
A Closer Look at AUROC and AUPRC under Class Imbalance

Matthew B. McDermott
Harvard Medical School
matthew_mcdermott@hms.harvard.edu

Haoran Zhang
Massachusetts Institute of Technology
haoranz@mit.edu

Lasse Hyldig Hansen
Aarhus University
201908623@post.au.dk

Giovanni Angelotti
IRCCS Humanitas Research Hospital
giovanni.angelotti@humanitas.it

Jack Gallifant
Massachusetts Institute of Technology
jgally@mit.edu

Abstract

In machine learning (ML), a widespread claim 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 tasks with class imbalance. This paper refutes this notion on two fronts. First, we theoretically characterize the behavior of AUROC and AUPRC in the presence of model mistakes, establishing clearly that AUPRC is not generally superior in cases of class imbalance. We further show that AUPRC can be a *harmful metric* as it can unduly favor model improvements in subpopulations with more frequent positive labels, heightening algorithmic disparities. Next, we empirically support our theory using experiments on both semi-synthetic and real-world fairness datasets. Prompted by these insights, we reviewed over 1.5 million scientific papers to understand the origin of this invalid claim—finding is often made without citation, misattributed to papers that do not argue this point, and aggressively overgeneralized from source arguments. Our findings represent a dual contribution: a significant technical advancement in understanding the relationship between AUROC and AUPRC and a stark warning about unchecked assumptions in the ML community.

1 Introduction

Machine learning (ML), especially in critical domains like healthcare, necessitates careful selection and application of evaluation metrics to guide appropriate model choices and understand performance nuances [150]. Model evaluation can happen in one of two settings: (1) a *methodological/model comparison* setting, which occurs outside of a specific deployment setting and target model usage workflows, optimal decision thresholds, or specific false-positive (FP) and false-negative (FN) costs are typically not known, or (2) an *application/deployment* setting, where reasonably specific estimates of model usage workflows and FP/FN costs can be made. In both settings, appropriate metric choice is critical—and inappropriate selection can hinder innovation when used for model comparison and lead to significant real-world costs (e.g., misdiagnosis in a medical setting) in deployment settings.

This study focuses on two widely used metrics for binary classification tasks across both evaluation contexts: Area Under the Precision-Recall Curve (AUPRC) and Area Under the Receiver Operating Characteristic (AUROC). Central to this paper is the following key claim:

Claim 1. Let f be a model which outputs continuous probabilistic predictions trained to solve a binary classification task for which the prevalence of negative labels is significantly higher than the prevalence of positive labels. For this problem, the AUPRC will yield a “better” or “more accurate” or “fairer” evaluation of f than the AUROC.

Claim 1 is made widely in both the scientific literature [399, 71, 159, 124], in ML educational content like notable textbooks [119, 141], and in popular press sources [80, 254]. It is so widespread that even basic search results for queries relating to AUROC and AUPRC¹ and large language model assistants like ChatGPT or Github Co-pilot will profess its veracity.² Throughout these sources, it has been justified on numerous, often imprecise grounds (see Section 5), but despite this extensive attention, *we show in this work that this claim is, in fact, wrong, and may be dangerous from a model fairness perspective*; further, many of its justifications are *invalid* or *misapplied* in common ML settings. More specifically, we show the following:

1) AUROC and AUPRC only differ with respect to model-dependent parameters in that AUROC weighs all false positives equally, whereas AUPRC weighs false positives at a threshold τ with the inverse of the model’s likelihood of outputting any scores greater than τ (Theorem 1). This result shows that we can reason about the suitability of optimizing or selecting AUROC vs. AUPRC based on whether we care more about reducing false positives above low thresholds or high thresholds. In particular,

2) AUROC favors model improvements uniformly over all positive samples, whereas AUPRC favors improvements for samples assigned higher scores over those assigned lower scores (Theorem 2). This indicates that *the key factor differentiating the utility of AUROC or AUPRC as an evaluation metric is not class imbalance at all, but it is rather based on the target use case of the model in question*. See Figure 1 for a visual explanation. It also reveals that *AUPRC can amplify algorithmic biases*. In particular,

3) AUPRC can unduly prioritize improvements to higher-prevalence subpopulations at the expense of lower-prevalence subpopulations, raising serious fairness concerns in any multi-population use cases (Theorem 3).

In this work, we establish these three claims theoretically via synthetic experiments and real-world validation on popular public fairness datasets. In addition, we demonstrate through an extensive, large-language model aided, literature review of over 1.5 million scientific papers that Claim 1 has been used to motivate numerous improper uses of AUPRC relative to AUROC across high-stakes domains like healthcare and in several established venues, including AAI, NeurIPS, ICML, ICLR, Cancer Cell, Nature Journals, PNAS, and more. Through this paper, we hope to shed light on the nuances of appropriate evaluation and provide key guidance to limit future misuse of evaluation metrics in the scientific and machine learning communities.

2 Theoretical Analyses

Please note that all notation used is defined in Appendix Section C.

2.1 Relationship between AUROC and AUPRC

In this section, we introduce Theorem 1, which is as follows:

Theorem 1. Let $\mathcal{X}, \mathcal{Y} = 0, 1$ represent a paired feature and binary classification label space from which i.i.d. samples $(x, y) \in \mathcal{X} \times \mathcal{Y}$ are drawn via the joint distribution over the random variables x, y . Let $f : \mathcal{X} \rightarrow (0, 1)$ be a binary classification model outputting continuous probability scores over this space. Then,

$$\begin{aligned} \text{AUROC}(f) &= 1 - \mathbb{E}_{t \sim f(x)|y=1} [\text{FPR}(f, t)] \\ \text{AUPRC}(f) &= 1 - P_y(y = 0) \mathbb{E}_{t \sim f(x)|y=1} \left[\frac{\text{FPR}(f, t)}{P_x(f(x) > t)} \right] \end{aligned}$$

¹See <https://archive.is/qXPKu>, which shows results dominated by those making Claim 1

²See <https://chat.openai.com/share/f8f7fddb-1553-41a5-976d-789e2f3a90d6>

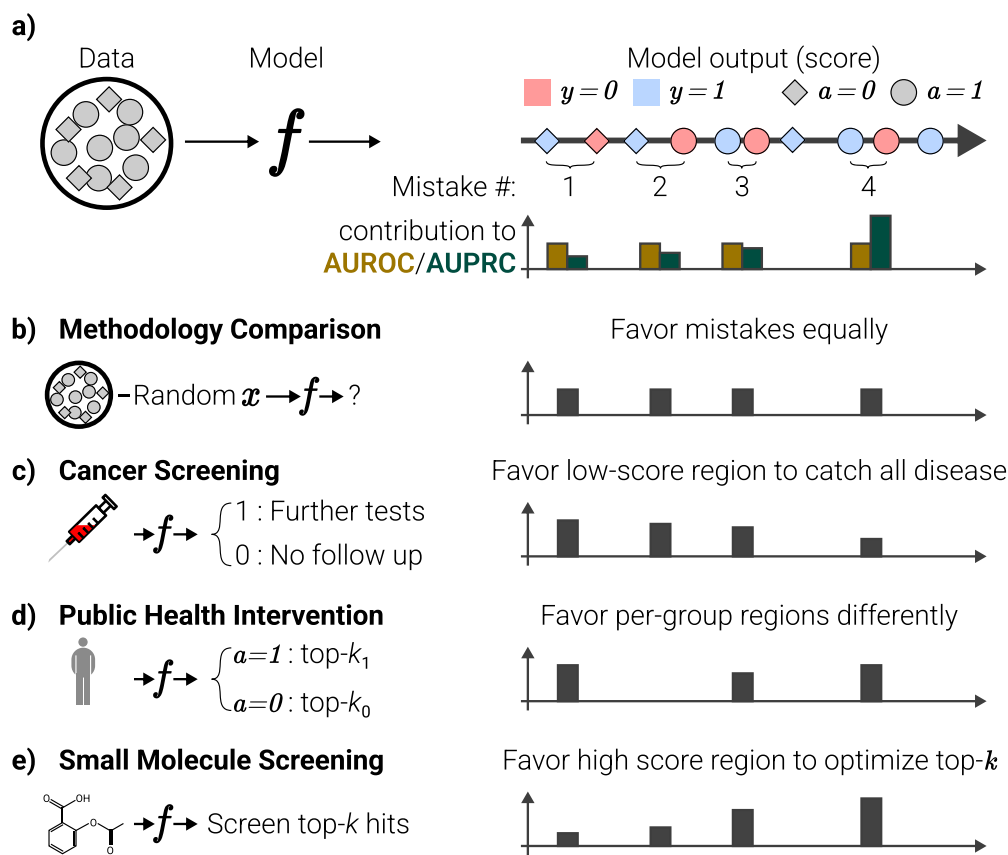


Figure 1: **a)** Consider a model f yielding continuous output scores for a binary classification task applied to a dataset consisting of two distinct subpopulations, $\mathcal{A} \in \{0, 1\}$. If we order samples in ascending order of output score, each misordered pair of samples (e.g., mistake 1-4) represents an opportunity for model improvement. Theorem 2 shows that a model's AUROC will improve by the same amount no matter which mistake you fix, while the model's AUPRC will improve by an amount correlated with the score of the sample. **b)** When comparing models absent a specific deployment scenario, we have no reason to value improving one mistake over another, and model evaluation metrics should therefore improve equally regardless of which mistake is corrected. **c)** When false negatives have a high cost relative to false positives, evaluation metrics should favor mistakes that have *lower scores*, regardless of any class imbalance. **d)** When limited resources will be distributed among a population according to model score, *in a manner that requires certain subpopulations to all be offered commensurate possible benefit from the intervention for ethical reasons*, evaluation metrics should prioritize the importance of within-group, high-score mistakes such that the highest risk members of all subgroups receive interventions. **e)** When false positives are expensive relative to false negatives and there are no fairness concerns, evaluation metrics should favor model improvements in decreasing order with score.

We provide the proof in Appendix Section D. The two key intuitions are that integrating over the TPR is equivalent to taking the expectation over the induced distribution of positive sample scores, and that via Bayes rule, $\text{Prec}(f, \tau) = 1 - P_y(y = 0) \frac{\text{FPR}(f, \tau)}{P_x(f(x) > \tau)}$.

Despite its simplicity, Theorem 1 has far-reaching implications. Namely, it reveals that the only difference between AUROC and AUPRC with respect to model dependent parameters (i.e., omitting the dependence of AUPRC on the fixed prevalence of the dataset, which is not model varying) is that optimizing AUROC equates to minimizing the expected false positive rate over all positive samples in an unweighted manner (equivalently, in expectation over the distribution of positive sample scores) whereas optimizing AUPRC equates to minimizing the expected false positive rate over all positive samples weighted by the inverse of the model’s “firing rate” ($P_x(f(x) > \tau)$) at the given positive sample score. This preference can be crystallized when we examine how AUROC vs. AUPRC would prioritize correcting indivisible units of model improvements, termed “mistakes” which we will discuss next.

2.2 AUPRC prioritizes high-score mistakes, AUROC treats all mistakes equally

Understanding how a given evaluation metric prioritizes correcting various kinds of model mistakes or errors offers significant insight into when that metric should be used for optimization or model selection. To examine this topic for AUROC and AUPRC, consider the following definition of an “incorrectly ranked adjacent pair”, which we will colloquially refer to as a “model mistake”:

Definition 2.1. Let $f, \mathcal{X}, \mathcal{Y}, x, y$ be defined as in Theorem 1. Further, let us suppose we have sampled a static dataset from x, y for evaluation which will be denoted $\mathbf{X}, \mathbf{y} = \{(x_1, y_1), \dots, (x_N, y_N)\}$, for $x_i \in \mathcal{X}, y_i \in \{0, 1\}$, and $N \in \mathbb{N}$. We assume for convenience that f is an injective map and all x_i are distinct (i.e., $\forall(i, j) : x_i \neq x_j$ which, by injectivity of f , implies that $f(x_i) \neq f(x_j)$).

We say that (x_i, x_j) are an *incorrectly ranked adjacent pair* and thus that the model makes a “mistake” at samples (x_i, x_j) if:

1. $y_i = 1$ and $y_j = 0$
2. $f(x_i) < f(x_j)$
3. $\nexists x_k$ such that $f(x_i) < f(x_k) < f(x_j)$.

Essentially, Definition 2.1 states that a *mistake* occurs when a model assigns adjacent probability scores to a pair of samples with discordant labels, as shown in Figure 1. With this in mind, we can then introduce Theorem 2 which states that AUROC improves by a constant amount regardless of which mistake is corrected for a given model and dataset whereas AUPRC improves more when the mistake corrected occurs at a higher score than when it occurs at a lower score:

Theorem 2. Define $f, \mathcal{X}, \mathbf{X}, \mathbf{y}$ and N as in Definition 2.1. Further, suppose without loss of generality that the dataset \mathbf{X} is ordered such that $f(x_i) < f(x_{i+1})$ for all i . Then, let us define $M = \{i | (x_i, x_{i+1}) \text{ is an incorrectly ranked adjacent pair for model } f\}$. Define f'_i to be a model that is identical to f except that the probabilities assigned to x_i and x_{i+1} are swapped:

$$f'_i : \begin{cases} f(x) & \text{if } x \notin \{x_i, x_{i+1}\} \\ f(x_{i+1}) & \text{if } x = x_i \\ f(x_i) & \text{if } x = x_{i+1}. \end{cases}$$

Then, $\text{AUROC}(f'_i) = \text{AUROC}(f'_j)$ for all $i, j \in M$, and $\text{AUPRC}(f'_i) < \text{AUPRC}(f'_j)$ for all $i, j \in M$ such that $i < j$.

The proof for Theorem 2 can be found in Appendix E. This proof simply stems from the fact that correcting a single mistake (x_i, x_j) (as defined in Definition 2.1) always changes the false positive rate by the same amount, and only changes it at the threshold $f(x_i)$. This, combined with the formalization of AUROC and AUPRC in Theorem 1, establishes the proof. Note that this Theorem can be trivially extended to include a case where ties are possible simply by noting that “swapping” two samples x_i and x_j in the manner of the theorem results in no change to either AUROC or AUPRC, and similarly by the same reasoning separating any tie in the appropriate direction will improve AUROC uniformly over samples and will improve AUPRC in a manner monotonic with model score.

2.3 AUPRC is explicitly discriminatory in favor of high-scoring subpopulations

The reliance on a model’s firing rate revealed in Theorem 1 and the optimization behavior in Theorem 2 reveals significant issues with the fairness of AUPRC. In particular, in this section we introduce Theorem 3:

Theorem 3. Let $f, \mathcal{X}, \mathbf{X}, \mathbf{y}, N, M$, and f'_j all be defined as in Theorem 2. Further, suppose that in this setting the domain \mathcal{X} now contains an attribute defining two subgroups, $\mathcal{A} = \{0, 1\}$, such that for any sample (x_i, y_i) , a_i denotes the subgroup to which that sample belongs. Let f be perfectly calibrated for samples in subgroup $a = 0$, such that $P_{y|a,x}(y = 1|a = 0, f(x) = t) = t$. Then,

$$P_{y|a}(y=1|a=0) \rightarrow 0 \lim P \left(a_i = a_{i+1} = 1 \mid i = \arg \max_{j \in M} (\text{AUPRC}(f'_j)) \right) = 1.$$

Essentially, Theorem 3 (proof provided in Appendix F) shows the following. Suppose we are training a model f over a dataset with two subpopulations: Population $a = 0$ and $a = 1$. If the model f is calibrated and the rate at which $y = 1$ for population $a = 0$ is sufficiently low relative to the rate at which $y = 1$ for population $a = 1$, then the mistake that, were it fixed, would maximally improve the AUPRC of f will be a mistake purely in population $a = 1$. This demonstrates that AUPRC provably favors higher prevalence subpopulations (those with a higher base rate at which $y = 1$) under sufficiently severe prevalence imbalance between subpopulations.

Note that this property is, generally speaking, not desirable. *In particular, this property establishes that in settings where model fairness among a set of subpopulations in the data is important, AUPRC should not be used as an evaluation metric due to the risk that it will introduce biases in favor of the highest prevalence subpopulations.* We validate this result empirically over both synthetic and real-world data in Section 3, demonstrating that the import of Theorem 3 is not merely limited to an analytical curiosity but can have real-world impact on algorithmic disparities in practice. Furthermore, note that this theorem does not indicate that AUPRC will be superior to AUROC for *differentiating* a low prevalence (or low risk) subpopulation relative to a high-risk subpopulation, a property that is sometimes attributed to AUPRC in the literature. Rather, Theorem 3 shows that maximizing AUPRC will be more likely to optimize solely within the high-risk subgroup, rather than optimizing to differentiate across subgroups, as low-risk subgroup samples will only occur in lower-score regions under severe class imbalance.

3 Experimental Validation

In this section, we establish via synthetic and real-world experiments that Theorem 3 is not merely an analytical effect but has real world consequences on the implications of optimizing or performing model selection via AUPRC.

3.1 Synthetic optimization experiments demonstrate AUPRC-induced disparities

In this section, we use a carefully constructed synthetic optimization procedure to demonstrate that, when all other factors are equal, optimizing by or performing model selection on the basis of AUPRC vs. AUROC risks exacerbating algorithmic disparities in the manner predicted by Theorem 3. For analyses under more realistic conditions with more standard models, see our real-world experiments in Section 3.2.

Experimental Setup. Let $y \in \{0, 1\}$ be the binary label, $s \in [0, 1]$ be the predicted score, and $a \in \{1, 2\}$ be the subpopulation. We fix $P_{y|a}(y = 1|a = 1) = 0.05$ and $P_{y|a}(y = 1|a = 2) = 0.01$. We sample a dataset for each group $\mathcal{D}_a = \{(s_1, y_1), \dots, (s_{n_a}, y_{n_a})\}$, such that $\text{AUROC}(\mathcal{D}_1) \approx \text{AUROC}(\mathcal{D}_2) \approx \text{AUROC}(\mathcal{D}_1 \cup \mathcal{D}_2) = 0.85$ (See Appendix G.1; A target AUROC of 0.65 was also profiled in Appendix Figure 5).

Our main experimental challenge is to determine how to simulate “optimizing” or “selecting” a model by AUROC or AUPRC. Simulating optimizing by these metrics allows us to explicitly assess how the use of either AUPRC or AUROC as an evaluation metric in model selection processes such as hyperparameter tuning or architecture search, can translate into model-induced inequities in dangerous ways. 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 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.

Next, we profile an optimization procedure that randomly permutes all the (sorted) model scores up to 3 positions (See Appendix G.3 for details). This has the effect of randomly adjusting all model scores, and can worsen model performance under some random permutations, but offers precisely the same “optimization 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.

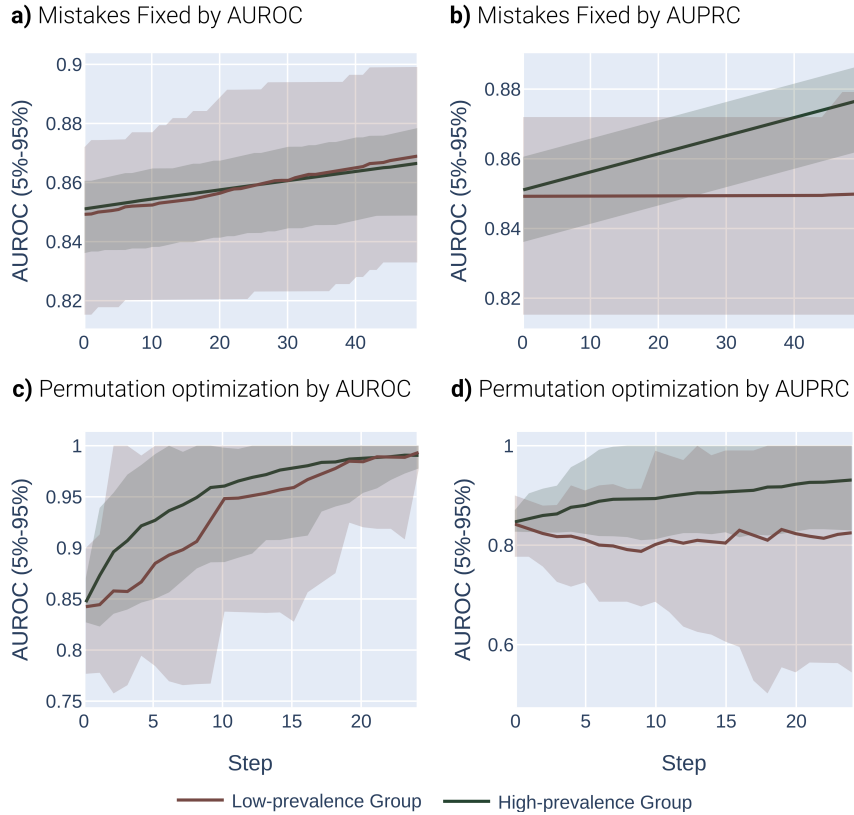


Figure 2: Synthetic experiment per-group AUROC, showing a confidence interval spanning the 5th to 95th percentile of results observed across all seeds, after successively either fixing individual mistakes, as defined in Definition 2.1, (a) and b)) or successively choosing the optimal score permutation (c) and d)) in order to optimize either AUROC (a) and c)) or AUPRC (b) and d)). It is clear across both forms of optimization that AUPRC definitively favors the higher prevalence subpopulation, whereas AUROC treats subgroups approximately equally. Similar patterns were observed when comparing per-group AUPRCs over the same experimental procedures, as shown in Appendix Figure 4.

