A Closer Look at AUROC and AUPRC under Class Imbalance

Matthew B. A. McDermott [1] Lasse Hyldig Hansen [*2] Haoran Zhang [*3] Giovanni Angelotti [4] Jack Gallifant [3]

Abstract

In machine learning (ML), a widespread adage is that the area under the precision-recall curve (AUPRC) is a superior metric for model comparison to the area under the receiver operating characteristic (AUROC) for binary classification tasks with class imbalance. This paper challenges this notion through novel mathematical analysis, illustrating that AUROC and AUPRC can be concisely related in probabilistic terms. We demonstrate that AUPRC, contrary to popular belief, is *not superior* in cases of class imbalance and might even be a *harmful metric*, given its inclination to unduly favor model improvements in subpopulations with more frequent positive labels. This bias can inadvertently heighten algorithmic disparities. Prompted by these insights, a thorough review of existing ML literature was conducted, utilizing large language models to analyze over 1.5 million papers from arXiv. Our investigation focused on the prevalence and substantiation of the purported AUPRC superiority. The results expose a significant deficit in empirical backing and a trend of misattributions that have fuelled the widespread acceptance of AUPRC's supposed advantages. Our findings represent a dual contribution: a significant technical advancement in understanding metric behaviors and a stark warning about unchecked assumptions in the ML community. All experiments are accessible at https://github.com/mmcdermott/AUC_is_all_you_need.

1. Introduction

Machine learning (ML), especially in critical domains like healthcare, necessitates carefully selecting and applying evaluation metrics to guide appropriate model choices and understand performance nuances (Hicks et al., 2022). This study focuses on two pivotal metrics for binary classification tasks: the Area Under the Precision-Recall Curve (AUPRC) and the Area Under the Receiver Operating Characteristic (AUROC). Within the ML community, it is widely believed that AUPRC is a "better" metric than AUROC for model comparison when positive instances are substantially rarer than negative ones. Literature justifies this claim on several grounds, many of which we contest:

- Precision-recall curves may be more visually indicative of real-world deployment objectives than the receiver operating characteristic (Cook & Ramadas, 2020; Leisman, 2018; Saito & Rehmsmeier, 2015; Ozenne et al., 2015; Yuan et al., 2015; Zhou et al., 2020).
- AUPRC is unaffected by the (large) number of true negatives, making it somehow "less optimistic" than AUROC (Albora & Zaccaria, 2022; Leisman, 2018; Czakon, 2022).
- In scenarios of low prevalence, AUPRC is often significantly lower compared to AUROC (Goadrich et al., 2006; Mazzanti, 2023; Bleakley et al., 2007).
- AUPRC's dependence on prevalence is argued to be an advantageous property (Saito & Rehmsmeier, 2015; Goadrich et al., 2006; Yuan et al., 2015).

In this work, we show through careful mathematical, logical, and epistemological reasoning that this claim and many associated reasons are *invalid* or *misapplied* in common ML settings. More specifically, we show the following:

AUROC and AUPRC are probabilistically interrelated First, we show the following theorem:

Theorem 1. Let f denote a model that outputs scores from distributions p_+, p_-, and p for samples with positive, negative, and arbitrary labels, respectively. Then,

$$\mathrm{AUROC}(f) = 1 - \mathbb{E}_{\mathsf{p}_+}\left[\mathrm{FPR}(p_+)\right]$$
$$\mathrm{AUPRC}(f) = 1 - P_{\mathsf{y}}(0)\mathbb{E}_{\mathsf{p}_+}\left[\frac{\mathrm{FPR}(p_+)}{P_{\mathsf{p}}(p > p_+)}\right]$$

Presented this way, it is clear that for a fixed dataset, *AUROC and AUPRC only differ w.r.t. model-dependent parameters in that AUROC weighs all false positives equally,*

[*] Equal contribution [1] Harvard Medical School, Department of Biomedical Informatics [2] Cognitive Science, Aarhus University, Denmark [3] Massachusetts Institute of Technology [4] IRCCS Humanitas Research Hospital, Artificial Intelligence Center, Milan, Italy. Correspondence to: Matthew B. A. McDermott <matthew_mcdermott@hms.harvard.edu>.

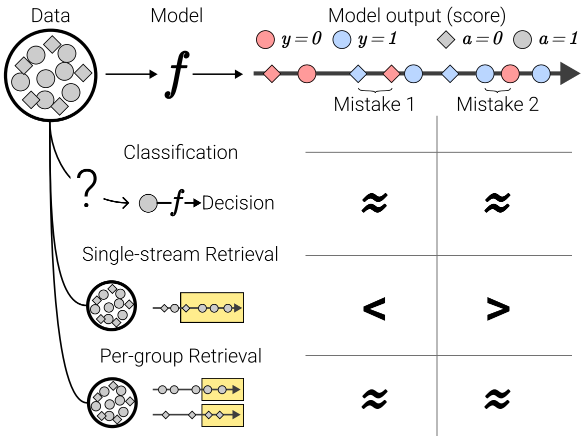


Figure 1. Atomic mistakes occur when neighboring samples, when ordered by model score, are out-of-order with respect to the classification label. AUROC improves by a constant amount no matter which atomic mistake is corrected; AUPRC improves in descending order with model score due to the dependence on model firing rate (Theorem 1). Which mistake you should prioritize fixing first depends on usage; in a classification setting, where you do not know whether the sample of interest is from a high-scoring or low-scoring region, you want to use a metric that optimizes scores in an unbiased manner, like AUROC. In a single-stream retrieval setting, where you choose the top-k samples, regardless of group membership and evaluate with those, a metric that favors mistakes in high-scoring regions like AUPRC will be most impactful. But, if you care about retrieving the top-k metrics from multiple distinct subpopulations within your dataset, *AUPRC will be dangerous as it will favor the high-prevalence sub-population.*

whereas AUPRC weighs false positives with the inverse of the model's likelihood of outputting a score greater than the given threshold (a quantity we will refer to as the "firing rate" of the model at a given threshold).

AUROC favors model improvements in an unbiased manner; AUPRC prioritizes high-score mistakes first Suppose we consider a static model over a finite dataset $\mathcal{X}$. Let us say that a model makes an *atomic mistake* with samples i, j if the model satisfies three properties over the two samples x_i, x_j: (1) $y_i = 1$ and $y_j = 0$, (2) $f(x_i) < f(x_j)$, and (3) there exists no sample x_k such that $f(x_i) < f(x_k) < f(x_j)$. Essentially, an *atomic mistake* occurs when a model assigns adjacent probability scores to a pair of samples with discordant labels (Figure 1).

With this concept of an atomic mistake, a pivotal question arises: *Given a fixed model f and dataset X, which such mistakes should be prioritized for correction?* We highlight two possible responses:

1. Do not prioritize any mistake over any other—all atomic improvements are equally valuable.
2. Prioritize fixing the atomic mistakes x_i, x_j in *descending* order of the assigned scores for $f(x_i)$.

The first strategy is well suited for *classification* settings, in which a user is given a model and a sample and must decide what action to take based on that sample. In such a setting, as we cannot guarantee that the given sample will be in any particular position of assigned model scores, an unbiased approach will optimally prepare the model to order possible positive samples amongst negative competitors. This is true regardless of the degree of class imbalance of the dataset.

The second strategy is *only* well suited for *single-group information retrieval* settings (which commonly features, but is *not* synonymous with class imbalance), where users use a model only by selecting the top k scored samples (for potentially unknown k) with the intent being to maximize the number of positively labeled samples among this subset (note that there are likely better, dedicated metrics for this task, such as precision at k).

We show that optimization or selection by AUROC corresponds exactly to this first strategy whereas optimization or selection by AUPRC corresponds to this second strategy (Theorem 2).

AUPRC overtly favors high-prevalence subpopulations The fact that AUPRC prioritizes fixing atomic mistakes for samples assigned the highest scores first also reveals a dangerous risk of this metric. Namely, if the underlying dataset can be decomposed into two separate subpopulations that have markedly different prevalences and our model is reasonably well calibrated, then *AUPRC will explicitly favor optimization for the higher-prevalence subpopulation first, whereas AUROC will optimize both subpopulations in an unbiased manner.* This is a significant concern in domains that frequently feature imbalanced classification problems with diverse patient populations, such as the medical domain (Fletcher et al., 2021; Chen et al., 2023). Thus, relying on AUPRC poses challenges to fairness and validity, even in information retrieval settings where fairness among possibly retrieved elements is a concern.

The literature regarding this claim is rife with misattributed citations, fallacies, and unchallenged assertions Through an automated review of 1.5M and a manual review of 128 arXiv papers, we profile the landscape regarding this claim in the academic literature and examine how frequently it is made without citation. We find that a significant portion of citations for this claim references papers *that do not make this claim in the first place* and that root arguments for this claim are *generally logically unsound or applied inappropriately or out of context in downstream citations and use*. This study addresses this significant gap in the scientific discourse on metric relationships and selection.

In particular, our literature review suggests that two primary phenomena have driven the widespread belief in AUPRC's superiority over AUROC for model comparison in cases of class imbalance. First, it is often observed that AUPRC values will be significantly lower in class imbalance situations, and precision-recall curves will appear less optimal compared to their AUROC and ROC curve counterparts, which is often taken as explicit or implicit evidence of AUROC's "optimism" and AUPRC's superiority. While this observation is accurate, its relevance in the context of considering metrics for *generic model comparison* is limited. The pivotal factor in such comparisons, *outside of a particular deployment scenario*, is not the absolute metric values but rather the relative rankings conferred upon models by these evaluation metrics. In cases of class imbalance, the lower AUPRC values relative to AUROC can be attributed to its inherently lower *chance* value in these contexts and the diminished firing rate for positive samples in models where AUROC is high, which means that AUPRC will weight the remaining false positives increasingly highly. However, in isolation, neither of these facts substantiates the claim that AUPRC produces a superior rank-ordering of candidate models than AUROC, regardless of its numerical value.

Secondly, it is true that AUPRC can be significantly more indicative of a model's precision or top-k retrieval performance compared to AUROC. This difference primarily arises from the inherent construction of these curves. Nevertheless, this again does not imply superiority in broader contexts of general model comparison. *If precision is known to be the primary metric of interest in a given scenario, then direct comparison of models based on precision is indeed more appropriate.* The advantage of AUROC over AUPRC lies in its capacity to offer a more generalizable, unbiased assessment of performance that is not anchored to precision. This characteristic emerges from and underscores the intrinsic value that AUROC favors model improvements equally across the entire range of model output scores. This trait becomes particularly crucial when the primary performance metric in downstream applications is indeterminate or when the goal is to optimize performance evenly across diverse subgroups with different prevalences within the dataset. In such cases, AUROC's ability to equitably value model improvements across the spectrum of output scores is more beneficial than a precision-anchored metric like AUPRC which favors improvements to high-prevalence groups at the expense of those to low prevalence groups.

Outline of the rest of the work Throughout the remainder of this paper, we will first present a formal establishment of these arguments, supported by a combination of theoretical reasoning and semi-synthetic empirical analyses. Next, we will comment more deeply on the methods and results of our literature search. Finally, we will conclude with closing thoughts. Overall, this work not only challenges prevailing assumptions regarding AUPRC and AUROC but also seeks to provide an impetus to elevate the discourse in machine learning toward more robust and evidence-based practices.

2. Theoretical Foundations of Our Assertions

2.1. Relationship Between AUROC and AUPRC

In this section, we will prove Theorem 1, restated below.

