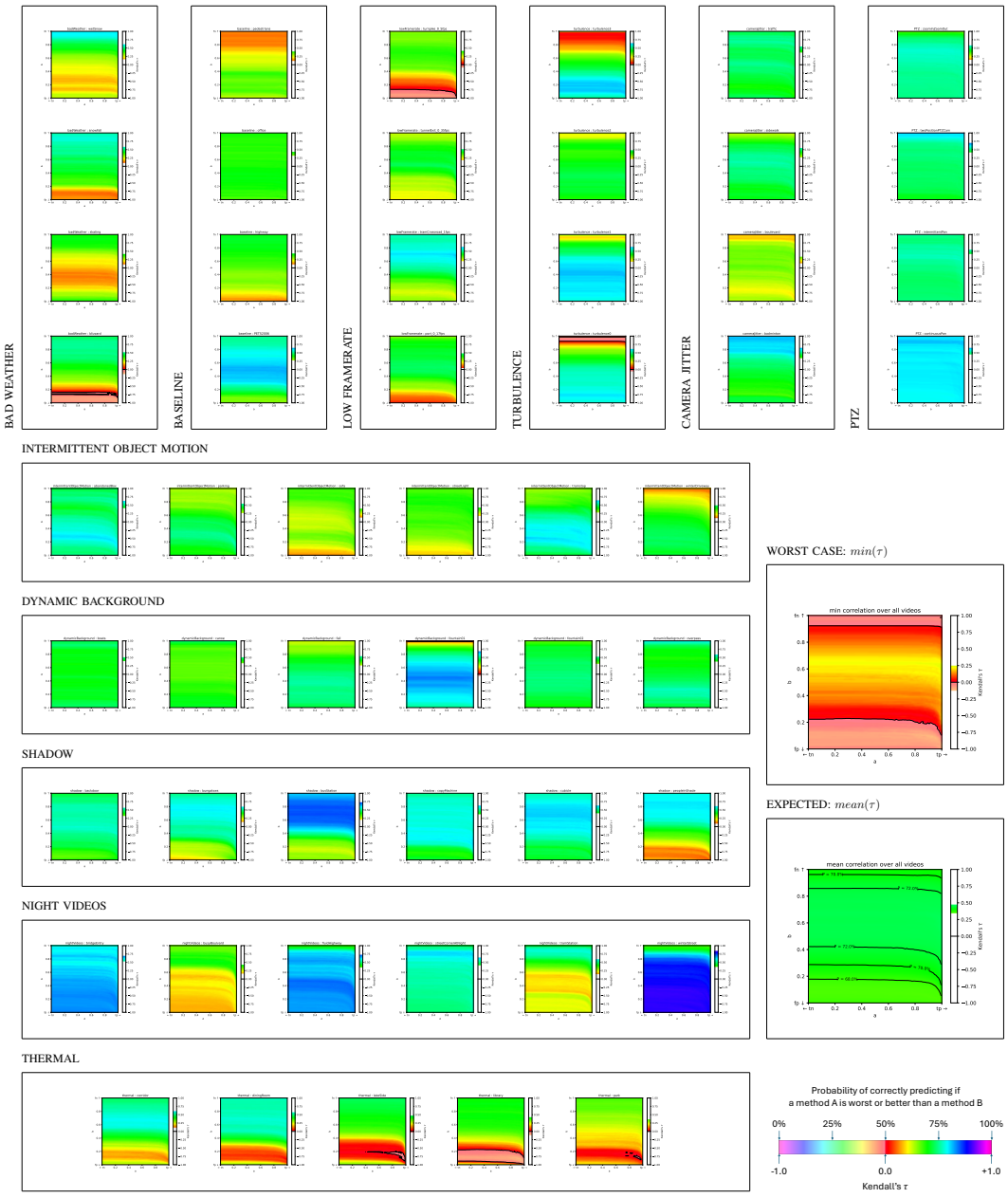


Fig. 13: Tiles showing the correlations (Kendall's τ), *w.r.t.* the application-specific preferences I, between the true ranking and the one predicted with the strategy "mean-P*", for 53 videos, as well as in the behavior in the worst case and in average. These rankings involve 40 BGS methods.

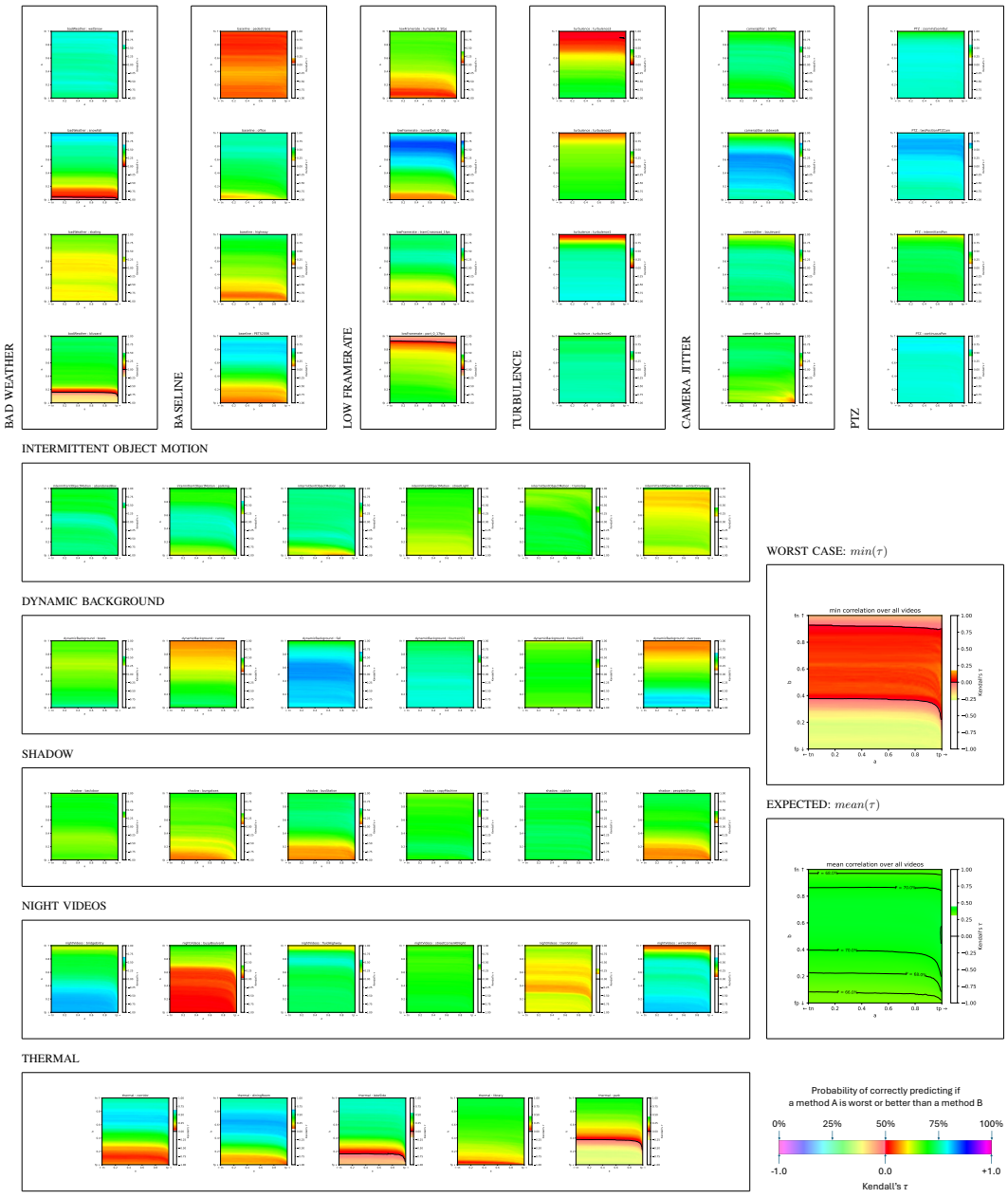


Fig. 14: Tiles showing the correlations (Kendall's τ), *w.r.t.* the application-specific preferences I, between the true ranking and the one predicted with the strategy "sem-P†", for 53 videos, as well as in the behavior in the worst case and in average. These rankings involve 40 BGS methods.

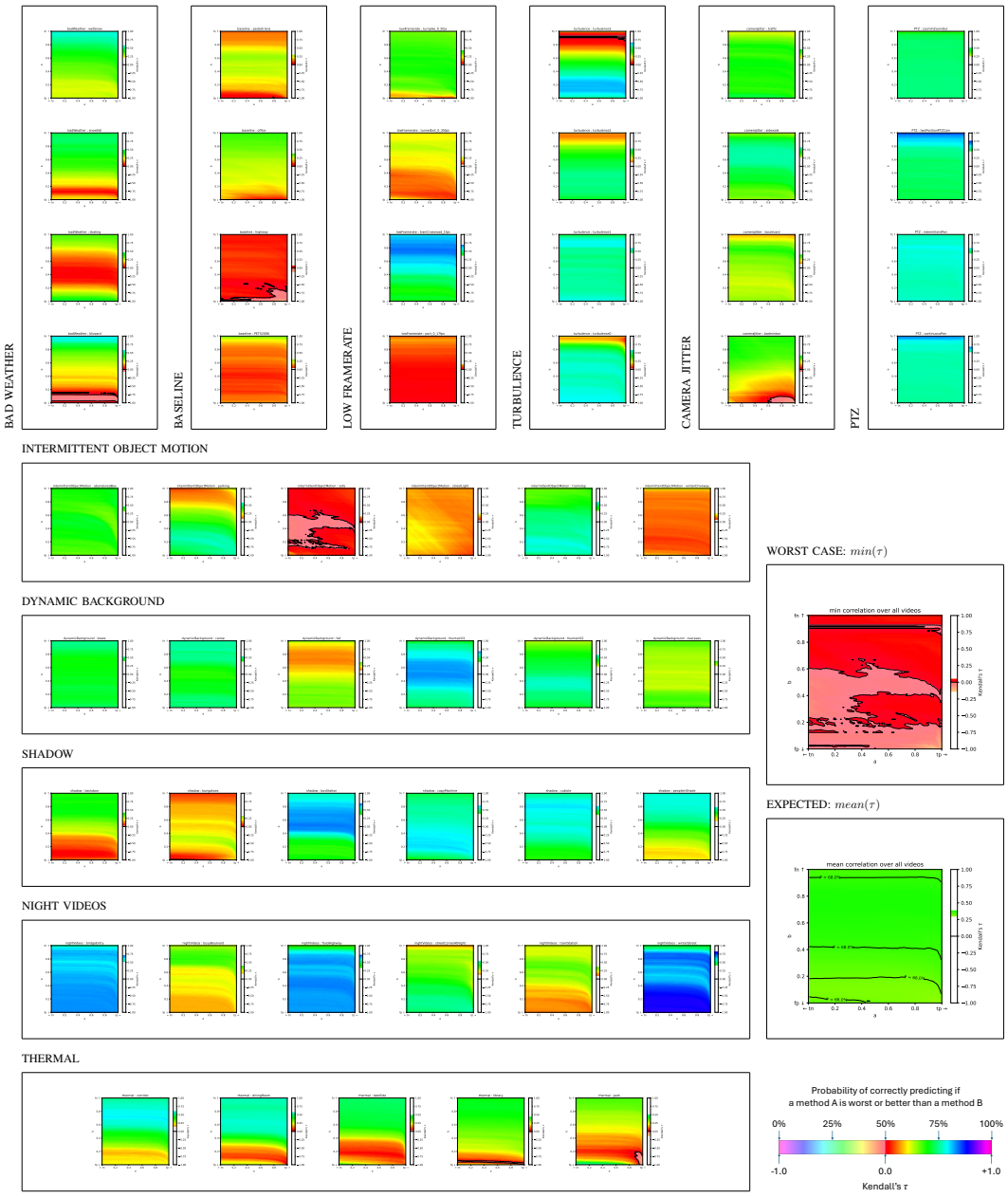


Fig. 15: Tiles showing the correlations (Kendall's τ), *w.r.t.* the application-specific preferences I, between the true ranking and the one predicted with the strategy "sem-P$^{\dagger *}$", for 53 videos, as well as in the behavior in the worst case and in average. These rankings involve 40 BGS methods.