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 a third random perturbation method, in Appendix G.3.

Results. Our results demonstrate the impact of the optimization metric on subpopulation disparity. In particular, in Figure 2, we observe a notable disparity introduced when optimizing under

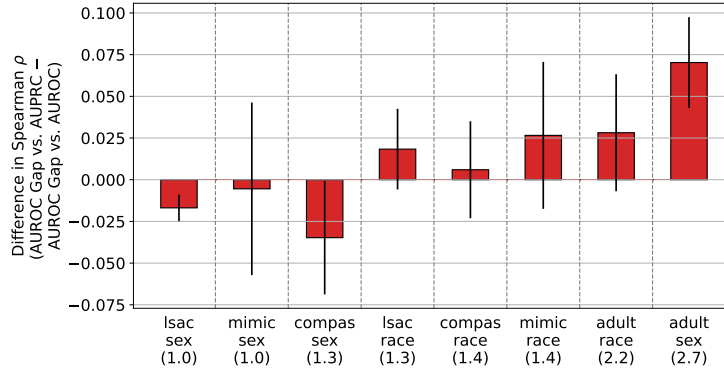


Figure 3: Difference in the Spearman’s ρ between the test-set signed AUROC gap versus the validation set overall AUPRC, and the AUROC gap versus the overall AUROC. Numbers in parentheses are the prevalence ratios between the two groups for the particular attribute, and datasets are sorted by this quantity. Error bars are 95% confidence intervals from 20 different random data splits.

the AUPRC metric regardless of the optimization procedure. 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 the low prevalence subgroup. In comparison, when optimizing for overall AUROC, the AUROC of both groups increase together. Note that we show the effect of this optimization on the AUPRC metric, which shows very similar trends, in Appendix Figure 4. These results demonstrate explicitly that not only does optimizing for AUPRC differ greatly than for AUROC, as has been noted historically by researchers developing explicit AUPRC optimization schemes [409], but it in fact does in an explicitly discriminatory way in very realistic scenarios.

3.2 Real-world experimental validation

To demonstrate the generalizability of our finding to the real world, we evaluate fairness gaps induced by AUROC and AUPRC selection on four common datasets in the fairness literature [441, 99, 205].

Datasets. We use the following four tabular binary classification datasets: `adult` [17], `compas` [14], `lsac` [413], and `mimic` [178]. In each dataset, we consider both sex and race as sensitive attributes. To mimic the setting of our theorems, we balance each dataset by the sensitive attribute during both training and test, by randomly subsampling the majority group. Further details about each dataset, as well as preprocessing steps, can be found in Appendix H.

Experimental setup. We train XGBoost models [65] on each dataset. For each task, we iterate over a grid of per-group weights in order to create a diverse set of models that favor different groups. For each setting of task and per-group weight, we conduct a random hyperparameter search [37] with 50 runs. Though more complex hyperparameter search methods such as BOHB [100] or TPE [36] might lead to better performance, random searches are far more popular and practical, and have been used in popular benchmarking libraries [131, 337].

We evaluate the validation set overall AUROC and AUPRC. We also evaluate the test set AUROC gap and AUPRC gap between groups, where gaps are defined as the value of the metric for the higher prevalence group minus the value for the lower prevalence group. Based on our theorems, our hypothesis is that overall AUPRC should be more positively correlated with the signed AUROC gap than overall AUROC, indicating that it better favors the higher prevalence group, especially when the prevalence ratio between groups is high. To test this hypothesis, we evaluate the Spearman correlation coefficient between these quantities. We repeat this experiment 20 times, with different random data splits, to obtain a 95% confidence interval.

Results. In Figure 3, we plot the difference in the Spearman correlation coefficient of the AUROC gap versus the overall AUPRC, and AUROC gap versus overall AUROC. We observe mixed results in datasets with low prevalence ratio. In dataset with higher prevalence ratio, we find that overall AUPRC

is more positively correlated with the AUROC gap than overall AUROC, indicating that AUPRC more aggressively favors the higher prevalence group. We emphasize that the prevalence ratios observed in these real-world datasets is much lower than the ratio of 5 used in our synthetic experiments, which may account for the mild effect observed. To see raw results from these experiments, see Appendix Figure 7. Similar results for neural network classifiers can be found in Appendix Section H.3.

Next, in Appendix Figure 8, we plot the difference in the Spearman’s ρ from Figure 3, versus the prevalence gap. We find that there is a statistically significant correlation between the two (Spearman’s $\rho = 0.905$, $p = 0.002$). Thus, while our power to detect a prevalence mediated AUPRC bias amplification effect is limited due to the limited prevalence disparities in these datasets, we nonetheless observe a strong positive correlation between the extent of the prevalence mismatch between the low and high prevalence group and the amount that AUPRC favors the high prevalence group over AUROC. *In other words, our results show that across these fairness datasets and attributes, as the prevalence disparity grows more extreme, we observe a statistically significant corresponding increase in the extent to which AUPRC introduces algorithmic bias, exactly in accordance with what Theorem 3 suggests.*

4 When Should One Use AUPRC vs. AUROC?

In Sections 2 and 3, we have shown that AUPRC is not universally superior in cases of class imbalance (and that instead, it merely preferentially optimizes high-score regions over low-score regions) and that it also poses serious risks to the model fairness in settings where subgroup prevalences differ. In light of this, how should we revise Claim 1 to reflect when we *actually* should use AUPRC instead of AUROC or vice versa? Below, we explore this question and provide practical guidance on metric selection for binary classification, building on our theoretical results and highlight specific contexts in which one may be favorable (Figure 1). Note that while we provide guidance below on situations in which AUROC vs. AUPRC is more or less favorable, this is not to suggest that authors should not report both metrics, or even larger sets of metrics or more nuanced analyses such as ROC or PR curves; rather this section is intended to offer guidance on what metrics should be seen as more or less appropriate for use in things like model selection, hyperparameter tuning, or being highlighted as the ‘critical’ metric in a given problem scenario.

For context-independent model evaluation, use AUROC: For model evaluations conducted outside of specific deployment contexts, where the differential costs of errors are undefined, the necessity for a metric that impartially values improvements across the entire model output space becomes paramount. As shown in Figure 1a. As it is not known in advance where samples of interest will live in the output space, nor are particular cost ratios known, there should be no preference for mistake correction. Therefore, in this setting AUROC is favorable as it uniformly accounts for every correction, offering a fair assessment irrespective of decision thresholds.

For deployment scenarios with elevated false negative costs, use AUROC: In applications where the consequences of false negatives are especially high, such as in the early screening for critical illnesses like cancer (Figure 1c), a primary goal of the model will be to achieve the fewest missed cancer cases; equating to prioritizing model recall. In such a scenario, the most important mistakes to correct occur at lower score thresholds, as high-score mistakes will not change which positive samples are missed in deployment settings (as chosen thresholds are likely to be low). This behavior is the *inverse* of what AUPRC prioritizes, demonstrating that in such situations, AUROC should be preferred over AUPRC.

For ethical resource distribution among diverse populations, use AUROC: When faced with the challenge of ethically allocating scarce resources across a broad population, necessitating equitable benefit distribution among subgroups (Figure 1d), one must avoid prioritizing model improvements that selectively favor one subpopulation. As AUPRC will target high-score regions selectively, it risks unduly favoring high-prevalence subpopulations, as shown in Theorem 3 and Figures 2 and 3. Even though in this resource distribution problem, high-score regions are selectively important compared to low-score regions, the fact that in this problem, we must prioritize across all subpopulations equally means that AUPRC’s global preference is untenable as it could induce bias.

For reducing false positives in high-cost, single-group intervention prioritization or information retrieval settings, use AUPRC: In scenarios where the cost associated with false positives significantly outweighs that of false negatives, absent of equity concerns—such as in selecting candidate molecules from a fragment library for drug development trials, where only the most promising molecules will proceed to costly experimental validation (Figure 1e)—the metric of choice should facilitate a reduction in high-score false positives. This necessitates a focus on correcting high-score errors, for which AUROC might not be ideal due to its uniform treatment of errors across the score spectrum, potentially obscuring improvements in critical high-stake decisions.

5 Literature Review: Examining how Claim 1 Became so Widespread

Claim 1 states that “AUPRC is better than AUROC in cases of class imbalance” and is widespread in the literature. Via both a manual literature search and an automated search of over 1.5M arXiv papers (see Appendix I for methodology), we observed 424 publications making this claim.³ This widespread dissemination of Claim 1 *has a significant impact on the validity of scientific discourse*. We observed examples of high-profile papers operating in medically critical settings where high-recall is a key priority evaluating ML systems via AUPRC due to their task’s underlying class imbalance [399]; papers focusing on fairness critical applications relying on AUPRC due to this claim, even while our results demonstrate AUPRC has major *problems* in the fairness regime [366, 306], and numerous other papers perpetuating this source of scientific misinformation.

Among the 424 papers we discovered referencing this claim, 167 did so with no associated citation. These papers were published in a wide range of venues, including high profile venues such as NeurIPS, ICML, and ICLR. 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. Further, Among the 257 that reference this claim and cite a source for this assertion, 135 *do not cite any papers that actually make this claim in the first place*. Most often, papers erroneously attribute this claim to [83], which was cited as a source for this claim 144 times. While [83] makes many interesting, meaningful claims about the ROC and PR curves, and *does argue that the precision-recall curve may be more informative than the ROC in cases of class imbalance* it never asserts that the *area under* the PR curve should be preferred over the *area under* the ROC in cases of class imbalance. This distinction is important, because while curves can be used to simultaneously communicate many different performance statistics to their viewers across different FPR/TPR or Precision/Recall criteria, and therefore should be assessed primarily as communication tools to deployment experts, *areas under these curves* are single-point summarizations which are primarily used for deployment-agnostic model comparison, which is a very different use case.

Even when appropriate papers are cited, the valid settings in which AUPRC should be preferred (see Section 4 for examples) are often over-shadowed by significant over-generalizations to preferring AUPRC in *all* settings featuring class imbalance. For example, claims such as that “precision-recall curves are more informative of deployment metrics” are often used to justify why AUPRC should be used in all cases of class imbalance, rather than just in cases where the relevant deployment metrics are most directly associated with the PR curve. Another class of arguments made in favor of Claim 1 is rooted in claims that AUROC is poor in cases of class imbalance because its scores are misleadingly high. While this argument can reflect a meaningful limitation of the communication value of the ROC or the AUROC, comments about singleton metric results (rather than model comparison through metric values) are inherently orthogonal to the goal of model evaluation. *In other words, what matters for model evaluation is not how high a given metric is, but rather the extent to which the metric meaningfully captures the right improvements in the model in the right ways*. The widespread nature of Claim 1 has also led researchers astray when exploring new optimization procedures for AUPRC, by advocating for the importance of AUPRC when processing skewed data, even in domains such as medical diagnoses that often have high false negative costs relative to false positive costs [409]. For a more extensive breakdown of the arguments we observed in the literature and the sources making them, see Appendix Tables 2 and 3.

³Note that throughout this section, full citations for all of the larger lists of papers we reference will be relegated to Appendix Section I for concision.

6 Limitations and Future Works

There are still a number of areas for further improvement and future work. Firstly, our theoretical findings can be refined and generalized to, e.g., take into account the difficulty of the target task (which may differ between subgroups), not require models to be calibrated (in the case of Theorem 3), or specifically take into account more than 2 subpopulations for more nuanced comparisons beyond what can be inferred through pairwise comparisons between subpopulations, where our results would naturally apply. Further, extending our real-world experiments to more fairness datasets and identifying more nuanced ways to probe the impact of metric choice on disparity measures would significantly strengthen this work. These analyses can also be extended to consider other metrics, such as the area under the precision-recall-gain curve [104], the area under the net benefit curve [384, 307], and single-threshold, deployment centric metrics as well. In addition, one of the largest limitations of Theorem 3 is its restrictive assumptions, in particular the requirement of perfect calibration. A ripe area of future work is thus to investigate how we can soften our analyses for models with imperfect calibration or to determine whether or not our results imply anything about the viability or safety of post-hoc calibration of models optimized either through AUPRC or AUROC.

7 Conclusion

This study interrogates the pervasive assumption within the ML community that AUPRC is a better evaluation metric than AUROC in class-imbalanced settings. Our empirical analyses and literature review reveal several concrete findings that challenge this notion. In particular, we show that while optimizing for AUROC equates to minimizing the model’s FPR in an unbiased manner over positive sample scores, optimizing for AUPRC equates to minimizing the FPR specifically for regions where the model outputs higher scores relative to lower scores. *Further, we show both theoretically and empirically over synthetic and real-world fairness datasets that AUPRC can be an explicitly discriminatory metric through favoring higher-prevalence subgroups.*

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.

Acknowledgements

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

References

- [1] Nasim Abdollahi, Seyed Ali Madani Tonekaboni, Jay Huang, Bo Wang, and Stephen MacKinnon. Nodecoder: a graph-based machine learning platform to predict active sites of modeled protein structures, 2023.
- [2] Avraham Adler. Using machine learning techniques to identify key risk factors for diabetes and undiagnosed diabetes, 2021.
- [3] B Thomas Adler, Luca De Alfaro, Santiago M Mola-Velasco, Paolo Rosso, and Andrew G West. Wikipedia vandalism detection: Combining natural language, metadata, and reputation features. In *Computational Linguistics and Intelligent Text Processing: 12th International Conference, CICLing 2011, Tokyo, Japan, February 20-26, 2011. Proceedings, Part II 12*, pages 277–288. Springer, 2011.
- [4] Sergey Afanasiev, Anastasiya Smirnova, and Diana Kotereva. Itsy bitsy spidernet: Fully connected residual network for fraud detection, 2021.
- [5] Karan Aggarwal, Onur Atan, Ahmed K Farahat, Chi Zhang, Kosta Ristovski, and Chetan Gupta. Two birds with one network: Unifying failure event prediction and time-to-failure

- modeling. In *2018 IEEE international conference on big data (Big Data)*, pages 1308–1317. IEEE, 2018.
- [6] Faruk Ahmed and Aaron Courville. Detecting semantic anomalies. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):3154–3162, Apr. 2020.
- [7] Sajid Ahmed, Farshid Rayhan, Asif Mahbub, Md Rafsan Jani, Swakkhar Shatabda, and Dewan Md Farid. Liubooost: locality informed under-boosting for imbalanced data classification. In *Emerging Technologies in Data Mining and Information Security: Proceedings of IEMIS 2018, Volume 2*, pages 133–144. Springer, 2019.
- [8] Giambattista Albora and Andrea Zaccaria. Machine learning to assess relatedness: the advantage of using firm-level data. *Complexity*, 2022, 2022.
- [9] Jr. Allam, Tarek and Jason D McEwen. Paying attention to astronomical transients: introducing the time-series transformer for photometric classification. *RAS Techniques and Instruments*, 3(1):209–223, 10 2023.
- [10] Maxime Alvarez, Jean-Charles Verdier, D’Jeff K. Nkashama, Marc Frappier, Pierre-Martin Tardif, and Froduald Kabanza. A revealing large-scale evaluation of unsupervised anomaly detection algorithms, 2022.
- [11] Ilya Amburg, Jon Kleinberg, and Austin R Benson. Planted hitting set recovery in hypergraphs. *Journal of Physics: Complexity*, 2(3):035004, 2021.
- [12] Shin Ando and Chun Yuan Huang. 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*, pages 770–785. Springer, 2017.
- [13] Renan Andrades and Mariana Recamonde-Mendoza. Machine learning methods for prediction of cancer driver genes: a survey paper. *Briefings in Bioinformatics*, 23(3):bbac062, 2022.
- [14] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. In *Ethics of data and analytics*, pages 254–264. Auerbach Publications, 2022.
- [15] Nino Antulov-Fantulin, Dijana Tolic, Matija Piskorec, Zhang Ce, and Irena Vodenska. Inferring short-term volatility indicators from the bitcoin blockchain. In Luca Maria Aiello, Chantal Cherifi, Hocine Cherifi, Renaud Lambiotte, Pietro Lió, and Luis M. Rocha, editors, *Complex Networks and Their Applications VII*, pages 508–520, Cham, 2019. Springer International Publishing.
- [16] Bruno Arcanjo, Bruno Ferrarini, Michael Milford, Klaus D McDonald-Maier, and Shoaib Ehsan. An efficient and scalable collection of fly-inspired voting units for visual place recognition in changing environments. *IEEE Robotics and Automation Letters*, 7(2):2527–2534, 2022.
- [17] Arthur Asuncion and David Newman. Uci machine learning repository, 2007.
- [18] Sezin Kircali Ata, Yuan Fang, Min Wu, Jiaqi Shi, Chee Keong Kwoh, and Xiaoli Li. Multi-view collaborative network embedding. *ACM Transactions on Knowledge Discovery from Data*, 15(3):1–18, April 2021.
- [19] Anand Avati, Kenneth Jung, Stephanie Harman, Lance Downing, Andrew Ng, and Nigam H Shah. Improving palliative care with deep learning. *BMC medical informatics and decision making*, 18(4):55–64, 2018.
- [20] Kleanthis Avramidis, Shanti Stewart, and Shrikanth Narayanan. On the role of visual context in enriching music representations. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.
- [21] Simon Axelrod and Rafael Gomez-Bombarelli. Molecular machine learning with conformer ensembles. *Machine Learning: Science and Technology*, 4(3):035025, 2023.

- [22] Kasra Babaei, Zhi Yuan Chen, and Tomas Maul. Aegr: a simple approach to gradient reversal in autoencoders for network anomaly detection. *Soft Computing*, 25(24):15269–15280, 2021.
- [23] Kasra Babaei, ZhiYuan Chen, and Tomas Maul. Data augmentation by autoencoders for unsupervised anomaly detection, 2019.
- [24] Van Bach Nguyen, Kanishka Ghosh Dastidar, Michael Granitzer, and Wissam Siblini. The importance of future information in credit card fraud detection. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 10067–10077. PMLR, 28–30 Mar 2022.
- [25] Bart Baesens, Sebastiaan H^oppner, Irene Ortner, and Tim Verdonck. robrose: A robust approach for dealing with imbalanced data in fraud detection. *Statistical Methods & Applications*, 30(3):841–861, 2021.
- [26] Leen De Baets, Joeri Ruysinck, Thomas Peiffer, Johan Decruyenaere, Filip De Turck, Femke Ongenaes, and Tom Dhaene. Positive blood culture detection in time series data using a bilstm network, 2016.
- [27] Vassileios Balntas, Karel Lenc, Andrea Vedaldi, and Krystian Mikolajczyk. Hpatches: A benchmark and evaluation of handcrafted and learned local descriptors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5173–5182, 2017.
- [28] Amitava Banerjee, Joseph D Hart, Rajarshi Roy, and Edward Ott. Machine learning link inference of noisy delay-coupled networks with optoelectronic experimental tests. *Physical Review X*, 11(3):031014, 2021.
- [29] Batuhan Bardak and Mehmet Tan. Using clinical drug representations for improving mortality and length of stay predictions. In *2021 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, pages 1–8. IEEE, 2021.
- [30] Ioannis Bargiotas, Argyris Kalogeratos, Myrto Limnios, Pierre-Paul Vidal, Damien Ricard, and Nicolas Vayatis. Revealing posturographic profile of patients with parkinsonian syndromes through a novel hypothesis testing framework based on machine learning. *PLOS ONE*, 16(2):1–22, 02 2021.
- [31] Jens Bayer, David M^unch, and Michael Arens. Image-based out-of-distribution-detector principles on graph-based input data in human action recognition. In *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part I*, page 26–40, Berlin, Heidelberg, 2021. Springer-Verlag.
- [32] Neslihan Bayramoglu, Miika T Nieminen, and Simo Saarakkala. Automated detection of patellofemoral osteoarthritis from knee lateral view radiographs using deep learning: data from the multicenter osteoarthritis study (most). *Osteoarthritis and Cartilage*, 29(10):1432–1447, 2021.
- [33] Andreas Beger. Precision-recall curves, April 2016.
- [34] Hafida Benhidour, Lama Almeshkhas, and Said Kerrache. An approach for link prediction in directed complex networks based on asymmetric similarity-popularity, 2022.
- [35] Austin R Benson and Jon Kleinberg. Found graph data and planted vertex covers. *Advances in Neural Information Processing Systems*, 31, 2018.
- [36] James Bergstra, R^emi Bardenet, Yoshua Bengio, and Bal^azs K^egl. Algorithms for hyper-parameter optimization. *Advances in neural information processing systems*, 24, 2011.
- [37] James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of machine learning research*, 13(2), 2012.
- [38] Victor Besnier, Andrei Bursuc, David Picard, and Alexandre Briot. Triggering failures: Out-of-distribution detection by learning from local adversarial attacks in semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15701–15710, 2021.