Theorem 1. Let f denote a model that outputs scores from distributions p_+, p_-, and p for samples with positive, negative, and arbitrary labels, respectively. Then,

$$\begin{aligned}\mathrm{AUROC}(f) &= 1 - \mathbb{E}_{\mathsf{p}_+}\left[\mathrm{FPR}(p_+)\right]\\ \mathrm{AUPRC}(f) &= 1 - P_{\mathsf{y}}(0)\mathbb{E}_{\mathsf{p}_+}\left[\frac{\mathrm{FPR}(p_+)}{P_{\mathsf{p}}(p > p_+)}\right]\end{aligned}$$

Proof. Recall that AUROC and AUPRC are as follows:

$$\begin{aligned}\mathrm{AUROC} &= \int_0^1 \mathrm{TPR}\, d\mathrm{FPR} = 1 - \int_0^1 \mathrm{FPR}\, d\mathrm{TPR}\\ \mathrm{AUPRC} &= \int_0^1 \mathrm{Prec}\, d\mathrm{TPR}\end{aligned}$$

However, we can further clarify these by leveraging the fact that $\mathrm{TPR}(\tau) = P_{\mathsf{p}_+}(p_+ > \tau) = \int_\tau^1 p_+(t)dt$, as below:

$$\begin{aligned}\int_0^1 g(\tau)d(\mathrm{TPR}(\tau)) &= \int_1^0 g(\tau)\frac{d\mathrm{TPR}(\tau)}{d\tau}d\tau\\ &= \int_1^0 g(\tau)\frac{d}{d\tau}(P_{\mathsf{p}_+}(p_+ > \tau))d\tau\\ &= \int_1^0 g(\tau)\frac{d}{d\tau}\left(\int_\tau^1 p_+(t)dt\right)d\tau\\ &= \int_1^0 g(\tau)(-p_+(\tau))d\tau\\ &= \mathbb{E}_{\mathsf{p}_+}\left[g\right]\end{aligned}$$

So, $\mathrm{AUROC} = 1 - \mathbb{E}_{\mathsf{p}_+}\left[\mathrm{FPR}\right]$ & $\mathrm{AUPRC} = \mathbb{E}_{\mathsf{p}_+}\left[\mathrm{Prec}\right]$. To further simplify, we expand Prec via Bayes rule:

$$\begin{aligned}\mathrm{Prec} &= 1 - P_{\mathsf{y}|\mathsf{p}>\tau}(y = 0)\\ &= 1 - \underbrace{P_{\mathsf{p}|\mathsf{y}=0}(p > \tau)}_{\mathrm{FPR}(\tau)}\frac{P_{\mathsf{y}}(y = 0)}{P_{\mathsf{p}}(p > \tau)}\end{aligned}$$

Thus,

$$\begin{aligned}\mathrm{AUROC}_{\boldsymbol{\theta}} &= 1 - \mathbb{E}_{p_+}\left[\mathrm{FPR}\right] \\ \mathrm{AUPRC}_{\boldsymbol{\theta}} &= \mathbb{E}_{p_+}\left[\mathrm{Prec}\right] \\ &= 1 - P_{\mathsf{y}}(y=0)\mathbb{E}_{p_+}\left[\frac{\mathrm{FPR}}{P_{\mathsf{p}}(p > p_+)}\right]\end{aligned}$$

□

2.2. Relative Prioritization of *Atomic Mistakes*

Understanding how a given evaluation metric prioritizes the correction of various kinds of model mistakes or errors offers significant insight into what properties that metric will promote when used for an optimization or model selection procedure. Here, we prove the claims offered in Section 1 regarding how AUROC and AUPRC prioritize the correction of *atomic mistakes* in a given model.

Theorem 2. Given our prior definition of an *atomic mistake*, the improvement in a model f's AUROC is invariant to the specific mistake corrected, whereas the improvement to AUPRC is positively correlated with the score assigned by f to the samples in the mistake being fixed.

Proof. Suppose f has a given, non-empty set M of atomic mistakes, such that, without loss of generality, $(x_i, x_{i+1}) \in M$. Suppose we construct a new model f' with empirical distributions p'_+ and p'_- by replicating the scores assigned by the model f with x_i and x_{i+1} swapped (i.e., we correct the mistake (x_i, x_{i+1}), so $x'_i = x_{i+1}$ and $x'_{i+1} = x_i$).

For which thresholds drawn from the original distribution p_+ will the number of false positives of f' differ from the number of false positives of f at that same threshold? For any threshold $\tau < x_i$, fixing the mistake (x_i, x_{i+1}) will not change the number of false positives with threshold τ, because both x_i and x_{i+1} are above τ. For any threshold $\tau > x_{i+1}$, the number will likewise not change as both x_i and x_{i+1} are below τ. The only τ that will have an impact is $\tau = x_i$ (recall that this is for an empirical distribution p_+ which contains x_i and by the definition of atomic mistakes, there are no samples in f with scores between x_i and x_{i+1}). In f, the fact that $x_{i+1} > x_i$ yet has a negative label means that there will be one false positive corresponding to sample $i+1$ greater than x_i in addition to all those that exist with scores greater than x_{i+1}. For f', however, the samples have swapped, so $x'_i > x'_{i+1}$ and thus there is no false positive corresponding to sample $i+1$ at the positive score threshold corresponding to x'_i. Therefore, the number of false positives will only change to decrease by one for the threshold x_i when the mistake (x_i, x_{i+1}) is corrected.

As AUROC weights the false positive rate at all positive samples equally and the false positive rate is proportional to the number of false positives, this shows that AUROC will improve by a constant amount no matter which atomic mistake is fixed. In contrast, as AUPRC weights false positives inversely by the model's firing rate, it will improve by an amount that is directly linearly correlated with the inverse of the model's firing rate, implying that it favors mistakes with higher scores and disfavors mistakes with lower scores. □

Theorem 2 shows that optimizing or selecting by AUPRC will implicitly value a model's discrimination in regions where the output score is high, whereas AUROC is unbiased in where it values model improvements. This naturally both clarifies the claim that "AUPRC should be preferred in cases of class imbalance" and raises serious concerns about the usage of AUPRC in general.

2.3. AUPRC is suited for single-group retrieval, not class-imbalanced binary classification

Theorem 2 not only refutes the conventional wisdom linking AUPRC's preference to class imbalance but rather indicates its suitability based on the intended usage of the model. It becomes apparent that the decision to use AUPRC over AUROC hinges not on class imbalance but on whether the metric should be optimized in a biased or unbiased manner. In information retrieval contexts, where a model exclusively selects the top k items for evaluation, prioritizing corrections in high-score regions aligns with the typical usage. This approach ensures that the metric focuses on the most relevant part of the model's output for the task at hand.[1] In contrast, in applications beyond information retrieval, such as diagnostic predictions in healthcare, the focus is on individual predictions rather than a ranked list. For instance, a clinician utilizing a model for patient diagnosis is concerned with the model's accuracy for the specific patient under examination, *not for the top k patients in the overall dataset on which the model was trained.* Therefore, in such settings, particularly when dealing with rare conditions, optimizing or selecting models based on AUPRC would be inappropriate.

2.4. AUPRC is explicitly discriminatory

Despite its potential relevance in information retrieval settings, the reliance of the AUPRC on a model's firing rate introduces significant fairness concerns. Especially in datasets comprising various subpopulations with differing prevalence rates, a *well-calibrated model's preference for high-scoring regions implies that AUPRC may inadvertently favor higher-prevalence subpopulations, disadvantaging those*

[1] It is worth noting that in such scenarios, it is possible that a different metric explicitly calculating the k-thresholded score (e.g., precision at k) might be more appropriate. However, this discussion falls beyond the scope of AUPRC versus AUROC comparison.

with lower prevalence. This issue is particularly acute in sectors like healthcare, where equitable model performance across diverse patient groups with varying disease incidences is paramount. The propensity of AUPRC to prioritize improvements in subgroups with higher prevalence constitutes an *unacceptable risk* and suggests that AUPRC should not be favored as an evaluation metric. This risk is not merely theoretical; our synthetic experiments in Section 3 demonstrate that optimizations driven by AUPRC markedly prefer high-prevalence groups. Such findings underscore the urgency of re-evaluating the use of AUPRC as an evaluation metric, considering its inherent biases that could perpetuate disparities in critical applications like healthcare.

3. Synthetic Experimental Validation

To substantiate our theoretical claims regarding AUROC and AUPRC, we validate each theorem and proof in turn; all experiments can be replicated and validated at `https://github.com/mmcdermott/AUC_is_all_you_need`.

3.1. Theorems 1 and 2

To validate Theorems 1 and 2, we sample synthetic scores for positive and negative labeled samples for models across a battery of possible AUROC values and prevalence values and confirm that the relationships we predict in said theorems hold with high numerical precision in all settings. As the details of these comparisons are largely technical, we relegate further details to Appendix B.

3.2. Synthetic Optimization Experiments Demonstrate AUPRC-induced Disparities

More critical than the straightforward validation of Theorems 1 and 2 is validating that these findings imply that AUPRC can be a discriminatory metric. To do so, we construct a synthetic dataset of model scores and labels for a population containing two distinct subpopulations with differing prevalences, denoted as G_1 and G_2. Then, under various procedures, we optimize the initialized scores over a series of iterations. We assume that subgroup labels are unknown to the model developer, and so the optimized score can only be selected by either overall AUROC or overall AUPRC. We track how the overall and per-subgroup AUROC and AUPRC metrics change through this process.

Experimental Setup. Let $Y \in \{0, 1\}$ be the binary label, $S \in [0, 1]$ be the predicted score, and $G \in \{G_1, G_2\}$ be the subpopulation. We fix $\mathbb{P}(Y = 1|G = G_1) = 0.05$ and $\mathbb{P}(Y = 1|G = G_2) = 0.01$. We sample a dataset for each group $\mathcal{D}_g = \{(S_1, Y_1), ..., (S_{n_g}, Y_{n_g})\}$, such that $\text{AUROC}(\mathcal{D}_1) \approx \text{AUROC}(\mathcal{D}_2) \approx \text{AUROC}(\mathcal{D}_1 \cup \mathcal{D}_2) = 0.85$.

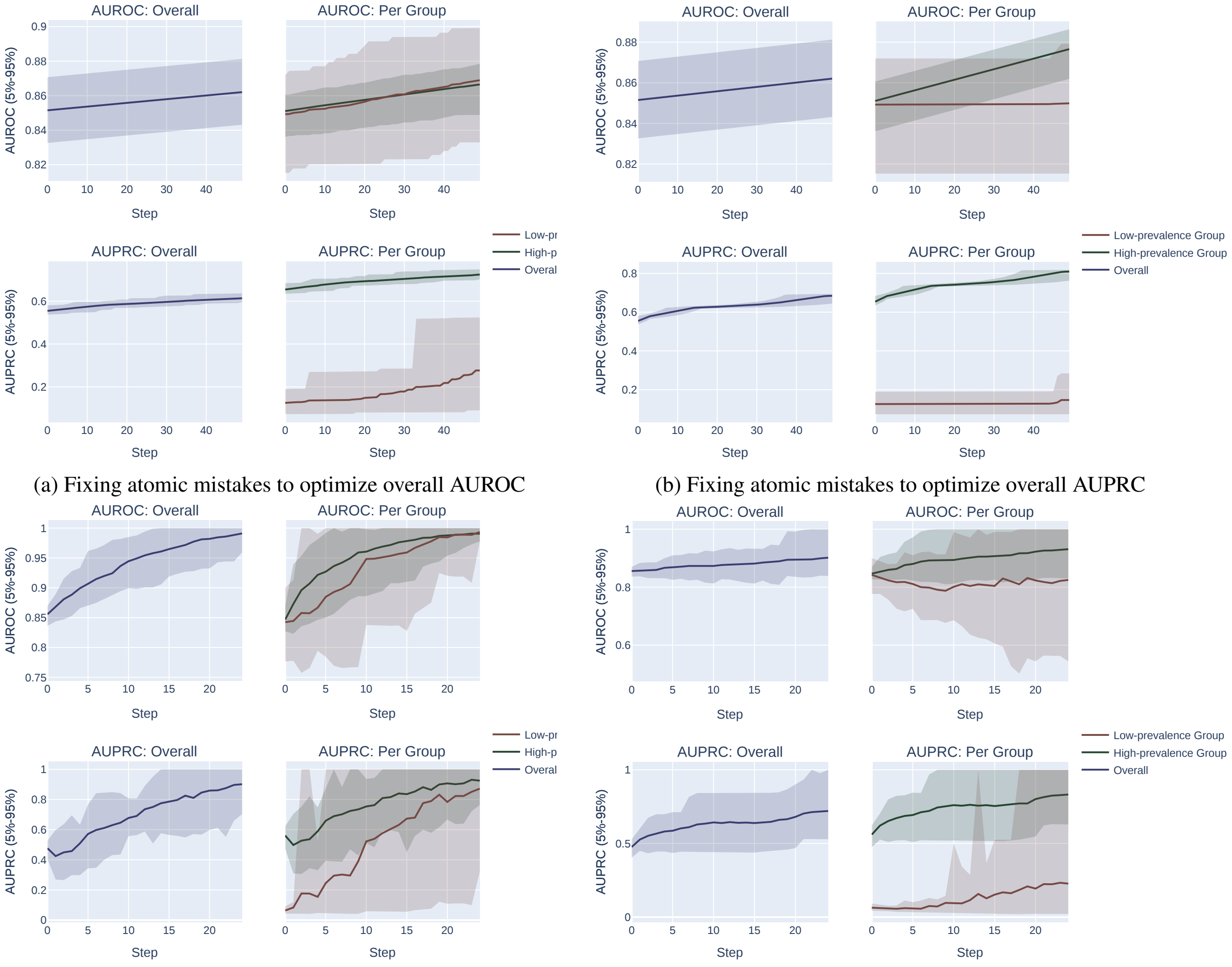
Our main experimental challenge is to determine how to simulate "optimizing" or "selecting" a model by AUROC or AUPRC. We explore two approaches here. First, we can simply correct the atomic mistake that maximally improves AUROC or AUPRC in each optimization iteration. In our experiments, we use $n_1 = n_2 = 200$ and optimize for 50 steps for this experiment. This is the most straightforward possible optimization procedure to analyze, but it is unrealistic. In real optimization scenarios, larger model changes will be made at once, and a model will have an opportunity to *degrade* performance in some regions in order to improve it in others.

However, simulating a more realistic optimization process is challenging. A natural idea would be to simply perturb the assigned model scores by a random noise vector of fixed magnitude a fixed number of times, and output the perturbed model scores that optimize the target metric. However, this approach is subtly biased in favor of the lower-prevalence group. In particular, because scores for the low-prevalence group tend to have greater density in the space of model scores,[2] a random perturbation of fixed magnitude will proportionally induce more score *permutations* in the low-prevalence group than the high-prevalence group, which affords the system greater capacity to improve the model for the low-prevalence group independent of the choice of AUROC or AUPRC.