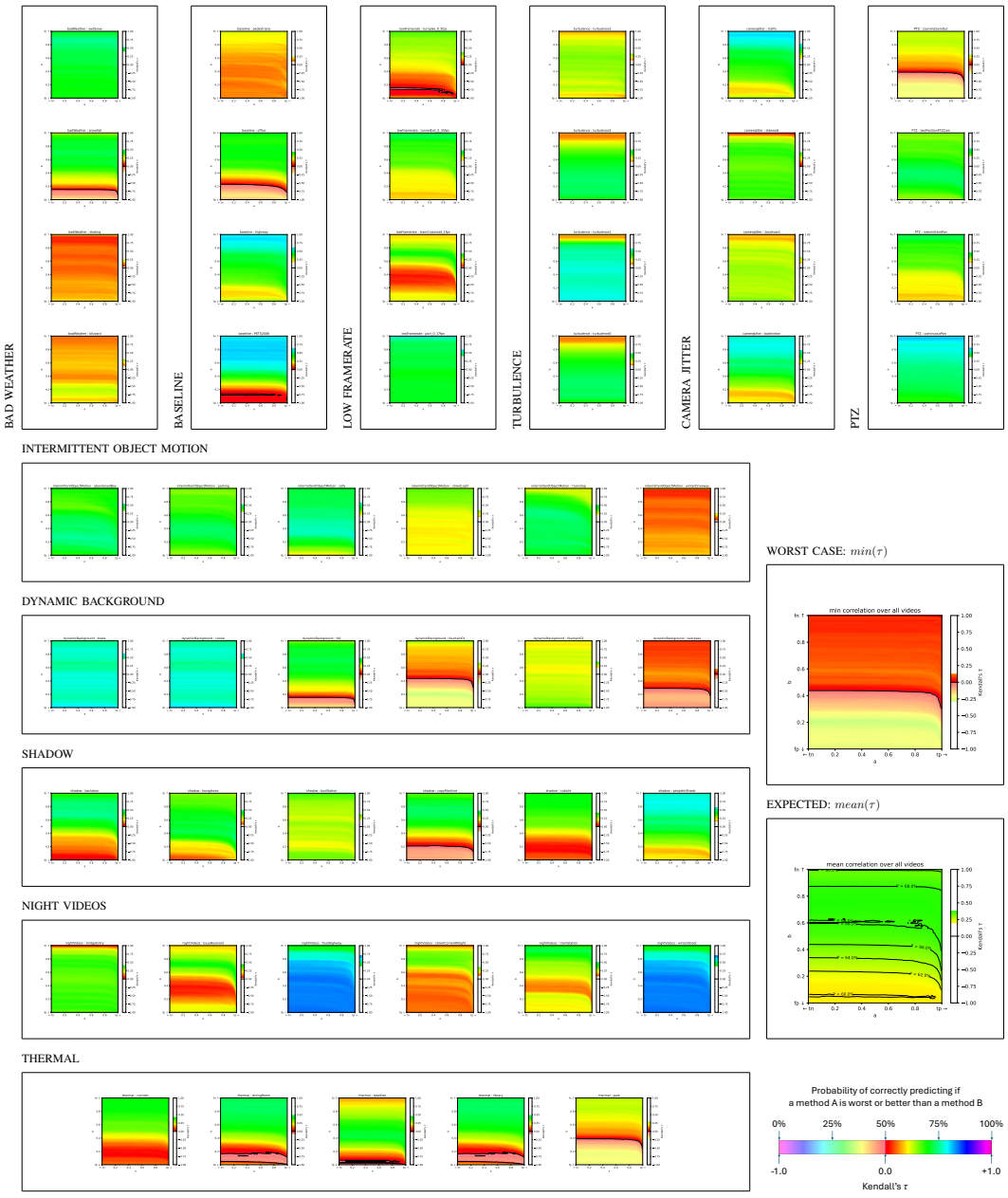


Fig. 16: Tiles showing the correlations (Kendall's τ), *w.r.t.* the application-specific preferences I, between the true ranking and the one predicted with the strategy "sem – $d^{\dagger}$", for 53 videos, as well as in the behavior in the worst case and in average. These rankings involve 40 BGS methods.

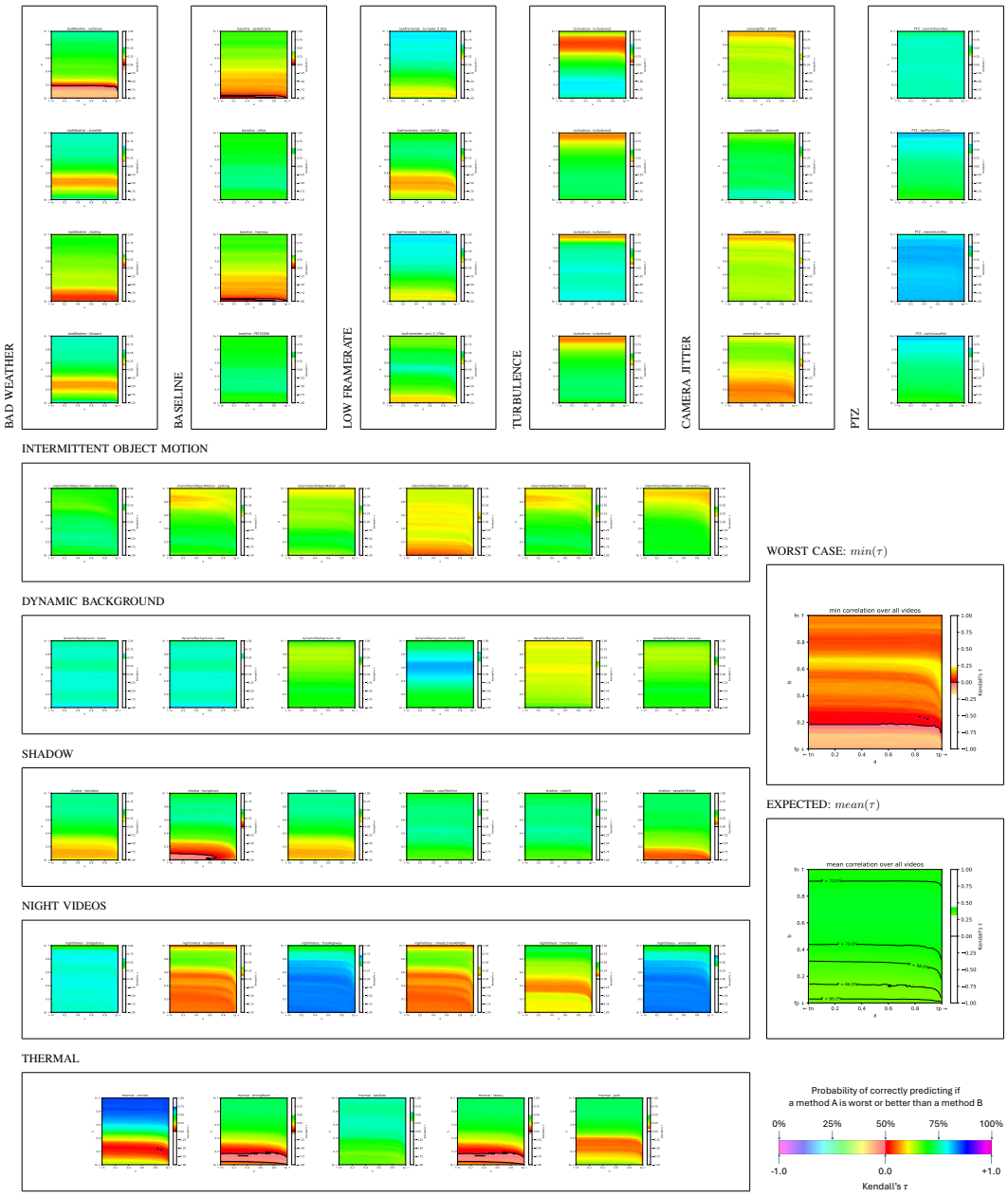


Fig. 17: Tiles showing the correlations (Kendall's τ), *w.r.t.* the application-specific preferences I, between the true ranking and the one predicted with the strategy "$\mathrm{sem} - \mathrm{d}^{\dagger *}$", for 53 videos, as well as in the behavior in the worst case and in average. These rankings involve 40 BGS methods.

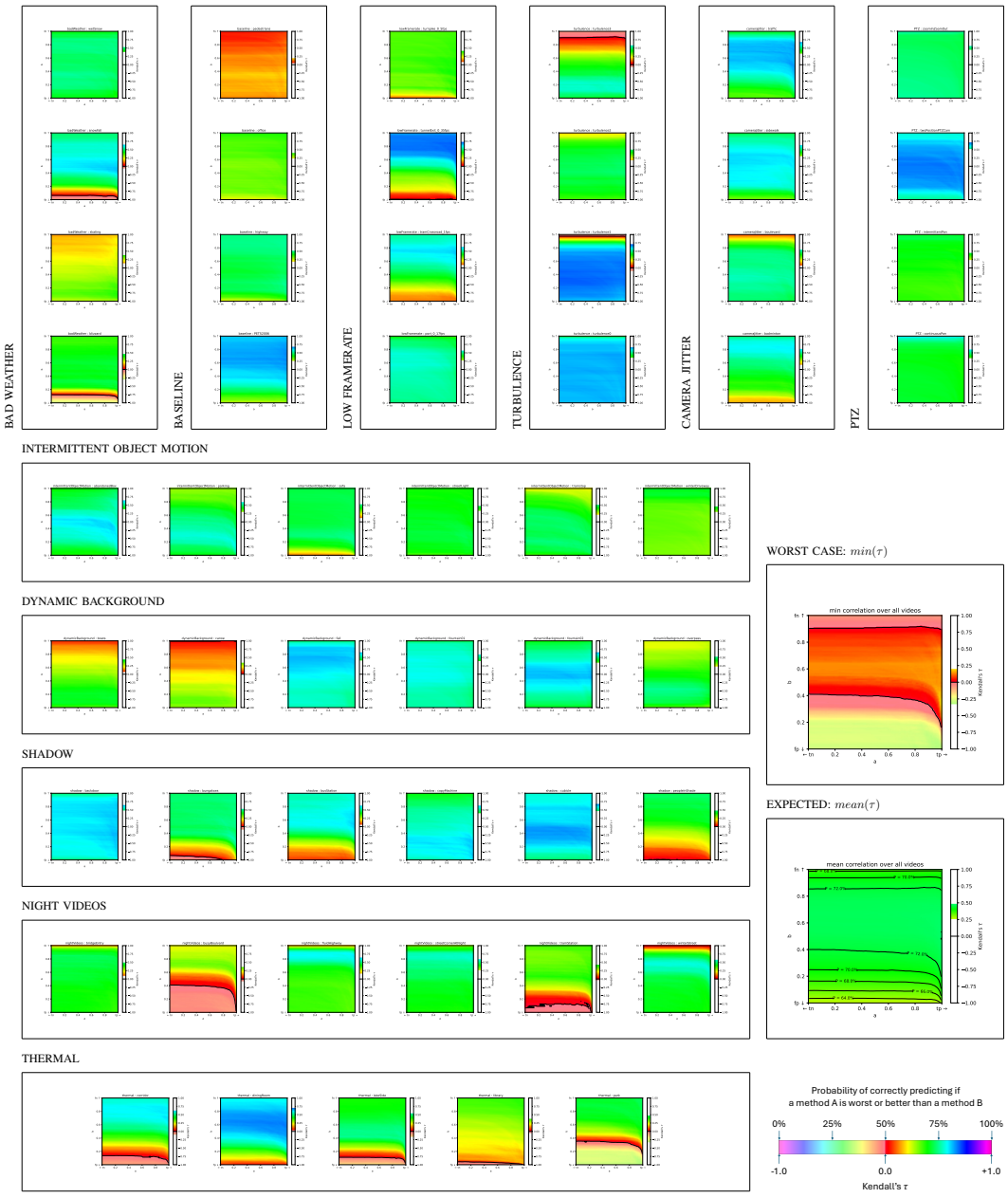


Fig. 18: Tiles showing the correlations (Kendall's τ), *w.r.t.* the application-specific preferences I, between the true ranking and the one predicted with the strategy "avg†", for 53 videos, as well as in the behavior in the worst case and in average. These rankings involve 40 BGS methods.

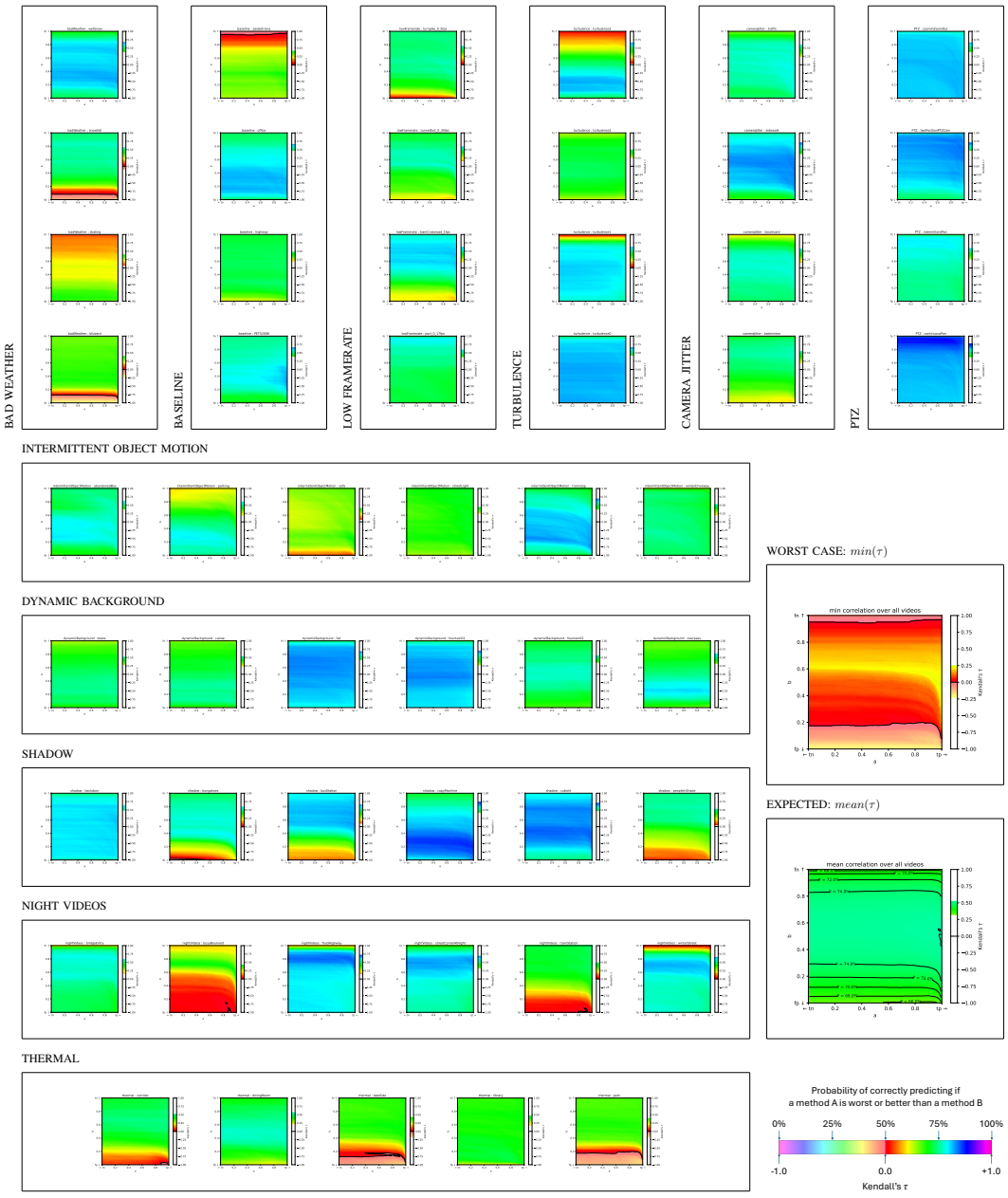


Fig. 19: Tiles showing the correlations (Kendall's τ), *w.r.t.* the application-specific preferences I, between the true ranking and the one predicted with the strategy "avg$^{\dagger *}$", for 53 videos, as well as in the behavior in the worst case and in average. These rankings involve 40 BGS methods.