- [39] Debayan Bhattacharya, Benjamin Tobias Becker, Finn Behrendt, Marcel Bengs, Dirk Beyersdorff, Dennis Eggert, Elina Petersen, Florian Jansen, Marvin Petersen, Bastian Cheng, et al. Supervised contrastive learning to classify paranasal anomalies in the maxillary sinus. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 429–438. Springer, 2022.
- [40] Kevin Bleakley, Gérard Biau, and Jean-Philippe Vert. Supervised reconstruction of biological networks with local models. *Bioinformatics*, 23(13):i57–i65, 2007.
- [41] D Blevins, P Moriano, R Bridges, M Verma, M Iannacone, and S Hollifield. Time-based can intrusion detection benchmark. In *Workshop on Automotive and Autonomous Vehicle Security (AutoSec)*, 2021.
- [42] Nathaniel J Bloomfield, Susan Wei, Bartholomew A. Woodham, Peter Wilkinson, and Andrew P Robinson. Automating the assessment of biofouling in images using expert agreement as a gold standard. *Scientific reports*, 11(1):2739, 2021.
- [43] Hermann Blum, Paul-Edouard Sarlin, Juan Nieto, Roland Siegwart, and Cesar Cadena. The fishyscapes benchmark: Measuring blind spots in semantic segmentation. *International Journal of Computer Vision*, 129(11):3119–3135, 2021.
- [44] Ali Borji, Ming-Ming Cheng, Huaizu Jiang, and Jia Li. Salient object detection: A benchmark. *IEEE transactions on image processing*, 24(12):5706–5722, 2015.
- [45] Kendrick Boyd, Vitor Santos Costa, Jesse Davis, and C David Page. Unachievable region in precision-recall space and its effect on empirical evaluation. In *Proceedings of the... International Conference on Machine Learning. International Conference on Machine Learning*, volume 2012, page 349. NIH Public Access, 2012.
- [46] Kendrick Boyd, Kevin H Eng, and C David Page. 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*, pages 451–466. Springer, 2013.
- [47] Jakob Božič, Domen Tabernik, and Danijel Škočaj. End-to-end training of a two-stage neural network for defect detection. In *2020 25th International conference on pattern recognition (ICPR)*, pages 5619–5626. IEEE, 2021.
- [48] Jan Brabec, Tomáš Komárek, Vojtěch Franc, and Lukáš Machlica. On model evaluation under non-constant class imbalance. In *Computational Science–ICCS 2020: 20th International Conference, Amsterdam, The Netherlands, June 3–5, 2020, Proceedings, Part IV 20*, pages 74–87. Springer, 2020.
- [49] Jonathan Brophy and Daniel Lowd. Eggs: A flexible approach to relational modeling of social network spam, 2020.
- [50] Nicolas Brosse, Carlos Riquelme, Alice Martin, Sylvain Gelly, and Éric Moulines. On last-layer algorithms for classification: Decoupling representation from uncertainty estimation, 2020.
- [51] Jason Brownlee. Tour of evaluation metrics for imbalanced classification, January 2020.
- [52] Jonathan Bryan and Pablo Moriano. Graph-based machine learning improves just-in-time defect prediction. *Plos one*, 18(4):e0284077, 2023.
- [53] Céline Budding, Fabian Eitel, Kerstin Ritter, and Stefan Haufe. Evaluating saliency methods on artificial data with different background types, 2021.
- [54] Marcin Budka, Akanda Wahid Ul Ashraf, Matthew Bennett, Scott Neville, and Alun Mackrill. Deep multilabel cnn for forensic footwear impression descriptor identification. *Applied Soft Computing*, 109:107496, 2021.
- [55] Tian Cai, Li Xie, Muge Chen, Yang Liu, Di He, Shuo Zhang, Cameron Mura, Philip E. Bourne, and Lei Xie. Exploration of dark chemical genomics space via portal learning: Applied to targeting the undruggable genome and covid-19 anti-infective polypharmacology, 2021.

- [56] Chen Chai, Juanwu Lu, Xuan Jiang, Xiupeng Shi, and Zeng Zeng. An automated machine learning (automl) method for driving distraction detection based on lane-keeping performance. *arXiv preprint arXiv:2103.08311*, 2021.
- [57] Satrajit Chakrabarty, Pamela LaMontagne, Joshua Shimony, Daniel S Marcus, and Aristeidis Sotiras. Mri-based classification of idh mutation and 1p/19q codeletion status of gliomas using a 2.5 d hybrid multi-task convolutional neural network. *Neuro-Oncology Advances*, 5(1):vdad023, 2023.
- [58] Neeloy Chakraborty, Aamir Hasan, Shuijing Liu, Tianchen Ji, Weihang Liang, D. Livingston McPherson, and Katherine Driggs-Campbell. 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*, page 1125–1134, Richland, SC, 2023. International Foundation for Autonomous Agents and Multiagent Systems.
- [59] Saptarshi Chakraborty, Colin B Begg, and Ronglai Shen. Using the “hidden” genome to improve classification of cancer types. *Biometrics*, 77(4):1445–1455, 2021.
- [60] Saptarshi Chakraborty, Zoe Guan, Colin B Begg, and Ronglai Shen. Topical hidden genome: Discovering latent cancer mutational topics using a bayesian multilevel context-learning approach. *arXiv preprint arXiv:2212.14567*, 2022.
- [61] Raghavendra Chalapathy, Aditya Krishna Menon, and Sanjay Chawla. Robust, deep and inductive anomaly detection. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2017, Skopje, Macedonia, September 18–22, 2017, Proceedings, Part I 10*, pages 36–51. Springer, 2017.
- [62] Robin Chan, Matthias Rottmann, and Hanno Gottschalk. Entropy maximization and meta classification for out-of-distribution detection in semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5128–5137, 2021.
- [63] Hugh Chen, Scott Lundberg, and Su-In Lee. Checkpoint ensembles: Ensemble methods from a single training process, 2017.
- [64] Hugh Chen, Scott Lundberg, and Su-In Lee. Hybrid gradient boosting trees and neural networks for forecasting operating room data, 2018.
- [65] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.
- [66] Yu Chen, Jiaqi Jin, Hui Zhao, Pengjie Wang, Guojun Liu, Jian Xu, and Bo Zheng. Asymptotically unbiased estimation for delayed feedback modeling via label correction. In *Proceedings of the ACM Web Conference 2022*, pages 369–379, 2022.
- [67] Zhuohao Chen, James Gibson, Ming-Chang Chiu, Qiaohong Hu, Tara K Knight, Daniella Meeker, James A Tulskey, Kathryn I Pollak, and Shrikanth Narayanan. Automated empathy detection for oncology encounters. In *2020 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 1–8. IEEE, 2020.
- [68] Davide Chicco. Ten quick tips for machine learning in computational biology. *BioData mining*, 10(1):35, 2017.
- [69] Julien Chiquet, Stephane Robin, and Mahendra Mariadassou. Variational inference for sparse network reconstruction from count data. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1162–1171. PMLR, 09–15 Jun 2019.
- [70] Brian Y Cho, Tucker Hermans, and Alan Kuntz. Planning sensing sequences for subsurface 3d tumor mapping. In *2021 International Symposium on Medical Robotics (ISMR)*, pages 1–7. IEEE, 2021.

- [71] Edward Choi, Cao Xiao, Walter F. Stewart, and Jimeng Sun. 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*, page 4552–4562, Red Hook, NY, USA, 2018. Curran Associates Inc.
- [72] Eunji Chong, Katha Chanda, Zhefan Ye, Audrey Southerland, Nataniel Ruiz, Rebecca M. Jones, Agata Rozga, and James M. Rehg. Detecting gaze towards eyes in natural social interactions and its use in child assessment. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 1(3), sep 2017.
- [73] Peter Christen. Application of advanced record linkage techniques for complex population reconstruction. *arXiv preprint arXiv:1612.04286*, 2016.
- [74] Xu Chu, Yang Lin, Jingyue Gao, Jiangtao Wang, Yasha Wang, and Leye Wang. Multi-label robust factorization autoencoder and its application in predicting drug-drug interactions, 2018.
- [75] Marc Claesen, Frank De Smet, Johan A.K. Suykens, and Bart De Moor. A robust ensemble approach to learn from positive and unlabeled data using svm base models. *Neurocomputing*, 160:73–84, July 2015.
- [76] Johanna C Clauser, Judith Maas, Jutta Arens, Thomas Schmitz-Rode, Ulrich Steinseifer, and Benjamin Berkels. Automation of hemocompatibility analysis using image segmentation and a random forest. *arXiv preprint arXiv:2010.06245*, 2020.
- [77] Felipe Colombelli, Thayne Woycinck Kowalski, and Mariana Recamonde-Mendoza. A hybrid ensemble feature selection design for candidate biomarkers discovery from transcriptome profiles. *Knowledge-Based Systems*, 254:109655, 2022.
- [78] Jonathan Cook and Vikram Ramadas. When to consult precision-recall curves. *The Stata Journal*, 20(1):131–148, 2020.
- [79] Skyler J. Cranmer and Bruce A. Desmarais. What can we learn from predictive modeling?, 2016.
- [80] Jakub Czakon. F1 score vs roc auc vs accuracy vs pr auc: Which evaluation metric should you choose?, July 2022.
- [81] Anusha Damodaran, Fabio Di Troia, Corrado Aaron Visaggio, Thomas H Austin, and Mark Stamp. A comparison of static, dynamic, and hybrid analysis for malware detection. *Journal of Computer Virology and Hacking Techniques*, 13:1–12, 2017.
- [82] Armin Danesh Pazho, Ghazal Alinezhad Noghre, Babak Rahimi Ardabili, Christopher Neff, and Hamed Tabkhi. *CHAD: Charlotte Anomaly Dataset*, page 50–66. Springer Nature Switzerland, 2023.
- [83] Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd International Conference on Machine Learning, ICML '06*, page 233–240, New York, NY, USA, 2006. Association for Computing Machinery.
- [84] Alice Del Vecchio, Andreea Deac, Pietro Liò, and Petar Veličković. Neural message passing for joint paratope-epitope prediction. *arXiv preprint arXiv:2106.00757*, 2021.
- [85] Jianyuan Deng, Zhibo Yang, Hehe Wang, Iwao Ojima, Dimitris Samaras, and Fusheng Wang. Unraveling key elements underlying molecular property prediction: A systematic study, 2023.
- [86] Amit G. Deshwar, Shankar Vembu, Christina K. Yung, Gun Ho Jang, Lincoln Stein, and Quaid Morris. Reconstructing subclonal composition and evolution from whole genome sequencing of tumors, 2015.
- [87] Giancarlo Di Biase, Hermann Blum, Roland Siegwart, and Cesar Cadena. Pixel-wise anomaly detection in complex driving scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16918–16927, 2021.

- [88] Daisy Yi Ding, Chloé Simpson, Stephen Pfohl, Dave C Kale, Kenneth Jung, and Nigam H Shah. The effectiveness of multitask learning for phenotyping with electronic health records data. In *BIOCOMPUTING 2019: Proceedings of the Pacific Symposium*, pages 18–29. World Scientific, 2018.
- [89] Jonas C. Ditz, Bernhard Reuter, and Nico Pfeifer. Convolutional motif kernel networks, 2023.
- [90] Rémi Domingues, Pietro Michiardi, Jérémie Barlet, and Maurizio Filippone. A comparative evaluation of novelty detection algorithms for discrete sequences. *Artificial Intelligence Review*, 53:3787–3812, 2020.
- [91] Shangjia Dong, Tianbo Yu, Hamed Farahmand, and Ali Mostafavi. A hybrid deep learning model for predictive flood warning and situation awareness using channel network sensors data. *Computer-Aided Civil and Infrastructure Engineering*, 36(4):402–420, 2021.
- [92] Wen Dong, Tong Guan, Bruno Lepri, and Chunming Qiao. Pocketcare: Tracking the flu with mobile phones using partial observations of proximity and symptoms. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 3(2):1–23, 2019.
- [93] Julia Dressel and Hany Farid. The accuracy, fairness, and limits of predicting recidivism. *Science advances*, 4(1):eaao5580, 2018.
- [94] Guodong Du, Liang Yuan, Kong Joo Shin, and Shunsuke Managi. Enhancement of land-use change modeling using convolutional neural networks and convolutional denoising autoencoders, 2018.
- [95] Marco Dueñas, Víctor Ortiz, Massimo Riccaboni, and Francesco Serti. Assessing the Impact of COVID-19 on Trade: a Machine Learning Counterfactual Analysis. Working papers 79, Red Investigadores de Economía, April 2021.
- [96] Witold Dyrka, Mateusz Pyzik, François Coste, and Hugo Talibart. Estimating probabilistic context-free grammars for proteins using contact map constraints. *PeerJ*, 7:e6559, March 2019.
- [97] Gabriel Erion, Hugh Chen, Scott M. Lundberg, and Su-In Lee. Anesthesiologist-level forecasting of hypoxemia with only spo2 data using deep learning, 2017.
- [98] Şirag Erkol, Satyaki Sikdar, Filippo Radicchi, and Santo Fortunato. Consistency pays off in science. *Quantitative Science Studies*, 4(2):491–500, 2023.
- [99] Alessandro Fabris, Stefano Messina, Gianmaria Silvello, and Gian Antonio Susto. Algorithmic fairness datasets: the story so far. *Data Mining and Knowledge Discovery*, 36(6):2074–2152, 2022.
- [100] Stefan Falkner, Aaron Klein, and Frank Hutter. Bohb: Robust and efficient hyperparameter optimization at scale. In *International conference on machine learning*, pages 1437–1446. PMLR, 2018.
- [101] Yang Feng, Min Zhou, and Xin Tong. Imbalanced classification: A paradigm-based review. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 14(5):383–406, 2021.
- [102] Alberto Fernández, Salvador García, Mikel Galar, Ronaldo C Prati, Bartosz Krawczyk, and Francisco Herrera. *Learning from imbalanced data sets*, volume 10. Springer, 2018.
- [103] Andres Ferraro, Dmitry Bogdanov, Xavier Serra Jay, Ho Jeon, and Jason Yoon. How low can you go? reducing frequency and time resolution in current cnn architectures for music auto-tagging. In *2020 28th European signal processing conference (EUSIPCO)*, pages 131–135. IEEE, 2021.
- [104] Peter Flach and Meelis Kull. Precision-recall-gain curves: Pr analysis done right. *Advances in neural information processing systems*, 28, 2015.
- [105] Carlos Floyd, Herbert Levine, Christopher Jarzynski, and Garegin A Papoian. Understanding cytoskeletal avalanches using mechanical stability analysis. *Proceedings of the National Academy of Sciences*, 118(41):e2110239118, 2021.

- [106] Marina Fomicheva, Piyawat Lertvittayakumjorn, Wei Zhao, Steffen Eger, and Yang Gao. The eval4nlp shared task on explainable quality estimation: Overview and results. In Yang Gao, Steffen Eger, Wei Zhao, Piyawat Lertvittayakumjorn, and Marina Fomicheva, editors, *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 165–178, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.
- [107] Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra. Fsd50k: An open dataset of human-labeled sound events. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 30:829–852, dec 2021.
- [108] Fabrizio Frasca, Diego Galeano, Guadalupe Gonzalez, Ivan Laponogov, Kirill Veselkov, Alberto Paccanaro, and Michael M. Bronstein. Learning interpretable disease self-representations for drug repositioning, 2019.
- [109] Marco Frasca and Nicolo Cesa Bianchi. Multitask protein function prediction through task dissimilarity. *IEEE/ACM transactions on computational biology and bioinformatics*, 16(5):1550–1560, 2017.
- [110] Marco Frasca and Nicolò Cesa-Bianchi. Positive and unlabeled learning through negative selection and imbalance-aware classification, 2019.
- [111] Yuming Fu, Xue-Bing Wu, Qian Yang, Anthony GA Brown, Xiaotong Feng, Qinchun Ma, and Shuyan Li. Finding quasars behind the galactic plane. i. candidate selections with transfer learning. *The Astrophysical Journal Supplement Series*, 254(1):6, 2021.
- [112] Silvio Galesso, Maria Alejandra Bravo, Mehdi Naouar, and Thomas Brox. Probing contextual diversity for dense out-of-distribution detection, 2022.
- [113] Varun Gangal, Abhinav Arora, Arash Einolghozati, and Sonal Gupta. Likelihood ratios and generative classifiers for unsupervised out-of-domain detection in task oriented dialog. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7764–7771, Apr. 2020.
- [114] Dario Garcia-Gasulla, Eduard Ayguadé, Jesús Labarta, Ulises Cortés, and Toyotaro Suzumura. Hierarchical hyperlink prediction for the www, 2016.
- [115] Matthieu Garcin and Samuel Stéphan. Credit scoring using neural networks and sure posterior probability calibration, 2021.
- [116] Jean-Gabriel Gaudreault, Paula Branco, and João Gama. An analysis of performance metrics for imbalanced classification. In *International Conference on Discovery Science*, pages 67–77. Springer, 2021.
- [117] Sachin Gavali, Chuming Chen, Julie Cowart, Xi Peng, Shanshan Ding, Cathy Wu, and Tammy Anderson. Understanding the factors related to the opioid epidemic using machine learning. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1309–1314. IEEE, 2021.
- [118] Isaac D Gerg and Vishal Monga. Structural prior driven regularized deep learning for sonar image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–16, 2021.
- [119] Aurélien Géron. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. "O'Reilly Media, Inc.", 2022.
- [120] Djordje Gligorijevic, Jelena Gligorijevic, and Aaron Flores. Prospective modeling of users for online display advertising via deep time-aware model. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management, CIKM '20*, page 2461–2468, New York, NY, USA, 2020. Association for Computing Machinery.
- [121] Mark Goadrich, Louis Oliphant, and Jude Shavlik. Gleaner: Creating ensembles of first-order clauses to improve recall-precision curves. *Machine Learning*, 64:231–261, 2006.
- [122] Izhak Golan and Ran El-Yaniv. Deep anomaly detection using geometric transformations. *Advances in neural information processing systems*, 31, 2018.

- [123] Samuel Goldman, Ria Das, Kevin K. Yang, and Connor W. Coley. Machine learning modeling of family wide enzyme-substrate specificity screens. *PLOS Computational Biology*, 18(2):e1009853, February 2022.
- [124] Hongyu Gong, Alberto Valido, Katherine M Ingram, Giulia Fanti, Suma Bhat, and Dorothy L Espelage. 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, pages 14804–14812, 2021.
- [125] Yuan Gong, Yu-An Chung, and James Glass. Psla: Improving audio tagging with pretraining, sampling, labeling, and aggregation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3292–3306, 2021.
- [126] Eric Goodman, Chase Zimmerman, and Corey Hudson. Packet2vec: Utilizing word2vec for feature extraction in packet data. Technical report, Sandia National Lab.(SNL-NM), Albuquerque, NM (United States); Sandia . . . , 2019.
- [127] Julia Sophia Gottfriedsen, Max Berrendorf, Pierre Gentine, Birgit Hassler, Markus Reichstein, Katja Weigel, and Veronika Eyring. On the generalization of agricultural drought classification from climate data. thirty-fifth conference on neural information processing systems (neurips) workshop 2021"" tackling climate change with machine learning"". 2021.
- [128] Anil Goyal and Jihed Khiari. Diversity-aware weighted majority vote classifier for imbalanced data. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2020.
- [129] Eric Granger, Madhu Kiran, Louis-Antoine Blais-Morin, et al. A comparison of cnn-based face and head detectors for real-time video surveillance applications. In *2017 seventh international conference on image processing theory, tools and applications (IPTA)*, pages 1–7. IEEE, 2017.
- [130] Zoe Guan, Giovanni Parmigiani, Danielle Braun, and Lorenzo Trippa. Prediction of hereditary cancers using neural networks. *The annals of applied statistics*, 16(1):495, 2022.
- [131] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020.
- [132] Ekaterina Gurina, Nikita Klyuchnikov, Alexey Zaytsev, Evgenya Romanenkova, Ksenia Antipova, Igor Simon, Victor Makarov, and Dmitry Koroteev. Application of machine learning to accidents detection at directional drilling. *Journal of Petroleum Science and Engineering*, 184:106519, 2020.
- [133] Guillaume Haben, Sarra Habchi, Mike Papadakis, Maxime Cordy, and Yves Le Traon. Discerning legitimate failures from false alerts: A study of chromium’s continuous integration, 2021.
- [134] Tom Rolandus Hagedoorn and Gerasimos Spanakis. Massive open online courses temporal profiling for dropout prediction. In *2017 IEEE 29th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 231–238. IEEE, 2017.
- [135] Alexander Hägele, Jonas Rothfuss, Lars Lorch, Vignesh Ram Somnath, Bernhard Schölkopf, and Andreas Krause. Bacadi: Bayesian causal discovery with unknown interventions. In *International Conference on Artificial Intelligence and Statistics*, pages 1411–1436. PMLR, 2023.
- [136] Melissa Hall, Bobbie Chern, Laura Gustafson, Denisse Ventura, Harshad Kulkarni, Candace Ross, and Nicolas Usunier. Towards reliable assessments of demographic disparities in multi-label image classifiers, 2023.
- [137] Jacob A. Harer, Louis Y. Kim, Rebecca L. Russell, Onur Ozdemir, Leonard R. Kosta, Akshay Rangamani, Lei H. Hamilton, Gabriel I. Centeno, Jonathan R. Key, Paul M. Ellingwood, Erik Antelman, Alan Mackay, Marc W. McConley, Jeffrey M. Opper, Peter Chin, and Tomo Lazovich. Automated software vulnerability detection with machine learning, 2018.