To accommodate for this bias while still providing a more realistic optimization procedure, we additionally profile an optimization procedure that randomly permutes all the (sorted) model scores up to 3 positions. This has the effect of randomly adjusting all model scores, and can worsen model performance under some random permutations, but offset precisely the same capacity to the low and high prevalence subgroups. To ensure the model is under some optimization constraint (and therefore does not always find the "perfect" permutation to maximize both metrics identically), we allow the model to sample only 15 possible permutations before choosing the best option. This means the system will be forced to navigate optimization trade-offs between which permutations improve the right regions of the score most effectively among its limited set. We use $n_1 = n_2 = 100$ for these experiments and optimize for 25 total steps.

Across both settings, we run these experiments across 20 randomly sampled datasets and show the mean and an empirical 90% confidence interval around the mean in Figure 2. We present a formal mathematical formulation of these perturbations, as well as profile some aspects of the random perturbation method of optimization (despite the possible biases it has) in Appendix B.3.

[2] By virtue of their lower prevalence, this group's scores are "squished" into a smaller region of the probability space



(a) Fixing atomic mistakes to optimize overall AUROC

(b) Fixing atomic mistakes to optimize overall AUPRC

(c) Randomly permuting scores to optimize overall AUROC

(d) Randomly permuting scores to optimize overall AUPRC

Figure 2. Comparison of the impact of optimizing for overall AUROC and overall AUPRC on the per-group AUROC and AUPRCs of two groups in a synthetic setting, using both the *sequentially fixing atomic mistakes* optimization procedure (M2; *top*) and the *sequentially permuting nearby scores* optimization procedure (M3; *bottom*) described in Section 3.2. Note that the prevalence of Y in the high-prevalence group and the low-prevalence group are 0.05 and 0.01 respectively.

Results. Our results demonstrate the impact of the optimization metric on subpopulation disparity. In particular, in Figures 2b and 2d, we observe a notable disparity introduced when optimizing under the AUPRC metric. This is evident in the performance metrics across the high and low prevalence subpopulations, which exhibit significant divergence as the optimization process favors the group with higher prevalence. In the more realistic, random-permutation optimization procedure (Figure 2d), this even results in a decrease in the AUROC for G_2. In comparison, when optimizing for overall AUROC (Figures 2a & 2c), the AUROC and AUPRC of both groups increase together.

4. If AUPRC is not better than AUROC under class imbalance, why did we think it was?

The claim that "AUPRC is better than AUROC in cases of class imbalance" is widespread in the literature. Via both a manual literature search and an automated search of over 1.5M arXiv papers (see Appendix D for methodology), we observed 128 publications making this claim. We analyzed these papers to understand better why this incorrect claim is so widespread in machine learning, and what correct insights we can glean from the surrounding scientific discourse.

4.1. Weaknesses in the current literature

This claim is frequently stated without any citation Among the 128 papers we discovered referencing this claim, 31 did so with no associated citation (Liu et al., 2023; Randl et al., 2023; Tusfiqur et al., 2022; Piermarini et al., 2023; Zhang & Bondell, 2018; Torfi et al., 2022; Wu et al., 2020; Navarro et al., 2022; Wagner et al., 2023; Herbach, 2021; Si & Roberts, 2021; Narayanan et al., 2022; Rayhan et al., 2017; Yang et al., 2022a;b; Harer et al., 2018; Lee et al., 2020; Zavrtanik et al., 2021; Rezvani et al., 2021; Prapas et al., 2023; Thambawita et al., 2020; Vijayan et al., 2017; Brophy & Lowd, 2020; Lyu et al., 2021; Chakraborty et al., 2023; Rajabi & He, 2021; Kim et al., 2022; Kiran et al., 2018; Mousavian et al., 2016; Rohani & Eslahchi, 2019; Rao et al., 2022). These papers were published in venues ranging from on arXiv only to Nature Scientific Reports, ICCV, ECCV, and Bioinformatics, among others. This reflects not only the widespread belief in this claim, but also that we may be too comfortable making seemingly "correct" assertions without appropriate attribution in ML today.

This claim is frequently attributed to papers that *do not make this claim* Among the 97 that reference this claim and cite a source for this assertion, 39 *do not cite any papers that make this claim in the first place* (Yang et al., 2015; Li et al., 2020; Kyono et al., 2018; Seo et al., 2021; Hong et al., 2019; Hagedoorn & Spanakis, 2017; Babaei et al., 2021; Zou et al., 2022; Mangolin et al., 2022; Mosteiro et al., 2021; Showalter & Wu, 2019; Cranmer & Desmarais, 2016; Bryan & Moriano, 2023; Zhang et al., 2017; Domingues et al., 2020; Shukla & Marlin, 2019; Blevins et al., 2021; Hsu et al., 2020; Smith et al., 2023; Chu et al., 2018; Deshwar et al., 2015; Mongia et al., 2021; Rubin et al., 2012; Ahmed & Courville, 2020; Gong et al., 2021; Shukla & Marlin, 2018; Ma et al., 2022; Lei Ba et al., 2015; Newby et al., 2022; Ando & Huang, 2017; Stolman et al., 2022; Won et al., 2019; Stephenson et al., 2022; Srivastava et al., 2019; Karadzhov et al., 2022; Vens et al., 2008; López et al., 2013; Hall et al., 2023; Goyal & Khiari, 2020). In total, 13 sources are cited that neither reference nor argue this claim (Davis & Goadrich, 2006; Branco et al., 2016; Provost & Fawcett, 1997; Sokolova & Lapalme, 2009; Wahid-Ul-Ashraf et al., 2019; Ezzat et al., 2017; Burez & Van den Poel, 2009; Flach et al., 2011; Krawczyk, 2016; He & Garcia, 2009; LCT14558, 2017; Lobo et al., 2008). Most often, papers erroneously attribute this claim to (Davis & Goadrich, 2006), which was cited as a source for this claim 47 times. While (Davis & Goadrich, 2006) makes many interesting, meaningful claims about the ROC and PR curves, and *does argue that the precision-recall curve is more informative than the ROC* it never asserts that the *area under* the PR curve should be preferred over the *area under* the ROC in cases of class imbalance. It references the emergence of the use of AUPRC instead of or in addition to the AUROC in this context, citing among those references a paper that would later be re-published as (Goadrich et al., 2006), which does make this claim, but (Davis & Goadrich, 2006) itself makes no claim about whether or not AUPRC should be preferred in this way, even by proxy to those prior references. The fact that, despite this, it receives so much citation volume for this claim reflects very poorly on the accuracy of our scientific discourse in ML today.

For clarity, the ROC Curve graphically represents the trade-off between the TPR and FPR at various thresholds, and the PR Curve focuses on the trade-off between Precision and Recall. AUROC and AUPRC are the respective areas under each respective curve, providing an overall reflection of the model *across all thresholds*. (Davis & Goadrich, 2006) rightly points out that PR curves can be more informative than ROC curves as they may more appropriately reflect deployment objectives. However, the informativeness of the curve itself is distinct from the informativeness of the Area under it. The PR curve's utility in reflecting target deployment objectives does not necessarily extend to AUPRC being superior to AUROC in all cases of class imbalance. Despite this, the claim is often extrapolated to assert, without caveat, that AUPRC is universally superior in such scenarios. This generalization overlooks the nuanced differences between the curves and their corresponding area metrics.

Papers that do make variations on this claim are frequently misunderstood, mis-cited, or are outright wrong While many papers reference this claim with either no or no appropriate citation, 97 do reference a paper with an associated citation to a work that makes either this claim explicitly or makes an apparently similar claim.[3] Unfortunately, those papers that do make a related claim are often misunderstood, mis-cited, or outright wrong. For example, consider Table 1, which shows the arguments commonly made among papers that argue this claim, with commentary on their validity. Of these claims, only two are logically sound, and those clearly do not imply that AUPRC is universally superior to AUROC in cases of class imbalance; rather, they emphasize that AUPRC may be better correlated with relevant deployment metrics for tasks whose use case is only single-stream retrieval. The other arguments are largely unfounded.

In general, many of these arguments seem to reflect the idea that "because AUPRC is often lower (and thus more clearly bad) for datasets with low prevalence, it has greater discriminatory power between poor and good models in this setting." This argument is not sound. Just because AUPRC tends to be lower (which is a consequence both of its chance value being equal to the prevalence of the positive label in the dataset and due to the greater importance placed on regions of low firing rate by Theorem 1) does not mean it will be more incisive in model comparison. The utility of a metric for model comparison is driven by how it separates models—*e.g.*, will the right kinds of models be ranked more highly under the evaluation metric than the wrong kinds of models. By our "mistake correction" framework in Section 2.2, it is clear that what differentiates AUROC and AUPRC from this perspective is that AUPRC favors improvements in regions of low firing rate, whereas AUROC is unbiased. This can be valuable in single-stream retrieval settings, but is far from indicative of general superiority in cases of class imbalance, and is dangerous from a fairness perspective, as discussed above.

Note that there are additional arguments often made in papers that merely *reference* the claim being made (rather than argue for it as a primary source), and those arguments differ from the ones in Table 1. See Appendix Table 2 for a breakdown of those arguments.

4.2. Valid insights from the current literature

Prior mathematical analyses have found complementary perspectives to our findings regarding the nature of the AUPRC Our mathematical findings are, while novel, highly concordant with prior findings in the literature. For example, (Yuan et al., 2015) showed that AUROC and AUPRC can both be seen to be weighted combinations of the same quantities, with AUROC having model-independent weights and AUPRC having task/model-dependent weights, which is fully revealed by our analyses. However, they argue this is a benefit, despite its clear ramifications on the utility and possible fairness implications of AUPRC. Furthermore, (Su et al., 2015) show that when rescaled to both lie between 0 and 1, AUPRC is approximately equal to AUROC scaled by the model's "initial precision rate", arguing that this justifies AUPRC's utility for retrieval tasks. Our framework shows much more clearly that AUPRC universally favors improving models from the highest scoring region down, justifying this preference more thoroughly.

There are other reasons why one may prefer AUROC over AUPRC (Hall et al., 2023) points out that the fact that AUPRC depends on prevalence can make it challenging to use as an evaluation metric to detect subgroup disparities, as different subgroups may have different prevalences.

5. Conclusion

This study rigorously interrogates the pervasive assumption within the machine learning community that AUPRC is a more appropriate evaluation metric than AUROC in class-imbalanced settings. Our empirical analyses, along with an exhaustive literature review, have revealed several important findings that critically challenge this belief.

Our empirical investigations highlight the robustness of AUROC to class imbalance, which may be desirable especially in domains like healthcare. In contrast, we find that selecting models using overall AUPRC can lead to disparities across subpopulations, particularly when these subpopulations have different outcome prevalence. Our work questions the reliability of AUPRC especially in settings where equity and fairness are imperative.

Our thorough literature review exposes a notable deficiency in the field: the advocacy of AUPRC over AUROC in imbalanced datasets often rests on shaky empirical grounds or stems from misconceptions and oversimplified interpretations. These unjustified endorsements in prior works also neglect the crucial implications of fairness in metric selection.

In summary, our research advocates for a more thoughtful and context-aware approach to selecting evaluation metrics in machine learning. This paradigm shift, favoring a balanced and conscientious approach to metric selection, is essential in advancing the field towards developing not only technically sound, but also equitable and just models.

[3] This number includes papers also reflected in the set of papers that reference an invalid citation, as some papers cite both a valid and an invalid source.

Claim	References	Valid?	Commentary
Precision-recall curves or other associated metrics *may* more appropriately reflect deployment objectives than the receiver operating characteristic.	(Cook & Ramadas, 2020; Leisman, 2018; Saito & Rehmsmeier, 2015; Yuan et al., 2015; Bleakley et al., 2007; Ozenne et al., 2015; Rosenberg, 2022; Zhou et al., 2020; Lichtnwalter & Chawla, 2012; Yang et al., 2015)	✓	While this claim is true, the informativeness of the PR curve for target deployment metrics is insufficient to conclude that the AUPRC is superior to the AUROC in all cases of class imbalance. Despite this, it is often taken to assert this more general claim without caveat.
AUPRC does not depend on the number of true negatives, so will be less optimistic than the AUROC	(Leisman, 2018; Goadrich et al., 2006; Cranmer & Desmarais, 2016)		As shown in Theorem 1, AUROC and AUPRC can both be naturally expressed as a function of the expectation of the model's false positive rate. More generally, the lack of dependence on one quadrant among the mutually dependent four quadrants of a confusion matrix is not an informative property for the AUROC and AUPRC metrics.
AUPRC will often be significantly lower, farther from optimality, and/or will grow more non-linearly as model performance improves than AUROC for low-prevalence tasks	(Leisman, 2018; Yuan et al., 2015; Goadrich et al., 2006; Mazzanti, 2023; Rosenberg, 2022; Zhou et al., 2020; Lichtnwalter & Chawla, 2012; Yang et al., 2015; Cranmer & Desmarais, 2016)	✓	Metric utility for model comparison depends on how appropriately it prioritizes model improvements. Therefore, it is less about the raw magnitude of the metric and more about the situations in which the order of a set of models will differ under one metric vs. another. One could easily make AUROC yield smaller values or grow more quickly near optimality by simply exponentiating it, but this would not yield a better metric.
AUPRC depends on prevalence, which is a desirable property	(Saito & Rehmsmeier, 2015; Goadrich et al., 2006; Yuan et al., 2015)		This statement is too vague to be formally evaluated; whether or not this dependence on prevalence is desirable depends on the context. For model comparison in general, we argue it is not desirable in this form as it induces the biases in AUPRC previously discussed.
AUPRC better captures differentiating a positive sample with high score from a "hard" negative sample ("hard" meaning one also with high score)	(Rosenberg, 2022)	✓	While this claim is true by Theorem 2, it is not clear why this would be desired in general; this implicitly favors comparing "hard" negatives against "easy" positives as opposed to "easy" negatives against "hard" positives.