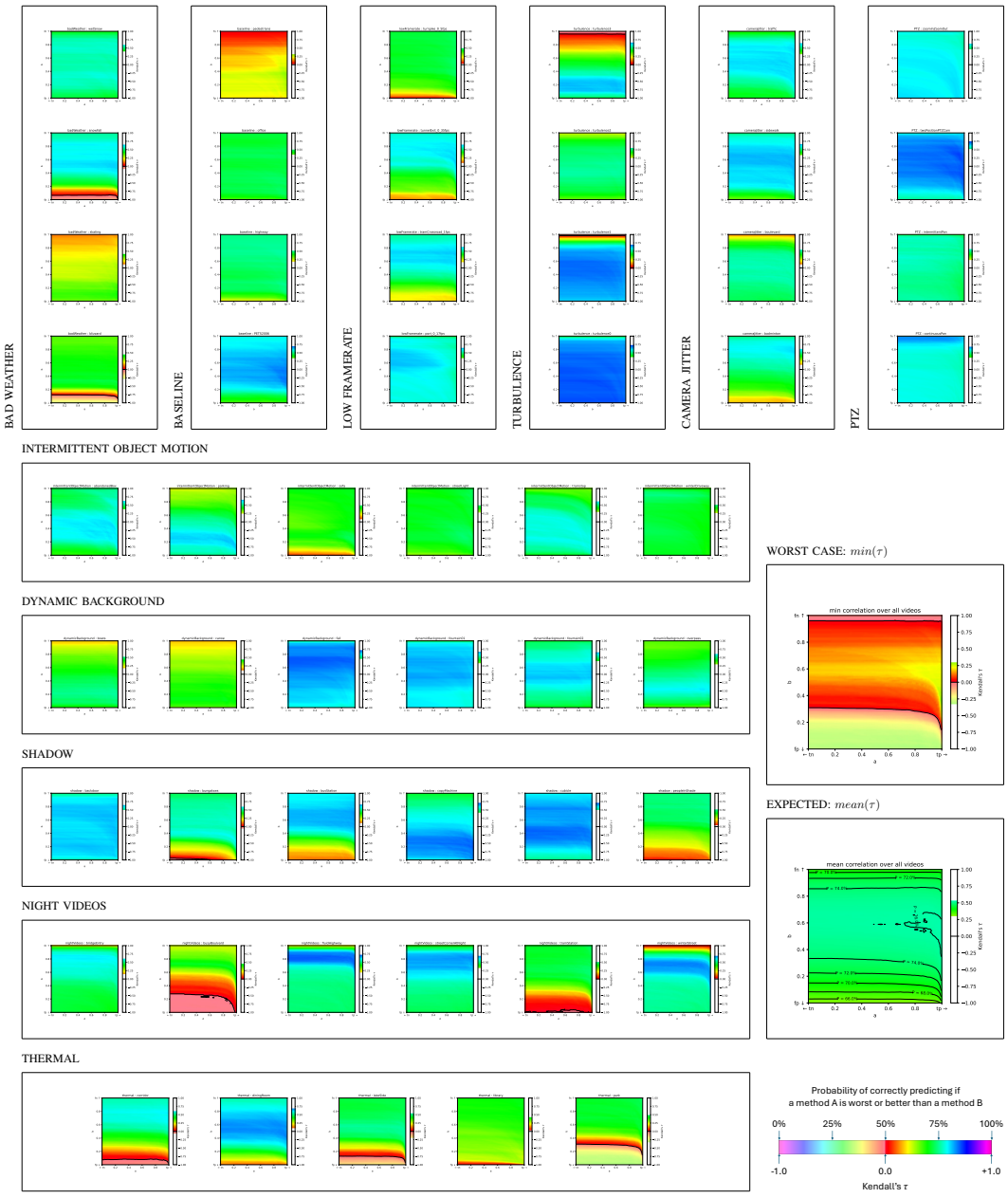


Fig. 20: Tiles showing the correlations (Kendall's τ), *w.r.t.* the application-specific preferences I, between the true ranking and the one predicted with the strategy "all†", for 53 videos, as well as in the behavior in the worst case and in average. These rankings involve 40 BGS methods.

F. Supplementary material for the 8th article

This appendix contains the supplementary material provided with my 8th selected article [8], presented in Chapter 8.

A. Supplementary material

A.1. Priors of the source domains

In this section, we provide further details about the priors of each source domain used in our experiments, we provide the histograms with respect to the semantic classes for all domains (each corresponding to a dataset) in Figure 9. As can be seen, there is a huge class imbalance with some of the classes. Some of them are especially underrepresented, like two wheels on the RainSnow, MTID/Drone, GRAM-RTM/M-30-HD and UT/Sherbrooke datasets.

A.2. Details about the networks

We use the TinyNet network for both the source and the domain discriminator models. This is a segmentation architecture first introduced in [203] for soccer player segmentation and later refined in [198]. Interestingly, due to its lightweight architecture, this network can be used for real-time inference on embedded devices, making it a great choice for real-world surveillance. Compared to the original architecture, we adapt it for the 720p input image size (compared to 1080p in the original publication). For that, only the values of the pooling layers applied in the pyramid pooling have to be adjusted. The architecture of the adapted network can be found in Figure 10.

A.3. Details about the domain discriminator model

Figure 11 shows an example of multi-domain images that we generated to train our domain discriminator model. As can be seen, four patches cropped from four randomly drawn images are combined to create training images, so that the domain discriminator model learns to recognise different domains within a same image. This is indeed an interesting feature since a target domain could combine some elements from different source domains (for example, the road of the target domain could look like the one from one source domain but the weather and lighting conditions may be those of another source domain).

A.4. Interpretation of posteriors

The interpretation given to the posteriors depends on the priors. In particular, $P(H \mid E) = P(H)$ when the evidence provides no information about the hypotheses. In fact, the priors can be seen as a reference, the posteriors being expressed w.r.t. it. Consider P' and P'', two probability measures sharing common likelihoods $P'(E \mid H) = P''(E \mid H)$. The posteriors $P'(H \mid E)$ and $P''(H \mid E)$ convey the same information, but are expressed w.r.t. different references $P'(H)$ and $P''(H)$. The change of reference is termed *prior shift* or *target shift*. Assuming non-zero priors, it is given by [94, 199]

$$P''(H \mid E) \propto \frac{P''(H)}{P'(H)} P'(H \mid E)$$

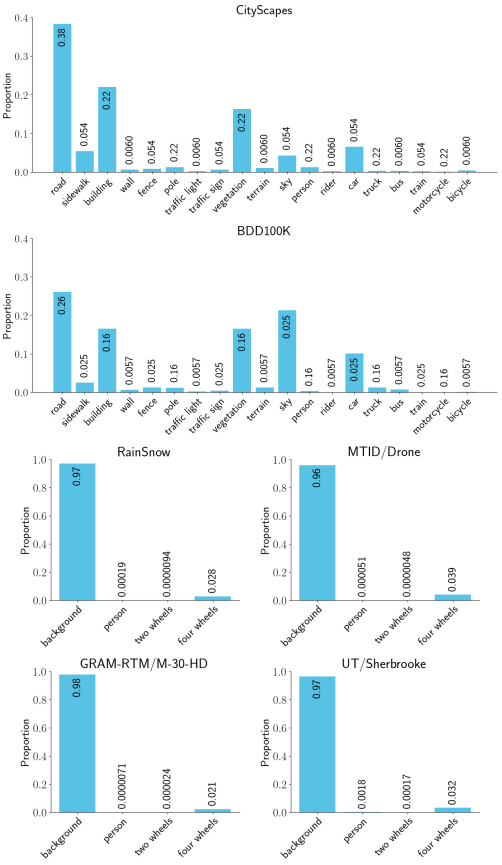


Figure 9. Histograms giving the priors, *i.e.* the proportion of each semantic class, for each domain (from top to bottom and from left to right: CityScapes, BDD100K, RainSnow, MTID/Drone, GRAM-RTM/M-30-HD and UT/Sherbrooke). All numbers are given up to 2 significant figures.

and, when $\mathbb{H}$ forms a partition of Ω, by

$$P''(H \mid E) = \frac{\frac{P''(H)}{P'(H)} P'(H \mid E)}{\sum_{H \in \mathbb{H}} \frac{P''(H)}{P'(H)} P'(H \mid E)}.$$

A.5. Validation of predicted posteriors

As stated in Section 4.1, it is important that the source and domain discriminator models output trustworthy posteriors. To verify that, we perform two tests: (1) we compare the class-by-class average of the posteriors estimated by our algorithm to the priors of the dataset—in particular, its test set (shown for one source model in Figure 12), and

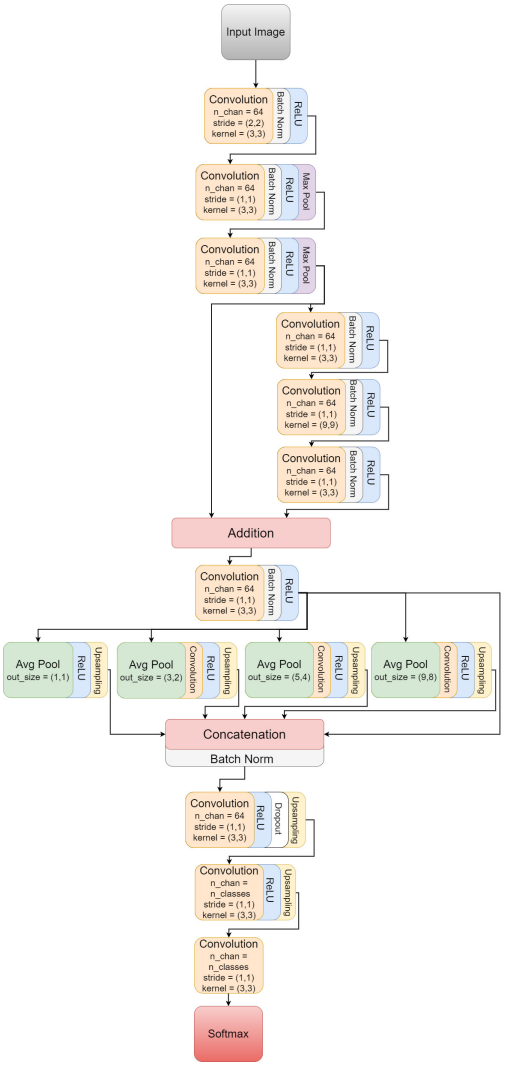


Figure 10. TinyNet architecture adapted to 720p images and an arbitrary number of output classes (either the semantic classes or the number of source domains).

(2) we verify that our model is well calibrated using calibration plots (shown for one source model and one semantic class in Figure 13).

Since both the average of the posteriors is comparable to the priors of the dataset and the calibration plots are close to the $y = x$ line, we are confident that the outputs of our algorithm correspond to posteriors. We perform these tests

Figure 11. Example of multi-domain image used to train our domain discriminator model. The image is obtained with a mosaic transform for images of the CityScapes and BDD100K datasets.

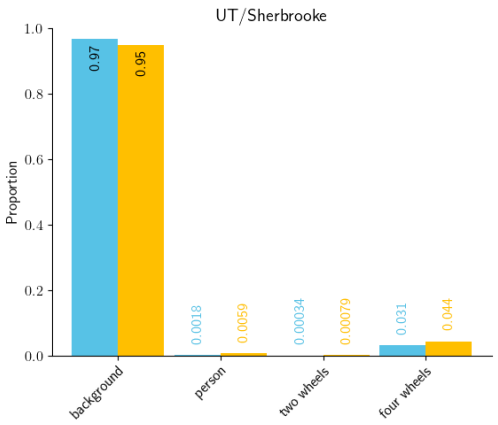


Figure 12. Histogram for the UT/Sherbrooke dataset comparing the priors of the test set (in blue) and the average of the posteriors obtained with our algorithm for each class (in orange). All numbers are given to 2 significant figures.

for all source and discriminator models.

A.6. Extra results

In this section, we provide extra results for both of our experiments. First, we provide the same graph as the one depicted in Figure 6 for the 3 remaining segmentation scores (*i.e.* balanced accuracy, mean IoU, and balanced mean IoU) in Figure 14. As can be seen, for all scores, our algorithm outperforms all other heuristics for each target domain.

Next, we provide the performance of our algorithm for our second experiment with 4 source domains for different λ_k in Table 1 for the accuracy and mean IOU, and in Table 2 for the balanced accuracy and balanced mean IOU. We compare our algorithm with the four source models (one model for each source domain), the random selection

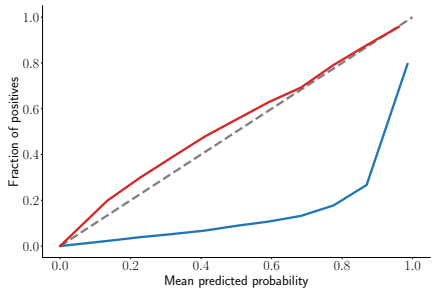


Figure 13. Calibration plot for the UT/Sherbrooke source model **before the target shift (in blue)** and **after the target shift (in red)** for the "four wheels" class. The dotted line corresponds to the perfect case where the output of the model correspond to the posteriors.

of source models, and the linear combination of posteriors. Note that the histograms presented in Figure 8 represent the column-by-column average values of these tables.

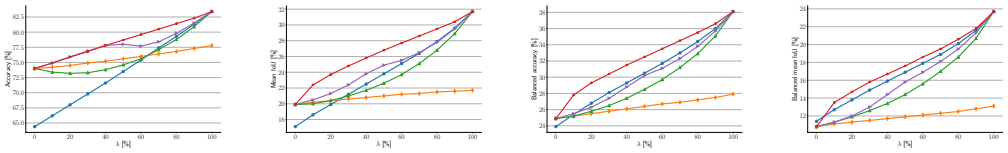


Figure 14. Results of our first experiment (all decisions made with the MAP strategy) showing from left to right: the accuracy, mean IoU, balanced accuracy, and balanced mean IoU. We compare the **source model trained on CityScapes (in blue)**, the **source model trained on BDD100K (in orange)**, the **random selection of source models (in green)**, the **linear combination of posteriors (in purple)**, and **our algorithm (in red)**. $\lambda = 0$ (resp. $= 100$) corresponds to BDD100K (resp. CityScapes) as target domain.