- [138] Hrayr Harutyunyan, Hrant Khachatryan, David C Kale, Greg Ver Steeg, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *Scientific data*, 6(1):96, 2019.
- [139] Hrayr Harutyunyan, Hrant Khachatryan, David C. Kale, Greg Ver Steeg, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *Scientific Data*, 6(1), June 2019.
- [140] Seyed Raein Hashemi, Seyed Sadegh Mohseni Salehi, Deniz Erdogmus, Sanjay P Prabhu, Simon K Warfield, and Ali Gholipour. 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.
- [141] Haibo He and Yunqian Ma. Imbalanced learning: foundations, algorithms, and applications. 2013.
- [142] Jieyue He, Xinxing Yang, Zhuo Gong, and Ibrahim Zamit. Hybrid attentional memory network for computational drug repositioning. *BMC bioinformatics*, 21:1–17, 2020.
- [143] Jörn Hees, Dayananda Herurkar, and Mario Meier. Recol: Reconstruction error columns for outlier detection, 2021.
- [144] Dan Hendrycks, Steven Basart, Mantas Mazeika, Andy Zou, Joseph Kwon, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling out-of-distribution detection for real-world settings. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 8759–8773. PMLR, 17–23 Jul 2022.
- [145] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017.
- [146] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations*, 2019.
- [147] Jens Henriksson, Christian Berger, Markus Borg, Lars Tornberg, Cristofer Englund, Sankar Raman Sathyamoorthy, and Stig Ursing. Towards structured evaluation of deep neural network supervisors. In *2019 IEEE International Conference On Artificial Intelligence Testing (AITest)*, pages 27–34. IEEE, 2019.
- [148] Ulysse Herbach. Gene regulatory network inference from single-cell data using a self-consistent proteomic field, 2021.
- [149] Justus I. Hibshman and Tim Weninger. Inherent limits on topology-based link prediction, 2023.
- [150] Steven A Hicks, Inga Strümke, Vajira Thambawita, Malek Hammou, Michael A Riegler, Pål Halvorsen, and Sravanthi Parasa. On evaluation metrics for medical applications of artificial intelligence. *Scientific reports*, 12(1):5979, 2022.
- [151] Anna Himmelhuber, Mitchell Joblin, Martin Ringsquandl, and Thomas Runkler. Demystifying graph neural network explanations. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 67–75. Springer, 2021.
- [152] David Hin, Andrey Kan, Huaming Chen, and M Ali Babar. Linevd: Statement-level vulnerability detection using graph neural networks. In *Proceedings of the 19th international conference on mining software repositories*, pages 596–607, 2022.
- [153] Kaoutar Daoud Hiri, Matjaž Hren, and Tomaž Curk. Nlp-based classification of software tools for metagenomics sequencing data analysis into edam semantic annotation, 2022.
- [154] Ryohei Hisano, Didier Sornette, and Takayuki Mizuno. Prediction of esg compliance using a heterogeneous information network. *Journal of Big Data*, 7(1), March 2020.

- [155] Charmgil Hong and Milos Hauskrecht. Mcode: Multivariate conditional outlier detection. *arXiv preprint arXiv:1505.04097*, 2015.
- [156] Shenda Hong, Cao Xiao, Trong Nghia Hoang, Tengfei Ma, Hongyan Li, and Jimeng Sun. Rdpd: Rich data helps poor data via imitation. In *28th International Joint Conference on Artificial Intelligence, IJCAI 2019*, pages 5895–5901. International Joint Conferences on Artificial Intelligence, 2019.
- [157] Shenda Hong, Cao Xiao, Tengfei Ma, Hongyan Li, and Jimeng Sun. Mina: Multilevel knowledge-guided attention for modeling electrocardiography signals. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5888–5894. International Joint Conferences on Artificial Intelligence Organization, 7 2019.
- [158] Julia Hornauer and Vasileios Belagiannis. Heatmap-based out-of-distribution detection. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, January 2023.
- [159] Chao-Chun Hsu, Shantanu Karnwal, Sendhil Mullainathan, Ziad Obermeyer, and Chenhao Tan. Characterizing the value of information in medical notes. In Trevor Cohn, Yulan He, and Yang Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2062–2072, Online, November 2020. Association for Computational Linguistics.
- [160] Stanley Bryan Z. Hua, Mandy Rickard, John Weaver, Alice Xiang, Daniel Alvarez, Kyla N. Velear, Kunj Sheth, Gregory E. Tasian, Armando J. Lorenzo, Anna Goldenberg, and Lauren Erdman. From single-visit to multi-visit image-based models: single-visit models are enough to predict obstructive hydronephrosis. In Jorge Brieva, Pamela Guevara, Natasha Lepore, Marius G. Linguraru, Letícia Rittner, and Eduardo Romero Castro M.D., editors, *18th International Symposium on Medical Information Processing and Analysis*, volume 12567, page 1256710. International Society for Optics and Photonics, SPIE, 2023.
- [161] Yiqing Hua, Armin Namavari, Kaishuo Cheng, Mor Naaman, and Thomas Ristenpart. Increasing adversarial uncertainty to scale private similarity testing. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 1777–1794, 2022.
- [162] Benjamin Hughes and Tilo Burghardt. Automated visual fin identification of individual great white sharks. *International Journal of Computer Vision*, 122:542–557, 2017.
- [163] Zepeng Huo, Xiaoning Qian, Shuai Huang, Zhangyang Wang, and Bobak J Mortazavi. Density-aware personalized training for risk prediction in imbalanced medical data. In *Machine Learning for Healthcare Conference*, pages 101–122. PMLR, 2022.
- [164] Fantine Huot, R. Lily Hu, Nita Goyal, Tharun Sankar, Matthias Ihme, and Yi-Fan Chen. Next day wildfire spread: A machine learning dataset to predict wildfire spreading from remote-sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13, 2022.
- [165] Vân Anh Huynh-Thu and Guido Sanguinetti. Gene regulatory network inference: an introductory survey. *Gene regulatory networks: Methods and protocols*, pages 1–23, 2019.
- [166] Zina M Ibrahim, Daniel Bean, Thomas Searle, Linglong Qian, Honghan Wu, Anthony Shek, Zeljko Kraljevic, James Galloway, Sam Norton, James TH Teo, et al. A knowledge distillation ensemble framework for predicting short-and long-term hospitalization outcomes from electronic health records data. *IEEE Journal of Biomedical and Health Informatics*, 26(1):423–435, 2021.
- [167] Yotam Intrator, Gilad Katz, and Asaf Shabtai. Mdgan: Boosting anomaly detection using multi-discriminator generative adversarial networks, 2018.
- [168] Mary Isangediok and Kelum Gajamannage. Fraud detection using optimized machine learning tools under imbalance classes. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 4275–4284. IEEE, 2022.
- [169] Olga Isupova, Danil Kuzin, and Lyudmila Mihaylova. Learning methods for dynamic topic modeling in automated behavior analysis. *IEEE transactions on neural networks and learning systems*, 29(9):3980–3993, 2017.

- [170] Jay Jacobs, Sasha Romanosky, Octavian Suci, Benjamin Edwards, and Armin Sarabi. Enhancing vulnerability prioritization: Data-driven exploit predictions with community-driven insights, 2023.
- [171] Lucas Jaffe, Michael Zelinski, and Wesam Sakla. Remote sensor design for visual recognition with convolutional neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, 57(11):9090–9108, November 2019.
- [172] Shantanu Jain, Martha White, and Predrag Radivojac. Recovering true classifier performance in positive-unlabeled learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1), Feb. 2017.
- [173] Dong-ju Jeong, Insung Hwang, and Nam Ik Cho. Co-salient object detection based on deep saliency networks and seed propagation over an integrated graph. *IEEE Transactions on Image Processing*, 27(12):5866–5879, 2018.
- [174] Biaobin Jiang, Kyle Kloster, David F Gleich, and Michael Gribskov. Aprank: an adaptive pagerank model for protein function prediction on bi-relational graphs. *Bioinformatics*, 33(12):1829–1836, 2017.
- [175] Wei Jiang, Jing-Hao Xue, and Weichuan Yu. Estimating reproducibility in genome-wide association studies, 2015.
- [176] Yunjiang Jiang, Yue Shang, Rui Li, Wen-Yun Yang, Guoyu Tang, Chaoyi Ma, Yun Xiao, and Eric Zhao. A unified neural network approach to e-commerce relevance learning. In *Proceedings of the 1st International Workshop on Deep Learning Practice for High-Dimensional Sparse Data*, pages 1–7, 2019.
- [177] Di Jin, Shuyang Gao, Seokhwan Kim, Yang Liu, and Dilek Hakkani-Tür. Towards textual out-of-domain detection without in-domain labels. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:1386–1395, 2022.
- [178] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.
- [179] Eric Jonas, Monica Bobra, Vaishaal Shankar, J Todd Hoeksema, and Benjamin Recht. Flare prediction using photospheric and coronal image data. *Solar Physics*, 293(3):48, 2018.
- [180] Cheng Ju, James Li, Bram Wasti, and Shengbo Guo. Semisupervised learning on heterogeneous graphs and its applications to facebook news feed, 2018.
- [181] Sanghun Jung, Jungsoo Lee, Daehoon Gwak, Sungha Choi, and Jaegul Choo. Standardized max logits: A simple yet effective approach for identifying unexpected road obstacles in urban-scene segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15425–15434, 2021.
- [182] Ruthwik R Junuthula, Kevin S Xu, and Vijay K Devabhaktuni. Evaluating link prediction accuracy in dynamic networks with added and removed edges. In *2016 IEEE international conferences on big data and cloud computing (BDCloud), social computing and networking (SocialCom), sustainable computing and communications (SustainCom)(BDCloud-SocialCom-SustainCom)*, pages 377–384. IEEE, 2016.
- [183] Georgi Karadzhov, Tom Stafford, and Andreas Vlachos. What makes you change your mind? an empirical investigation in online group decision-making conversations, 2022.
- [184] Md Rezaul Karim, Michael Cochez, Joao Bosco Jares, Mamtaz Uddin, Oya Beyan, and Stefan Decker. Drug-drug interaction prediction based on knowledge graph embeddings and convolutional-lstm network. In *Proceedings of the 10th ACM international conference on bioinformatics, computational biology and health informatics*, pages 113–123, 2019.
- [185] Mostafa Karimi, Di Wu, Zhangyang Wang, and Yang Shen. Explainable deep relational networks for predicting compound–protein affinities and contacts. *Journal of chemical information and modeling*, 61(1):46–66, 2020.

- [186] Kleomenis Katevas, Katrin Hⁱⁱansel, Richard Clegg, Ilias Leontiadis, Hamed Haddadi, and Laurissa Tokarchuk. Finding dory in the crowd: Detecting social interactions using multi-modal mobile sensing. In *Proceedings of the 1st Workshop on Machine Learning on Edge in Sensor Systems*, pages 37–42, 2019.
- [187] Michal Kazmierski, Mattea Welch, Sejin Kim, Chris McIntosh, Princess Margaret Head, Neck Cancer Group, Katrina Rey-McIntyre, Shao Hui Huang, Tirth Patel, Tony Tadic, Michael Milosevic, Fei-Fei Liu, Andrew Hope, Scott Bratman, and Benjamin Haibe-Kains. A machine learning challenge for prognostic modelling in head and neck cancer using multi-modal data, 2021.
- [188] Said Kerrache and Hafida Benhidour. A complex network based graph embedding method for link prediction, 2022.
- [189] Kathleen R Kerwin and Nathaniel D Bastian. Stacked generalizations in imbalanced fraud data sets using resampling methods. *The Journal of Defense Modeling and Simulation*, 18(3):175–192, 2021.
- [190] Refilwe Kgoadi, Chris Engelbrecht, Ian Whittingham, and Andrew Tkachenko. General classification of light curves using extreme boosting, 2019.
- [191] Jang-Hyun Kim, Sangdoo Yun, and Hyun Oh Song. Neural relation graph: A unified framework for identifying label noise and outlier data. *Advances in Neural Information Processing Systems*, 36, 2024.
- [192] Minkyung Kim, Junsik Kim, Jongmin Yu, and Jun Kyun Choi. v. In *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 39–46, 2022.
- [193] B. Ravi Kiran, Dilip Mathew Thomas, and Ranjith Parakkal. An overview of deep learning based methods for unsupervised and semi-supervised anomaly detection in videos. *Journal of Imaging*, 4(2), 2018.
- [194] Ivan Kiskin, Marianne Sinka, Adam D. Cobb, Waqas Rafique, Lawrence Wang, Davide Zilli, Benjamin Gutteridge, Rinita Dam, Theodoros Marinos, Yunpeng Li, Dickson Msaky, Emmanuel Kaindoa, Gerard Killeen, Eva Herreros-Moya, Kathy J. Willis, and Stephen J. Roberts. Humbugdb: A large-scale acoustic mosquito dataset, 2021.
- [195] Ross S. Kleiman, Paul S. Bennett, Peggy L. Peissig, Richard L. Berg, Zhaobin Kuang, Scott J. Hebbing, Michael D. Caldwell, and David Page. High-throughput machine learning from electronic health records, 2019.
- [196] Nikita Klyuchnikov, Alexey Zaytsev, Arseniy Gruzdev, Georgiy Ovchinnikov, Ksenia Antipova, Leyla Ismailova, Ekaterina Muravleva, Evgeny Burnaev, Artyom Semenikhin, Alexey Cherepanov, et al. Data-driven model for the identification of the rock type at a drilling bit. *Journal of Petroleum science and Engineering*, 178:506–516, 2019.
- [197] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanus Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. Wilds: A benchmark of in-the-wild distribution shifts. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5637–5664. PMLR, 18–24 Jul 2021.
- [198] Yunchuan Kong and Tianwei Yu. forgenet: a graph deep neural network model using tree-based ensemble classifiers for feature graph construction. *Bioinformatics*, 36(11):3507–3515, 03 2020.
- [199] Anna-Kathrin Kopetzki, Bertrand Charpentier, Daniel Zⁱⁱugner, Sandhya Giri, and Stephan Gⁱⁱunnemann. Evaluating robustness of predictive uncertainty estimation: Are dirichlet-based models reliable? In *International Conference on Machine Learning*, pages 5707–5718. PMLR, 2021.

- [200] Denis Krompaß, Stephan Baier, and Volker Tresp. Type-constrained representation learning in knowledge graphs. In *The Semantic Web-ISWC 2015: 14th International Semantic Web Conference, Bethlehem, PA, USA, October 11-15, 2015, Proceedings, Part I 14*, pages 640–655. Springer, 2015.
- [201] Sofia Ira Ktena, Alykhan Tejani, Lucas Theis, Pranay Kumar Myana, Deepak Dilipkumar, Ferenc Huszár, Steven Yoo, and Wenzhe Shi. Addressing delayed feedback for continuous training with neural networks in ctr prediction. In *Proceedings of the 13th ACM Conference on Recommender Systems, RecSys '19*, page 187–195, New York, NY, USA, 2019. Association for Computing Machinery.
- [202] Viraj Kulkarni, Manish Gawali, and Amit Kharat. Key technology considerations in developing and deploying machine learning models in clinical radiology practice. *JMIR Med Inform*, 9(9):e28776, Sep 2021.
- [203] Aditya Kunar. Effective and privacy preserving tabular data synthesizing, 2021.
- [204] Trent Kyono, Fiona J. Gilbert, and Mihaela van der Schaar. Mammo: A deep learning solution for facilitating radiologist-machine collaboration in breast cancer diagnosis, 2018.
- [205] Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed Chi. Fairness without demographics through adversarially reweighted learning. *Advances in neural information processing systems*, 33:728–740, 2020.
- [206] Conrad Lee, Bobo Nick, Ulrik Brandes, and Pádraig Cunningham. Link prediction with social vector clocks. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 784–792, 2013.
- [207] Hao-Chih Lee, Matteo Danieletto, Riccardo Miotto, Sarah T Cherng, and Joel T Dudley. Scaling structural learning with no-bears to infer causal transcriptome networks. In *Pacific Symposium on Biocomputing 2020*, pages 391–402. World Scientific, 2019.
- [208] I-Ta Lee, Manish Marwah, and Martin Arlitt. Attention-based self-supervised feature learning for security data, 2020.
- [209] Joongoo Lee and Min-Su Shin. Estimation of photometric redshifts. ii. identification of out-of-distribution data with neural networks. *The Astronomical Journal*, 163(2):98, 2022.
- [210] Jimmy Lei Ba, Kevin Swersky, Sanja Fidler, et al. Predicting deep zero-shot convolutional neural networks using textual descriptions. In *Proceedings of the IEEE international conference on computer vision*, pages 4247–4255, 2015.
- [211] Jussi Leinonen, Ulrich Hamann, and Urs Germann. Seamless lightning nowcasting with recurrent-convolutional deep learning. *Artificial Intelligence for the Earth Systems*, 1(4):e220043, 2022.
- [212] Daniel E Leisman. Rare events in the icu: an emerging challenge in classification and prediction. *Critical care medicine*, 46(3):418–424, 2018.
- [213] Jake Lever. Classification evaluation: It is important to understand both what a classification metric expresses and what it hides. *Nature methods*, 13(8):603–605, 2016.
- [214] Jingyi Jessica Li and Xin Tong. Statistical hypothesis testing versus machine learning binary classification: Distinctions and guidelines. *Patterns*, 1(7):100115, October 2020.
- [215] Longyuan Li, Junchi Yan, Haiyang Wang, and Yaohui Jin. Anomaly detection of time series with smoothness-inducing sequential variational auto-encoder. *IEEE transactions on neural networks and learning systems*, 32(3):1177–1191, 2020.
- [216] Qiuqia Li, PM Ness, Anton Ragni, and Mark JF Gales. Bi-directional lattice recurrent neural networks for confidence estimation. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6755–6759. IEEE, 2019.

- [217] Qiujia Li, Yu Zhang, David Qiu, Yanzhang He, Liangliang Cao, and Philip C Woodland. 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)*, pages 6537–6541. IEEE, 2022.
- [218] Wenyuan Li, Yunlong Wang, Yong Cai, Corey Arnold, Emily Zhao, and Yilian Yuan. Semi-supervised rare disease detection using generative adversarial network, 2018.
- [219] Xiaoxiao Li, Rabah Al-Zaidy, Amy Zhang, Stefan Baral, Le Bao, and C. Lee Giles. Automating document classification with distant supervision to increase the efficiency of systematic reviews, 2020.
- [220] Xiaoya Li, Jiwei Li, Xiaofei Sun, Chun Fan, Tianwei Zhang, Fei Wu, Yuxian Meng, and Jun Zhang. k folden: k -fold ensemble for out-of-distribution detection. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3102–3115, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.
- [221] Yufei Li, Simin Chen, and Wei Yang. Estimating predictive uncertainty under program data distribution shift. *arXiv preprint arXiv:2107.10989*, 2021.
- [222] Ryan Lichtnwalter and Nitesh V Chawla. Link prediction: fair and effective evaluation. In *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 376–383. IEEE, 2012.
- [223] Bryan Lim and Mihaela van der Schaar. Disease-atlas: Navigating disease trajectories using deep learning. In *Machine Learning for Healthcare Conference*, pages 137–160. PMLR, 2018.
- [224] Krzysztof Lis, Sina Honari, Pascal Fua, and Mathieu Salzmann. Detecting road obstacles by erasing them. *IEEE transactions on pattern analysis and machine intelligence*, 2023.
- [225] Bin Liu, Konstantinos Blekas, and Grigorios Tsoumakas. Multi-label sampling based on local label imbalance. *Pattern Recognition*, 122:108294, 2022.
- [226] Bin Liu, Dimitrios Papadopoulos, Fragkiskos D Malliaros, Grigorios Tsoumakas, and Apostolos N Papadopoulos. Multiple similarity drug–target interaction prediction with random walks and matrix factorization. *Briefings in Bioinformatics*, 23(5):bbac353, 2022.
- [227] Bin Liu, Jin Wang, Kaiwei Sun, and Grigorios Tsoumakas. Fine-grained selective similarity integration for drug–target interaction prediction. *Briefings in Bioinformatics*, 24(2):bbad085, 2023.
- [228] Hanyang Liu, Michael Montana, Dingwen Li, Chase Renfro, Thomas Kannampallil, and Chenyang Lu. Predicting intraoperative hypoxemia with hybrid inference sequence autoencoder networks. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 1269–1278, 2022.
- [229] Yang Liu, Dingkan Yang, Yan Wang, Jing Liu, Jun Liu, Azzedine Boukerche, Peng Sun, and Liang Song. Generalized video anomaly event detection: Systematic taxonomy and comparison of deep models, 2023.
- [230] Zhining Liu, Pengfei Wei, Jing Jiang, Wei Cao, Jiang Bian, and Yi Chang. Mesa: boost ensemble imbalanced learning with meta-sampler. *Advances in neural information processing systems*, 33:14463–14474, 2020.
- [231] Artyom Lobanov, Egor Bogomolov, Yaroslav Golubev, Mikhail Mirzayanov, and Timofey Bryksin. Predicting tags for programming tasks by combining textual and source code data, 2023.
- [232] Daniel Lopez-Martinez, Christina Chen, and Ming-Jun Chen. Machine learning for dynamically predicting the onset of renal replacement therapy in chronic kidney disease patients using claims data. In *International Workshop on Applications of Medical AI*, pages 18–28. Springer, 2022.