Table 1. Various arguments and our responses to them present for this claim in the literature.

Acknowledgements

MBAM is supported by a Berkowitz Postdoctoral Fellowship. JG is funded by the National Institute of Health through DS-I Africa U54 TW012043-01 and Bridge2AI OT2OD032701.

References

Adler, A. Using machine learning techniques to identify key risk factors for diabetes and undiagnosed diabetes, 2021.

Afanasiev, S., Smirnova, A., and Kotereva, D. Itsy bitsy spidernet: Fully connected residual network for fraud detection, 2021.

Ahmed, F. and Courville, A. Detecting semantic anomalies. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):3154–3162, Apr. 2020. doi: 10.1609/aaai.v34i04.5712. URL https://ojs.aaai.org/index.php/AAAI/article/view/5712.

Albora, G. and Zaccaria, A. Machine learning to assess relatedness: the advantage of using firm-level data. *Complexity*, 2022, 2022.

Alvarez, M., Verdier, J.-C., Nkashama, D. K., Frappier, M., Tardif, P.-M., and Kabanza, F. A revealing large-scale evaluation of unsupervised anomaly detection algorithms, 2022.

Ando, S. and Huang, C. Y. Deep over-sampling framework for classifying imbalanced data. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2017, Skopje, Macedonia, September 18–22, 2017, Proceedings, Part I 10*, pp. 770–785. Springer, 2017.

Axelrod, S. and Gomez-Bombarelli, R. Molecular machine learning with conformer ensembles. *Machine Learning: Science and Technology*, 4(3):035025, 2023.

Babaei, K., Chen, Z. Y., and Maul, T. Aegr: a simple approach to gradient reversal in autoencoders for network anomaly detection. *Soft Computing*, 25(24):15269–15280, 2021.

Bach Nguyen, V., Ghosh Dastidar, K., Granitzer, M., and Siblini, W. The importance of future information in credit card fraud detection. In Camps-Valls, G., Ruiz, F. J. R., and Valera, I. (eds.), *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 10067–10077. PMLR, 28–30 Mar 2022. URL https://proceedings.mlr.press/v151/bach-nguyen22a.html.

Bleakley, K., Biau, G., and Vert, J.-P. Supervised reconstruction of biological networks with local models. *Bioinformatics*, 23(13):i57–i65, 2007.

Blevins, D., Moriano, P., Bridges, R., Verma, M., Iannacone, M., and Hollifield, S. Time-based can intrusion detection benchmark. In *Workshop on Automotive and Autonomous Vehicle Security (AutoSec)*, 2021.

Boyd, K., Eng, K. H., and Page, C. D. Area under the precision-recall curve: point estimates and confidence intervals. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2013, Prague, Czech Republic, September 23-27, 2013, Proceedings, Part III 13*, pp. 451–466. Springer, 2013.

Branco, P., Torgo, L., and Ribeiro, R. P. A survey of predictive modeling on imbalanced domains. *ACM computing surveys (CSUR)*, 49(2):1–50, 2016.

Brophy, J. and Lowd, D. Eggs: A flexible approach to relational modeling of social network spam, 2020.

Bryan, J. and Moriano, P. Graph-based machine learning improves just-in-time defect prediction. *Plos one*, 18(4): e0284077, 2023.

Budka, M., Ashraf, A. W. U., Bennett, M., Neville, S., and Mackrill, A. Deep multilabel cnn for forensic footwear impression descriptor identification. *Applied Soft Computing*, 109:107496, 2021.

Burez, J. and Van den Poel, D. Handling class imbalance in customer churn prediction. *Expert Systems with Applications*, 36(3, Part 1):4626–4636, 2009. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2008.05.027. URL https://www.sciencedirect.com/science/article/pii/S0957417408002121.

Chakraborty, N., Hasan, A., Liu, S., Ji, T., Liang, W., McPherson, D. L., and Driggs-Campbell, K. Structural attention-based recurrent variational autoencoder for highway vehicle anomaly detection. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '23, pp. 1125–1134, Richland, SC, 2023. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9781450394321.

Chen, R. J., Wang, J. J., Williamson, D. F., Chen, T. Y., Lipkova, J., Lu, M. Y., Sahai, S., and Mahmood, F. Algorithmic fairness in artificial intelligence for medicine and healthcare. *Nature biomedical engineering*, 7(6): 719–742, 2023.

Cho, B. Y., Hermans, T., and Kuntz, A. Planning sensing sequences for subsurface 3d tumor mapping. In *2021 International Symposium on Medical Robotics (ISMR)*, pp. 1–7. IEEE, 2021.

Choi, E., Xiao, C., Stewart, W. F., and Sun, J. Mime: Multilevel medical embedding of electronic health records for predictive healthcare. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pp. 4552–4562, Red Hook, NY, USA, 2018. Curran Associates Inc.

Chu, X., Lin, Y., Gao, J., Wang, J., Wang, Y., and Wang, L. Multi-label robust factorization autoencoder and its application in predicting drug-drug interactions, 2018.

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

Cranmer, S. J. and Desmarais, B. A. What can we learn from predictive modeling?, 2016.

Czakon, J. F1 Score vs ROC AUC vs Accuracy vs PR AUC: Which Evaluation Metric Should You Choose?, July 2022. URL https://neptune.ai/blog/f1-score-accuracy-roc-auc-pr-auc.

Danesh Pazho, A., Alinezhad Noghre, G., Rahimi Ardabili, B., Neff, C., and Tabkhi, H. *CHAD: Charlotte Anomaly Dataset*, pp. 50–66. Springer Nature Switzerland, 2023. ISBN 9783031314353. doi: 10.1007/978-3-031-31435-3_4. URL http://dx.doi.org/10.1007/978-3-031-31435-3_4.

Davis, J. and Goadrich, M. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pp. 233–240, New York, NY, USA, 2006. Association for Computing Machinery. ISBN 1595933832. doi: 10.1145/1143844.1143874. URL https://doi.org/10.1145/1143844.1143874.

Deng, J., Yang, Z., Wang, H., Ojima, I., Samaras, D., and Wang, F. Unraveling key elements underlying molecular property prediction: A systematic study, 2023.

Deshwar, A. G., Vembu, S., Yung, C. K., Jang, G. H., Stein, L., and Morris, Q. Reconstructing subclonal composition and evolution from whole genome sequencing of tumors, 2015.

Ding, D. Y., Simpson, C., Pfohl, S., Kale, D. C., Jung, K., and Shah, N. H. The effectiveness of multitask learning for phenotyping with electronic health records data. In *BIOCOMPUTING 2019: Proceedings of the Pacific Symposium*, pp. 18–29. World Scientific, 2018.

Domingues, R., Michiardi, P., Barlet, J., and Filippone, M. A comparative evaluation of novelty detection algorithms for discrete sequences. *Artificial Intelligence Review*, 53: 3787–3812, 2020.

Ezzat, A., Zhao, P., Wu, M., Li, X.-L., and Kwoh, C.-K. Drug-target interaction prediction with graph regularized matrix factorization. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 14(3):646–656, 2017. doi: 10.1109/TCBB.2016.2530062.

Flach, P., Hernández-Orallo, J., and Ferri, C. A coherent interpretation of auc as a measure of aggregated classification performance. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, pp. 657–664, Madison, WI, USA, 2011. Omnipress. ISBN 9781450306195.

Fletcher, R. R., Nakeshimana, A., and Olubeko, O. Addressing fairness, bias, and appropriate use of artificial intelligence and machine learning in global health, 2021.

Fu, Y., Wu, X.-B., Yang, Q., Brown, A. G., Feng, X., Ma, Q., and Li, S. Finding quasars behind the galactic plane. i. candidate selections with transfer learning. *The Astrophysical Journal Supplement Series*, 254(1):6, 2021.

Garcin, M. and Stéphan, S. Credit scoring using neural networks and sure posterior probability calibration, 2021.

Gaudreault, J.-G., Branco, P., and Gama, J. An analysis of performance metrics for imbalanced classification. In *International Conference on Discovery Science*, pp. 67–77. Springer, 2021.

Goadrich, M., Oliphant, L., and Shavlik, J. Gleaner: Creating ensembles of first-order clauses to improve recall-precision curves. *Machine Learning*, 64:231–261, 2006.

Gong, H., Valido, A., Ingram, K. M., Fanti, G., Bhat, S., and Espelage, D. L. Abusive language detection in heterogeneous contexts: Dataset collection and the role of supervised attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35(17), pp. 14804–14812, 2021.

Goyal, A. and Khiari, J. Diversity-aware weighted majority vote classifier for imbalanced data. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, 2020. doi: 10.1109/IJCNN48605.2020.9207261.

Hagedoorn, T. R. and Spanakis, G. Massive open online courses temporal profiling for dropout prediction. In *2017 IEEE 29th International Conference on Tools with Artificial Intelligence (ICTAI)*, pp. 231–238. IEEE, 2017.

Hall, M., Chern, B., Gustafson, L., Ventura, D., Kulkarni, H., Ross, C., and Usunier, N. Towards reliable assessments of demographic disparities in multi-label image classifiers, 2023.

Harer, J. A., Kim, L. Y., Russell, R. L., Ozdemir, O., Kosta, L. R., Rangamani, A., Hamilton, L. H., Centeno, G. I., Key, J. R., Ellingwood, P. M., Antelman, E., Mackay, A., McConley, M. W., Opper, J. M., Chin, P., and Lazovich, T. Automated software vulnerability detection with machine learning, 2018.

Hashemi, S. R., Salehi, S. S. M., Erdogmus, D., Prabhu, S. P., Warfield, S. K., and Gholipour, A. Asymmetric loss functions and deep densely-connected networks for highly-imbalanced medical image segmentation: Application to multiple sclerosis lesion detection. *IEEE Access*, 7:1721–1735, 2018.

He, H. and Garcia, E. A. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, 2009. doi: 10.1109/TKDE.2008.239.

Herbach, U. Gene regulatory network inference from single-cell data using a self-consistent proteomic field, 2021.

Hibshman, J. I. and Weninger, T. Inherent limits on topology-based link prediction, 2023.

Hicks, S. A., Strümke, I., Thambawita, V., Hammou, M., Riegler, M. A., Halvorsen, P., and Parasa, S. On evaluation metrics for medical applications of artificial intelligence. *Scientific reports*, 12(1):5979, 2022.

Hiri, K. D., Hren, M., and Curk, T. Nlp-based classification of software tools for metagenomics sequencing data analysis into edam semantic annotation, 2022.

Hong, S., Xiao, C., Hoang, T. N., Ma, T., Li, H., and Sun, J. Rdpd: Rich data helps poor data via imitation. In *28th International Joint Conference on Artificial Intelligence, IJCAI 2019*, pp. 5895–5901. International Joint Conferences on Artificial Intelligence, 2019.

Hsu, C.-C., Karnwal, S., Mullainathan, S., Obermeyer, Z., and Tan, C. Characterizing the value of information in medical notes. In Cohn, T., He, Y., and Liu, Y. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 2062–2072, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.187. URL https://aclanthology.org/2020.findings-emnlp.187.

Isupova, O., Kuzin, D., and Mihaylova, L. Learning methods for dynamic topic modeling in automated behavior analysis. *IEEE transactions on neural networks and learning systems*, 29(9):3980–3993, 2017.

Ju, C., Li, J., Wasti, B., and Guo, S. Semisupervised learning on heterogeneous graphs and its applications to facebook news feed, 2018.

Karadzhov, G., Stafford, T., and Vlachos, A. What makes you change your mind? an empirical investigation in online group decision-making conversations, 2022.

Kim, M., Kim, J., Yu, J., and Choi, J. K. Unsupervised deep one-class classification with adaptive threshold based on training dynamics. In *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*, pp. 39–46, 2022. doi: 10.1109/ICDMW58026.2022.00014.

Kiran, B. R., Thomas, D. M., and Parakkal, R. An overview of deep learning based methods for unsupervised and semi-supervised anomaly detection in videos. *Journal of Imaging*, 4(2), 2018. ISSN 2313-433X. doi: 10.3390/jimaging4020036. URL https://www.mdpi.com/2313-433X/4/2/36.

Krawczyk, B. Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, 5(4):221–232, 2016.

Kulkarni, V., Gawali, M., and Kharat, A. Key technology considerations in developing and deploying machine learning models in clinical radiology practice. *JMIR Med Inform*, 9(9):e28776, Sep 2021. ISSN 2291-9694. doi: 10.2196/28776. URL https://medinform.jmir.org/2021/9/e28776.