Table 1. **Results of our second experiment (all decisions made with the MAP strategy) for various target domains obtained from 4 source domains for various λ_k for the accuracy/mean IoU scores. We compare the different source models (SM) trained on their source domain, the random selection of source models, the linear combination of posteriors, and our algorithm.**

Parameters				Methods						
λ_1	λ_2	λ_3	λ_4	SM_1 RainSnow	SM_2 MTID/Drone	SM_3 GRAM-RTM/M-30-HD	SM_4 UT/Sherbrooke	Random selection of source models	Linear combination of posteriors	**Ours**
1	0	0	0	97.9/35.1	96.4/26.3	96.0/27.6	96.1/28.3	97.9/35.1	97.9/35.1	97.9/35.1
0	1	0	0	96.8/33.3	98.2/40.4	95.5/31.3	95.0/29.4	98.2/40.4	98.2/40.4	98.2/40.4
0	0	1	0	98.2/34.6	97.8/31.8	99.5/55.7	96.1/28.2	99.5/55.7	99.5/55.7	99.5/55.7
0	0	0	1	96.2/32.9	95.7/27.4	96.5/32.3	99.0/67.1	99.0/67.1	99.0/67.1	99.0/67.1
½	½	0	0	97.4/**35.0**	97.2/33.7	95.7/29.6	95.5/28.9	97.3/34.0	**97.7**/34.1	97.0/31.7
½	0	½	0	98.0/35.3	97.2/28.8	97.8/38.0	96.1/28.4	97.9/36.9	**98.5**/36.4	98.0/**39.3**
½	0	0	½	97.1/33.5	96.0/27.0	96.2/30.0	97.5/52.6	97.3/42.5	**98.2**/42.8	**98.2**/**57.4**
0	½	½	0	97.4/33.7	98.0/36.8	97.6/**43.1**	95.6/28.9	97.8/39.7	**98.4**/39.2	97.4/33.7
0	½	0	½	96.5/33.6	96.9/34.1	96.0/31.9	97.1/54.6	97.0/43.0	97.8/42.8	**98.0**/**60.4**
0	0	½	½	97.2/33.7	96.8/29.6	98.1/38.3	97.5/54.1	97.8/46.8	**98.7**/43.5	98.5/**62.0**
⅓	⅓	⅓	0	97.6/34.9	97.4/33.2	97.1/**36.4**	95.8/28.8	97.3/34.8	**98.0**/34.5	97.1/31.3
⅓	⅓	0	⅓	97.0/33.8	96.7/31.7	96.0/30.5	96.7/48.4	96.8/37.3	97.6/35.4	**97.7**/**55.3**
⅓	0	⅓	⅓	97.4/33.8	96.7/28.4	97.4/34.4	97.0/47.9	97.3/38.7	**98.2**/36.4	98.0/**56.3**
0	⅓	⅓	⅓	97.0/33.7	97.2/33.5	97.3/35.8	96.7/49.5	97.1/39.3	97.7/35.5	**97.8**/**57.3**
¼	¼	¼	¼	97.2/33.9	97.0/31.7	96.9/33.7	96.5/45.4	96.9/35.9	**97.8**/34.1	97.6/**53.9**

Table 2. **Results of our second experiment (all decisions made with the MAP strategy) for various target domains obtained from 4 source domains for various λ_k for the balanced accuracy/balanced mean IoU scores. We compare the different source models (SM) trained on their source domain, the random selection of source models, the linear combination of posteriors, and our algorithm.**

Parameters				Methods						
λ_1	λ_2	λ_3	λ_4	SM_1 RainSnow	SM_2 MTID/Drone	SM_3 GRAM-RTM/M-30-HD	SM_4 UT/Sherbrooke	Random selection of source models	Linear combination of posteriors	**Ours**
1	0	0	0	38.7/21.0	28.0/9.5	30.8/12.2	31.9/13.4	38.7/21.0	38.7/21.0	38.7/21.0
0	1	0	0	37.0/18.8	46.7/29.6	37.0/18.6	34.0/15.6	46.7/29.6	46.7/29.6	46.7/29.6
0	0	1	0	39.1/18.8	35.2/15.7	58.8/39.0	33.8/15.4	58.8/39.0	58.8/39.0	58.8/39.0
0	0	0	1	38.7/19.7	30.1/11.9	36.7/17.1	75.9/61.1	75.9/61.1	75.9/61.1	75.9/61.1
½	½	0	0	**38.5/20.8**	38.1/19.4	34.2/15.5	33.1/14.5	38.3/20.1	36.0/17.7	34.2/15.9
½	0	½	0	39.2/20.4	31.3/12.8	43.1/25.2	32.8/14.3	41.1/22.8	37.7/19.2	**43.4/25.3**
½	0	0	½	38.0/19.0	29.3/10.9	33.9/14.7	62.2/45.4	50.1/32.0	45.1/27.5	**66.5/50.4**
0	½	½	0	37.6/18.0	41.7/22.2	**48.3/30.3**	33.7/15.1	45.0/26.2	41.3/22.6	37.6/18.3
0	½	0	½	38.5/19.3	38.6/20.1	36.9/17.2	63.4/46.7	51.0/33.0	45.3/27.8	**70.1/54.5**
0	0	½	½	39.0/19.8	32.3/13.9	42.4/22.7	64.6/48.2	53.5/35.2	45.0/27.1	**74.0/58.8**
⅓	⅓	⅓	0	38.5/19.7	37.2/18.2	**41.7/23.6**	33.2/14.7	39.1/20.5	36.2/17.5	34.3/15.6
⅓	⅓	0	⅓	38.1/19.0	35.5/17.0	35.1/15.7	56.5/39.1	43.3/24.9	37.4/19.1	**63.6/47.3**
⅓	0	⅓	⅓	38.3/19.2	30.9/12.5	38.8/19.4	56.8/39.7	44.6/26.0	37.9/19.5	**66.1/49.9**
0	⅓	⅓	⅓	38.4/19.1	37.6/19.1	40.6/20.9	57.8/40.7	45.3/26.7	38.2/19.9	**66.3/49.9**
¼	¼	¼	¼	38.1/18.9	35.3/16.8	38.4/19.0	53.0/35.5	41.2/22.5	35.8/17.4	**62.0/45.1**

Bibliography

[1] S. Piérard and M. Van Droogenbroeck, "Summarizing the performances of a background subtraction algorithm measured on several videos," in *IEEE International Conference on Image Processing (ICIP)*, (Abu Dhabi, United Arab Emirates), pp. 3234–3238, October 2020.

[2] S. Piérard, A. Halin, A. Cioppa, A. Deliège, and M. Van Droogenbroeck, "Foundations of the theory of performance-based ranking," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (Nashville, Tennessee, USA), pp. 14293–14302, IEEE, June 2025.

[3] S. Piérard, A. Deliège, and M. 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), IEEE, June 2026.

[4] S. Piérard, A. Halin, A. Cioppa, A. Deliège, and M. Van Droogenbroeck, "The Tile: A 2D map of ranking scores for two-class classification," *arXiv*, vol. abs/2412.04309, 2024.

[5] A. Halin, S. Piérard, A. Cioppa, and M. Van Droogenbroeck, "A hitchhiker's guide to understanding performances of two-class classifiers," *arXiv*, vol. abs/2412.04377, 2024.

[6] S. Piérard, A. Deliège, and M. Van Droogenbroeck, "Multi-domain performance analysis with scores tailored to user preferences," *arXiv*, vol. abs/2512.08715, 2025.

[7] S. Piérard, A. Deliège, A. Halin, and M. 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)*, (Nantes, France), pp. 1–6, IEEE, June 2025.

[8] S. Piérard, A. Cioppa, A. Halin, R. Vandeghen, M. Zanella, B. Macq, S. Mahmoudi, and M. Van Droogenbroeck, "Mixture domain adaptation to improve semantic segmentation in real-world surveillance," in *IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, (Waikoloa, Hawaii, USA), pp. 22–31, IEEE, January 2023.

[9] A. Halin, S. Piérard, R. Vandeghen, B. Gérin, M. Zanella, M. Colot, J. Held, A. Cioppa, E. Jean, G. Bontempi, S. Mahmoudi, B. Macq, and M. Van Droogenbroeck, "Physically interpretable probabilistic domain characterization," in *Asian*

Conference on Computer Vision Workshops (ACCV Workshops), vol. 15482 of *Lecture Notes in Computer Science*, (Hanoi, Vietnam), pp. 17–35, Springer Nature Singapore, 2025.

[10] O. Barnich and M. Van Droogenbroeck, "ViBe: A universal background subtraction algorithm for video sequences," *IEEE Transactions on Image Processing*, vol. 20, pp. 1709–1724, June 2011.

[11] A. N. Kolmogorov, "Sur la notion de la moyenne," *Atti della Academia Nazionale dei Lincei*, vol. 12, no. 9, pp. 388–391, 1930.

[12] M. Fréchet, "Sur diverses définitions de la moyenne d'un élément aléatoire de nature quelconque," *Giornale dell'Istituto Italiano degli Attuari*, vol. 19, pp. 1–15, 1956.

[13] H. Karcher, "Riemannian center of mass and mollifier smoothing," *Communications on Pure and Applied Mathematics*, vol. 30, pp. 509–541, September 1977.

[14] T. Bouwmans, F. El Baf, and B. Vachon, "Statistical background modeling for foreground detection: A survey," in *Handbook of Pattern Recognition and Computer Vision (volume 4)*, ch. 3, pp. 181–199, World Scientific Publishing, January 2010.

[15] T. Bouwmans, "Traditional and recent approaches in background modeling for foreground detection: An overview," *Computer Science Review*, vol. 11-12, pp. 31–66, May 2014.

[16] T. Bouwmans, S. Javed, M. Sultana, and S. K. Jung, "Deep neural network concepts for background subtraction:a systematic review and comparative evaluation," *Neural Networks*, vol. 117, pp. 8–66, September 2019.

[17] S. Elhabian, K. El-Sayed, and S. Ahmed, "Moving object detection in spatial domain using background removal techniques – State-of-art," *Recent Patents on Computer Science*, vol. 1, pp. 32–54, January 2008.

[18] P.-M. Jodoin, S. Piérard, Y. Wang, and M. Van Droogenbroeck, "Overview and benchmarking of motion detection methods," in *Background Modeling and Foreground Detection for Video Surveillance*, ch. 24, Chapman and Hall/CRC, July 2014.

[19] A. Sobral and A. Vacavant, "A comprehensive review of background subtraction algorithms evaluated with synthetic and real videos," *Computer Vision and Image Understanding*, vol. 122, pp. 4–21, May 2014.

[20] N. Vaswani, T. Bouwmans, S. Javed, and P. Narayanamurthy, "Robust subspace learning: Robust PCA, robust subspace tracking, and robust subspace recovery," *IEEE Signal Processing Magazine*, vol. 35, pp. 32–55, July 2018.

[21] N. Goyette, P.-M. Jodoin, F. Porikli, J. Konrad, and P. Ishwar, "Changedetection.net: A new change detection benchmark dataset," in *IEEE International*

Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), (Providence, RI, USA), pp. 1–8, Institute of Electrical and Electronics Engineers (IEEE), June 2012.

[22] N. Goyette, P.-M. Jodoin, F. Porikli, J. Konrad, and P. Ishwar, "A novel video dataset for change detection benchmarking," *IEEE Transactions on Image Processing*, vol. 23, pp. 4663–4679, November 2014.

[23] C. Cuevas, E. M. Yáñez, and N. García, "Labeled dataset for integral evaluation of moving object detection algorithms: LASIESTA," *Computer Vision and Image Understanding*, vol. 152, pp. 103–117, November 2016.

[24] J. Davis and M. Goadrich, "The relationship between precision-recall and ROC curves," in *International Conference on Machine Learning (ICML)*, (Pittsburgh, Pennsylvania), pp. 233–240, ML Research Press, June 2006.

[25] G. Forman and M. Scholz, "Apples-to-apples in cross-validation studies: Pitfalls in classifier performance measurement," *ACM SIGKDD Explorations Newsletter*, vol. 12, pp. 49–57, November 2010.

[26] K. Boyd, V. Costa, J. Davis, and D. Page, "Unachievable region in precision-recall space and its effect on empirical evaluation," in *International Conference on Machine Learning (ICML)*, (Edinburgh, UK), pp. 639–646, June 2012.

[27] C. Goutte and E. Gaussier, "A probabilistic interpretation of precision, recall and F-score, with implication for evaluation," in *Advances in Information Retrieval (Proceedings of ECIR)*, vol. 3408 of *Lecture Notes in Computer Science*, pp. 345–359, Springer, March 2005.

[28] J. Bertrand, *Calcul des probabilités*. Gauthier-Villars et fils, 1889.

[29] P. A. Flach, "The geometry of ROC space: Understanding machine learning metrics through ROC isometrics," in *International Conference on Machine Learning (ICML)*, (Washington, District of Columbia, USA), pp. 194–201, ML Research Press, August 2003.

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

[31] J. D. Marshall, H. Aguinis, and J. R. Beltran, "Theories of performance: A review and integration," *Academy of Management Annals*, vol. 18, pp. 600–625, July 2024.

[32] A. N. Kolmogorov, *Grundbegriffe der Wahrscheinlichkeitsrechnung*, vol. 2 of *Ergebnisse der Mathematik und Ihrer Grenzgebiete*. Berlin, Germany: Springer Berlin Heidelberg, 1933.

[33] A. N. Kolmogorov, *Foundations of the theory of probability*. New York City, New York, USA: Chelsea Publishing Company, 1950.