- [233] Daniel Lopez-Martinez, Alex Yakubovich, Martin Seneviratne, Adam D. Lelkes, Akshit Tyagi, Jonas Kemp, Ethan Steinberg, N. Lance Downing, Ron C. Li, Keith E. Morse, Nigam H. Shah, and Ming-Jun Chen. Instability in clinical risk stratification models using deep learning. In Antonio Parziale, Monica Agrawal, Shalmali Joshi, Irene Y. Chen, Shengpu Tang, Luis Oala, and Adarsh Subbaswamy, editors, *Proceedings of the 2nd Machine Learning for Health symposium*, volume 193 of *Proceedings of Machine Learning Research*, pages 552–565. PMLR, 28 Nov 2022.
- [234] Guan-Rong Lu, Yueh-Cheng Liu, Tung-I Chen, Hung-Ting Su, Tsung-Han Wu, and Winston H Hsu. Anomaly-aware semantic segmentation by leveraging synthetic-unknown data. *arXiv preprint arXiv:2111.14343*, 2021.
- [235] Yvan Lucas, Pierre-Edouard Portier, Léa Laporte, Sylvie Calabretto, Olivier Caelen, Liyun He-Guelton, and Michael Granitzer. Multiple perspectives hmm-based feature engineering for credit card fraud detection. In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing, SAC '19*, page 1359–1361, New York, NY, USA, 2019. Association for Computing Machinery.
- [236] Jeffrey Lund, Piper Armstrong, Wilson Fearn, Stephen Cowley, Emily Hales, and Kevin Seppi. Cross-referencing using fine-grained topic modeling. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *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)*, pages 3978–3987, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.
- [237] Julia Lust and Alexandru Paul Condurache. A survey on assessing the generalization envelope of deep neural networks: Predictive uncertainty, out-of-distribution and adversarial samples, 2021.
- [238] Yingzhe Lyu, Gopi Krishnan Rajbahadur, Dayi Lin, Boyuan Chen, and Zhen Ming (Jack) Jiang. Towards a consistent interpretation of aiops models. *ACM Trans. Softw. Eng. Methodol.*, 31(1), nov 2021.
- [239] Victoria López, Alberto Fernández, Salvador García, Vasile Palade, and Francisco Herrera. An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information Sciences*, 250:113–141, 2013.
- [240] Liantao Ma, Junyi Gao, Yasha Wang, Chaohe Zhang, Jiangtao Wang, Wenjie Ruan, Wen Tang, Xin Gao, and Xinyu Ma. Adacare: Explainable clinical health status representation learning via scale-adaptive feature extraction and recalibration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 825–832, 2020.
- [241] Liantao Ma, Chaohe Zhang, Junyi Gao, Xianfeng Jiao, Zhihao Yu, Yinghao Zhu, Tianlong Wang, Xinyu Ma, Yasha Wang, Wen Tang, et al. Mortality prediction with adaptive feature importance recalibration for peritoneal dialysis patients. *Patterns*, 4(12), 2023.
- [242] Liantao Ma, Chaohe Zhang, Yasha Wang, Wenjie Ruan, Jiangtao Wang, Wen Tang, Xinyu Ma, Xin Gao, and Junyi Gao. Concare: Personalized clinical feature embedding via capturing the healthcare context. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 833–840, 2020.
- [243] Xinyu Ma, Xu Chu, Yasha Wang, Hailong Yu, Liantao Ma, Wen Tang, and Junfeng Zhao. Medfact: Modeling medical feature correlations in patient health representation learning via feature clustering, 2022.
- [244] Yiming Ma, Victor Sanchez, Soodeh Nikan, Devesh Upadhyay, Bhushan Atote, and Tanaya Guha. Real-time driver monitoring systems through modality and view analysis. *arXiv preprint arXiv:2210.09441*, 2022.
- [245] Brielen Madureira and David Schlangen. Instruction clarification requests in multimodal collaborative dialogue games: Tasks, and an analysis of the CoDraw dataset. In Andreas Vlachos and Isabelle Augenstein, editors, *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2303–2319, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics.

- [246] Lena Maier-Hein, Annika Reinke, Patrick Godau, Minu D Tizabi, Florian Buettner, Evangelia Christodoulou, Ben Glocker, Fabian Isensee, Jens Kleesiek, Michal Kozubek, et al. Metrics reloaded: recommendations for image analysis validation. *Nature methods*, pages 1–18, 2024.
- [247] Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.
- [248] Andrey Malinin, Bruno Mlodozienec, and Mark Gales. Ensemble distribution distillation. In *International Conference on Learning Representations*, 2020.
- [249] Sunil Mallya, Marc Overhage, Sravan Bodapati, Navneet Srivastava, and Sahika Genc. Savehr: self attention vector representations for ehr based personalized chronic disease onset prediction and interpretability. *arXiv preprint arXiv:1911.05370*, 2019.
- [250] Saurav Manchanda, Pranjul Yadav, Khoa Doan, and S Sathiya Keerthi. Targeted display advertising: the case of preferential attachment. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 1868–1877. IEEE, 2019.
- [251] Rafael B Mangolin, Rodolfo M Pereira, Alceu S Britto Jr, Carlos N Silla Jr, Valéria D Feltrim, Diego Bertolini, and Yandre MG Costa. A multimodal approach for multi-label movie genre classification. *Multimedia Tools and Applications*, 81(14):19071–19096, 2022.
- [252] Johan Markdahl, Nicolo Colombo, Johan Thunberg, and Jorge Gonçalves. 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)*, pages 423–428. IEEE, 2017.
- [253] Mansour Zoubairou A Mayaki and Michel Riveill. Multiple inputs neural networks for fraud detection. In *2022 International Conference on Machine Learning, Control, and Robotics (MLCR)*, pages 8–13, 2022.
- [254] Samuele Mazzanti. Why you should stop using the roc curve, September 2023.
- [255] John McKay, Isaac Gerg, and Vishal Monga. Bridging the gap: Simultaneous fine tuning for data re-balancing. In *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*, pages 7062–7065. IEEE, 2018.
- [256] Matúš Medo, Daniel M Aebersold, and Michaela Medová. Protrank: bypassing the imputation of missing values in differential expression analysis of proteomic data. *BMC bioinformatics*, 20:1–12, 2019.
- [257] Aryan Mehboodi, Shrawan Singhal, and S. V. Sreenivasan. Squeeze flow of micro-droplets: convolutional neural network with trainable and tunable refinement, 2022.
- [258] Julia A Meister, Khuong An Nguyen, and Zhiyuan Luo. Audio feature ranking for sound-based covid-19 patient detection. In *EPIA Conference on Artificial Intelligence*, pages 146–158. Springer, 2022.
- [259] David Merrell and Anthony Gitter. Inferring signaling pathways with probabilistic programming. *Bioinformatics*, 36(Supplement 2):i822–i830, December 2020.
- [260] Mika A Merrill and Tim Althoff. Self-supervised pretraining and transfer learning enable flu and covid-19 predictions in small mobile sensing datasets. In Bobak J. Mortazavi, Tasmie Sarker, Andrew Beam, and Joyce C. Ho, editors, *Proceedings of the Conference on Health, Inference, and Learning*, volume 209 of *Proceedings of Machine Learning Research*, pages 191–206. PMLR, 22 Jun–24 Jun 2023.
- [261] Jiaju Miao and Wei Zhu. Precision–recall curve (prc) classification trees. *Evolutionary intelligence*, 15(3):1545–1569, 2022.
- [262] Kai Middlebrook, Shyam Sudhakaran, and David Guy Brizan. Muslcat: Multi-scale multi-level convolutional attention transformer for discriminative music modeling on raw waveforms, 2021.

- [263] Dimity Miller, Feras Dayoub, Michael Milford, and Niko Sünderhauf. Evaluating merging strategies for sampling-based uncertainty techniques in object detection. In *2019 international conference on robotics and automation (icra)*, pages 2348–2354. IEEE, 2019.
- [264] Shahryar Minhas, Peter D Hoff, and Michael D Ward. Inferential approaches for network analysis: Amen for latent factor models. *Political Analysis*, 27(2):208–222, 2019.
- [265] Amanda J Minnich, Kevin McLoughlin, Margaret Tse, Jason Deng, Andrew Weber, Neha Murad, Benjamin D Madej, Bharath Ramsundar, Tom Rush, Stacie Calad-Thomson, et al. Ampl: a data-driven modeling pipeline for drug discovery. *Journal of chemical information and modeling*, 60(4):1955–1968, 2020.
- [266] Yisroel Mirsky, Tomer Golomb, and Yuval Elovici. Lightweight collaborative anomaly detection for the iot using blockchain. *Journal of Parallel and Distributed Computing*, 145:75–97, 2020.
- [267] Pratik K Mishra, Alex Mihailidis, and Shehroz S Khan. Skeletal video anomaly detection using deep learning: Survey, challenges, and future directions. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2024.
- [268] Ali Mollahosseini, Behzad Hasani, and Mohammad H. Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, January 2019.
- [269] Lena Mondrejevski, Ioanna Miliou, Annaclaudia Montanino, David Pitts, Jaakko Hollmén, and Panagiotis Papapetrou. Flicu: A federated learning workflow for intensive care unit mortality prediction. In *2022 IEEE 35th International Symposium on Computer-Based Medical Systems (CBMS)*, pages 32–37. IEEE, 2022.
- [270] Aanchal Mongia, Sanjay Kr Saha, Emilie Chouzenoux, and Angshul Majumdar. A computational approach to aid clinicians in selecting anti-viral drugs for covid-19 trials. *Scientific reports*, 11(1):9047, 2021.
- [271] Michael Moor, Max Horn, Bastian Rieck, Damian Roqueiro, and Karsten Borgwardt. Early recognition of sepsis with gaussian process temporal convolutional networks and dynamic time warping. In Finale Doshi-Velez, Jim Fackler, Ken Jung, David Kale, Rajesh Ranganath, Byron Wallace, and Jenna Wiens, editors, *Proceedings of the 4th Machine Learning for Healthcare Conference*, volume 106 of *Proceedings of Machine Learning Research*, pages 2–26. PMLR, 09–10 Aug 2019.
- [272] José Morano, Álvaro S Hervella, N Barreira, Jorge Novo, and José Rouco. Multimodal transfer learning-based approaches for retinal vascular segmentation. In *ECAI 2020*, pages 1866–1873. IOS Press, 2020.
- [273] José Morano, Álvaro S. Hervella, Jorge Novo, and José Rouco. Simultaneous segmentation and classification of the retinal arteries and veins from color fundus images. *Artificial Intelligence in Medicine*, 118:102116, August 2021.
- [274] Candelaria Mosquera, Luciana Ferrer, Diego Milone, Daniel Luna, and Enzo Ferrante. Impact of class imbalance on chest x-ray classifiers: towards better evaluation practices for discrimination and calibration performance, 2022.
- [275] Pablo Mosteiro, Emil Rijcken, Kalliopi Zervanou, Uzay Kaymak, Floortje Scheepers, and Marco Spruit. Machine learning for violence risk assessment using dutch clinical notes. *Journal of Artificial Intelligence for Medical Sciences*, 2(1-2):44–54, 2021.
- [276] Zaynab Mousavian, Sahand Khakabimamaghani, Kaveh Kavousi, and Ali Masoudi-Nejad. Drug–target interaction prediction from pssm based evolutionary information. *Journal of pharmacological and toxicological methods*, 78:42–51, 2016.
- [277] Faezeh Movahedi, Rema Padman, and James F Antaki. Limitations of receiver operating characteristic curve on imbalanced data: assist device mortality risk scores. *The Journal of Thoracic and Cardiovascular Surgery*, 165(4):1433–1442, 2023.

- [278] Christopher W Murphy. Class imbalance techniques for high energy physics. *SciPost Physics*, 7(6):076, 2019.
- [279] Daniel Muthukrishna, Gautham Narayan, Kaisey S Mandel, Rahul Biswas, and Renée Hložek. Rapid: early classification of explosive transients using deep learning. *Publications of the Astronomical Society of the Pacific*, 131(1005):118002, 2019.
- [280] Anubhav Reddy Nallabasannagari, Madhu Reddiboina, Ryan Seltzer, Trevor Zeffiro, Ajay Sharma, and Mahendra Bhandari. All data inclusive, deep learning models to predict critical events in the medical information mart for intensive care iii database (mimic iii), 2020.
- [281] Santhosh Narayanan, Carsten Maple, and Mark Hooper. A point process model for rare event detection, 2022.
- [282] Jose M Navarro, Alexis Huet, and Dario Rossi. Human readable network troubleshooting based on anomaly detection and feature scoring. *Computer Networks*, 219:109447, 2022.
- [283] Bijay Neupane, Torben Bach Pedersen, and Bo Thiesson. Utilizing device-level demand forecasting for flexibility markets-full version. *arXiv preprint arXiv:1805.00702*, 2018.
- [284] Eli Newby, Jorge Gómez Tejeda Zañudo, and Réka Albert. 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.
- [285] Minh-Nghia Nguyen and Ngo Anh Vien. Scalable and interpretable one-class svms with deep learning and random fourier features. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2018, Dublin, Ireland, September 10–14, 2018, Proceedings, Part I 18*, pages 157–172. Springer, 2019.
- [286] Maximilian Nickel, Kevin Murphy, Volker Tresp, and Evgeniy Gabrilovich. A review of relational machine learning for knowledge graphs. *Proceedings of the IEEE*, 104(1):11–33, January 2016.
- [287] Zhale Nowroozilarki, Arash Pakbin, James Royalty, Donald KK Lee, and Bobak J Mortazavi. Real-time mortality prediction using mimic-iv icu data via boosted nonparametric hazards. In *2021 IEEE EMBS international conference on biomedical and health informatics (BHI)*, pages 1–4. IEEE, 2021.
- [288] Antonios Ntroumpogiannis, Michail Giannoulis, Nikolaos Myrtakis, Vassilis Christophides, Eric Simon, and Ioannis Tsamardinos. A meta-level analysis of online anomaly detectors. *The VLDB Journal*, pages 1–42, 2023.
- [289] Chris J Oates, Richard Amos, and Simon EF Spencer. Quantifying the multi-scale performance of network inference algorithms. *Statistical Applications in Genetics and Molecular Biology*, 13(5):611–631, 2014.
- [290] Philipp Oberdiek, Matthias Rottmann, and Hanno Gottschalk. Classification uncertainty of deep neural networks based on gradient information. In *Artificial Neural Networks in Pattern Recognition: 8th IAPR TC3 Workshop, ANNPR 2018, Siena, Italy, September 19–21, 2018, Proceedings 8*, pages 113–125. Springer, 2018.
- [291] María Óskarsdóttir, Waqas Ahmed, Katrien Antonio, Bart Baesens, Rémi Dendievel, Tom Donas, and Tom Reynkens. Social network analytics for supervised fraud detection in insurance. *Risk Analysis*, 42(8):1872–1890, 2022.
- [292] Brice Ozenne, Fabien Subtil, and Delphine Maucourt-Boulch. 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.
- [293] Ozan Ozyegen, Devika Kabe, and Mucahit Cevik. Word-level text highlighting of medical texts for telehealth services. *Artificial Intelligence in Medicine*, 127:102284, 2022.
- [294] José A. Padrón-Hidalgo, Valero Laparra, and Gustau Camps-Valls. Unsupervised anomaly and change detection with multivariate gaussianization. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–10, 2022.

- [295] Aniello Panariello, Angelo Porrello, Simone Calderara, and Rita Cucchiara. Consistency-based self-supervised learning for temporal anomaly localization. In *European Conference on Computer Vision*, pages 338–349. Springer, 2022.
- [296] Guansong Pang, Chunhua Shen, Huidong Jin, and Anton van den Hengel. Deep weakly-supervised anomaly detection. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD ’23, page 1795–1807, New York, NY, USA, 2023. Association for Computing Machinery.
- [297] Guansong Pang, Chunhua Shen, and Anton Van Den Hengel. Deep anomaly detection with deviation networks. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 353–362, 2019.
- [298] Guansong Pang, Anton van den Hengel, Chunhua Shen, and Longbing Cao. Toward deep supervised anomaly detection: Reinforcement learning from partially labeled anomaly data. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, KDD ’21. ACM, August 2021.
- [299] Fragkiskos Papadopoulos and Kaj-Kolja Kleineberg. Link persistence and conditional distances in multiplex networks. *Physical Review E*, 99(1):012322, 2019.
- [300] Min Sue Park, Hyeontae Jo, Haeun Lee, Se Young Jung, and Hyung Ju Hwang. Machine learning-based covid-19 patients triage algorithm using patient-generated health data from nationwide multicenter database. *Infectious Diseases and Therapy*, 11(2):787–805, February 2022.
- [301] Ilya N Pashchenko, Kirill V Sokolovsky, and Panagiotis Gavras. Machine learning search for variable stars. *Monthly Notices of the Royal Astronomical Society*, 475(2):2326–2343, 2018.
- [302] Ali Payani and Faramarz Fekri. Inductive logic programming via differentiable deep neural logic networks. *arXiv preprint arXiv:1906.03523*, 2019.
- [303] Xueping Peng, Guodong Long, Tao Shen, Sen Wang, and Jing Jiang. Self-attention enhanced patient journey understanding in healthcare system. In Frank Hutter, Kristian Kersting, Jeffrey Lijffijt, and Isabel Valera, editors, *Machine Learning and Knowledge Discovery in Databases*, pages 719–735, Cham, 2021. Springer International Publishing.
- [304] Xueping Peng, Guodong Long, Tao Shen, Sen Wang, Jing Jiang, and Michael Blumenstein. Temporal self-attention network for medical concept embedding. In *2019 IEEE international conference on data mining (ICDM)*, pages 498–507. IEEE, 2019.
- [305] Xueping Peng, Guodong Long, Tao Shen, Sen Wang, Jing Jiang, and Chengqi Zhang. Bitenet: bidirectional temporal encoder network to predict medical outcomes. In *2020 IEEE International Conference on Data Mining (ICDM)*, pages 412–421. IEEE, 2020.
- [306] Eike Petersen, Melanie Ganz, Sune Holm, and Aasa Feragen. On (assessing) the fairness of risk score models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT ’23, page 817–829, New York, NY, USA, 2023. Association for Computing Machinery.
- [307] Stephen Pfohl, Yizhe Xu, Agata Foryciarz, Nikolaos Ignatiadis, Julian Genkins, and Nigam Shah. Net benefit, calibration, threshold selection, and training objectives for algorithmic fairness in healthcare. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1039–1052, 2022.
- [308] Dario Piermarini, Antonio M. Sudoso, and Veronica Piccialli. Predicting municipalities in financial distress: a machine learning approach enhanced by domain expertise, 2023.
- [309] Walter Hugo Lopez Pinaya, Petru-Daniel Tudosiu, Robert Gray, Geraint Rees, Parashkev Nachev, Sébastien Ourselin, and M. Jorge Cardoso. Unsupervised brain anomaly detection and segmentation with transformers. In Mattias Heinrich, Qi Dou, Marleen de Bruijne, Jan Lellmann, Alexander Schläfer, and Floris Ernst, editors, *Proceedings of the Fourth Conference on Medical Imaging with Deep Learning*, volume 143 of *Proceedings of Machine Learning Research*, pages 596–617. PMLR, 07–09 Jul 2021.