Kyono, T., Gilbert, F. J., and van der Schaar, M. Mammo: A deep learning solution for facilitating radiologist-machine collaboration in breast cancer diagnosis, 2018.

LCT14558. Imbalanced data & why you should NOT use ROC curve, March 2017. URL https://kaggle.com/code/lct14558/imbalanced-data-why-you-should-not-use-roc-curve.

Lee, C., Nick, B., Brandes, U., and Cunningham, P. Link prediction with social vector clocks. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 784–792, 2013.

Lee, I.-T., Marwah, M., and Arlitt, M. Attention-based self-supervised feature learning for security data, 2020.

Lei Ba, J., Swersky, K., Fidler, S., et al. Predicting deep zero-shot convolutional neural networks using textual descriptions. In *Proceedings of the IEEE international conference on computer vision*, pp. 4247–4255, 2015.

Leisman, D. E. Rare events in the icu: an emerging challenge in classification and prediction. *Critical care medicine*, 46(3):418–424, 2018.

Li, Q., Zhang, Y., Qiu, D., He, Y., Cao, L., and Woodland, P. C. Improving confidence estimation on out-of-domain data for end-to-end speech recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6537–6541. IEEE, 2022.

Li, X., Al-Zaidy, R., Zhang, A., Baral, S., Bao, L., and Giles, C. L. Automating document classification with distant supervision to increase the efficiency of systematic reviews, 2020.

Lichtnwalter, R. and Chawla, N. V. Link prediction: fair and effective evaluation. In *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pp. 376–383. IEEE, 2012.

Lim, B. and van der Schaar, M. Disease-atlas: Navigating disease trajectories using deep learning. In *Machine Learning for Healthcare Conference*, pp. 137–160. PMLR, 2018.

Liu, Y., Yang, D., Wang, Y., Liu, J., Liu, J., Boukerche, A., Sun, P., and Song, L. Generalized video anomaly event detection: Systematic taxonomy and comparison of deep models, 2023.

Lobo, J. M., Jiménez-Valverde, A., and Real, R. Auc: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, 17(2):145–151, 2008. doi: https://doi.org/10.1111/j.1466-8238.2007.00358.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1466-8238.2007.00358.x.

Lopez-Martinez, D., Yakubovich, A., Seneviratne, M., Lelkes, A. D., Tyagi, A., Kemp, J., Steinberg, E., Downing, N. L., Li, R. C., Morse, K. E., Shah, N. H., and Chen, M.-J. Instability in clinical risk stratification models using deep learning. In Parziale, A., Agrawal, M., Joshi, S., Chen, I. Y., Tang, S., Oala, L., and Subbaswamy, A. (eds.), *Proceedings of the 2nd Machine Learning for Health symposium*, volume 193 of *Proceedings of Machine Learning Research*, pp. 552–565. PMLR, 28 Nov 2022. URL https://proceedings.mlr.press/v193/lopez-martinez22a.html.

Lund, J., Armstrong, P., Fearn, W., Cowley, S., Hales, E., and Seppi, K. Cross-referencing using fine-grained topic modeling. In Burstein, J., Doran, C., and Solorio, T. (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3978–3987, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1399. URL https://aclanthology.org/N19-1399.

Lyu, Y., Rajbahadur, G. K., Lin, D., Chen, B., and Jiang, Z. M. J. Towards a consistent interpretation of aiops models. *ACM Trans. Softw. Eng. Methodol.*, 31(1), nov 2021. ISSN 1049-331X. doi: 10.1145/3488269. URL https://doi.org/10.1145/3488269.

López, V., Fernández, A., García, S., Palade, V., and Herrera, F. An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information Sciences*, 250:113–141, 2013. ISSN 0020-0255. doi: https://doi.org/10.1016/j.ins.2013.07.007. URL https://www.sciencedirect.com/science/article/pii/S0020025513005124.

Ma, L., Zhang, C., Wang, Y., Ruan, W., Wang, J., Tang, W., Ma, X., Gao, X., and Gao, J. Concare: Personalized clinical feature embedding via capturing the healthcare context. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34(01), pp. 833–840, 2020.

Ma, X., Chu, X., Wang, Y., Yu, H., Ma, L., Tang, W., and Zhao, J. Medfact: Modeling medical feature correlations in patient health representation learning via feature clustering, 2022.

Mangolin, R. B., Pereira, R. M., Britto Jr, A. S., Silla Jr, C. N., Feltrim, V. D., Bertolini, D., and Costa, Y. M. A multimodal approach for multi-label movie genre classification. *Multimedia Tools and Applications*, 81(14): 19071–19096, 2022.

Markdahl, J., Colombo, N., Thunberg, J., and Gonçalves, J. Experimental design trade-offs for gene regulatory network inference: An in silico study of the yeast saccharomyces cerevisiae cell cycle. In *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, pp. 423–428. IEEE, 2017.

Mayaki, M. Z. A. and Riveill, M. Multiple inputs neural networks for fraud detection. In *2022 International Conference on Machine Learning, Control, and Robotics (MLCR)*, pp. 8–13, 2022. doi: 10.1109/MLCR57210.2022.00011.

Mazzanti, S. Why you should stop using the ROC curve, September 2023. URL https://towardsdatascience.com/why-you-should-stop-using-the-roc-curve-a46a9adc7

Mehboudi, A., Singhal, S., and Sreenivasan, S. V. Squeeze flow of micro-droplets: convolutional neural network with trainable and tunable refinement, 2022.

Meister, J. A., Nguyen, K. A., and Luo, Z. Audio feature ranking for sound-based covid-19 patient detection. In *EPIA Conference on Artificial Intelligence*, pp. 146–158. Springer, 2022.

Miao, J. and Zhu, W. Precision–recall curve (prc) classification trees. *Evolutionary intelligence*, 15(3):1545–1569, 2022.

Mongia, A., Saha, S. K., Chouzenoux, E., and Majumdar, A. A computational approach to aid clinicians in selecting anti-viral drugs for covid-19 trials. *Scientific reports*, 11 (1):9047, 2021.

Moor, M., Horn, M., Rieck, B., Roqueiro, D., and Borgwardt, K. Early recognition of sepsis with gaussian process temporal convolutional networks and dynamic time warping. In Doshi-Velez, F., Fackler, J., Jung, K., Kale, D., Ranganath, R., Wallace, B., and Wiens, J. (eds.), *Proceedings of the 4th Machine Learning for Healthcare Conference*, volume 106 of *Proceedings of Machine Learning Research*, pp. 2–26. PMLR, 09–10 Aug 2019. URL https://proceedings.mlr.press/v106/moor19a.html.

Mosquera, C., Ferrer, L., Milone, D., Luna, D., and Ferrante, E. Impact of class imbalance on chest x-ray classifiers: towards better evaluation practices for discrimination and calibration performance, 2022.

Mosteiro, P., Rijcken, E., Zervanou, K., Kaymak, U., Scheepers, F., and Spruit, M. Machine learning for violence risk assessment using dutch clinical notes. *Journal of Artificial Intelligence for Medical Sciences*, 2(1-2): 44–54, 2021.

Mousavian, Z., Khakabimamaghani, S., Kavousi, K., and Masoudi-Nejad, A. Drug–target interaction prediction from pssm based evolutionary information. *Journal of pharmacological and toxicological methods*, 78:42–51, 2016.

Muthukrishna, D., Narayan, G., Mandel, K. S., Biswas, R., and Hložek, R. Rapid: early classification of explosive transients using deep learning. *Publications of the Astronomical Society of the Pacific*, 131(1005):118002, 2019.

Narayanan, S., Maple, C., and Hooper, M. A point process model for rare event detection, 2022.

Navarro, J. M., Huet, A., and Rossi, D. Human readable network troubleshooting based on anomaly detection and feature scoring. *Computer Networks*, 219:109447, 2022.

Newby, E., Tejeda Zañudo, J. G., and Albert, R. Structure-based approach to identifying small sets of driver nodes in biological networks. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 32(6):063102, 06 2022. ISSN 1054-1500. doi: 10.1063/5.0080843. URL https://doi.org/10.1063/5.0080843.

Ntroumpogiannis, A., Giannoulis, M., Myrtakis, N., Christophides, V., Simon, E., and Tsamardinos, I. A meta-level analysis of online anomaly detectors. *The VLDB Journal*, pp. 1–42, 2023.

Ozenne, B., Subtil, F., and Maucort-Boulch, D. The precision–recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases. *Journal of clinical epidemiology*, 68(8):855–859, 2015.

Ozyegen, O., Kabe, D., and Cevik, M. Word-level text highlighting of medical texts for telehealth services. *Artificial Intelligence in Medicine*, 127:102284, 2022.

Pang, G., Shen, C., Jin, H., and van den Hengel, A. Deep weakly-supervised anomaly detection. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, pp. 1795–1807, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701030. doi: 10.1145/3580305.3599302. URL https://doi.org/10.1145/3580305.3599302.

Pashchenko, I. N., Sokolovsky, K. V., and Gavras, P. Machine learning search for variable stars. *Monthly Notices of the Royal Astronomical Society*, 475(2):2326–2343, 2018.

Piermarini, D., Sudoso, A. M., and Piccialli, V. Predicting municipalities in financial distress: a machine learning approach enhanced by domain expertise, 2023.

Prapas, I., Ahuja, A., Kondylatos, S., Karasante, I., Panagiotou, E., Alonso, L., Davalas, C., Michail, D., Carvalhais, N., and Papoutsis, I. Deep learning for global wildfire forecasting, 2023.

Provost, F. and Fawcett, T. Analysis and visualization of classifier performance with nonuniform class and cost distributions. In *Proceedings of AAAI-97 Workshop on AI Approaches to Fraud Detection & Risk Management*, pp. 57–63, 1997.

Rajabi, F. and He, J. S. Click-through rate prediction using graph neural networks and online learning, 2021.

Randl, K., Armengol, N. L., Mondrejevski, L., and Miliou, I. Early prediction of the risk of icu mortality with deep federated learning. In *2023 IEEE 36th International Symposium on Computer-Based Medical Systems (CBMS)*, pp. 706–711. IEEE, 2023.

Rao, S. X., Lanfranchi, C., Zhang, S., Han, Z., Zhang, Z., Min, W., Cheng, M., Shan, Y., Zhao, Y., and Zhang, C. Modelling graph dynamics in fraud detection with "attention", 2022.

Rayhan, F., Ahmed, S., Shatabda, S., Farid, D. M., Mousavian, Z., Dehzangi, A., and Rahman, M. S. idti-esboost: identification of drug target interaction using evolutionary and structural features with boosting. *Scientific reports*, 7 (1):17731, 2017.

Rayhan, F., Ahmed, S., Mousavian, Z., Farid, D. M., and Shatabda, S. Frnet-dti: Deep convolutional neural network for drug-target interaction prediction. *Heliyon*, 6(3), 2020.

Rezvani, R., Kouchaki, S., Nilforooshan, R., Sharp, D. J., and Barnaghi, P. Semi-supervised learning for identifying the likelihood of agitation in people with dementia, 2021.

Rohani, N. and Eslahchi, C. Drug-drug interaction predicting by neural network using integrated similarity. *Scientific reports*, 9(1):13645, 2019.

Romero, M., Ramírez, O., Finke, J., and Rocha, C. Feature extraction with spectral clustering for gene function prediction using hierarchical multi-label classification. *Applied Network Science*, 7(1):28, 2022.

Rosenberg, D. Imbalanced Data? Stop Using ROC-AUC and Use AUPRC Instead, June 2022. URL https://towardsdatascience.com/imbalanced-data-stop-using-roc-auc-and-use-auprc-instead-46af4910a494.

Rubin, T. N., Chambers, A., Smyth, P., and Steyvers, M. Statistical topic models for multi-label document classification. *Machine learning*, 88:157–208, 2012.

Ruff, L., Kauffmann, J. R., Vandermeulen, R. A., Montavon, G., Samek, W., Kloft, M., Dietterich, T. G., and Müller, K.-R. A unifying review of deep and shallow anomaly detection. *Proceedings of the IEEE*, 109(5):756–795, 2021. doi: 10.1109/JPROC.2021.3052449.

Sahiner, B., Chen, W., Pezeshk, A., and Petrick, N. Comparison of two classifiers when the data sets are imbalanced: the power of the area under the precision-recall curve as the figure of merit versus the area under the ROC curve. In Kupinski, M. A. and Nishikawa, R. M. (eds.), *Medical Imaging 2017: Image Perception, Observer Performance, and Technology Assessment*, volume 10136, pp. 101360G. International Society for Optics and Photonics, SPIE, 2017. doi: 10.1117/12.2254742. URL https://doi.org/10.1117/12.2254742.

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

Sarvari, H., Domeniconi, C., Prenkaj, B., and Stilo, G. Unsupervised boosting-based autoencoder ensembles for outlier detection. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 91–103. Springer, 2021.

Schwarz, K., Allam, A., Perez Gonzalez, N. A., and Krauthammer, M. Attentionddi: Siamese attention-based deep learning method for drug–drug interaction predictions. *BMC bioinformatics*, 22(1):1–19, 2021.

Seo, E., Hutchinson, R. A., Fu, X., Li, C., Hallman, T. A., Kilbride, J., and Robinson, W. D. Stateconet: Statistical ecology neural networks for species distribution modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 513–521, 2021.