[34] M. Yin, Z. Yao, Y. Cao, X. Li, Z. Zhang, S. Lin, and H. Hu, "Disentangled non-local neural networks," in *European Conference on Computer Vision (ECCV)*, vol. 12360 of *Lecture Notes in Computer Science*, pp. 191–207, Springer International Publishing, 2020.

[35] E. Romera, J. M. Alvarez, L. M. Bergasa, and R. Arroyo, "ERFNet: Efficient residual factorized ConvNet for real-time semantic segmentation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, pp. 263–272, January 2018.

[36] R. C. Prati, G. E. A. P. A. Batista, and M. C. Monard, "A survey on graphical methods for classification predictive performance evaluation," *IEEE Transactions on Knowledge And Data Engineering*, vol. 23, pp. 1601–1618, November 2011.

[37] S.-S. Choi, S.-H. Cha, and C. C. a. Tappert, "A survey of binary similarity and distance measures," *Journal of Systemics, Cybernetics and Informatics*, vol. 8, no. 1, pp. 43–48, 2010.

[38] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing & Management*, vol. 45, pp. 427–437, July 2009.

[39] M. Gösgens, A. Zhiyanov, A. Tikhonov, and L. Prokhorenkova, "Good classification measures and how to find them," in *Advances in Neural Information Processing Systems (NeurIPS)*, (Vancouver, Canada), pp. 1–12, Curran Associates, Inc., 2021.

[40] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge And Data Engineering*, vol. 22, pp. 1345–1359, October 2010.

[41] A. Cioppa, S. Giancola, V. Somers, F. Magera, X. Zhou, H. Mkhallati, A. Deliège, J. Held, C. Hinojosa, A. M. Mansourian, P. Miralles, O. Barnich, C. De Vleeschouwer, A. Alahi, B. Ghanem, M. Van Droogenbroeck, A. Kamal, A. Maglo, A. Clapés, A. Abdelaziz, A. Xarles, A. Orcesi, A. Scott, B. Liu, B. Lim, C. Chen, F. Deuser, F. Yan, F. Yu, G. Shitrit, G. Wang, G. Choi, H. Kim, H. Guo, H. Fahrudin, H. Koguchi, H. Ardö, I. Salah, I. Yerushalmy, I. Muhammad, I. Uchida, I. Be'ery, J. Rabarisoa, J. Lee, J. Fu, J. Yin, J. Xu, J. Nang, J. Denize, J. Li, J. Zhang, J. Kim, K. Synowiec, K. Kobayashi, K. Zhang, K. Habel, K. Nakajima, L. Jiao, L. Ma, L. Wang, L. Wang, M. Li, M. Zhou, M. Nasr, M. Abdelwahed, M. Liashuha, N. Falaleev, N. Oswald, Q. Jia, Q.-C. Pham, R. Song, R. Hérault, R. Peng, R. Chen, R. Liu, R. Baikulov, R. Fukushima, S. Escalera, S. Lee, S. Chen, S. Ding, T. Someya, T. B. Moeslund, T. Li, W. Shen, W. Zhang, W. Li, W. Dai, W. Luo, W. Zhao, W. Zhang, X. Yang, Y. Ma, Y. Joo, Y. Zeng, Y. Gan, Y. Zhu, Y. Zhong, Z. Ruan, Z. Li, Z. Huang, and Z. Meng, "SoccerNet 2023 challenges results," *Sports Engineering*, vol. 27, pp. 1–18, July 2024.

[42] M. Kristan, J. Matas, A. Leonardis, M. Felsberg, L. Cehovin, G. Fernandez, T. Vojir, G. Hager, G. Nebehay, R. Pflugfelder, A. Gupta, A. Bibi, A. Lukezic, A. Garcia-Martin, A. Saffari, A. Petrosino, and A. Solis Montero, "The visual object track-

ing VOT2015 challenge results," in *IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, (Santiago, Chile), pp. 564–586, December 2015.

[43] M. Kristan, J. Matas, P. Tokmakov, M. Felsberg, L. Čehovin Zajc, A. Lukežič, K.-T. Tran, X.-S. Vu, J. Bjorklund, H. J. Chang, and G. Fernandez, "The second visual object tracking segmentation VOTS2024 challenge results," in *European Conference on Computer Vision Workshops (ECCV Workshops)*, (Milan, Italy), October 2024.

[44] L. Maier-Hein, M. Eisenmann, A. Reinke, S. Onogur, M. Stankovic, P. Scholz, T. Arbel, H. Bogunovic, A. P. Bradley, A. Carass, C. Feldmann, A. F. Frangi, P. M. Full, B. van Ginneken, A. Hanbury, K. Honauer, M. Kozubek, B. A. Landman, K. März, O. Maier, K. Maier-Hein, B. H. Menze, H. Müller, P. F. Neher, W. Niessen, N. Rajpoot, G. C. Sharp, K. Sirinukunwattana, S. Speidel, C. Stock, D. Stoyanov, A. A. Taha, F. van der Sommen, C.-W. Wang, M.-A. Weber, G. Zheng, P. Jannin, and A. Kopp-Schneider, "Why rankings of biomedical image analysis competitions should be interpreted with care," *Nature Communications*, vol. 9, pp. 1–13, December 2018.

[45] T. T. D. Nguyen, H. Rezatofighi, B.-N. Vo, B.-T. Vo, S. Savarese, and I. Reid, "How trustworthy are performance evaluations for basic vision tasks?," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, pp. 8538–8552, July 2023.

[46] P. P. Texel, "Measure, metric, and indicator: An object-oriented approach for consistent terminology," *Proc. IEEE Southeastcon*, pp. 1–5, April 2013.

[47] G. Canbek, S. Sagiroglu, T. T. Temizel, and N. Baykal, "Binary classification performance measures/metrics: A comprehensive visualized roadmap to gain new insights," in *International Conference on Computer Science and Engineering (UBMK)*, (Antalya, Turkey), pp. 821–826, IEEE, October 2017.

[48] B. A. Davey and H. A. Priestley, *Introduction to Lattices and Order*. Cambridge University Press, second ed., April 2002.

[49] G. Grätzer, *General Lattice Theory*. Birkhäuser Basel, second ed., 2003.

[50] M. G. Kendall, "A new measure of rank correlation," *Biometrika*, vol. 30, p. 81, June 1938.

[51] C. Spearman, "The proof and measurement of association between two things," *The American Journal of Psychology*, vol. 15, pp. 72–101, January 1904.

[52] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M.

Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, A. Vijaykumar, A. P. Bardelli, A. Rothberg, A. Hilboll, A. Kloeckner, A. Scopatz, A. Lee, A. Rokem, C. N. Woods, C. Fulton, C. Masson, C. Häggström, C. Fitzgerald, D. A. Nicholson, D. R. Hagen, D. V. Pasechnik, E. Olivetti, E. Martin, E. Wieser, F. Silva, F. Lenders, F. Wilhelm, G. Young, G. A. Price, G.-L. Ingold, G. E. Allen, G. R. Lee, H. Audren, I. Probst, J. P. Dietrich, J. Silterra, J. T. Webber, J. Slavič, J. Nothman, J. Buchner, J. Kulick, J. L. Schönberger, J. V. de Miranda Cardoso, J. Reimer, J. Harrington, J. L. C. Rodríguez, J. Nunez-Iglesias, J. Kuczynski, K. Tritz, M. Thoma, M. Newville, M. Kümmerer, M. Bolingbroke, M. Tartre, M. Pak, N. J. Smith, N. Nowaczyk, N. Shebanov, O. Pavlyk, P. A. Brodtkorb, P. Lee, R. T. McGibbon, R. Feldbauer, S. Lewis, S. Tygier, S. Sievert, S. Vigna, S. Peterson, S. More, T. Pudlik, T. Oshima, T. J. Pingel, T. P. Robitaille, T. Spura, T. R. Jones, T. Cera, T. Leslie, T. Zito, T. Krauss, U. Upadhyay, Y. O. Halchenko, and Y. Vázquez-Baeza, "SciPy 1.0: fundamental algorithms for scientific computing in Python," *Nature Methods*, vol. 17, pp. 261–272, Feb. 2020.

[53] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, pp. 37–46, April 1960.

[54] E. B. Fowlkes and C. L. Mallows, "A method for comparing two hierarchical clusterings," *Journal of the American Statistical Association*, vol. 78, pp. 553–569, September 1983.

[55] D. Ballabio, F. Grisoni, and R. Todeschini, "Multivariate comparison of classification performance measures," *Chemometrics and Intelligent Laboratory Systems*, vol. 174, pp. 33–44, March 2018.

[56] A. R. Conn, K. Scheinberg, and L. N. Vicente, *Introduction to Derivative-Free Optimization*, vol. 8. Philadelphia, Pennsylvania, USA: Society for Industrial and Applied Mathematics, January 2009.

[57] P. Christen, D. J. Hand, and N. Kirielle, "A review of the F-measure: Its history, properties, criticism, and alternatives," *ACM Computing Surveys*, vol. 56, pp. 1–24, October 2023.

[58] L. Ferrer, "Analysis and comparison of classification metrics," *arXiv*, vol. abs/2209.05355, 2022.

[59] C. Ferri, J. Hernández-Orallo, and R. Modroiu, "An experimental comparison of performance measures for classification," *Pattern Recognition Letters*, vol. 30, pp. 27–38, January 2009.

[60] M. Hossin and M. N. Sulaiman, "A review on evaluation metrics for data classification evaluations," *International Journal of Data Mining & Knowledge Management Process*, vol. 5, pp. 1–11, March 2015.

[61] N. Japkowicz and M. Shah, *Evaluating Learning Algorithms: A Classification Perspective*. Cambridge University Press, January 2011.

[62] D. M. W. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness & correlation," *Journal of Machine Learning Technologies*, vol. 2, pp. 37–63, December 2011.

[63] Y. Liu, Y. Zhou, S. Wen, and C. Tang, "A strategy on selecting performance metrics for classifier evaluation," *International Journal of Mobile Computing and Multimedia Communications*, vol. 6, pp. 20–35, October 2014.

[64] M. Fréchet, "Les éléments aléatoires de nature quelconque dans un espace distancié," *Annales de l'institut Henri Poincaré*, vol. 10, no. 4, pp. 215–310, 1948.

[65] B. Garcia-Garcia, T. Bouwmans, and A. J. Rosales Silva, "Background subtraction in real applications: Challenges, current models and future directions," *Computer Science Review*, vol. 35, pp. 1–42, Feb. 2020.

[66] Y. Wang, P.-M. Jodoin, F. Porikli, J. Konrad, Y. Benezeth, and P. Ishwar, "CDnet 2014: An expanded change detection benchmark dataset," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, (Columbus, Ohio, USA), pp. 393–400, Institute of Electrical and Electronics Engineers (IEEE), June 2014.

[67] S. Piérard, A. Halin, F. Marelli, S. Pernas, and J. 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 2025.

[68] R. J. Hogan, C. A. T. Ferro, I. T. Jolliffe, and D. B. Stephenson, "Equitability revisited: Why the "equitable threat score" is not equitable," *Weather and Forecasting*, vol. 25, pp. 710–726, April 2010.

[69] M. J. Warrens, "The effect of combining categories on Bennett, Alpert and Goldstein's S," *Statistical Methodology*, vol. 9, pp. 341–352, May 2012.

[70] T. F. Heston, "Standardizing predictive values in diagnostic imaging research," *Journal of Magnetic Resonance Imaging*, vol. 33, pp. 505–505, January 2011.

[71] I. A. Gardner and M. Greiner, "Receiver-operating characteristic curves and likelihood ratios: improvements over traditional methods for the evaluation and application of veterinary clinical pathology tests," *Veterinary Clinical Pathology*, vol. 35, pp. 8–17, March 2006.

[72] A. S. Glas, J. Lijmer, M. H. Prins, G. Bonsel, and P. M. M. Bossuyt, "The diagnostic odds ratio: a single indicator of test performance," *Journal of Clinical Epidemiology*, vol. 56, pp. 1129–1135, November 2003.

[73] C. D. Brown and H. T. Davis, "Receiver operating characteristics curves and related decision measures: A tutorial," *Chemometrics and Intelligent Laboratory Systems*, vol. 80, pp. 24–38, January 2006.

[74] D. G. Altman and M. Bland, "Diagnostic tests 2: Predictive values," *British Medical Journal*, vol. 309, pp. 102–102, July 1994.

[75] U. Kaymak, A. Ben-David, and R. Potharst, "The AUK: A simple alternative to the AUC," *Engineering Applications of Artificial Intelligence*, vol. 25, pp. 1082–1089, August 2012.

[76] D. S. Wilks, *Statistical methods in the atmospheric sciences*. Elsevier, fourth ed., 2020.

[77] T. Byrt, J. Bishop, and J. B. Carlin, "Bias, prevalence and kappa," *Journal of Clinical Epidemiology*, vol. 46, pp. 423–429, May 1993.

[78] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, "The balanced accuracy and its posterior distribution," in *IEEE International Conference on Pattern Recognition (ICPR)*, (Istanbul, Turkey), pp. 3121–3124, Institute of Electrical and Electronics Engineers (IEEE), August 2010.

[79] M. Kubat, "Addressing the curse of imbalanced training sets: one-sided selection," in *International Conference on Machine Learning (ICML)*, pp. 179–186, Morgan Kaufmann, 1997.

[80] R. Barandela, J. S. Sánchez, V. García, and E. Rangel, "Strategies for learning in class imbalance problems," *Pattern Recognition*, vol. 36, pp. 849–851, March 2003.

[81] X. Guo, Y. Yin, C. Dong, G. Yang, and G. Zhou, "On the class imbalance problem," in *International Conference on Natural Computation*, (Jinan, China), pp. 192–201, Institute of Electrical and Electronics Engineers (IEEE), October 2008.

[82] W. J. Youden, "Index for rating diagnostic tests," *Cancer*, vol. 3, pp. 32–35, January 1950.

[83] R. Fluss, D. Faraggi, and B. Reiser, "Estimation of the Youden index and its associated cutoff point," *Biometrical Journal*, vol. 47, pp. 458–472, August 2005.

[84] M. Wimmer, B. Radig, and M. Beetz, "A person and context specific approach for skin color classification," in *International Conference on Pattern Recognition (ICPR)*, (Hong Kong, China), pp. 39–42, IEEE, 2006.