- [310] Adrian Alan Pol, Thea Aarrestad, Katya Govorkova, Roi Halily, Tal Kopetz, Anat Klempner, Vladimir Loncar, Jennifer Ngadiuba, Maurizio Pierini, Olya Sirkin, et al. Jet single shot detection. In *EPJ Web of Conferences*, volume 251, page 04027. EDP Sciences, 2021.
- [311] Jordi Pons, Oriol Nieto, Matthew Prockup, Erik Schmidt, Andreas Ehmann, and Xavier Serra. End-to-end learning for music audio tagging at scale. In *Proceedings of the 19th International Society for Music Information Retrieval Conference*, 2018.
- [312] Ioannis Prapas, Akanksha Ahuja, Spyros Kondylatos, Ilektra Karasante, Eleanna Panagiotou, Lazaro Alonso, Charalampos Davalas, Dimitrios Michail, Nuno Carvalhais, and Ioannis Papoutsis. Deep learning for global wildfire forecasting, 2023.
- [313] Sam Preston, Mu Wei, Rajesh Rao, Robert Tinn, Naoto Usuyama, Michael Lucas, Roshanthi Weerasinghe, Soohee Lee, Brian Piening, Paul Tittel, Naveen Valluri, Tristan Naumann, Carlo Bifulco, and Hoifung Poon. Towards structuring real-world data at scale: Deep learning for extracting key oncology information from clinical text with patient-level supervision, 2022.
- [314] Qi Qi, Youzhi Luo, Zhao Xu, Shuiwang Ji, and Tianbao Yang. Stochastic optimization of areas under precision-recall curves with provable convergence. *Advances in neural information processing systems*, 34:1752–1765, 2021.
- [315] Gwenolé Quéllec, Hassan Al Hajj, Mathieu Lamard, Pierre-Henri Conze, Pascale Massin, and Béatrice Cochener. Explain: Explanatory artificial intelligence for diabetic retinopathy diagnosis. *Medical Image Analysis*, 72:102118, August 2021.
- [316] Mahmudur Rahman, Tanay Kumar Saha, Mohammad Al Hasan, Kevin S. Xu, and Chandan K. Reddy. Dylink2vec: Effective feature representation for link prediction in dynamic networks, 2018.
- [317] Farzaneh Rajabi and Jack Siyuan He. Click-through rate prediction using graph neural networks and online learning, 2021.
- [318] Korbinian Randl, Núria Lladós Armengol, Lena Mondrejevski, and Ioanna Miliou. Early prediction of the risk of icu mortality with deep federated learning. In *2023 IEEE 36th International Symposium on Computer-Based Medical Systems (CBMS)*, pages 706–711. IEEE, 2023.
- [319] Susie Xi Rao, Clémence Lanfranchi, Shuai Zhang, Zhichao Han, Zitao Zhang, Wei Min, Mo Cheng, Yinan Shan, Yang Zhao, and Ce Zhang. Modelling graph dynamics in fraud detection with Attention, 2022.
- [320] Gopikrishna Rathinavel, Nikhil Muralidhar, Timothy O’Shea, and Naren Ramakrishnan. Detecting irregular network activity with adversarial learning and expert feedback. In *2022 IEEE International Conference on Data Mining (ICDM)*, pages 1161–1166. IEEE, 2022.
- [321] Farshid Rayhan, Sajid Ahmed, Zaynab Mousavian, Dewan Md Farid, and Swakkhar Shatabda. Frnet-dti: Deep convolutional neural network for drug-target interaction prediction. *Heliyon*, 6(3), 2020.
- [322] Farshid Rayhan, Sajid Ahmed, Swakkhar Shatabda, Dewan Md Farid, Zaynab Mousavian, Abdollah Dehzangi, and M Sohel Rahman. idti-esboost: identification of drug target interaction using evolutionary and structural features with boosting. *Scientific reports*, 7(1):17731, 2017.
- [323] Haroldas Razvadauskas, Evaldas Vaiciukynas, Kazimieras Buskus, Lukas Drukteinis, Lukas Arlauskas, Saulius Sadauskas, and Albinas Naudziunas. Exploring traditional machine learning for identification of pathological auscultations, 2022.
- [324] Ismat Ara Reshma, Sylvain Cussat-Blanc, Radu Tudor Ionescu, Hervé Luga, and Josiane Mothe. Natural vs balanced distribution in deep learning on whole slide images for cancer detection. In *Proceedings of the 36th Annual ACM Symposium on Applied Computing*, pages 18–25, 2021.

- [325] Roonak Rezvani, Samaneh Kouchaki, Ramin Nilforooshan, David J. Sharp, and Payam Barnaghi. Semi-supervised learning for identifying the likelihood of agitation in people with dementia, 2021.
- [326] Georgios Rizos, Jenna Lawson, Simon Mitchell, Pranay Shah, Xin Wen, Cristina Banks-Leite, Robert Ewers, and Björn W. Schuller. Propagating variational model uncertainty for bioacoustic call label smoothing. *Patterns*, 5(3):100932, 2024.
- [327] Narjes Rohani and Changiz Eslahchi. Drug-drug interaction predicting by neural network using integrated similarity. *Scientific reports*, 9(1):13645, 2019.
- [328] Michał Romaszewski, Przemysław Głomb, Arkadiusz Sochan, and Michał Cholewa. A dataset for evaluating blood detection in hyperspectral images. *Forensic science international*, 320:110701, 2021.
- [329] Miguel Romero, Oscar Ramírez, Jorge Finke, and Camilo Rocha. Feature extraction with spectral clustering for gene function prediction using hierarchical multi-label classification. *Applied Network Science*, 7(1):28, 2022.
- [330] Caitlin Rose, Jeyhan S Kartaltepe, Gregory F Snyder, Vicente Rodriguez-Gomez, LY Aaron Yung, Pablo Arrabal Haro, Micaela B Bagley, Antonello Calabró, Nikko J Cleri, MC Cooper, et al. Identifying galaxy mergers in simulated ceers nircam images using random forests. *The Astrophysical Journal*, 942(1):54, 2023.
- [331] Daniel Rosenberg. Imbalanced data? stop using roc-auc and use auprc instead, June 2022.
- [332] Sankardas Roy, Jordan DeLoach, Yuping Li, Nic Herndon, Doina Caragea, Xinming Ou, Venkatesh Prasad Ranganath, Hongmin Li, and Nicolais Guevara. Experimental study with real-world data for android app security analysis using machine learning. In *Proceedings of the 31st Annual Computer Security Applications Conference*, pages 81–90, 2015.
- [333] Timothy N Rubin, America Chambers, Padhraic Smyth, and Mark Steyvers. Statistical topic models for multi-label document classification. *Machine learning*, 88:157–208, 2012.
- [334] Clemente Rubio-Manzano, Tomás Lermada, Claudia Martínez-Araneda, Christian Vidal, and Alejandra Segura. Teach me to play, gamer! imitative learning in computer games via linguistic description of complex phenomena and decision trees. *Soft Computing*, 27(6):3023–3035, 2023.
- [335] Lukas Ruff, Jacob R. Kauffmann, Robert A. Vandermeulen, Grégoire Montavon, Wojciech Samek, Marius Kloft, Thomas G. Dietterich, and Klaus-Robert Müller. A unifying review of deep and shallow anomaly detection. *Proceedings of the IEEE*, 109(5):756–795, 2021.
- [336] Victor Saase, Holger Wenz, Thomas Ganslandt, Christoph Groden, and Máté E. Maros. Simple statistical methods for unsupervised brain anomaly detection on mri are competitive to deep learning methods, 2020.
- [337] Shiori Sagawa, Pang Wei Koh, Tony Lee, Irena Gao, Sang Michael Xie, Kendrick Shen, Ananya Kumar, Weihua Hu, Michihiro Yasunaga, Henrik Marklund, et al. Extending the wilds benchmark for unsupervised adaptation. *arXiv preprint arXiv:2112.05090*, 2021.
- [338] Berkman Sahiner, Weijie Chen, Aria Pezeshk, and Nicholas Petrick. 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 Matthew A. Kupinski and Robert M. Nishikawa, editors, *Medical Imaging 2017: Image Perception, Observer Performance, and Technology Assessment*, volume 10136, page 101360G. International Society for Optics and Photonics, SPIE, 2017.
- [339] Takaya Saito and Marc Rehmsmeier. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, 10(3):e0118432, 2015.
- [340] Takaya Saito and Marc Rehmsmeier. Precrec: fast and accurate precision–recall and ROC curve calculations in R. *Bioinformatics*, 33(1):145–147, 09 2016.

- [341] Amir Sarabadani, Aaron Halfaker, and Dario Taraborelli. Building automated vandalism detection tools for wikidata. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 1647–1654, 2017.
- [342] Mehrzad Saremi and Maryam Amirmazlaghani. Reconstruction of gene regulatory networks using multiple datasets. *IEEE/ACM transactions on computational biology and bioinformatics*, 19(3):1827–1839, 2021.
- [343] Hamed Sarvari, Carlotta Domeniconi, Bardh Prenkaj, and Giovanni Stilo. Unsupervised boosting-based autoencoder ensembles for outlier detection. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 91–103. Springer, 2021.
- [344] Hamed Sarvari, Carlotta Domeniconi, and Giovanni Stilo. Graph-based selective outlier ensembles. In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, pages 518–525, 2019.
- [345] Thomas Schlegl, Hrvoje Bogunovic, Sophie Klimescha, Philipp Seeböck, Amir Sadeghipour, Bianca Gerendas, Sebastian M. Waldstein, Georg Langs, and Ursula Schmidt-Erfurth. Fully automated segmentation of hyperreflective foci in optical coherence tomography images, 2018.
- [346] Thomas Schlegl, Heiko Stino, Michael Niederleithner, Andreas Pollreisz, Ursula Schmidt-Erfurth, Wolfgang Drexler, Rainer A. Leitgeb, and Tilman Schmolz. Data-centric ai approach to improve optic nerve head segmentation and localization in oct en face images, 2022.
- [347] Oliver Schulte, Zhensong Qian, Arthur E Kirkpatrick, Xiaoqian Yin, and Yan Sun. Fast learning of relational dependency networks. *Machine Learning*, 103:377–406, 2016.
- [348] Kyriakos Schwarz, Ahmed Allam, Nicolas Andres Perez Gonzalez, and Michael Krauthammer. Attentionddi: Siamese attention-based deep learning method for drug–drug interaction predictions. *BMC bioinformatics*, 22(1):1–19, 2021.
- [349] Kyriakos Schwarz, Alicia Pliego-Mendieta, Amina Mollaysa, Lara Planas-Paz, Chantal Pauli, Ahmed Allam, and Michael Krauthammer. Ddos: A graph neural network based drug synergy prediction algorithm, 2024.
- [350] Eugene Seo, Rebecca A Hutchinson, Xiao Fu, Chelsea Li, Tyler A Hallman, John Kilbride, and W Douglas Robinson. Stateconet: Statistical ecology neural networks for species distribution modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 513–521, 2021.
- [351] Hrishikesh Sharma, Prakhar Pradhan, and Balamuralidhar P. Scnet: a generalized attention-based model for crack fault segmentation. In *Proceedings of the twelfth indian conference on computer vision, graphics and image processing*, pages 1–9, 2021.
- [352] Mradul Sharma, Jitadeepa Nayak, Maharaj Krishna Koul, Smarajit Bose, and Abhas Mitra. Gamma/hadron segregation for a ground based imaging atmospheric cherenkov telescope using machine learning methods: Random forest leads. *Research in Astronomy and Astrophysics*, 14(11):1491, 2014.
- [353] Hongda Shen and Eren Kursun. Label augmentation via time-based knowledge distillation for financial anomaly detection, 2021.
- [354] Xiupeng Shi, Yiik Diew Wong, Michael Zhi-Feng Li, Chandrasekar Palanisamy, and Chen Chai. A feature learning approach based on xgboost for driving assessment and risk prediction. *Accident Analysis & Prevention*, 129:170–179, 2019.
- [355] Yishai Shimoni, Ehud Karavani, Sivan Ravid, Peter Bak, Tan Hung Ng, Sharon Hensley Alford, Denise Meade, and Yaara Goldschmidt. An evaluation toolkit to guide model selection and cohort definition in causal inference, 2019.
- [356] Samuel Showalter and Zhixin Wu. Minimizing the societal cost of credit card fraud with limited and imbalanced data, 2019.

- [357] Hai Shu and Hongtu Zhu. Sensitivity analysis of deep neural networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):4943–4950, July 2019.
- [358] Satya Narayan Shukla and Benjamin Marlin. Interpolation-prediction networks for irregularly sampled time series. In *International Conference on Learning Representations*, 2019.
- [359] Satya Narayan Shukla and Benjamin M. Marlin. Modeling irregularly sampled clinical time series, 2018.
- [360] Yaniv Shulman. Unsupervised contextual anomaly detection using joint deep variational generative models, 2019.
- [361] Yuqi Si and Kirk Roberts. Three-level hierarchical transformer networks for long-sequence and multiple clinical documents classification, 2021.
- [362] Amit Kumar Sikder, Hidayet Aksu, and A Selcuk Uluagac. A context-aware framework for detecting sensor-based threats on smart devices. *IEEE Transactions on Mobile Computing*, 19(2):245–261, 2019.
- [363] Marília Costa Rosendo Silva, Felipe Alves Siqueira, João Pedro Mantovani Tarrega, João Victor Pataca Beinotti, Augusto Sousa Nunes, Miguel de Mattos Gardini, Vinícius Adolfo Pereira da Silva, Nádia Félix Felipe da Silva, and André Carlos Ponce de Leon Ferreira de Carvalho. No pattern, no recognition: a survey about reproducibility and distortion issues of text clustering and topic modeling, 2022.
- [364] Edgar Simo-Serra, Eduard Trulls, Luis Ferraz, Iasonas Kokkinos, and Francesc Moreno-Noguer. Fracking deep convolutional image descriptors, 2015.
- [365] Aditya Kumar Singh and B. Uma Shankar. Multi-label classification on remote-sensing images, 2022.
- [366] Harvineet Singh, Rina Singh, Vishwali Mhasawade, and Rumi Chunara. Fairness violations and mitigation under covariate shift. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 3–13, 2021.
- [367] Joakim Skarding, Bogdan Gabrys, and Katarzyna Musial. Foundations and modeling of dynamic networks using dynamic graph neural networks: A survey. *IEEE Access*, 9:79143–79168, 2021.
- [368] Johannes De Smedt and Jochen De Weerd. Predictive process model monitoring using recurrent neural networks, 2023.
- [369] Anna L Smith, Tian Zheng, and Andrew Gelman. Prediction scoring of data-driven discoveries for reproducible research. *Statistics and Computing*, 33(1):11, 2023.
- [370] Artur Sokolovsky, Luca Arnaboldi, Jaume Bacardit, and Thomas Gross. Volume-centred range bars: Novel interpretable representation of financial markets designed for machine learning applications, 2022.
- [371] Roghayeh Soleymani, Eric Granger, and Giorgio Fumera. Progressive boosting for class imbalance and its application to face re-identification. *Expert Systems with Applications*, 101:271–291, 2018.
- [372] Janne Spijkervet and John Ashley Burgoyne. Contrastive learning of musical representations. In *Proceedings of the 22nd International Society for Music Information Retrieval Conference*, pages 673–681. ISMIR, October 2021.
- [373] Saurabh Srivastava, Vinay P. Namboodiri, and T. V. Prabhakar. Putworkbench: Analysing privacy in ai-intensive systems, 2019.
- [374] Benjamin Stadnick, Jan Witowski, Vishwaesh Rajiv, Jakub Chłędowski, Farah E. Shamout, Kyunghyun Cho, and Krzysztof J. Geras. Meta-repository of screening mammography classifiers, 2022.

- [375] Christoph Stanik, Marlo Haering, and Walid Maalej. Classifying multilingual user feedback using traditional machine learning and deep learning. In *2019 IEEE 27th international requirements engineering conference workshops (REW)*, pages 220–226. IEEE, 2019.
- [376] Georg Steinbuss and Klemens Böhm. Benchmarking unsupervised outlier detection with realistic synthetic data. *ACM Trans. Knowl. Discov. Data*, 15(4), apr 2021.
- [377] Oliver L. Stephenson, Tobias Köhne, Eric Zhan, Brent E. Cahill, Sang-Ho Yun, Zachary E. Ross, and Mark Simons. Deep learning-based damage mapping with insar coherence time series. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–17, 2022.
- [378] Callum L. Stewart, Amos Folarin, and Richard Dobson. Personalized acute stress classification from physiological signals with neural processes, 2020.
- [379] Andrew Stolman, Caleb Levy, C Seshadhri, and Aneesh Sharma. 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)*, pages 388–396. SIAM, 2022.
- [380] Nils Strodthoff, Baris Goktepe, Thomas Schierl, Cornelius Hellge, and Wojciech Samek. Enhanced machine learning techniques for early harq feedback prediction in 5g. *IEEE Journal on Selected Areas in Communications*, 37(11):2573–2587, November 2019.
- [381] Olly Styles, Tanaya Guha, and Victor Sanchez. Multi-camera trajectory forecasting with trajectory tensors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):8482–8491, 2021.
- [382] Sujanya Suresh, Savitha Ramasamy, P. N. Suganthan, and Cheryl Sze Yin Wong. Incremental knowledge tracing from multiple schools, 2022.
- [383] Lionel Tabourier, Daniel F Bernardes, Anne-Sophie Libert, and Renaud Lambiotte. Rankmerging: a supervised learning-to-rank framework to predict links in large social networks. *Machine Learning*, 108:1729–1756, 2019.
- [384] Rajesh Talluri and Sanjay Shete. Using the weighted area under the net benefit curve for decision curve analysis. *BMC medical informatics and decision making*, 16:1–9, 2016.
- [385] Fei Tan, Zhi Wei, Jun He, Xiang Wu, Bo Peng, Haoran Liu, and Zhenyu Yan. A blended deep learning approach for predicting user intended actions. In *2018 IEEE international conference on data mining (ICDM)*, pages 487–496. IEEE, 2018.
- [386] Federico Tavella, Alberto Giarretta, Mauro Conti, and Sasitharan Balasubramaniam. A machine learning-based approach to detect threats in bio-cyber dna storage systems. *Computer Communications*, 187:59–70, 2022.
- [387] Vajira Thambawita, Debesh Jha, Hugo Lewi Hammer, Håvard D Johansen, Dag Johansen, Pål Halvorsen, and Michael A Riegler. 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.
- [388] Santosh Thoduka, Juergen Gall, and Paul G. Ploger. Using visual anomaly detection for task execution monitoring. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, September 2021.
- [389] Maud Thomas and Holger Rootzén. Real-time prediction of severe influenza epidemics using extreme value statistics. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 71(2):376–394, 2022.
- [390] Aleksei Tiulpin, Stefan Klein, Sita MA Bierma-Zeinstra, Jérôme Thevenot, Esa Rahtu, Joyce van Meurs, Edwin HG Oei, and Simo Saarakkala. Multimodal machine learning-based knee osteoarthritis progression prediction from plain radiographs and clinical data. *Scientific reports*, 9(1):20038, 2019.

- [391] Amir sina Torfi, Edward A Fox, and Chandan K Reddy. Differentially private synthetic medical data generation using convolutional gans. *Information Sciences*, 586:485–500, 2022.
- [392] Meredith V. Trotter, Cuong Q. Nguyen, Stephen Young, Rob T. Woodruff, and Kim M. Branson. Epigenomic language models powered by cerebras, 2021.
- [393] Hasan Md Tusfiqur, Duy M. H. Nguyen, Mai T. N. Truong, Triet A. Nguyen, Binh T. Nguyen, Michael Barz, Hans-Juergen Profitlich, Ngoc T. T. Than, Ngan Le, Pengtao Xie, and Daniel Sonntag. Drg-net: Interactive joint learning of multi-lesion segmentation and classification for diabetic retinopathy grading, 2022.
- [394] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: A good closed-set classifier is all you need. In *International Conference on Learning Representations*, 2022.
- [395] Celine Vens, Jan Struyf, Leander Schietgat, Sašo Džeroski, and Hendrik Blockeel. Decision trees for hierarchical multi-label classification. *Machine learning*, 73:185–214, 2008.
- [396] Pim Verschuuren, Serena Palazzo, Tom Powell, Steve Sutton, Alfred Pilgrim, and Michele Faucci Giannelli. Supervised machine learning techniques for data matching based on similarity metrics. *arXiv preprint arXiv:2007.04001*, 2020.
- [397] V Vijayan, D Critchlow, and T Milenković. Alignment of dynamic networks. *Bioinformatics*, 33(14):i180–i189, 07 2017.
- [398] Gaurav Vishwakarma, Aditya Sonpal, and Johannes Hachmann. Metrics for benchmarking and uncertainty quantification: Quality, applicability, and best practices for machine learning in chemistry. *Trends in Chemistry*, 3(2):146–156, 2021.
- [399] Sophia J. Wagner, Daniel Reisenbüchler, Nicholas P. West, Jan Moritz Niehues, Jiefu Zhu, Sebastian Foersch, Gregory Patrick Veldhuizen, Philip Quirke, Heike I. Grabsch, Piet A. van den Brandt, Gordon G. A. Hutchins, Susan D. Richman, Tanwei Yuan, Rupert Langer, Josien C. A. Jenniskens, Kelly Offermans, Wolfram Mueller, Richard Gray, Stephen B. Gruber, Joel K. Greenson, Gad Rennert, Joseph D. Bonner, Daniel Schmolze, Jitendra Jonnagaddala, Nicholas J. Hawkins, Robyn L. Ward, Dion Morton, Matthew Seymour, Laura Magill, Marta Nowak, Jennifer Hay, Viktor H. Koelzer, David N. Church, David Church, Enric Domingo, Joanne Edwards, Bengt Glimelius, Ismail Gogenur, Andrea Harkin, Jen Hay, Timothy Iveson, Emma Jaeger, Caroline Kelly, Rachel Kerr, Noori Maka, Hannah Morgan, Karin Oien, Clare Orange, Claire Palles, Campbell Roxburgh, Owen Sansom, Mark Saunders, Ian Tomlinson, Christian Matek, Carol Geppert, Chaolong Peng, Cheng Zhi, Xiaoming Ouyang, Jacqueline A. James, Maurice B. Loughrey, Manuel Salto-Tellez, Hermann Brenner, Michael Hoffmeister, Daniel Truhn, Julia A. Schnabel, Melanie Boxberg, Tingying Peng, and Jakob Nikolas Kather. Transformer-based biomarker prediction from colorectal cancer histology: A large-scale multicentric study, September 2023. Publisher: Elsevier.
- [400] Guanghui Wang, Ming Yang, Lijun Zhang, and Tianbao Yang. Momentum accelerates the convergence of stochastic auprc maximization. In *International Conference on Artificial Intelligence and Statistics*, pages 3753–3771. PMLR, 2022.
- [401] Ruoyu Wang, Syed Ali Khurram, Amina Asif, Lawrence Young, and Nasir Rajpoot. Rank the triplets: A ranking-based multiple instance learning framework for detecting hpv infection in head and neck cancers using routine h&e images, 2022.
- [402] Shirley Wang, Matthew BA McDermott, Geeticka Chauhan, Marzyeh Ghassemi, Michael C Hughes, and Tristan Naumann. Mimic-extract: A data extraction, preprocessing, and representation pipeline for mimic-iii. In *Proceedings of the ACM conference on health, inference, and learning*, pages 222–235, 2020.
- [403] Siyue Wang, Giles R. S. Atkinson, and Wayne B. Hayes. Sana: cross-species prediction of gene ontology go annotations via topological network alignment. *npj Systems Biology and Applications*, 8(1), July 2022.