Shen, H. and Kursun, E. Label augmentation via time-based knowledge distillation for financial anomaly detection, 2021.

Showalter, S. and Wu, Z. Minimizing the societal cost of credit card fraud with limited and imbalanced data, 2019.

Shukla, S. N. and Marlin, B. Interpolation-prediction networks for irregularly sampled time series. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=r1efr3C9Ym.

Shukla, S. N. and Marlin, B. M. Modeling irregularly sampled clinical time series, 2018.

Si, Y. and Roberts, K. Three-level hierarchical transformer networks for long-sequence and multiple clinical documents classification, 2021.

Silva, M. C. R., Siqueira, F. A., Tarrega, J. P. M., Beinotti, J. V. P., Nunes, A. S., de Mattos Gardini, M., da Silva, V. A. P., da Silva, N. F. F., and de Leon Ferreira de Carvalho, A. C. P. No pattern, no recognition: a survey about reproducibility and distortion issues of text clustering and topic modeling, 2022.

Skarding, J., Gabrys, B., and Musial, K. Foundations and modeling of dynamic networks using dynamic graph neural networks: A survey. *IEEE Access*, 9:79143–79168, 2021.

Smith, A. L., Zheng, T., and Gelman, A. Prediction scoring of data-driven discoveries for reproducible research. *Statistics and Computing*, 33(1):11, 2023.

Sokolova, M. and Lapalme, G. A systematic analysis of performance measures for classification tasks. *Information processing & management*, 45(4):427–437, 2009.

Srivastava, S., Namboodiri, V. P., and Prabhakar, T. V. Putworkbench: Analysing privacy in ai-intensive systems, 2019.

Steinbuss, G. and Böhm, K. Benchmarking unsupervised outlier detection with realistic synthetic data. *ACM Trans. Knowl. Discov. Data*, 15(4), apr 2021. ISSN 1556-4681. doi: 10.1145/3441453. URL https://doi.org/10.1145/3441453.

Stephenson, O. L., Köhne, T., Zhan, E., Cahill, B. E., Yun, S.-H., Ross, Z. E., and Simons, M. Deep learning-based damage mapping with insar coherence time series. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–17, 2022. doi: 10.1109/TGRS.2021.3084209.

Stolman, A., Levy, C., Seshadhri, C., and Sharma, A. Classic graph structural features outperform factorization-based graph embedding methods on community labeling. In *Proceedings of the 2022 SIAM International Conference on Data Mining (SDM)*, pp. 388–396. SIAM, 2022.

Su, W., Yuan, Y., and Zhu, M. A relationship between the average precision and the area under the roc curve. In *Proceedings of the 2015 International Conference on The Theory of Information Retrieval*, ICTIR '15, pp. 349–352, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450338332. doi: 10.1145/2808194.2809481. URL https://doi-org.ezp-prod1.hul.harvard.edu/10.1145/2808194.2809481.

Thambawita, V., Jha, D., Hammer, H. L., Johansen, H. D., Johansen, D., Halvorsen, P., and Riegler, M. A. An extensive study on cross-dataset bias and evaluation metrics interpretation for machine learning applied to gastrointestinal tract abnormality classification. *ACM Transactions on Computing for Healthcare*, 1(3):1–29, 2020.

Tiulpin, A., Klein, S., Bierma-Zeinstra, S. M., Thevenot, J., Rahtu, E., Meurs, J. v., Oei, E. H., and Saarakkala, S. Multimodal machine learning-based knee osteoarthritis progression prediction from plain radiographs and clinical data. *Scientific reports*, 9(1):20038, 2019.

Torfi, A., Fox, E. A., and Reddy, C. K. Differentially private synthetic medical data generation using convolutional gans. *Information Sciences*, 586:485–500, 2022.

Tusfiqur, H. M., Nguyen, D. M. H., Truong, M. T. N., Nguyen, T. A., Nguyen, B. T., Barz, M., Profitlich, H.-J., Than, N. T. T., Le, N., Xie, P., and Sonntag, D. Drg-net: Interactive joint learning of multi-lesion segmentation and classification for diabetic retinopathy grading, 2022.

Vens, C., Struyf, J., Schietgat, L., Džeroski, S., and Blockeel, H. Decision trees for hierarchical multi-label classification. *Machine learning*, 73:185–214, 2008.

Vijayan, V., Critchlow, D., and Milenković, T. Alignment of dynamic networks. *Bioinformatics*, 33(14): i180–i189, 07 2017. ISSN 1367-4803. doi: 10.1093/bioinformatics/btx246. URL https://doi.org/10.1093/bioinformatics/btx246.

Wagner, S. J., Reisenbüchler, D., West, N. P., Niehues, J. M., Veldhuizen, G. P., Quirke, P., Grabsch, H. I., van den Brandt, P. A., Hutchins, G. G. A., Richman, S. D., Yuan, T., Langer, R., Jenniskens, J. C. A., Offermans, K., Mueller, W., Gray, R., Gruber, S. B., Greenson, J. K., Rennert, G., Bonner, J. D., Schmolze, D., James, J. A., Loughrey, M. B., Salto-Tellez, M., Brenner, H., Hoffmeister, M., Truhn, D., Schnabel, J. A., Boxberg, M., Peng, T., and Kather, J. N. Fully transformer-based biomarker prediction from colorectal cancer histology: a large-scale multicentric study, 2023.

Wahid-Ul-Ashraf, A., Budka, M., and Musial, K. How to predict social relationships — physics-inspired approach to link prediction. *Physica A: Statistical Mechanics and its Applications*, 523:1110–1129, 2019. ISSN 0378-4371. doi: https://doi.org/10.1016/j.physa.2019.04.246. URL https://www.sciencedirect.com/science/article/pii/S0378437119306193.

Weiss, M. and Tonella, P. Fail-safe execution of deep learning based systems through uncertainty monitoring. In *2021 14th IEEE conference on software testing, verification and validation (ICST)*, pp. 24–35. IEEE, 2021.

Weiss, M. and Tonella, P. Uncertainty quantification for deep neural networks: An empirical comparison and usage guidelines. *Software Testing, Verification and Reliability*, 33(6):e1840, 2023. doi: https://doi.org/10.1002/stvr.1840. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/stvr.1840.

Won, M., Chun, S., and Serra, X. Toward interpretable music tagging with self-attention, 2019.

Wu, P., Liu, J., Shi, Y., Sun, Y., Shao, F., Wu, Z., and Yang, Z. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pp. 322–339. Springer, 2020.

Yang, T. Deep auc maximization for medical image classification: Challenges and opportunities, 2021.

Yang, X., Yang, G., and Chu, J. The computational drug repositioning without negative sampling. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 20(2):1506–1517, 2022a.

Yang, Y., Lichtenwalter, R. N., and Chawla, N. V. Evaluating link prediction methods. *Knowledge and Information Systems*, 45:751–782, 2015.

Yang, Z.-Y., Ye, Z.-F., Xiao, Y.-J., Hsieh, C.-Y., and Zhang, S.-Y. Spldextratrees: robust machine learning approach for predicting kinase inhibitor resistance. *Briefings in Bioinformatics*, 23(3):bbac050, 2022b.

Yuan, Y., Su, W., and Zhu, M. Threshold-free measures for assessing the performance of medical screening tests. *Frontiers in public health*, 3:57, 2015.

Zavrtanik, V., Kristan, M., and Skočaj, D. Draem-a discriminatively trained reconstruction embedding for surface anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8330–8339, 2021.

Zhang, D., Fu, H., Han, J., Borji, A., and Li, X. A review of co-saliency detection technique: Fundamentals, applications, and challenges, 2017.

Zhang, W., Hisano, R., Ohnishi, T., and Mizuno, T. Non-diagonal mixture of dirichlet network distributions for analyzing a stock ownership network. In *Complex Networks & Their Applications IX: Volume 1, Proceedings of the Ninth International Conference on Complex Networks and Their Applications COMPLEX NETWORKS 2020*, pp. 75–86. Springer, 2021.

Zhang, Y. and Bondell, H. D. Variable Selection via Penalized Credible Regions with Dirichlet–Laplace Global-Local Shrinkage Priors. *Bayesian Analysis*, 13(3):823 – 844, 2018. doi: 10.1214/17-BA1076. URL https://doi.org/10.1214/17-BA1076.

Zhou, Q. M., Lu, Z., Brooke, R. J., Hudson, M. M., and Yuan, Y. Is the new model better? one metric says yes, but the other says no. which metric do i use?, 2020.

Zou, Y., Jeong, J., Pemula, L., Zhang, D., and Dabeer, O. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In *European Conference on Computer Vision*, pp. 392–408. Springer, 2022.

A. Notation

Let $f_{\boldsymbol{\theta}} : \mathcal{X} \to (0,1)$ denote a model, with $\mathcal{X}$ its input domain and $(0,1) = \{r \in \mathbb{R} | 0 < r < 1\}$ its output domain for a binary classification model. Let $\boldsymbol{x}_i \in \mathcal{X}$ be an input sample, $p_i^{(\boldsymbol{\theta})} = f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)$ its predicted probability under the model with parameters $\boldsymbol{\theta}$ ($\boldsymbol{\theta}$ may be omitted in context) and $y_i \in \{0,1\}$ its true label.

Let $\mathrm{N_P}$ be the number of data points with a positive label and $\mathrm{N_N}$ the number with a negative label. Further, given a threshold τ, define

$$\begin{aligned}
\mathrm{TP}_{\boldsymbol{\theta}}(\tau) &= \left|\{x_i \in \boldsymbol{X} | p_i^{(\boldsymbol{\theta})} \geq \tau, y_i = 1\}\right| \\
\mathrm{FN}_{\boldsymbol{\theta}}(\tau) &= \left|\{x_i \in \boldsymbol{X} | p_i^{(\boldsymbol{\theta})} < \tau, y_i = 1\}\right| \\
\mathrm{TN}_{\boldsymbol{\theta}}(\tau) &= \left|\{x_i \in \boldsymbol{X} | p_i^{(\boldsymbol{\theta})} < \tau, y_i = 0\}\right| \\
\mathrm{FP}_{\boldsymbol{\theta}}(\tau) &= \left|\{x_i \in \boldsymbol{X} | p_i^{(\boldsymbol{\theta})} \geq \tau, y_i = 0\}\right| \\
\mathrm{TPR}_{\boldsymbol{\theta}}(\tau) &= \frac{\mathrm{TP}_{\boldsymbol{\theta}}(\tau)}{\mathrm{N_P}} \\
&= P_{\mathsf{p|y=1}}(p > \tau) \\
&= P(p_+ > \tau) \\
\mathrm{FPR}_{\boldsymbol{\theta}}(\tau) &= \frac{\mathrm{FP}_{\boldsymbol{\theta}}(\tau)}{\mathrm{N_P}} \\
&= P_{\mathsf{p|y=0}}(p > \tau) \\
&= P(p_- > \tau) \\
\mathrm{Prec}_{\boldsymbol{\theta}}(\tau) &= \frac{\mathrm{TP}_{\boldsymbol{\theta}}(\tau)}{\mathrm{TP}_{\boldsymbol{\theta}}(\tau) + \mathrm{FP}_{\boldsymbol{\theta}}(\tau)} \\
&= P_{\mathsf{y|p>\tau}}(y = 1) \\
\mathrm{AUROC}_{\boldsymbol{\theta}} &= \int_0^1 \mathrm{TPR}_{\boldsymbol{\theta}} \frac{d\mathrm{FPR}_{\boldsymbol{\theta}}}{d\tau} d\tau \\
&= \int_0^1 \mathrm{TPR}_{\boldsymbol{\theta}} d\mathrm{FPR}_{\boldsymbol{\theta}} \\
&= 1 - \int_0^1 \mathrm{FPR}_{\boldsymbol{\theta}} d\mathrm{TPR}_{\boldsymbol{\theta}} \\
\mathrm{AUPRC}_{\boldsymbol{\theta}} &= \int_0^1 \mathrm{Prec}_{\boldsymbol{\theta}} \frac{d\mathrm{TPR}_{\boldsymbol{\theta}}}{d\tau} d\tau \\
&= \int_0^1 \mathrm{Prec}_{\boldsymbol{\theta}} d\mathrm{TPR}_{\boldsymbol{\theta}}
\end{aligned}$$

B. Details for Synthetic Experiments

B.1. Sampling a random model with a given AUROC

A key component of our synthetic experiments is the ability to sample a set of model scores and labels randomly that will have a target AUROC. To do this, we use the following procedure (which may or may not be previously known; we derived it from scratch for this work, but make no claim about its novelty). Let N be the number of points we are sampling overall, and N_+ be the number of positive points being sampled (which is dictated by the user given prevalence).