[85] D. M. W. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," *arXiv*, vol. abs/2010.16061, 2020.

[86] M. Burset and R. Guigó, "Evaluation of gene structure prediction programs," *Genomics*, vol. 34, pp. 353–367, June 1996.

[87] M. Sitarz, "Extending F1 metric, probabilistic approach," *Advances in Artificial Intelligence and Machine Learning*, vol. 3, no. 2, pp. 1025–1038, 2023.

[88] I. Z. Batyrshin, N. Kubysheva, V. Solovyev, and L. A. Villa-Vargas, "Visualization of similarity measures for binary data and 2x2 tables," *Computación y Sistemas*, vol. 20, pp. 345–353, September 2016.

[89] M. Brusco, J. D. Cradit, and D. Steinley, "A comparison of 71 binary similarity coefficients: The effect of base rates," *PLOS ONE*, vol. 16, p. e0247751, April 2021.

[90] J. C. Gower and P. Legendre, "Metric and euclidean properties of dissimilarity coefficients," *Journal of Classification*, vol. 3, pp. 5–48, March 1986.

[91] I. R. Mejia and I. Batyrshin, "Towards a classification of binary similarity measures," in *Advances in Soft Computing*, vol. 10632 of *Lecture Notes in Computer Science*, pp. 325–335, Springer International Publishing, 2018.

[92] M. J. Warrens, "A comparison of multi-way similarity coefficients for binary sequences," *International Journal of Research and Reviews in Applied Sciences*, vol. 16, pp. 64–75, July 2013.

[93] F. B. Baulieu, "A classification of presence/absence based dissimilarity coefficients," *Journal of Classification*, vol. 6, pp. 233–246, December 1989.

[94] T. Sipka, M. Sulc, and J. Matas, "The hitchhiker's guide to prior-shift adaptation," in *IEEE Winter Conference on Applications of Computer Vision (WACV)*, (Waikoloa, Hawaii, USA), pp. 2031–2039, IEEE, January 2022.

[95] F. Provost and T. Fawcett, "Robust classification for imprecise environments," *Machine Learning*, vol. 42, pp. 203–231, March 2001.

[96] W. A. Scott, "Reliability of content analysis: The case of nominal scale coding," *Public Opinion Quarterly*, vol. 19, no. 3, pp. 321–325, 1955.

[97] J. L. Fleiss, "Measuring nominal scale agreement among many raters," *Psychological Bulletin*, vol. 76, pp. 378–382, November 1971.

[98] Y. Liu, Y. Xu, N. B. Shah, and A. Singh, "Integrating rankings into quantized scores in peer review," *Transactions on Machine Learning Research*, pp. 1–28, October 2022.

[99] A. Tharwat, "Classification assessment methods," *Applied Computing and Informatics*, vol. 17, no. 1, pp. 168–192, 2021.

[100] F. S. Nahm, "Receiver operating characteristic curve: overview and practical use for clinicians," *Korean Journal of Anesthesiology*, vol. 75, pp. 25–36, Feb. 2022.

[101] T. Fawcett, "ROC graphs: Notes and practical considerations for researchers," *Machine Learning*, vol. 31, pp. 1–38, 2004.

[102] D. Bamber, "The area above the ordinal dominance graph and the area below the receiver operating characteristic graph," *Journal of Mathematical Psychology*, vol. 12, pp. 387–415, November 1975.

[103] W. M. Briggs and R. Zaretzki, "The skill plot: A graphical technique for evaluating continuous diagnostic tests," *Biometrics*, vol. 64, pp. 250–256, Feb. 2008.

[104] A. M. Carrington, D. G. Manuel, P. W. Fieguth, T. Ramsay, V. Osmani, B. Wernly, C. Bennett, S. Hawken, O. Magwood, Y. Sheikh, M. McInnes, and A. Holzinger, "Deep ROC analysis and AUC as balanced average accuracy, for improved classifier selection, audit and explanation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, pp. 329–341, January 2023.

[105] A. Ahmadzadeh, D. J. Kempton, P. C. Martens, and R. A. Angryk, "Contingency space: A semimetric space for classification evaluation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, pp. 1501–1513, Feb. 2023.

[106] J. M. Hughes-Oliver, "Population and empirical PR curves for assessment of ranking algorithms," *arXiv*, vol. abs/1810.08635, 2018.

[107] A. Menon, H. Narasimhan, S. Agarwal, and S. Chawla, "On the statistical consistency of algorithms for binary classification under class imbalance," in *International Conference on Machine Learning (ICML)*, vol. 28 of *Proceedings of Machine Learning Research*, (Atlanta, Georgia, USA), pp. 603–611, June 2013.

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

[109] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (Las Vegas, Nevada, USA), pp. 3213–3223, Institute of Electrical and Electronics Engineers (IEEE), June 2016.

[110] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ADE20K dataset," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (Honolulu, Hawaii, USA), pp. 5122–5130, Institute of Electrical and Electronics Engineers (IEEE), July 2017.

[111] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes (VOC) challenge," *International Journal of Computer Vision*, vol. 88, pp. 303–338, June 2010.

[112] H. Caesar, J. Uijlings, and V. Ferrari, "COCO-stuff: Thing and stuff classes in context," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (Salt Lake City, Utah, USA), pp. 1209–1218, Institute of Electrical and Electronics Engineers (IEEE), June 2018.

[113] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (Boston, Massachusetts, USA), pp. 3431–3440, Institute of Electrical and Electronics Engineers (IEEE), June 2015.

[114] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, pp. 640–651, April 2017.

[115] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv*, vol. abs/1706.05587, 2017.

[116] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (Honolulu, Hawaii, USA), pp. 6230–6239, Institute of Electrical and Electronics Engineers (IEEE), July 2017.

[117] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *European Conference on Computer Vision (ECCV)*, vol. 11211 of *Lecture Notes in Computer Science*, (Munich, Germany), pp. 833–851, Springer International Publishing, September 2018.

[118] H. Zhang, K. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi, and A. Agrawal, "Context encoding for semantic segmentation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (Salt Lake City, Utah, USA), pp. 7151–7160, Institute of Electrical and Electronics Engineers (IEEE), June 2018.

[119] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (Salt Lake City, Utah, USA), pp. 7794–7803, Institute of Electrical and Electronics Engineers (IEEE), June 2018.

[120] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "BiSeNet: Bilateral segmentation network for real-time semantic segmentation," in *European Conference on Computer Vision (ECCV)*, vol. 11217 of *Lecture Notes in Computer Science*, pp. 325–341, Springer International Publishing, 2018.

[121] H. Zhao, X. Qi, X. Shen, J. Shi, and J. Jia, "ICNet for real-time semantic segmentation on high-resolution images," in *European Conference on Computer Vision (ECCV)*, vol. 11207 of *Lecture Notes in Computer Science*, pp. 418–434, Springer International Publishing, 2018.

[122] H. Zhao, Y. Zhang, S. Liu, J. Shi, C. C. Loy, D. Lin, and J. Jia, "PSANet: Point-wise spatial attention network for scene parsing," in *European Conference on Computer Vision (ECCV)*, vol. 11213 of *Lecture Notes in Computer Science*, pp. 270–286, Springer International Publishing, 2018.

[123] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun, "Unified perceptual parsing for scene understanding," in *European Conference on Computer Vision (ECCV)*, vol. 11209 of *Lecture Notes in Computer Science*, pp. 432–448, Springer International Publishing, 2018.

[124] H. Wu, J. Zhang, K. Huang, K. Liang, and Y. Yu, "FastFCN: Rethinking dilated convolution in the backbone for semantic segmentation," *arXiv*, vol. abs/1903.11816, 2019.

[125] R. P. K. Poudel, S. Liwicki, and R. Cipolla, "Fast-SCNN: Fast semantic segmentation network," *arXiv*, vol. abs/1902.04502, 2019.

[126] L. Huang, Y. Yuan, J. Guo, C. Zhang, X. Chen, and J. Wang, "Interlaced sparse self-attention for semantic segmentation," *arXiv*, vol. abs/1907.12273, 2019.

[127] J. He, Z. Deng, L. Zhou, Y. Wang, and Y. Qiao, "Adaptive pyramid context network for semantic segmentation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (Long Beach, California, USA), pp. 7511–7520, Institute of Electrical and Electronics Engineers (IEEE), June 2019.

[128] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (Long Beach, California, USA), pp. 3141–3149, Institute of Electrical and Electronics Engineers (IEEE), June 2019.

[129] A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollar, "Panoptic segmentation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (Long Beach, California, USA), pp. 9396–9405, Institute of Electrical and Electronics Engineers (IEEE), June 2019.

[130] Z. Zhu, M. Xu, S. Bai, T. Huang, and X. Bai, "Asymmetric non-local neural networks for semantic segmentation," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, (Seoul, South Korea), pp. 593–602, Institute of Electrical and Electronics Engineers (IEEE), October 2019.

[131] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "CCNet: Criss-cross attention for semantic segmentation," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, (Seoul, South Korea), pp. 603–612, Institute of Electrical and Electronics Engineers (IEEE), October 2019.

[132] J. He, Z. Deng, and Y. Qiao, "Dynamic multi-scale filters for semantic segmentation," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, (Seoul, South Korea), pp. 3561–3571, Institute of Electrical and Electronics Engineers (IEEE), October 2019.

[133] X. Li, Z. Zhong, J. Wu, Y. Yang, Z. Lin, and H. Liu, "Expectation-maximization attention networks for semantic segmentation," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, (Seoul, South Korea), pp. 9166–9175, Institute of Electrical and Electronics Engineers (IEEE), October 2019.

[134] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "GCNet: Non-local networks meet squeeze-excitation networks and beyond," in *IEEE/CVF International Conference on Computer Vision Workshops (ICCV Workshops)*, (Seoul, South Korea),

pp. 1971–1980, Institute of Electrical and Electronics Engineers (IEEE), October 2019.

[135] A. Kirillov, Y. Wu, K. He, and R. Girshick, "PointRend: Image segmentation as rendering," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (Seattle, Washington, USA), pp. 9796–9805, Institute of Electrical and Electronics Engineers (IEEE), June 2020.

[136] Y. Yuan, X. Chen, and J. Wang, "Object-contextual representations for semantic segmentation," in *European Conference on Computer Vision (ECCV)*, vol. 12351 of *Lecture Notes in Computer Science*, pp. 173–190, Springer International Publishing, 2020.

[137] T. Wu, S. Tang, R. Zhang, J. Cao, and Y. Zhang, "CGNet: A light-weight context guided network for semantic segmentation," *IEEE Transactions on Image Processing*, vol. 30, pp. 1169–1179, 2021.

[138] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. Torr, and L. Zhang, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (Nashville, Tennessee, USA), pp. 6877–6886, Institute of Electrical and Electronics Engineers (IEEE), June 2021.

[139] C. Yu, C. Gao, J. Wang, G. Yu, C. Shen, and N. Sang, "BiSeNet V2: Bilateral network with guided aggregation for real-time semantic segmentation," *International Journal of Computer Vision*, vol. 129, pp. 3051–3068, September 2021.

[140] M. Fan, S. Lai, J. Huang, X. Wei, Z. Chai, J. Luo, and X. Wei, "Rethinking BiSeNet for real-time semantic segmentation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (Nashville, Tennessee, USA), pp. 9711–9720, Institute of Electrical and Electronics Engineers (IEEE), June 2021.

[141] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, (Montréal, Canada), pp. 12159–12168, Institute of Electrical and Electronics Engineers (IEEE), October 2021.

[142] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for semantic segmentation," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, (Montréal, Canada), pp. 7242–7252, Institute of Electrical and Electronics Engineers (IEEE), October 2021.

[143] W. Zhang, J. Pang, K. Chen, and C. C. Loy, "K-Net: Towards unified image segmentation," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, pp. 10326–10338, Curran Associates, Inc., December 2021.

[144] B. Cheng, A. G. Schwing, and A. Kirillov, "Per-pixel classification is not all you need for semantic segmentation," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, pp. 17864–17875, Curran Associates, Inc., December 2021.

[145] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, (Virtual conference), pp. 12077–12090, Curran Associates, Inc., December 2021.

[146] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (New Orleans, Louisiana, USA), pp. 1280–1289, Institute of Electrical and Electronics Engineers (IEEE), June 2022.

[147] M. Xu, Z. Zhang, F. Wei, H. Hu, and X. Bai, "Side adapter network for open-vocabulary semantic segmentation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (New Orleans, Louisiana, USA), pp. 2945–2954, Institute of Electrical and Electronics Engineers (IEEE), June 2023.

[148] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell, "BDD100K: A diverse driving dataset for heterogeneous multitask learning," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, (Seattle, Washington, USA), pp. 2633–2642, Institute of Electrical and Electronics Engineers (IEEE), June 2020.

[149] G. Wilson and D. J. Cook, "A survey of unsupervised deep domain adaptation," *ACM Transactions on Intelligent Systems and Technology*, vol. 11, pp. 1–46, July 2020.

[150] I. Gulrajani and D. Lopez-Paz, "In search of lost domain generalization," *arXiv*, vol. abs/2007.01434, 2020.

[151] S. Piérard, A. Halin, F. Marelli, S. Pernas, and J. Pierre, "Sorbetto: a Python library for producing classification tiles," 2025. https://github.com/uliege-performance/sorbetto, DOI: 10.5281/zenodo.17591788.

[152] Y. Ding, B. Jiang, A. Yu, A. Zheng, and J. Liang, "Which model to transfer? a survey on transferability estimation," *arXiv*, vol. abs/2402.15231, 2024.

[153] C. V. Nguyen, T. Hassner, M. W. Seeger, and C. Archambeau, "LEEP: a new measure to evaluate transferability of learned representations," in *International Conference on Machine Learning (ICML)*, vol. 119 of *Proceedings of Machine Learning Research*, pp. 7294–7305, July 2020.

[154] L. Chaves, A. Bissoto, E. Valle, and S. Avila, "The performance of transferability metrics does not translate to medical tasks," in *MICCAI Workshop on Domain Adaptation and Representation Transfer*, vol. 14293 of *Lecture Notes in Computer Science*, pp. 105–114, Springer Nature Switzerland, October 2023.