- [404] Wei-Chen Wang and Jonas Mueller. Detecting label errors in token classification data. *arXiv preprint arXiv:2210.03920*, 2022.
- [405] Xuhong Wang, Ying Du, Shijie Lin, Ping Cui, Yuntian Shen, and Yupu Yang. advae: A self-adversarial variational autoencoder with gaussian anomaly prior knowledge for anomaly detection. *Knowledge-Based Systems*, 190:105187, 2020.
- [406] Yaqing Wang, Yifan Ethan Xu, Xian Li, Xin Luna Dong, and Jing Gao. Automatic validation of textual attribute values in e-commerce catalog by learning with limited labeled data. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2533–2541, 2020.
- [407] Michael Weiss and Paolo Tonella. Fail-safe execution of deep learning based systems through uncertainty monitoring. In *2021 14th IEEE conference on software testing, verification and validation (ICST)*, pages 24–35. IEEE, 2021.
- [408] Michael Weiss and Paolo Tonella. Uncertainty quantification for deep neural networks: An empirical comparison and usage guidelines. *Software Testing, Verification and Reliability*, 33(6):e1840, 2023.
- [409] Peisong Wen, Qianqian Xu, Zhiyong Yang, Yuan He, and Qingming Huang. Exploring the algorithm-dependent generalization of auprc optimization with list stability. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 28335–28349. Curran Associates, Inc., 2022.
- [410] Fiorella Wever, T Anderson Keller, Laura Symul, and Victor Garcia. As easy as apc: overcoming missing data and class imbalance in time series with self-supervised learning. *2nd Workshop on Self-Supervised Learning: Theory and Practice of the 35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.*, 2021.
- [411] Julian Wiederer, Julian Schmidt, Ulrich Kressel, Klaus Dietmayer, and Vasileios Belagiannis. A benchmark for unsupervised anomaly detection in multi-agent trajectories. In *2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)*, pages 130–137. IEEE, 2022.
- [412] Sarah Wiegrefe, Edward Choi, Sherry Yan, Jimeng Sun, and Jacob Eisenstein. Clinical concept extraction for document-level coding. In Dina Demner-Fushman, Kevin Bretonnel Cohen, Sophia Ananiadou, and Junichi Tsujii, editors, *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 261–272, Florence, Italy, August 2019. Association for Computational Linguistics.
- [413] Linda F Wightman. Isac national longitudinal bar passage study. Isac research report series. 1998.
- [414] Minz Won, Keunwoo Choi, and Xavier Serra. Semi-supervised music tagging transformer, 2021.
- [415] Minz Won, Sanghyuk Chun, and Xavier Serra. Toward interpretable music tagging with self-attention, 2019.
- [416] Minz Won, Andres Ferraro, Dmitry Bogdanov, and Xavier Serra. Evaluation of cnn-based automatic music tagging models. *arXiv preprint arXiv:2006.00751*, 2020.
- [417] Peng Wu, Jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, and Zhiwei Yang. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In *Proceedings of the European Conference on Computer Vision*, pages 322–339. Springer, 2020.
- [418] Yifan Wu, Min Zeng, Ying Yu, Yaohang Li, and Min Li. A pseudo label-wise attention network for automatic icd coding. *IEEE Journal of Biomedical and Health Informatics*, 26(10):5201–5212, 2022.

- [419] Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9(2):513–530, 2018.
- [420] Chengyuan Xu, Curtis McCully, Boning Dong, D Andrew Howell, and Pradeep Sen. Cosmicconn: A cosmic-ray detection deep-learning framework, data set, and toolkit. *The Astrophysical Journal*, 942(2):73, 2023.
- [421] Nuo Xu, Pinghui Wang, Long Chen, Jing Tao, and Junzhou Zhao. Mr-gnn: multi-resolution and dual graph neural network for predicting structured entity interactions. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence, IJCAI'19*, page 3968–3974. AAAI Press, 2019.
- [422] Weihuang Xu, Guohao Yu, Alina Zare, Brendan Zurweller, Diane L. Rowland, Joel Reyes-Cabrera, Felix B. Fritschi, Roser Matamala, and Thomas E. Juenger. Overcoming small minirhizotron datasets using transfer learning. *Computers and Electronics in Agriculture*, 175:105466, August 2020.
- [423] Yanhua Xu and Dominik Wojtczak. Multi-channel neural networks for predicting influenza a virus hosts and antigenic types. *arXiv preprint arXiv:2206.03823*, 2022.
- [424] Tianbao Yang. Deep auc maximization for medical image classification: Challenges and opportunities, 2021.
- [425] Xinxing Yang, Genke Yang, and Jian Chu. The computational drug repositioning without negative sampling. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 20(2):1506–1517, 2022.
- [426] Xinxing Yang, Genke Yang, and Jian Chu. The neural metric factorization for computational drug repositioning. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 20(1):731–741, 2022.
- [427] Xinxing Yang, Genke Yang, and Jian Chu. Self-supervised learning for label sparsity in computational drug repositioning. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2023.
- [428] Yang Yang, Ryan N Lichtenwalter, and Nitesh V Chawla. Evaluating link prediction methods. *Knowledge and Information Systems*, 45:751–782, 2015.
- [429] Zi-Yi Yang, Zhao-Feng Ye, Yi-Jia Xiao, Chang-Yu Hsieh, and Sheng-Yu Zhang. Spldextratrees: robust machine learning approach for predicting kinase inhibitor resistance. *Briefings in Bioinformatics*, 23(3):bbac050, 2022.
- [430] Shota Yasui, Gota Morishita, Fujita Komei, and Masashi Shibata. A feedback shift correction in predicting conversion rates under delayed feedback. In *Proceedings of The Web Conference 2020*, pages 2740–2746, 2020.
- [431] Chun-Kit Yeung and Dit-Yan Yeung. Incorporating features learned by an enhanced deep knowledge tracing model for stem/non-stem job prediction. *International Journal of Artificial Intelligence in Education*, 29(3):317–341, 2019.
- [432] Xin Yi, Scott J Adams, Robert DE Henderson, and Paul Babyn. Computer-aided assessment of catheters and tubes on radiographs: How good is artificial intelligence for assessment? *Radiology: Artificial Intelligence*, 2(1):e190082, 2020.
- [433] Yilin Yin and Chun-An Chou. Early icu mortality prediction and survival analysis for respiratory failure, 2021.
- [434] Shuhan Yuan and Xintao Wu. Deep learning for insider threat detection: Review, challenges and opportunities. *Computers & Security*, 104:102221, 2021.
- [435] Yan Yuan, Wanhua Su, and Mu Zhu. Threshold-free measures for assessing the performance of medical screening tests. *Frontiers in public health*, 3:57, 2015.

- [436] Mubariz Zaffar, Sourav Garg, Michael Milford, Julian Kooij, David Flynn, Klaus McDonald-Maier, and Shoaib Ehsan. Vpr-bench: An open-source visual place recognition evaluation framework with quantifiable viewpoint and appearance change. *International Journal of Computer Vision*, 129(7):2136–2174, May 2021.
- [437] Roei Zamir, Shai Bagon, David Samocha, Yael Yagil, Ronen Basri, Miri Sklair-Levy, and Meirav Galun. Segmenting microcalcifications in mammograms and its applications. In *Medical Imaging 2021: Image Processing*, volume 11596, pages 788–795. SPIE, 2021.
- [438] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Draem—a discriminatively trained reconstruction embedding for surface anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8330–8339, 2021.
- [439] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Dsr—a dual subspace re-projection network for surface anomaly detection. In *European conference on computer vision*, pages 539–554. Springer, 2022.
- [440] Ido Zehori, Nevo Itzhak, Yuval Shahar, and Mia Dor Schiller. Towards a user privacy-aware mobile gaming app installation prediction model, 2023.
- [441] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340, 2018.
- [442] Chaohe Zhang, Xu Chu, Liantao Ma, Yinghao Zhu, Yasha Wang, Jiangtao Wang, and Junfeng Zhao. M3care: Learning with missing modalities in multimodal healthcare data. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22. ACM, August 2022.
- [443] Chaohe Zhang, Xin Gao, Liantao Ma, Yasha Wang, Jiangtao Wang, and Wen Tang. Grasp: Generic framework for health status representation learning based on incorporating knowledge from similar patients. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(1):715–723, May 2021.
- [444] Dingwen Zhang, Huazhu Fu, Junwei Han, Ali Borji, and Xuelong Li. A review of co-saliency detection technique: Fundamentals, applications, and challenges, 2017.
- [445] Weijia Zhang. Non-i.i.d. multi-instance learning for predicting instance and bag labels using variational auto-encoder, 2021.
- [446] Weijia Zhang, Xuanhui Zhang, hanwen deng, and Min-Ling Zhang. Multi-instance causal representation learning for instance label prediction and out-of-distribution generalization. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 34940–34953. Curran Associates, Inc., 2022.
- [447] Wencan Zhang, Mariella Dimiccoli, and Brian Y Lim. Debiased-cam to mitigate image perturbations with faithful visual explanations of machine learning. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–32, 2022.
- [448] Wenning Zhang, Ryohei Hisano, Takaaki Ohnishi, and Takayuki Mizuno. Nondiagonal 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*, pages 75–86. Springer, 2021.
- [449] Xuan Zhang, Shiyu Li, Xi Li, Ping Huang, Jiulong Shan, and Ting Chen. Destseg: Segmentation guided denoising student-teacher for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3914–3923, 2023.
- [450] Yan Zhang and Howard D. Bondell. Variable Selection via Penalized Credible Regions with Dirichlet–Laplace Global-Local Shrinkage Priors. *Bayesian Analysis*, 13(3):823 – 844, 2018.

- [451] Qian M. Zhou, Zhe Lu, Russell J. Brooke, Melissa M Hudson, and Yan Yuan. Is the new model better? one metric says yes, but the other says no. which metric do i use?, 2020.
- [452] Yi Zhou, Boyang Wang, Lei Huang, Shanshan Cui, and Ling Shao. A benchmark for studying diabetic retinopathy: Segmentation, grading, and transferability. *IEEE Transactions on Medical Imaging*, 40(3):818–828, March 2021.
- [453] Qianqian Zou, Claus Brenner, and Monika Sester. Gaussian process mapping of uncertain building models with gmm as prior. *IEEE Robotics and Automation Letters*, 8(10):6579–6586, 2023.
- [454] Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In *European Conference on Computer Vision*, pages 392–408. Springer, 2022.
- [455] Ali Burak Ünal, Nico Pfeifer, and Mete Akgün. ppaurora: Privacy preserving area under receiver operating characteristic and precision-recall curves, 2023.

A Broader Impact and Ethical Considerations

This research paper challenges the conventional wisdom regarding the superiority of the AUPRC over AUROC in binary classification tasks with class imbalance and has several ethical implications and impacts.

Our analysis reveals that the preference for AUPRC in certain ML applications may not be empirically justified and could inadvertently amplify algorithmic biases. This calls for a re-examination of prevalent metrics within ML, especially in high-stakes domains like healthcare, finance, and criminal justice where biased models can have profound societal repercussions. The tendency of AUPRC to disproportionately favor models with higher prevalence of positive labels could exacerbate existing disparities, underscoring the ethical need for rigorous validation and scrutiny of evaluation metrics.

Additionally, our use of large language models for literature analysis demonstrates a novel approach in scrutinizing and re-evaluating long-standing assumptions in ML. This method could set a precedent for more comprehensive and robust scientific investigations in the field, fostering a culture of empirical rigor and ethical awareness.

The ethical dimension of our work lies in the spotlight it casts on metric selection in ML model evaluation. The potential of metrics like AUPRC to skew model performance favoring certain groups raises pressing concerns about fairness in algorithmic decision-making. This is particularly critical when algorithms influence key decisions affecting individuals and communities.

While we use the COMPAS dataset for recidivism prediction in this work, we recognize the many societal issues with automated predictions of recidivism [93]. We utilize this dataset as it is a commonly used dataset in the fairness literature, but do not advocate for deployment of these models in any way.

Our study contributes to the technical discourse on metric behaviors in ML and serves as a cautionary tale against uncritically embracing established norms. It underscores the imperative for careful metric selection aligned with ethical principles and fairness objectives in ML, highlighting the far-reaching consequences of these choices in shaping societal outcomes and advancing the field of ML.

B Code Availability

All code is available at https://github.com/hzhang0/auc_bias and <https://github.com/Lassehansen/ArxivMLClaimSearch>.

C Notation

Let $\mathcal{X}, \mathcal{Y} = 0, 1$ represent a paired feature and binary classification label space from which i.i.d. samples $(x, y) \in \mathcal{X} \times \mathcal{Y}$ are drawn via the joint distribution over the random variables x, y . Let $f_{\theta} : \mathcal{X} \rightarrow (0, 1)$ be a binary classification model parametrized by $\theta \in \mathbb{R}^d$ for some $d \in \mathbb{N}$ outputting continuous probability scores over this space.

We define random variable $s = f_{\theta}(x)$ to be the distribution of scores output by the model over input samples. Throughout the paper, θ may be omitted if it is clear from context. We will occasionally also use the notation s_+ and s_- to reflect the conditional distributions of model outputs conditioned on the label being 1 or 0, respectively:

$$\begin{aligned} s_+ &= f(x)|y = 1 \\ s_- &= f(x)|y = 0. \end{aligned}$$

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

$$\begin{aligned}
\text{TP}_\theta(\tau) &= \left| \{x_i \in \mathbf{X} \mid p_i^{(\theta)} \geq \tau, y_i = 1\} \right| \\
\text{FN}_\theta(\tau) &= \left| \{x_i \in \mathbf{X} \mid p_i^{(\theta)} < \tau, y_i = 1\} \right| \\
\text{TN}_\theta(\tau) &= \left| \{x_i \in \mathbf{X} \mid p_i^{(\theta)} < \tau, y_i = 0\} \right| \\
\text{FP}_\theta(\tau) &= \left| \{x_i \in \mathbf{X} \mid p_i^{(\theta)} \geq \tau, y_i = 0\} \right| \\
\text{FR}(f, \tau) &= P_p(p > \tau) \\
\text{TPR}_\theta(\tau) &= \frac{\text{TP}_\theta(\tau)}{\text{TP}_\theta(\tau) + \text{FN}_\theta(\tau)} \\
&= P_{x|y=1}(f(x) > \tau) \\
&= P_{s|y=1}(s > \tau) \\
&= P(s_+ > \tau) \\
\text{FPR}_\theta(\tau) &= \frac{\text{FP}_\theta(\tau)}{\text{FP}_\theta(\tau) + \text{TN}_\theta(\tau)} \\
&= P_{x|y=0}(f(x) > \tau) \\
&= P_{s|y=0}(s > \tau) \\
&= P(s_- > \tau) \\
\text{Prec}_\theta(\tau) &= \frac{\text{TP}_\theta(\tau)}{\text{TP}_\theta(\tau) + \text{FP}_\theta(\tau)} \\
&= P_{y|f(x) > \tau}(y = 1) \\
&= P_{y|s > \tau}(y = 1)
\end{aligned}$$

Lastly, recall

$$\begin{aligned}
\text{AUROC}_\theta &= \int_0^1 \text{TPR}_\theta \frac{d\text{FPR}_\theta}{d\tau} d\tau \\
&= \int_0^1 \text{TPR}_\theta d\text{FPR}_\theta \\
&= 1 - \int_0^1 \text{FPR}_\theta d\text{TPR}_\theta \\
\text{AUPRC}_\theta &= \int_0^1 \text{Prec}_\theta \frac{d\text{TPR}_\theta}{d\tau} d\tau \\
&= \int_0^1 \text{Prec}_\theta d\text{TPR}_\theta
\end{aligned}$$

D Proof of Theorem 1

Recall that all notation is defined formally in Appendix C.

Here, we prove Theorem 1, which states

Theorem 1. Let $\mathcal{X}, \mathcal{Y} = 0, 1$ represent a paired feature and binary classification label space from which i.i.d. samples $(x, y) \in \mathcal{X} \times \mathcal{Y}$ are drawn via the joint distribution over the random variables x, y . Let $f : \mathcal{X} \rightarrow (0, 1)$ be a binary classification model outputting continuous probability scores

over this space. Then,

$$\begin{aligned}\text{AUROC}(f) &= 1 - \mathbb{E}_{t \sim f(x)|y=1} [\text{FPR}(f, t)] \\ \text{AUPRC}(f) &= 1 - P_y(y = 0) \mathbb{E}_{t \sim f(x)|y=1} \left[\frac{\text{FPR}(f, t)}{P_x(f(x) > t)} \right]\end{aligned}$$

Proof. Recall that AUROC and AUPRC are as follows:

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

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

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

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

$$\begin{aligned}\text{Prec} &= 1 - P_{y|s>\tau}(y = 0) \\ &= 1 - \underbrace{P_{s|y=0}(s > \tau)}_{\text{FPR}(\tau)} \frac{P_y(y = 0)}{P_s(s > \tau)}\end{aligned}$$

Thus,

$$\begin{aligned}\text{AUROC}(f) &= 1 - \mathbb{E}_{t \sim s_+} [\text{FPR}(f, t)] \\ &= 1 - \mathbb{E}_{t \sim f(x)|y=1} [\text{FPR}(f, t)] \\ \text{AUPRC}(f) &= \mathbb{E}_{t \sim s_+} [\text{Prec}(f, t)] \\ &= 1 - P_y(y = 0) \mathbb{E}_{t \sim s_+} \left[\frac{\text{FPR}(f, t)}{P_{s \sim s}(s > t)} \right] \\ &= 1 - P_y(y = 0) \mathbb{E}_{t \sim f(x)|y=1} \left[\frac{\text{FPR}(f, t)}{P_x(f(x) > t)} \right]\end{aligned}$$

as desired. \square

Synthetic validation of Theorem 1 can also be found in our public code. Note that this formulation of AUPRC reflects earlier, different formulations of AUPRC, such as those found in the AUPRC optimization literature [409].

E Proof of Theorem 2

Here, we prove Theorem 2, which states

Theorem 2. Define f , \mathcal{X} , \mathbf{X} , \mathbf{y} and N as in Definition 2.1. Further, suppose without loss of generality that the dataset \mathbf{X} is ordered such that $f(x_i) < f(x_{i+1})$ for all i . Then, let us define $M = \{i | (x_i, x_{i+1}) \text{ is an incorrectly ranked adjacent pair for model } f\}$. Define f'_i to be a model that is identical to f except that the probabilities assigned to x_i and x_{i+1} are swapped:

$$f'_i : \begin{cases} f(x) & \text{if } x \notin \{x_i, x_{i+1}\} \\ f(x_{i+1}) & \text{if } x = x_i \\ f(x_i) & \text{if } x = x_{i+1}. \end{cases}$$

Then, $\text{AUROC}(f'_i) = \text{AUROC}(f'_j)$ for all $i, j \in M$, and $\text{AUPRC}(f'_i) < \text{AUPRC}(f'_j)$ for all $i, j \in M$ such that $i < j$.

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.

Note that as we use strict inequalities in our definition of the decision rule underlying the FPR here, a pair of scores that are tied but have different labels will not induce a false positive at the corresponding positively labeled sample's threshold, so separating such ties will have no impact on AUROC whatsoever. It would similarly not impact AUPRC as neither the FPR nor the model firing rate will decrease when the negative sample within the tie is perturbed to be strictly below the positive sample. \square

Synthetic empirical validation of Theorem 2 can also be found in our public code.

F Proof of Theorem 3

In this section, we formally prove Theorem 2. We begin by establishing Lemma 1 and 2.

Lemma 1. *Let a model f be perfectly calibrated and yield score distributions for positive and negative samples from probability density functions p_+ and p_- . Then $p_+(t) = \frac{t}{1-t} \frac{P_y(y=0)}{P_y(y=1)} p_-(t)$*

Proof. As this model is calibrated perfectly, we have that

$$\begin{aligned}
p_+(t) &= P_{s|y=1}(s=t) \\
&= \frac{P_{y|s=t}(y=1)p_s(t)}{P_y(y=1)} \\
&= t \frac{P_y(y=1)p_+(t) + P_y(y=0)p_-(t)}{P_y(y=1)} \\
&= tp_+(t) + t \frac{P_y(y=0)}{P_y(y=1)} p_-(t).
\end{aligned}$$

Thus, $p_+(t) = \frac{t}{1-t} \frac{P_y(y=0)}{P_y(y=1)} p_-(t)$ as desired. \square

Lemma 2. Let a model f be perfectly calibrated and yield score distributions for positive and negative samples from probability density functions p_+ and p_- , with overall distribution given by $p(t) = P_y(y=1)p_+(t) + P_y(y=0)p_-(t)$. Then for all $\tau \in (0, 1)$, $\text{FR}(f, \tau) \leq \frac{P_y(y=1)}{\tau}$.