1. Uniformly sample a random collection of positive-label sample scores between zero and one.
2. Between each (ascending) model positive score indexed from 1 $p_+^{(i)}$ and $p_+^{(i+1)}$, we can count the number of positive samples that have scores less than any value in this window (i) and the number that have scores greater than any value in this window (which will be $N_+ - i$).
3. As the target AUROC is the probability that a randomly sampled negative will be ranked more highly than a randomly sampled positive, we can leverage the number of less-than positive scores i and greater than positive scores $N_+ - i$ to compute the probability that a randomly sampled negative score will live in the window $(p_+^{(i)}, p_+^{(i+1)})$ via the binomial distribution.
4. Now, to sample a random negative, we simply first sample a random window $(p_+^{(i)}, p_+^{(i+1)})$ with the probabilities assigned above, then uniformly sample a value p_- within that window. We can repeat this process to the target number of negative samples $N - N_+$ to form our final set of scores.
5. If desired, the output scores can further be scaled to have expectation given by the dataset's prevalence or can be adjusted via a calibration method to be calibrated given the assigned labels. Both procedures can be done without affecting the AUROC. Note that as any calibrated model will have expected probability given by the label's prevalence (See Appendix B.2), the former condition is strictly weaker than the latter.

B.2. Calibration includes prevalence matching

Let p be a random variable describing the probabilities output by the model over the input distribution defined by the data generative function. If a model is calibrated, this means that $P_{\mathsf{y}|\mathsf{p}}(y = 1|p = q) = q$ — that the probability that the label for a given point is 1 is given precisely by the models output probability for that sample. With that in mind, we have:

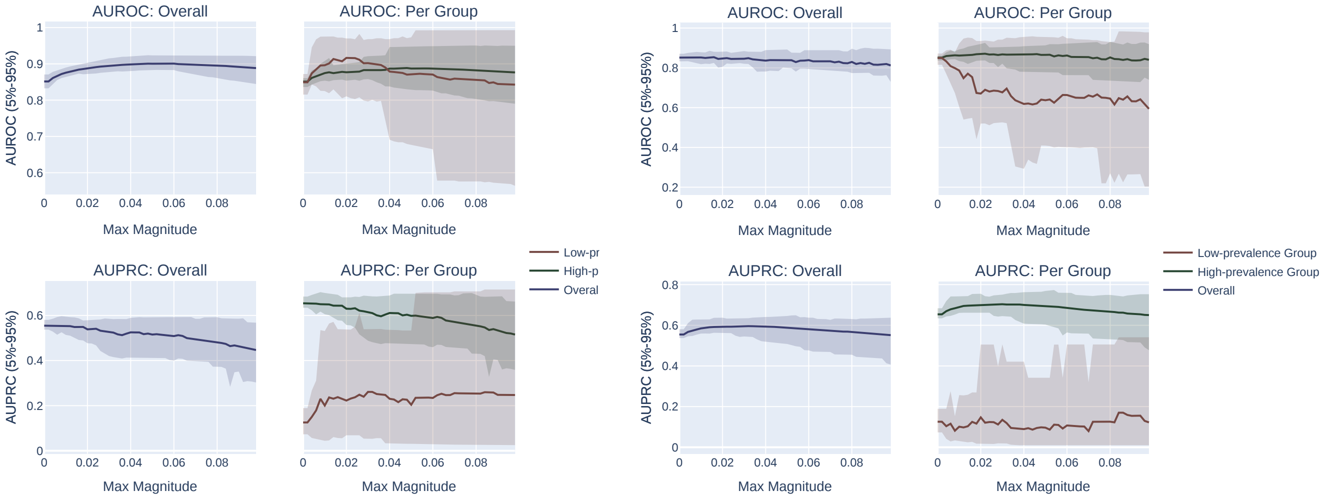
$$\begin{aligned}
\mathbb{E}_{\mathsf{p}}\left[q\right] &= \mathbb{E}_{\mathsf{p}}\left[P_{\mathsf{y}|\mathsf{p}}(y = 1|p = q)\right] \\
&= \int_0^1 P_{\mathsf{y}|\mathsf{p}}(y = 1|p = q)p_{\mathsf{p}}(q)dq \\
&= \int_0^1 P_{\mathsf{y},\mathsf{p}}(y = 1, p = q)dq \\
&= P_{\mathsf{y}}(y = 1)
\end{aligned}$$

B.3. Details on optimization procedures

M1. **Adding Random Noise.** We sample a vector $\epsilon \in \mathbb{R}^n$, where each element is uniformly drawn from $[-\delta, \delta]$. We compute the selection metric for $S' = S + \epsilon$. We repeat this procedure 100 times, and return the S' that achieves the maximum value for the selection metric. We vary the maximum magnitude of the perturbation $\delta \in [0, 0.1]$ in a grid. Results for this setting are shown in Figure 3.

M2. **Sequentially Fixing Atomic Mistakes.** We sequentially correct atomic mistakes, as defined in Figure 1. At each step, we first discover the set of all atomic mistakes M. To maximize AUROC, we randomly select a pair $(S_i, S_j) \in M$, and swap their scores in S, i.e. $S'_i = S_j, S'_j = S_i$. To maximize AUPRC, we swap the scores for the pair $(S_i, S_j) = \arg\max_{(s_i, s_j) \in M} s_j$. We repeat this process for 50 steps, with each one sequentially fixing another atomic mistake in S. Results for this setting are shown in Figures 2b and 2a.

M3. **Sequentially Permuting Nearby Scores.** We first sort S and Y such that S is in ascending order. We apply a random permutation to S by re-indexing it using a random ordering, but such that scores are not shuffled too far from their original index. Let σ be the ordered sequence $(1, 2, ..., n)$. Define Ω to be the set of all permutations of σ, such that for all $\omega \in \Omega$, $|\omega_i - \sigma_i| \leq \gamma$ for $i \in \{1, ..., n\}$. At each step, we sample $\omega \in \Omega$ with $\gamma = 3$ twenty times, where each ω corresponds to a new candidate ordering of S. We compute the selection metric for each of the twenty orderings, and return S' to be the score permutation that achieves the maximum value for the selection metric. We repeat this procedure for 25 steps, setting S at each step to be the S' output from the previous step. Results for this setting are shown in Figures 2d and 2c.



(a) Optimizing for Overall AUROC

(b) Optimizing for Overall AUPRC

Figure 3. Comparison of the impact of optimizing for overall AUROC and overall AUPRC on the per-group AUROC and AUPRCs of two groups in a synthetic setting, using the *adding random noise* optimization procedure (M1) described in Section 3.2. Note that the prevalence of Y in G_1 and G_2 are 0.05 and 0.01 respectively.

C. Synthetic Analytical Example

To further reinforce our findings, we develop an analytical model, assuming simple uniform conditional score distributions, to derive the equations of the ROC curve and PR curve, and, consequently, the AUROC and AUPRC. This simple model provides interesting insights into the behavior of AUROC and AUPRC under varying conditions. In this setting, our results show that optimizing for overall AUPRC can favor the group that already exhibits better performance, further exacerbating the performance gap, while optimizing for overall AUROC always favors the group that is in the minority.

Define random variables $Y \in \{0, 1\}$, $\hat{Y} \in [0, 1]$, $G \in \{G_0, G_1\}$.

Let $p_{0,-}(\cdot)$ denote $\mathbb{P}(\hat{Y}|G = G_0, Y = 0)$, and similarly for $p_{0,+}(\cdot)$, $p_{1,-}(\cdot)$, and $p_{1,-}(\cdot)$.

Suppose that a model $f : \mathcal{X} \to [0, 1]$ outputs the following conditional predicted score distributions for each group:

- $p_{0,-} = \text{Unif}[0, 1]$
- $p_{1,-} = \text{Unif}[0, 1]$
- $p_{0,+} = \text{Unif}[k_0, 1]$
- $p_{1,+} = \text{Unif}[k_1, 1]$

In addition, define:

- $\mathbb{P}(G = G_1) = \beta$
- $\mathbb{P}(Y = 1|G = G_g) = \alpha_g$
- $\mathbb{P}(Y = 1) = \alpha = \alpha_1\beta + \alpha_0(1 - \beta)$

Without loss of generality, assume $k_1 > k_0$, and so the model achieves a better prediction for G_1 over G_0. This may result from differences in prevalence, i.e. $\alpha_1 > \alpha_0$, or from differences in task difficulties (i.e. positive samples in G_1 are easier to identify).

Then, we have, at each threshold $\tau \in [0, 1]$:

$$
\begin{aligned}
\text{FPR}_0(\tau) &= \text{FPR}_1(\tau) = 1 - \tau \\
\text{TPR}_g(\tau) &= \begin{cases} 1, & \tau \leq k_g \\ 1 - \frac{\tau - k_g}{1 - k_g}, & k_g < \tau \end{cases} = \mathbb{1}_{\tau \leq k_g} + (1 - \frac{\tau - k_g}{1 - k_g})\mathbb{1}_{\tau > k_g} \\
p_+(\hat{y}) &= \frac{\beta}{1 - k_0}\mathbb{1}_{k_0 \leq \hat{y} \leq k_1} + (\frac{\beta}{1 - k_0} + \frac{1 - \beta}{1 - k_1})\mathbb{1}_{k_1 \leq \hat{y} \leq 1} \\
p_-(\hat{y}) &= 1 \\
\text{FPR}(\tau) &= 1 - \tau \\
\text{TPR}(\tau) &= \mathbb{1}_{\tau \leq k_0} + \mathbb{1}_{k_0 \leq \tau \leq k_1}(1 - \frac{\beta(\tau - k_0)}{1 - k_0}) + \mathbb{1}_{k_1 < \tau \leq 1}\left((1 - \tau)(\frac{\beta}{1 - k_0} + \frac{1 - \beta}{1 - k_1})\right)
\end{aligned}
$$

C.1. Optimizing AUROC

Deriving the expression for AUROC:

$$
\begin{aligned}
\text{TPR}(\text{FPR}) &= \mathbb{1}_{\text{FPR} \geq 1 - k_0} + \mathbb{1}_{1 - k_1 \leq \text{FPR} \leq 1 - k_0}(1 - \frac{\beta(1 - \text{FPR} - k_0)}{1 - k_0}) + \mathbb{1}_{\text{FPR} < 1 - k_1}\left((\text{FPR})(\frac{\beta}{1 - k_0} + \frac{1 - \beta}{1 - k_1})\right) \\
\text{AUROC} &= k_0 + \frac{1}{2}(k_1 - k_0)(2 - \frac{\beta(k_1 - k_0)}{1 - k_0}) + \frac{1}{2}(1 - k_1)(1 - \frac{\beta(k_1 - k_0)}{1 - k_0})
\end{aligned}
$$

Suppose that we now want to improve f, and we have two mutually exclusive options to choose from: (1) Improve k_0, e.g. by setting $k_0 := k_0 + \epsilon$. (2) Improve k_1, e.g. by setting $k_1 := k_1 + \epsilon$. To explore the impact of these options, let's compute $\frac{d\text{AUROC}}{dk_g}$.

$$\begin{aligned} \frac{d\text{AUROC}}{dk_0} &= 0.5\beta \\ \frac{d\text{AUROC}}{dk_1} &= 0.5 - 0.5\beta \end{aligned}$$

Surprisingly, this is independent of k_0 and k_1. Therefore, to maximize overall AUROC, we should choose to improve the *minority group*, i.e. improve k_0 when $\beta \geq 0.5$ and improve k_1 otherwise.

C.2. Optimizing AUPRC

Deriving the Expression for AUPRC:

$$\begin{aligned} \text{FPR}(\text{TPR}) &= \mathbb{1}_{0 \leq \text{TPR} \leq 1 - \frac{\beta(k_1 - k_0)}{1 - k_0}} \left(\frac{\text{TPR}(k_0 k_1 - k_0 - k_1 + 1)}{\beta k_0 - \beta k_1 - k_0 + 1} \right) + \mathbb{1}_{1 - \frac{\beta(k_1 - k_0)}{1 - k_0} \leq \text{TPR} \leq 1} \left(\frac{-\beta k_0 + \beta - \text{TPR} k_0 + \text{TPR} + k_0 - 1}{\beta} \right) \\ \text{PPV}(\tau) &= \frac{\text{TPR}(\tau)\alpha}{\text{TPR}(\tau)\alpha + \text{FPR}(\tau)(1 - \alpha)} \\ \text{PPV}(\text{TPR}) &= \frac{\text{TPR}\alpha}{\text{TPR}\alpha + \text{FPR}(\text{TPR})(1 - \alpha)} = \mathbb{1}_{0 \leq \text{TPR} \leq 1 - \frac{\beta(k_1 - k_0)}{1 - k_0}} \frac{\alpha\left(\beta k_0 - \beta k_1 - k_0 + 1\right)}{\alpha\left(\beta k_0 - \beta k_1 - k_0 + 1\right) - (\alpha - 1)\left(k_0 k_1 - k_0 - k_1 + 1\right)} \\ &\quad + \mathbb{1}_{1 - \frac{\beta(k_1 - k_0)}{1 - k_0} \leq \text{TPR} \leq 1} \frac{\alpha\beta\text{TPR}}{\alpha\beta\text{TPR} + (\alpha - 1)\left(\beta k_0 - \beta + k_0\text{TPR} - k_0 - \text{TPR} + 1\right)} \end{aligned}$$

Unfortunately, AUPRC $= \int_0^1 \text{PPV}(\text{TPR})\, d\text{TPR}$ is intractable. Instead, we compute the numeric derivative $\frac{d\text{AUPRC}}{dk_g}$ evaluated at a grid of k_g values satisfying $k_1 > k_0$. We plot a heatmap of $\frac{d\text{AUPRC}}{dk_1} - \frac{d\text{AUPRC}}{dk_0}$, where a positive value indicates that maximizing AUPRC should focus on G_1, and a negative value indicates that maximizing AUPRC should focus on G_0. Note that the model currently achieves a better predicted score distribution for G_1 than G_0, and so a fair optimization procedure should choose to improve k_0.