[155] Y. Wang, P.-M. Jodoin, F. Porikli, J. Konrad, Y. Benezeth, and P. Ishwar, "CDnet 2014: An expanded change detection benchmark dataset," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, (Columbus, Ohio, USA), pp. 393–400, Institute of Electrical and Electronics Engineers (IEEE), June 2014.

[156] K. Cranmer, J. Brehmer, and G. Louppe, "The frontier of simulation-based inference," *Proceedings of the National Academy of Sciences (PNAS)*, vol. 117, pp. 30055–30062, May 2020.

[157] T. Bouwmans and B. 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 2019.

[158] T. Bouwmans and B. 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 2019.

[159] M. Braham and M. Van Droogenbroeck, "Deep background subtraction with scene-specific convolutional neural networks," in *IEEE International Conference on Systems, Signals and Image Processing (IWSSIP)*, (Bratislava, Slovakia), pp. 1–4, May 2016.

[160] A. Cioppa, M. V. Droogenbroeck, and M. Braham, "Real-time semantic background subtraction," in *IEEE International Conference on Image Processing (ICIP)*, (Abu Dhabi, United Arab Emirates), pp. 3214–3218, Institute of Electrical and Electronics Engineers (IEEE), October 2020.

[161] J. Leens, S. Piérard, O. Barnich, M. Van Droogenbroeck, and J.-M. Wagner, "Combining color, depth, and motion for video segmentation," in *Computer Vision Systems*, vol. 5815 of *Lecture Notes in Computer Science*, pp. 104–113, Springer, 2009.

[162] S. Piérard, M. Braham, and M. Van Droogenbroeck, "An exploration of the performances achievable by combining unsupervised background subtraction algorithms," *arXiv*, vol. abs/2202.12563, 2022.

[163] M. Braham, S. Piérard, and M. Van Droogenbroeck, "Semantic background subtraction," in *IEEE International Conference on Image Processing (ICIP)*, (Beijing, China), pp. 4552–4556, September 2017.

[164] S. Piérard and M. Van Droogenbroeck, "A perfect estimation of a background image does not lead to a perfect background subtraction: analysis of the upper bound on the performance," in *International Conference on Image Analysis and Processing (ICIAP), Workshop on Scene Background Modeling and Initialization (SBMI)*, vol. 9281 of *Lecture Notes in Computer Science*, pp. 527–534, Springer, 2015.

[165] D. Cossock and T. Zhang, "Subset ranking using regression," in *Learning Theory*, vol. 4005 of *Lecture Notes in Computer Science*, pp. 605–619, Springer Berlin Heidelberg, 2006.

[166] P. Li, Q. Wu, and C. Burges, "McRank: Learning to rank using multiple classification and gradient boosting," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 20, (Vancouver, Canada), pp. 1–8, Curran Associates, Inc., December 2007.

[167] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender, "Learning to rank using gradient descent," in *International Conference on Machine Learning (ICML)*, (Bonn, Germany), pp. 89–96, ACM Press, 2005.

[168] Z. Cao, T. Qin, T.-Y. Liu, M.-F. Tsai, and H. Li, "Learning to rank: from pairwise approach to listwise approach," in *International Conference on Machine Learning (ICML)*, (Corvalis, Oregon, USA), pp. 129–136, ACM, June 2007.

[169] T. Liu, "Learning to rank for ir," *Foundation and Trends in IR*, vol. 3, no. 3, pp. 225–331, 2009.

[170] N. Tax, S. Bockting, and D. Hiemstra, "A cross-benchmark comparison of 87 learning to rank methods," *Information Processing & Management*, vol. 51, pp. 757–772, November 2015.

[171] A. Farahani, S. Voghoei, K. Rasheed, and H. R. Arabnia, "A brief review of domain adaptation," *Advances in Data Science and Information Engineering*, pp. 877–894, 2021.

[172] X. Liu, C. Yoo, F. Xing, H. Oh, G. El Fakhri, J.-W. Kang, and J. Woo, "Deep unsupervised domain adaptation: A review of recent advances and perspectives," *APSIPA Transactions on Signal and Information Processing*, vol. 11, pp. 1–51, August 2022.

[173] M. Kirchmeyer, A. Rakotomamonjy, E. de Bezenac, and P. Gallinari, "Mapping conditional distributions for domain adaptation under generalized target shift," in *International Conference on Learning Representations (ICLR)*, (Virtual conference), pp. 1–22, September 2022.

[174] I. Redko, E. Morvant, A. Habrard, M. Sebban, and Y. Bennani, *Advances in domain adaptation theory*. Elsevier, 2019.

[175] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *International Conference on Machine Learning (ICML)*, vol. 37 of *Proceedings of Machine Learning Research*, (Lille, France), pp. 1180–1189, ML Research Press, July 2015.

[176] K. Saito, K. Watanabe, Y. Ushiku, and T. Harada, "Maximum classifier discrepancy for unsupervised domain adaptation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (Salt Lake City, Utah, USA), pp. 3723–3732, IEEE, June 2018.

[177] J. Liang, D. Hu, and J. Feng, "Do we really need to access the source data? Source hypothesis transfer for unsupervised domain adaptation," in *International Conference on Machine Learning (ICML)*, vol. 119 of *Proceedings of Machine Learning Research*, pp. 6028–6039, ML Research Press, July 2020.

[178] D. Wang, E. Shelhamer, S. Liu, B. A. Olshausen, and T. Darrell, "Tent: Fully test-time adaptation by entropy minimization," in *International Conference on Learning Representations (ICLR)*, (Virtual conference), pp. 1–15, International Conference on Learning Representations (ICLR), May 2021.

[179] M. Boudiaf, R. Mueller, I. B. Ayed, and L. Bertinetto, "Parameter-free online test-time adaptation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (New Orleans, Louisiana, USA), pp. 8334–8343, Institute of Electrical and Electronics Engineers (IEEE), June 2022.

[180] Y. Zhang, P. David, and B. Gong, "Curriculum domain adaptation for semantic segmentation of urban scenes," in *IEEE International Conference on Computer Vision (ICCV)*, (Venice, Italy), pp. 2039–2049, IEEE, October 2017.

[181] Y. Li, L. Yuan, and N. Vasconcelos, "Bidirectional learning for domain adaptation of semantic segmentation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (Long Beach, California, USA), pp. 6929–6938, IEEE, June 2019.

[182] S. Zhao, B. Li, X. Yue, Y. Gu, P. Xu, R. Hu, H. Chai, and K. Keutzer, "Multi-source domain adaptation for semantic segmentation," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, (Vancouver, Canada), pp. 1–14, Curran Associates, Inc., December 2019.

[183] S. Zhao, B. Li, P. Xu, X. Yue, G. Ding, and K. Keutzer, "MADAN: Multi-source adversarial domain aggregation network for domain adaptation," *International Journal of Computer Vision*, vol. 129, pp. 2399–2424, May 2021.

[184] X. Peng, Q. Bai, X. Xia, Z. Huang, K. Saenko, and B. Wang, "Moment matching for multi-source domain adaptation," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, (Seoul, South Korea), pp. 1406–1415, IEEE, October 2019.

[185] J. He, X. Jia, S. Chen, and J. Liu, "Multi-source domain adaptation with collaborative learning for semantic segmentation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (Nashville, Tennessee, USA), pp. 11003–11012, IEEE, June 2021.

[186] S.-L. Sun and H.-L. Shi, "Bayesian multi-source domain adaptation," in *International Conference on Machine Learning and Cybernetics (ICMLC)*, (Tianjin, China), pp. 24–28, IEEE, July 2013.

[187] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, and J. Garcia-Rodriguez, "A review on deep learning techniques applied to semantic segmentation," *arXiv*, vol. abs/1704.06857, April 2017.

[188] X. Liu, Z. Deng, and Y. Yang, "Recent progress in semantic image segmentation," *Artificial Intelligence Review*, vol. 52, pp. 1089–1106, June 2018.

[189] S. Minaee, Y. Y. Boykov, F. Porikli, A. J. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, pp. 3523–3542, July 2022.

[190] B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, and A. Torralba, "Semantic understanding of scenes through the ADE20K dataset," *International Journal of Computer Vision*, vol. 127, pp. 302–321, December 2018.

[191] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common Objects in Context," in *European Conference on Computer Vision (ECCV)*, vol. 8693 of *Lecture Notes in Computer Science*, pp. 740–755, Springer, September 2014.

[192] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," in *IEEE International Conference on Computer Vision (ICCV)*, (Venice, Italy), pp. 2980–2988, IEEE, October 2017.

[193] W. Zhou, J. Liu, J. Lei, L. Yu, and J.-N. Hwang, "GMNet: Graded-feature multilabel-learning network for RGB-thermal urban scene semantic segmentation," *IEEE Transactions on Image Processing*, vol. 30, pp. 7790–7802, 2021.

[194] T. Cane and J. Ferryman, "Evaluating deep semantic segmentation networks for object detection in maritime surveillance," in *IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, (Auckland, New Zealand), pp. 1–6, IEEE, November 2018.

[195] N. A. Muhadi, A. F. Abdullah, S. K. Bejo, M. R. Mahadi, and A. Mijic, "Deep learning semantic segmentation for water level estimation using surveillance camera," *Applied Sciences*, vol. 11, pp. 1–25, October 2021.

[196] S. Zheng and G. Gupta, "Semantic-guided zero-shot learning for low-light image/video enhancement," in *IEEE/CVF Winter Conference on Applications of*

Computer Vision Workshops (WACVW), (Waikoloa, Hawaii, USA), pp. 581–590, IEEE, January 2022.

[197] A. Aakerberg, A. S. Johansen, K. Nasrollahi, and T. B. Moeslund, "Semantic segmentation guided real-world super-resolution," in *IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, (Waikoloa, Hawaii, USA), pp. 449–458, IEEE, January 2022.

[198] A. Cioppa, A. Deliège, M. Istasse, C. De Vleeschouwer, and M. Van Droogenbroeck, "ARTHuS: Adaptive real-time human segmentation in sports through online distillation," in *IEEE International Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), CVsports*, (Long Beach, California, USA), pp. 2505–2514, Institute of Electrical and Electronics Engineers (IEEE), June 2019.

[199] S. Piérard, A. Marcos Alvarez, A. Lejeune, and M. Van Droogenbroeck, "On-the-fly domain adaptation of binary classifiers," in *Proceedings of the BENELEARN Conference*, (Brussels, Belgium), pp. 20–28, Belgian-Dutch Conference on Machine Learning (BENELEARN), June 2014.

[200] Y. Mansour, M. Mohri, and R. Rostamizadeh, "Domain adaptation with multiple sources," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 21, (Vancouver, Canada), pp. 1041–1048, Curran Associates, Inc., December 2008.

[201] J. W. Miller, R. Goodman, and P. Smyth, "On loss functions which minimize to conditional expected values and posterior probabilities," *IEEE Transactions on Information Theory*, vol. 39, pp. 1404–1408, July 1993.

[202] D. Ramos, J. Franco-Pedroso, A. Lozano-Diez, and J. Gonzalez-Rodriguez, "Deconstructing cross-entropy for probabilistic binary classifiers," *Entropy*, vol. 20, pp. 1–20, March 2018.

[203] A. Cioppa, A. Deliège, and M. Van Droogenbroeck, "A bottom-up approach based on semantics for the interpretation of the main camera stream in soccer games," in *IEEE International Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), CVsports*, (Salt Lake City, Utah, USA), pp. 1846–1855, June 2018.

[204] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, (San Diego, California, USA), pp. 1–15, May 2015.

[205] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," *arXiv*, vol. abs/2004.10934, 2020.

[206] A. Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," in *International Conference on Machine Learning (ICML)*, (Bonn, Germany), pp. 625–632, ACM Press, 2005.

[207] 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*, vol. 12, pp. 2825–2830, November 2011.

[208] C. H. Bahnsen and T. B. Moeslund, "Rain removal in traffic surveillance: Does it matter?," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, pp. 2802–2819, August 2019.

[209] M. B. Jensen, A. Mogelmose, and T. B. Moeslund, "Presenting the multi-view traffic intersection dataset (MTID): A detailed traffic-surveillance dataset," in *IEEE International Conference on Intelligent Transportation Systems (ITSC)*, (Rhodes, Greece), pp. 1–6, IEEE, September 2020.

[210] R. Guerrero-Gómez-Olmedo, R. J. López-Sastre, S. Maldonado-Bascón, and A. Fernández-Caballero, "Vehicle tracking by simultaneous detection and viewpoint estimation," in *Natural and Artificial Computation in Engineering and Medical Applications*, vol. 7931 of *Lecture Notes in Computer Science*, pp. 306–316, Springer Berlin Heidelberg, 2013.

[211] J.-P. Jodoin, G.-A. Bilodeau, and N. Saunier, "Urban tracker: Multiple object tracking in urban mixed traffic," in *IEEE Winter Conference on Applications of Computer Vision (WACV)*, (Steamboat Springs, Colorado, USA), pp. 885–892, IEEE, March 2014.

[212] SAE International, "Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles," Tech. Rep. SAE Standard J3016 202104, Society of Automobile Engineers, Warrendale, PA, USA, April 2021.

[213] A. Perrels, A. Votsis, V. Nurmi, and K. Pilli-Sihvola, "Weather conditions, weather information and car crashes," *ISPRS International Journal of Geo-Information*, vol. 4, pp. 1–23, November 2015.

[214] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "CARLA: An open urban driving simulator," in *Annual Conference on Robot Learning*, vol. 78 of *Proceedings of Machine Learning Research*, (Mountain View, California, USA), pp. 1–16, ML Research Press, November 2017.

[215] M. Ibrahim, J. Haworth, and T. Cheng, "WeatherNet: Recognising weather and visual conditions from street-level images using deep residual learning," *ISPRS International Journal of Geo-Information*, vol. 8, pp. 1–18, November 2019.