We present our results in Figure 4. We find, when both groups have the same prevalence and proportion (Figure 4a), that the optimization procedure favors G_1, especially when G_1 is already well-off (i.e. k_1 is large). This trend remains the same when the proportion of G_1 is decreased, or when the prevalence of G_1 changes. The optimization procedure does begin to favor G_0 when it's in the minority, but the preference towards G_1 remains when k_1 is large.

Hence, our results show that optimizing for overall AUPRC may result in actions that discriminate towards particular groups – favoring the better performing group, especially when the performance gap is large, and when it is in the minority. This is in contrast with AUROC optimization, where the minority group is always preferred.

D. Literature Review Methodology

D.1. Paper Acquisition

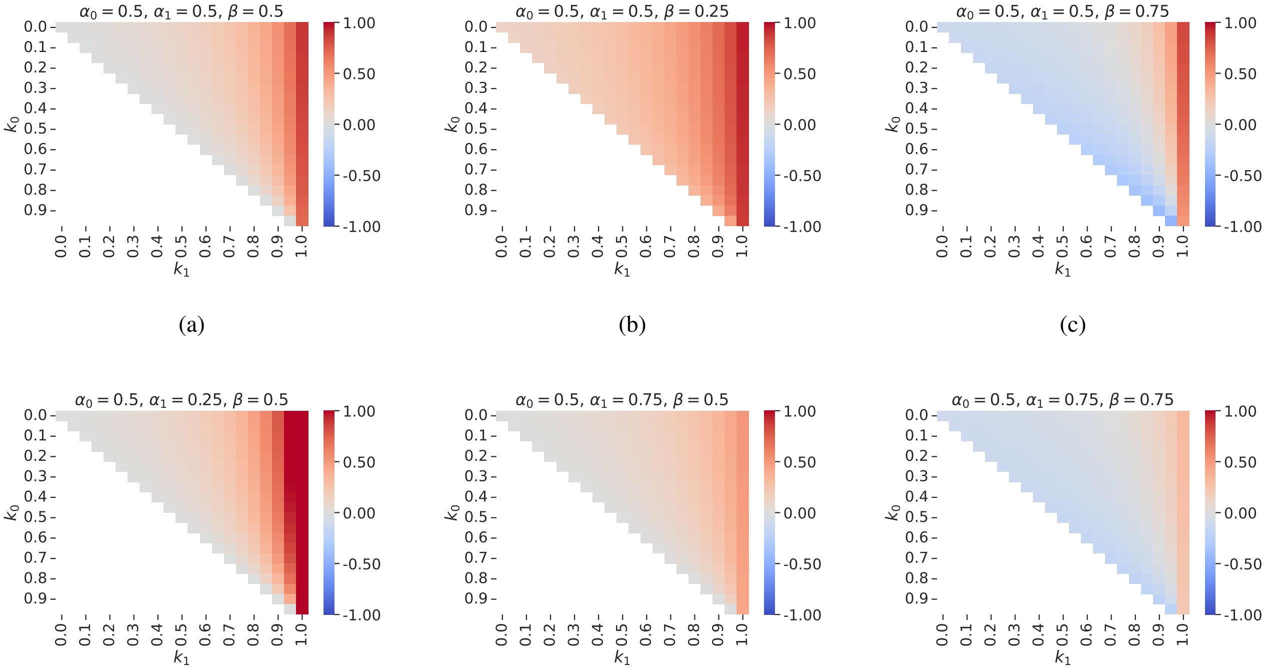
The initial phase of our comprehensive literature search involved the acquisition of a comprehensive dataset from the Arxiv repository. Utilizing the RedPajama dataset available through Hugging Face, we specifically targeted papers from the Arxiv database. The dataset, approximately 93.8 GB in size, encompassed over 1.5 million texts in JSONL format.

D.2. Keyword-Driven Filtering Process

1. **Keyword List Development:** We developed two distinct keyword lists to systematically identify papers relevant to our research on AUROC (Area Under the Receiver Operating Characteristic) and AUPRC (Area Under the Precision-Recall Curve) in our initial screening phase. The keyword lists can be accessed here for AUPRC and here for AUROC.

2. **Automated Script-Based Search:** Python scripts were employed to traverse the Arxiv dataset. These scripts detected occurrences of our predefined keywords, allowing efficient parsing of a vast number of texts.

3. **Dual Mention Selection Criterion:** We focused on papers discussing both AUROC and AUPRC. This criterion



(a)

(b)

(c)

(d)

(e)

(f)

Figure 4. Heatmaps of $\frac{d\text{AUPRC}}{dk_1} - \frac{d\text{AUPRC}}{dk_0}$ evaluated at a grid of k_0 and k_1, with various values of α_0, α_1, and β.

ensured the relevance of the papers to our research question. Through this process, we narrowed the pool from 16,022 texts (containing either set of keywords) to 8,244 texts mentioning both.

D.3. AI-Assisted Screening and Refinement

1. **Preliminary Analysis with GPT-3.5:** We utilized OpenAI's GPT-3.5 model for an initial round of AI-assisted analysis. This model identified and extracted papers making explicit claims regarding the comparative effectiveness of AUPRC over AUROC in scenarios of class imbalance, reducing our dataset to 2,728 papers.

2. **Further Refinement Using GPT-4.0 Turbo:** To refine our dataset further, we employed the GPT-4.0 Turbo model. Approximately 73% of the 2,728 papers were scrutinized using this model, leading to a distilled list of 197 highly relevant papers.

D.4. Manual Review

- **Shared Document for Collaborative Analysis:** We compiled all pertinent papers, along with their respective Arxiv IDs and the claims identified by GPT-4.0 Turbo, into a shared Google document for team review. Claims made in papers were found manually, and the specific quote of the claim they made was highlighted along with whether or not they had a citation for this claim.

D.4.1. Final Papers

After manual review, we identified 128 papers that make or reference some version of the claim that "AUPRC is better than AUROC in cases of class imbalance." (Cook & Ramadas, 2020; Leisman, 2018; Yang et al., 2015; Gaudreault et al., 2021; Albora & Zaccaria, 2022; Lim & van der Schaar, 2018; Liu et al., 2023; Randl et al., 2023; Tusfiqur et al., 2022; Piermarini et al., 2023; Zhang & Bondell, 2018; Weiss & Tonella, 2021; Afanasiev et al., 2021; Li et al., 2022; Torfi et al., 2022; Wu et al., 2020; Miao & Zhu, 2022; Navarro et al., 2022; Cho et al., 2021; Wagner et al., 2023; Isupova et al., 2017; Sarvari et al., 2021; Hiri et al., 2022; Herbach, 2021; Si & Roberts, 2021; Narayanan et al., 2022; Li et al., 2020; Lee et al., 2013; Rayhan et al., 2017; Kyono et al., 2018; Adler, 2021; Seo et al., 2021; Hong et al., 2019; Hagedoorn & Spanakis, 2017; Yang et al., 2022a; Babaei et al., 2021; Garcin & Stéphan, 2021; Mehboudi et al., 2022; Yang et al., 2022b; Shen & Kursun, 2021; Muthukrishna et al., 2019; Deng et al., 2023; Yang, 2021; Harer et al., 2018; Meister et al., 2022; Skarding et al., 2021; Alvarez et al., 2022; Zou et al., 2022; Mangolin et al., 2022; Mosteiro et al., 2021; Hashemi et al., 2018; Lee et al., 2020; Zavrtanik et al., 2021; Showalter & Wu, 2019; Cranmer & Desmarais, 2016; Bryan & Moriano, 2023; Zhang et al., 2017; Domingues et al., 2020; Markdahl et al., 2017; Fu et al., 2021; Pang et al., 2023; Rezvani et al., 2021; Ozyegen et al., 2022; Prapas et al., 2023; Rayhan et al., 2020; Thambawita et al., 2020; Shukla & Marlin, 2019; Blevins et al., 2021; Vijayan et al., 2017; Budka et al., 2021; Hsu et al., 2020; Smith et al., 2023; Choi et al., 2018; Ju et al., 2018; Pashchenko et al., 2018; Chu et al., 2018; Silva et al., 2022; Bach Nguyen et al., 2022; Deshwar et al., 2015; Brophy & Lowd, 2020; Mayaki & Riveill, 2022; Mongia et al., 2021; Tiulpin et al., 2019; Romero et al., 2022; Rubin et al., 2012; Schwarz et al., 2021; Lyu et al., 2021; Lopez-Martinez et al., 2022; Ahmed & Courville, 2020; Gong et al., 2021; Zhang et al., 2021; Shukla & Marlin, 2018; Lund et al., 2019; Ma et al., 2022; Ruff et al., 2021; Lei Ba et al., 2015; Chakraborty et al., 2023; Rajabi & He, 2021; Newby et al., 2022; Axelrod & Gomez-Bombarelli, 2023; Kim et al., 2022; Ando & Huang, 2017; Stolman et al., 2022; Mosquera et al., 2022; Kulkarni et al., 2021; Won et al., 2019; Stephenson et al., 2022; Srivastava et al., 2019; Moor et al., 2019; Danesh Pazho et al., 2023; Kiran et al., 2018; Steinbuss & Böhm, 2021; Ma et al., 2020; Karadzhov et al., 2022; Ding et al., 2018; Mousavian et al., 2016; Rayhan et al., 2017; Vens et al., 2008; Rohani & Eslahchi, 2019; López et al., 2013; Sahiner et al., 2017; Rao et al., 2022; Hibshman & Weninger, 2023; Ntroumpogiannis et al., 2023; Weiss & Tonella, 2023; Hall et al., 2023; Goyal & Khiari, 2020; Boyd et al., 2013).

All papers identified, manual screening results, and extracted quotes can be found here: `https://docs.google.com/spreadsheets/d/1NjDpwoj_8EkIwtGZzwM6w2nbst-LlGJPAqUcVINmPEk/edit?usp=sharing`.

D.5. Code Availability

All code pertaining to the Arxiv search can be found in the following GitHub repository:

`https://github.com/Lassehhansen/Arxiv_search/tree/main`

Claim	References	Commentary
Precision-recall curves or other associated metrics *may* more appropriately reflect deployment objectives than the receiver operating characteristic.	(Cook & Ramadas, 2020; Leisman, 2018; Yang et al., 2015; Muthukrishna et al., 2019; Deng et al., 2023; Harer et al., 2018; Ahmed & Courville, 2020)	While this claim is true, the informativeness of the PR curve for target deployment metrics is not sufficient to conclude that the AUPRC is superior to the AUROC in all cases of class imbalance. Despite this, it is often taken to assert this more general claim without caveat.
AUPRC does not depend on the number of true-negatives, so will be less optimistic than the AUROC	(Leisman, 2018; Kyono et al., 2018; Adler, 2021; Meister et al., 2022; Mosteiro et al., 2021; Showalter & Wu, 2019; Cranmer & Desmarais, 2016; Domingues et al., 2020; Rezvani et al., 2021; Hsu et al., 2020; Ju et al., 2018; Pashchenko et al., 2018; Romero et al., 2022; Vens et al., 2008)	As shown in Theorem 1, AUROC and AUPRC can both be naturally expressed as a function of the expectation of the model's false positive rate. More generally, lack of dependence on one quadrant among the mutually dependent four quadrants of a confusion matrix is not an informative property for the AUROC and AUPRC metrics.
AUPRC will often be significantly lower, farther from optimality, and/or will grow more non-linearly as model performance improves than AUROC for low-prevalence tasks	(Leisman, 2018; Yang et al., 2015; Mehboudi et al., 2022; Cranmer & Desmarais, 2016)	Metric utility for model comparison depends on how appropriately it prioritizes model improvements, and is therefore less about the raw magnitude of the metric and more about the situations in which the order of a set of models will differ under one metric vs. another. One could easily make AUROC yield smaller values or grow more quickly near optimality by simply exponentiating it, but this would not yield a better metric.
AUPRC depends on prevalence, which is a desirable property	(Navarro et al., 2022)	This statement is too vague to be formally evaluated; whether or not this dependence on prevalence is desirable depends on the context. For model comparison in general, we argue it is not desirable in this form as it induces the biases inhere in AUPRC previously discussed.
AUPRC better captures differentiating a positive sample with high score from a "hard" negative sample ("hard" meaning one also with high score)	(Kiran et al., 2018)	While this claim is true by Theorem 2, it is not clear why this would be desired in general; this implicitly favors comparing "hard" negatives against "easy" positives as opposed to "easy" negatives against "hard" positives.
AUROC is otherwise "optimistic" in low-prevalence settings	(Cook & Ramadas, 2020; Afanasiev et al., 2021; Wu et al., 2020; Miao & Zhu, 2022; Cho et al., 2021; Hagedoorn & Spanakis, 2017; Yang et al., 2022a; Mangolin et al., 2022; Silva et al., 2022; Lyu et al.,	This claim is underspecified, and un-true. AUROC always means the same thing, probabilistically, and that meaning independent from class imbalance.