[216] M. Introvigne, A. Ramazzina, S. Walz, D. Scheuble, and M. Bijelic, "Real-time environment condition classification for autonomous vehicles," in *IEEE Intelligent Vehicles Symposium (IV)*, (Jeju Island, Republic of Korea), pp. 1527–1533, IEEE, June 2024.

[217] X. Li, Z. Wang, and X. Lu, "A multi-task framework for weather recognition," in *ACM International Conference on Multimedia (MM)*, (Mountain View, California, USA), pp. 1318–1326, ACM, October 2017.

[218] I. Kobyzev, S. J. D. Prince, and M. A. Brubaker, "Normalizing flows: An introduction and review of current methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, pp. 3964–3979, November 2021.

[219] G. Papamakarios, E. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan, "Normalizing flows for probabilistic modeling and inference," *Journal of Machine Learning Research*, vol. 22, no. 57, pp. 1–64, 2021.

[220] F. Rozet, F. Divo, and S. Schnake, "Zuko: Normalizing flows in PyTorch." Software, 2022.

[221] A. Tonk and A. Boussif, "Operational design domain or operational envelope; seeking a suitable concept for autonomous railway systems," in *Proceedings of the European Safety and Reliability Conference (ESREL)*, (Dublin, Ireland), pp. 2104–2111, Research Publishing Services, August 2022.

[222] B. Gérin, M. Zanella, M. Wynen, S. Mahmoudi, B. Macq, and C. De Vleeschouwer, "Exploring viability of test-time training: Application to 3D segmentation in multiple sclerosis," in *IEEE Conference on Artificial Intelligence (CAI)*, vol. 34, (Singapore), pp. 557–562, Institute of Electrical and Electronics Engineers (IEEE), June 2024.

[223] B. Gérin, A. Halin, A. Cioppa, M. Henry, B. Ghanem, B. Macq, C. De Vleeschouwer, and M. Van Droogenbroeck, "Multi-stream cellular test-time adaptation of real-time models evolving in dynamic environments," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, vol. 33, (Seattle, Washington, USA), pp. 4472–4482, Institute of Electrical and Electronics Engineers (IEEE), June 2024.

[224] Q. Wang, O. Fink, L. Van Gool, and D. Dai, "Continual test-time domain adaptation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (New Orleans, Louisiana, USA), pp. 7191–7201, Institute of Electrical and Electronics Engineers (IEEE), June 2022.

[225] T. Gong, J. Jeong, T. Kim, Y. Kim, J. Shin, and S.-J. Lee, "NOTE: Robust continual test-time adaptation against temporal correlation," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 35, (New Orleans, Louisiana, USA), pp. 27253–27266, Curran Associates, Inc., November 2022.

[226] L. Yuan, B. Xie, and S. Li, "Robust test-time adaptation in dynamic scenarios," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (Vancouver, Canada), pp. 15922–15932, Institute of Electrical and Electronics Engineers (IEEE), June 2023.

[227] W. Wang, Z. Zhong, W. Wang, X. Chen, C. Ling, B. Wang, and N. Sebe, "Dynamically instance-guided adaptation: A backward-free approach for test-time domain adaptive semantic segmentation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (Vancouver, Canada), pp. 24090–24099, Institute of Electrical and Electronics Engineers (IEEE), June 2023.

[228] J. Houyon, A. Cioppa, Y. Ghunaim, M. Alfarra, A. Halin, M. Henry, B. Ghanem, and M. Van Droogenbroeck, "Online distillation with continual learning for cyclic domain shifts," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, vol. abs 2211 16234, (Vancouver, Canada), pp. 2437–2446, Institute of Electrical and Electronics Engineers (IEEE), June 2023.

[229] H. Tang, K. Chen, and K. Jia, "Unsupervised domain adaptation via structurally regularized deep clustering," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (Seattle, WA, USA), pp. 8722–8732, Institute of Electrical and Electronics Engineers (IEEE), June 2020.

[230] M. Zanella and I. B. Ayed, "On the test-time zero-shot generalization of vision-language models: Do we really need prompt learning?," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (Seattle, Washington, USA), pp. 23783–23793, IEEE, June 2024.

[231] M. Zanella, B. Gérin, and I. B. Ayed, "Boosting vision-language models with transduction," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 37, (Vancouver, Canada), pp. 62223–62256, Curran Associates, Inc., December 2024.

[232] C. Lu, D. Lin, J. Jia, and C.-K. Tang, "Two-class weather classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, pp. 2510–2524, December 2017.

[233] J. C. Villarreal Guerra, Z. Khanam, S. Ehsan, R. Stolkin, and K. McDonald-Maier, "Weather classification: A new multi-class dataset, data augmentation approach and comprehensive evaluations of convolutional neural networks," in *NASA/ESA Conference on Adaptive Hardware and Systems (AHS)*, (Edinburgh, UK), pp. 305–310, IEEE, August 2018.

[234] Z. Li, Y. Li, J. Zhong, and Y. Chen, "Multi-class weather classification based on multi-feature weighted fusion method," in *Earth and Environmental Science*, vol. 558 of *Journal of Physics: Conference Series*, (Dalian, China), IOP Publishing, August 2020.

[235] Y. Zhang, J. Sun, M. Chen, Q. Wang, Y. Yuan, and R. Ma, "Multi-weather classification using evolutionary algorithm on EfficientNet," in *Proceedings of the IEEE International Conference on Pervasive Computing and Communications Workshops and Other Affiliated Events (PerCom Workshops)*, (Kassel, Germany), pp. 546–551, Institute of Electrical and Electronics Engineers (IEEE), March 2021.

[236] J. Rebut, A. Bursuc, and P. Perez, "StyleLess layer: Improving robustness for real-world driving," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, (Prague, Czech Republic), pp. 8992–8999, Institute of Electrical and Electronics Engineers (IEEE), September 2021.

[237] J. Li, R. Xu, J. Ma, Q. Zou, J. Ma, and H. Yu, "Domain adaptive object detection for autonomous driving under foggy weather," in *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, (Waikoloa, Hawaii, USA), pp. 612–622, Institute of Electrical and Electronics Engineers (IEEE), January 2023.

[238] S. Lee, D. Kim, N. Kim, and S.-G. Jeong, "Drop to adapt: Learning discriminative features for unsupervised domain adaptation," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, (Seoul, South Korea), pp. 91–100, IEEE, October 2019.

[239] M. Jeon, J. Seo, and J. Min, "DA-RAW: Domain adaptive object detection for real-world adverse weather conditions," in *IEEE International Conference on Robotics and Automation (ICRA)*, vol. 17, (Yokohama, Japan), pp. 2013–2020, Institute of Electrical and Electronics Engineers (IEEE), May 2024.

[240] G. Pappalardo, R. Caponetto, R. Varrica, and S. Cafiso, "Assessing the operational design domain of lane support system for automated vehicles in different weather and road conditions," *Journal of Traffic and Transportation Engineering*, vol. 9, pp. 631–644, August 2022.

[241] M. Gyllenhammar, R. Johansson, F. Warg, D. Chen, H.-M. Heyn, M. Sanfridson, J. Söderberg, A. Thorsén, and S. Ursing, "Towards an operational design domain that supports the safety argumentation of an automated driving system," in *European congress on embedded real time systems (ERTS)*, (Toulouse, France), pp. 1–10, Society of Electricity, Electronics and Information and Communication Technologies (SEE), January 2020.

[242] I. Colwell, B. Phan, S. Saleem, R. Salay, and K. Czarnecki, "An automated vehicle safety concept based on runtime restriction of the operational design domain," in *IEEE Intelligent Vehicles Symposium*, (Changshu, China), pp. 1910–1917, IEEE, June 2018.

[243] C. Sun, Z. Deng, W. Chu, S. Li, and D. Cao, "Acclimatizing the operational design domain for autonomous driving systems," *IEEE Intelligent Transportation Systems Magazine*, vol. 14, pp. 10–24, March 2022.

[244] C. W. Lee, N. Nayeer, D. E. Garcia, A. Agrawal, and B. Liu, "Identifying the operational design domain for an automated driving system through assessed risk," in *IEEE Intelligent Vehicles Symposium*, vol. IV, (Las Vegas, Nevada, USA), pp. 1317–1322, IEEE, October 2020.

[245] M. Vasist, F. Rozet, O. Absil, P. Mollière, E. Nasedkin, and G. Louppe, "Neural posterior estimation for exoplanetary atmospheric retrieval," *Astronomy & Astrophysics*, vol. 672, pp. 1–10–, April 2023.

[246] C. Durkan, A. Bekasov, I. Murray, and G. Papamakarios, "Neural spline flows," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, (Vancouver, Canada), pp. 1–12, Curran Associates, Inc., December 2019.

[247] J. Lueckmann, P. J. Gonçalves, G. Bassetto, K. Öcal, M. Nonnenmacher, and J. H. Macke, "Flexible statistical inference for mechanistic models of neural dynamics," in *Advances in Neural Information Processing Systems (NeurIPS)*, (Long Beach, California, USA), pp. 1289–1299, Curran Associates, Inc., 2017.

[248] D. Greenberg, M. Nonnenmacher, and J. H. Macke, "Automatic posterior transformation for likelihood-free inference," in *International Conference on Machine Learning (ICML)*, vol. 97 of *Proceedings of Machine Learning Research*, (New York City, New York, USA), pp. 2404–2414, ML Research Press, 2019.

[249] J. Hermans, A. Delaunoy, F. Rozet, A. Wehenkel, V. Begy, and G. Louppe, "A crisis in simulation-based inference? beware, your posterior approximations can be unfaithful," *Transactions on Machine Learning Research*, vol. 11, pp. 2835–8856, November 2022.

[250] C. Robertson, J. A. Long, F. S. Nathoo, T. A. Nelson, and C. C. F. Plouffe, "Assessing quality of spatial models using the structural similarity index and posterior predictive checks," *Geographical Analysis*, vol. 46, pp. 53–74, January 2014.

[251] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (Las Vegas, Nevada, USA), pp. 770–778, IEEE, June 2016.

[252] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski, "DINOv2: Learning robust visual features without supervision," *Transactions on Machine Learning Research*, vol. 1, pp. 1–32, 2024.

[253] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning (ICML)*, vol. 139 of *Proceedings of Machine Learning Research*, (Virtual conference), pp. 8748–8763, ML Research Press, July 2021.

[254] F. Rozet, B. K. Miller, and A. Delaunoy, "LAMPE: Likelihood-free amortized posterior estimation." Software, 2021.

[255] R. J. Hyndman, "Computing and graphing highest density regions," *The American Statistician*, vol. 50, pp. 120–126, May 1996.

[256] A. Shoker, R. Yasmin, and P. Esteves-Verissimo, "WIP: Savvy: Trustworthy autonomous vehicles architecture," in *Symposium on Vehicle Security Privacy (VehicleSec)*, (San Diego, California, USA), pp. 1–7, Internet Society, Feb. 2024.

[257] S. Diamond and S. Boyd, "CVXPY: A Python-embedded modeling language for convex optimization," *Journal of Machine Learning Research*, vol. 17, no. 83, pp. 1–5, 2016.

[258] A. Agrawal, R. Verschueren, S. Diamond, and S. Boyd, "A rewriting system for convex optimization problems," *Journal of Control and Decision*, vol. 5, pp. 42–60, January 2018.

[259] S. Piérard, R. Phan-Ba, and M. Van Droogenbroeck, "Machine learning techniques to assess the performance of a gait analysis system," in *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, (Bruges, Belgium), pp. 419–424, April 2014.

[260] S. Piérard and M. Van Droogenbroeck, "Techniques to improve the foreground segmentation with a 3D camera and a color camera," in *Workshop on Circuits, Systems and Signal Processing (ProRISC)*, (Veldhoven, The Netherlands), pp. 247–250, November 2009.

[261] M. Van Droogenbroeck, M. Braham, and S. Piérard, "Foreground and background detection method." European Patent Office, EP 3438929 B1, July 2020.

[262] S. Piérard, A. Lejeune, and M. Van Droogenbroeck, "3D information is valuable for the detection of humans in video streams," in *Proceedings of 3D Stereo MEDIA*, (Liège, Belgium), pp. 1–4, December 2010.

[263] S. Piérard, A. Lejeune, and M. Van Droogenbroeck, "A probabilistic pixel-based approach to detect humans in video streams," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (Prague, Czech Republic), pp. 921–924, May 2011.

[264] Y. Wang, S. Piérard, S. Su, and P. Jodoin, "Improving pedestrian detection using motion-guided filtering," *Pattern Recognition Letters*, vol. 96, pp. 106–112, September 2017.

[265] O. Barnich, S. Piérard, and M. Van Droogenbroeck, "A virtual curtain for the detection of humans and access control," in *Advanced Concepts for Intelligent Vision Systems (ACIVS), Part II*, (Sydney, Australia), pp. 98–109, December 2010.

[266] T. Bouwmans, A. Greco, S. Piérard, A. V. Ricciardi, C. Sansone, M. Van Droogenbroeck, and B. Vento, "Illegal waste dumping detection." Technical note, 2026.

[267] F. Cantalloube, C. A. G. Gonzalez, O. Absil, C. Cantero, R. Bacher, M. Bonse, M. Bottom, C.-H. Dahlqvist, C. Desgrange, O. Flasseur, T. Fuhrmann, T. H. Henning, R. Jensen-Clem, M. Kenworthy, D. Mawet, D. Mesa, T. Meshkat, D. Mouillet, A. Müller, E. Nasedkin, B. Pairet, S. Piérard, J.-B. Ruffio, M. Samland, J. Stone, and M. Van Droogenbroeck, "Exoplanet imaging data challenge: benchmarking the various image processing methods for exoplanet detection," in *Adaptive Optics Systems VII*, vol. 11448 of *Proceedings of SPIE*, pp. 1–36, SPIE, December 2020.

[268] J. Yao and M. Shepperd, "The impact of using biased performance metrics on software defect prediction research," *Information and Software Technology*, vol. 139, pp. 1–14, November 2021.

[269] N. Lavrač, P. Flach, and B. Zupan, "Rule evaluation measures: A unifying view," in *Inductive Logic Programming*, vol. 1634 of *Lecture Notes in Computer Science*, pp. 174–185, Springer Berlin Heidelberg, 1999.