Proof. By definition, we have

$$\begin{aligned}
\text{FR}(f, \tau) &= \int_{\tau}^1 P_y(y=1)p_+(t) + P_y(y=0)p_-(t) dt \\
&= \int_{\tau}^1 P_y(y=1)p_+(t) + P_y(y=1) \frac{1-t}{t} p_+(t) dt \\
&= P_y(y=1) \int_{\tau}^1 \frac{1}{t} p_+(t) dt,
\end{aligned}$$

where step two leverages the fact that f is perfectly calibrated and the result in Lemma 1.

As $t \geq \tau$, $\frac{1}{t} \leq \frac{1}{\tau}$. Then, as $p_+(t) \geq 0$, $\int_{\tau}^1 \frac{1}{t} p_+(t) dt \leq \frac{1}{\tau} \int_{\tau}^1 p_+(t) dt$. Finally, as $\int_0^1 p_+(t) dt = 1$, we see that $\int_{\tau}^1 p_+(t) dt \leq 1$. Therefore,

$$\begin{aligned}
\text{FR}(f, \tau) &= P_y(y=1) \int_{\tau}^1 \frac{1}{t} p_+(t) dt \\
&\leq P_y(y=1) \cdot \frac{1}{\tau} \cdot 1 \\
&= \frac{P_y(y=1)}{\tau}.
\end{aligned}$$

\square

Theorem 3. Let $f, \mathcal{X}, \mathbf{X}, \mathbf{y}, N, M$, and f'_j all be defined as in Theorem 2. Further, suppose that in this setting the domain \mathcal{X} now contains an attribute defining two subgroups, $\mathcal{A} = \{0, 1\}$, such that for any sample (x_i, y_i) , a_i denotes the subgroup to which that sample belongs. Let f be perfectly calibrated for samples in subgroup $a = 0$, such that $P_{y|a,x}(y=1|a=0, f(x) = t) = t$. Then,

$$\lim_{P_{y|a}(y=1|a=0) \rightarrow 0} P \left(a_i = a_{i+1} = 1 \mid i = \arg \max_{j \in M} (\text{AUPRC}(f'_j)) \right) = 1.$$

Proof. Given Theorem 2, the atomic mistake that would, upon correction, result in the largest improvement to AUPRC is the mistake which occurs at maximal score (as this minimizes the firing rate, which is the denominator in the weighting term for AUPRC). Suppose that at threshold τ , the probability that a mistake will occur above score τ in subgroup 1 with N samples drawn is at least $\delta \in (0, 1]$. As the parameters for subgroup 1 are fixed as we vary the prevalence for subgroup 2, τ can be seen as a constant with respect to the limit we are taking.

But, by Lemma 2 and by the fact that f is perfectly calibrated for subgroup 2, we know that the probability that f will output a score for sample 2 regardless of its label that exceeds τ is upper bounded by $\frac{p_y^{(2)}}{\tau}$. In the limit as $p_y^{(2)}$ tends to zero, the probability that any probabilities will be observed at our greater than τ from subgroup 2 likewise tends to zero.

This means that while the probability that we observe a mistake from subgroup 1 stays fixed at at least $\delta > 0$, the probability that we could observe any mistake that involves any sample from subgroup 2 (either a cross-group mistake or a purely subgroup 2 mistake) tends to zero, establishing the claim. \square

G Details for Synthetic Experiments

G.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 G.2), the former condition is strictly weaker than the latter.

The procedure outlined above guarantees that, in expectation, the AUROC of the generated set of scores and labels will be precisely the target AUROC. However, if you apply this procedure independently across different sample subpopulations, this guarantee can only be applied on each subpopulation individually, and not necessarily on the overall population due to the unspecified xAUC term. However, in practice, for the experiments we ran here, that impact neither meaningfully impacts our experiments nor were the joint AUROCs sufficiently different from the target AUROC to warrant a more complex methodology.

G.2 Calibration includes prevalence matching

Let p be a random variable describing the probabilities (not scores) output by the model over the input distribution defined by the data generative function. If a model is calibrated, this means that $P_{y|p}(y = 1|p = q) = q$ — 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:

$$\begin{aligned}\mathbb{E}_p [q] &= \mathbb{E}_p [P_{y|p}(y = 1|p = q)] \\ &= \int_0^1 P_{y|p}(y = 1|p = q)p_p(q) dq \\ &= \int_0^1 P_{y,p}(y = 1, p = q) dq \\ &= P_y(y = 1)\end{aligned}$$

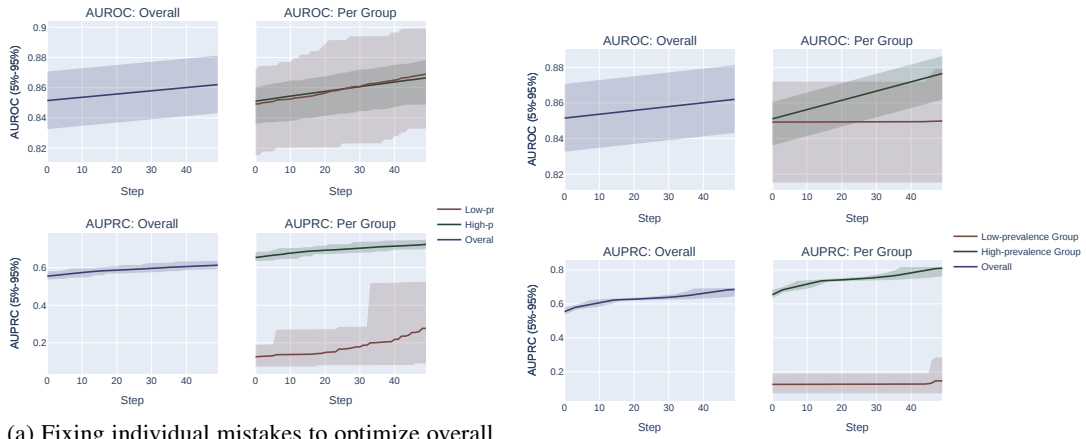
G.3 Details on optimization procedures

- 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 6.

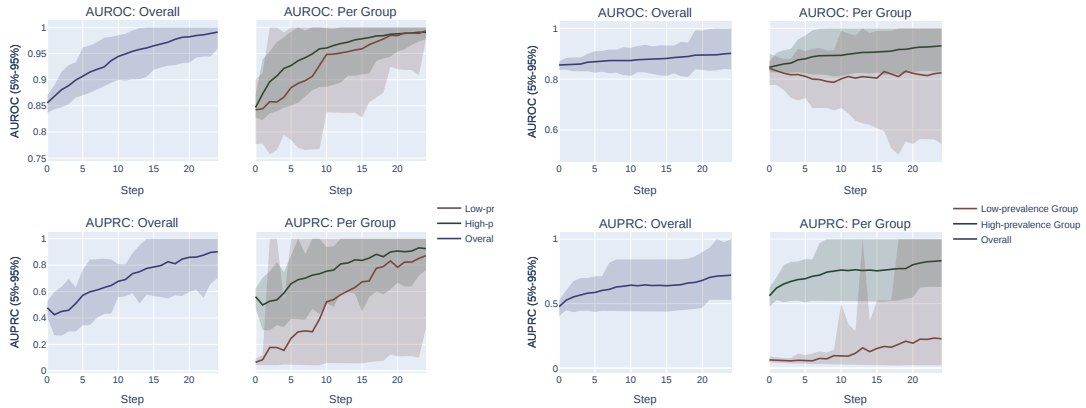
We note that this approach is subtly biased in favor of the lower-prevalence group. In particular, because scores for the low-prevalence group tend to be “squished” into a smaller region of the probability space, a random perturbation of fixed magnitude will proportionally induce more score *permutations* 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.

- 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 4b and 4a.
- 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 $(1, 2, \dots, n)$. 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 $i \in \{1, \dots, n\}$. 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 4d and 4c.



(a) Fixing individual mistakes to optimize overall AUROC

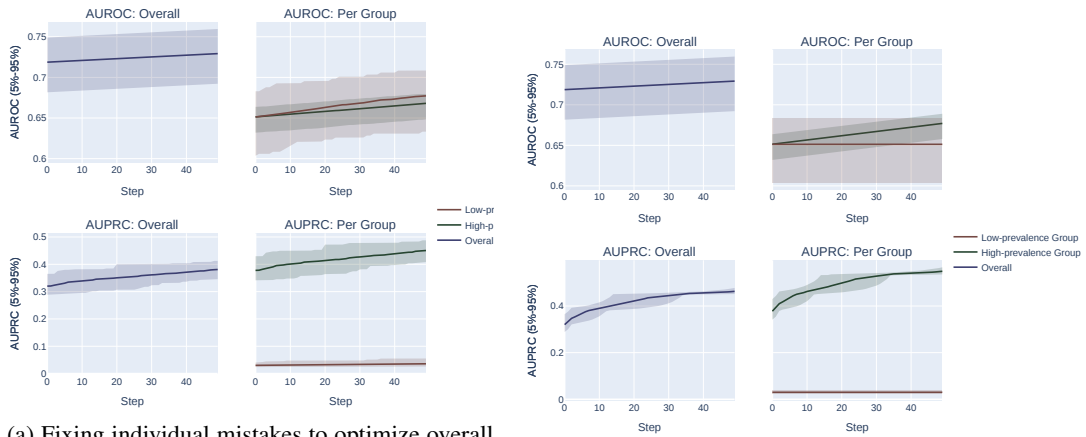
(b) Fixing individual mistakes to optimize overall AUPRC



(c) Randomly permuting scores to optimize overall AUROC

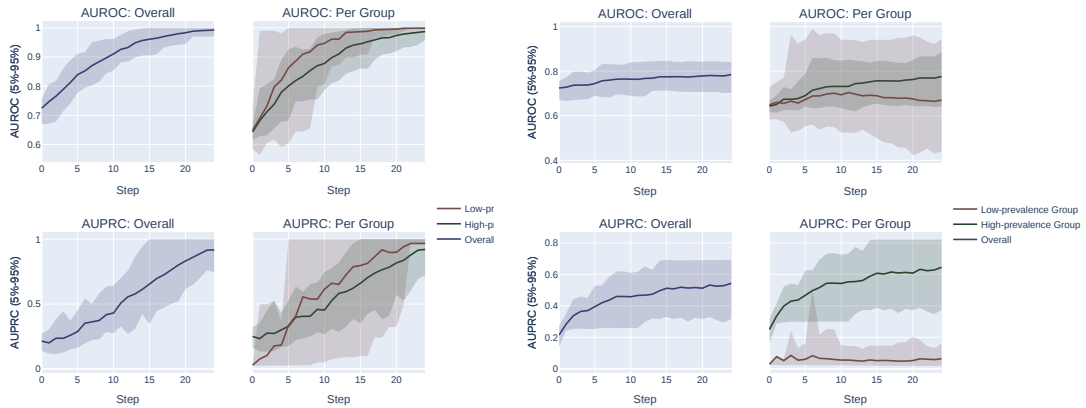
(d) Randomly permuting scores to optimize overall AUPRC

Figure 4: 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 individual mistakes* optimization procedure (M2; *top*) and the *sequentially permuting nearby scores* optimization procedure (M3; *bottom*) described in Section 3.1. Note that the prevalence of Y in the high-prevalence group and the low-prevalence group are 0.05 and 0.01 respectively.



(a) Fixing individual mistakes to optimize overall AUROC

(b) Fixing individual mistakes to optimize overall AUPRC



(c) Randomly permuting scores to optimize overall AUROC

(d) Randomly permuting scores to optimize overall AUPRC

Figure 5: 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 where the initial AUROC was set to 0.65 rather than 0.85, using both the *sequentially fixing individual mistakes* optimization procedure (M2; *top*) and the *sequentially permuting nearby scores* optimization procedure (M3; *bottom*) described in Section 3.1. Note that the prevalence of Y in the high-prevalence group and the low-prevalence group are 0.05 and 0.01 respectively.

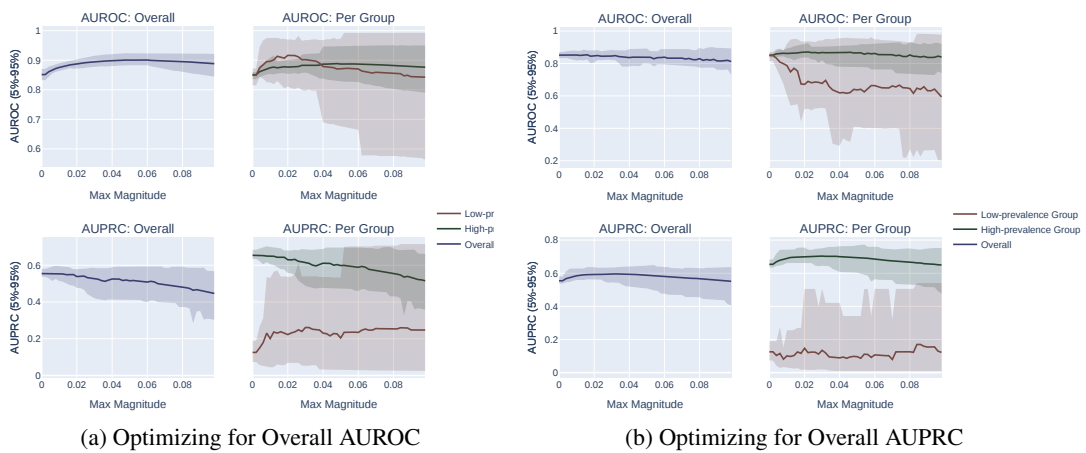


Figure 6: Comparison of the impact of optimizing for overall AUROC and overall AUPRC on the per-group AUROC and AUPRCs of two groups in a synthetic setting, using the *adding random noise* optimization procedure (M1) described in Section 3.1. Note that the prevalence of Y in G_1 and G_2 are 0.05 and 0.01 respectively.

H Additional Details on Real World Experiments

H.1 Dataset Details

We use the following four datasets. In all datasets, we use sex and race as protected attributes.

- `adult` [17]: The UCI Adult dataset, where the goal is to predict whether an individual’s income is $> \$50k$.
- `compas` [14]: The task to predict two-year recidivism. We only select samples belonging to “African-American” and “Caucasian”, leading to a binary race variable.
- `lsac` [413]: The task is to predict whether a law school applicant will pass the bar. We only select samples belonging to White and Black applicants.
- `mimic` [178]: We use the in-hospital mortality task proposed by [138], where the goal is to predict whether a patient will die in the ICU given labs and vitals from the first 48 hours of their hospital stay. We only select samples belonging to White and Black patients.

In each dataset, we balance the groups by subsampling the majority group. We then split each dataset into 50% training, 25% validation, 25% test sets, stratified by the group. Dataset statistics can be found in Table 1.

Table 1: Dataset statistics for the four binary classification datasets used in this study. Note that n refers to the number of samples *after* balancing by the corresponding attribute. Here, “Prevalence (Higher)” refers to the rate at which the prediction label $y = 1$ for the subpopulation with a higher such rate, and “Prevalence (Lower)” refers to the same rate but over the subpopulation of the dataset with a lower rate of $y = 1$.

Dataset	Attribute	n	# Features	Prevalence (Higher)	Prevalence (Lower)
<code>adult</code>	Sex	20,394	12	30.1%	10.7%
<code>adult</code>	Race	6,248	12	24.6%	12.2%
<code>compas</code>	Sex	2,438	6	52.0%	39.4%
<code>compas</code>	Race	4,908	6	55.4%	42.3%
<code>lsac</code>	Sex	15,906	8	96.0%	95.0%
<code>lsac</code>	Race	2,396	8	96.5%	77.2%
<code>mimic</code>	Sex	15,632	49	12.5%	11.9%
<code>mimic</code>	Race	4,030	49	13.9%	9.3%

H.2 Hyperparameter Grid

We use the following hyperparameter grid for our experiments:

- max depth: {1, 2, ..., 9}
- learning rate: [0.01, 0.3]
- number of estimators: [50, 1000]
- min child weight: {1, 2, ..., 9}
- use protected attribute as input feature: {yes, no}
- group weight of higher prevalence group: {1, 2, 3, 4, 5, 10, 15, 20, 25, 50}

H.3 Additional Results

All raw AUC results can be found in the code repository for these experiments, at https://github.com/hzhang0/auc_bias. Additionally, summarized results in different views can be found in Figures 7 and 8.

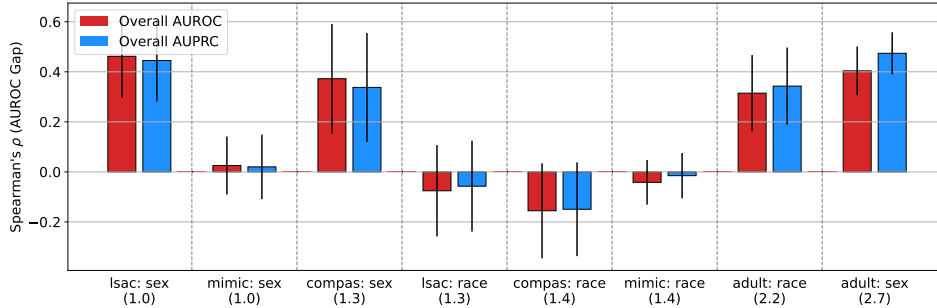


Figure 7: Spearman’s ρ between the test-set signed AUROC gap versus the validation set overall AUPRC, and the AUROC gap versus the overall AUROC. Numbers in parentheses are the prevalence ratios between the two groups for the particular attribute, and datasets are sorted by this quantity. Error bars are 95% confidence intervals from 20 different random data splits.

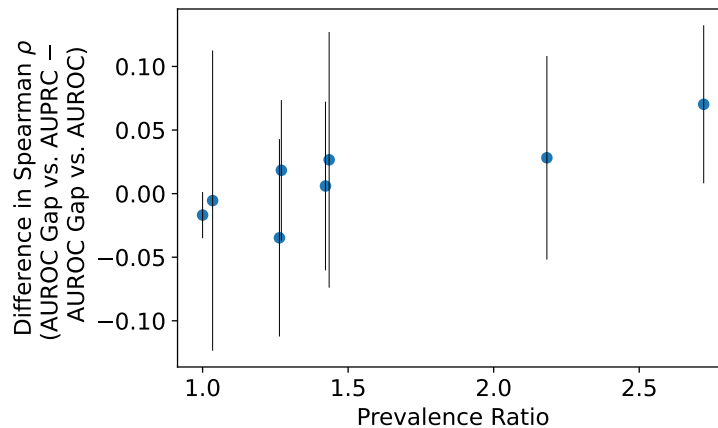


Figure 8: Correlation between the prevalence ratio, and the difference between the Spearman’s ρ of the AUROC gap versus AUROC and the AUROC gap versus AUPRC. Each point represents a dataset and attribute combination. *This correlation itself has a Spearman’s ρ of 0.905 ($p = 0.002$).*

I Literature Review Methodology

I.1 Paper Acquisition

The initial phase of our comprehensive literature search involved the acquisition of datasets from both the ArXiv preprint server (through the RedPajama dataset on Hugging Face), as well as from a subset of years of NeurIPS, ICML, ICLR, ACL, and CVPR conference proceedings (all scraped manually). The ArXiv dataset, approximately 93.8 GB in size, encompassed over 1.5 million texts in JSONL format. For NeurIPS, we developed a script to scrape conference papers from 1987 to 2019 (9680 texts), aiming to enrich our search. Other venues contributed fewer papers to our assessment process.

I.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 and NeurIPS datasets. These scripts detected occurrences of our predefined keywords, allowing efficient parsing of a vast number of texts from both sources.
3. **Dual Mention Selection Criterion:** We focused on papers discussing both AUROC and AUPRC. This criterion 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 in the Arxiv dataset. In the NeurIPS dataset, out of 9,680 texts reviewed, 78 were found to contain keywords from AUPRC and AUROC.

I.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 for the arXiv dataset. 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 from Arxiv 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. Out of the 2,728 papers scrutinized from Arxiv using this model, 201 were found to be relevant. For NeurIPS, our focused search with GPT-4 resulted in identifying 2 papers of particular relevance to our thesis from the initial set that contained keywords related to both AUPRC and AUROC.

I.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.

I.4.1 Final Papers

After manual review, we identified 424 papers that make or reference some version of the claim that “AUPRC is better than AUROC in cases of class imbalance.” [78, 212, 428, 116, 8, 223, 229, 318, 393, 308, 450, 407, 4, 217, 391, 417, 261, 282, 70, 399, 169, 343, 153, 148, 361, 281, 219, 206, 204, 2, 350, 156, 134, 425, 22, 115, 257, 429, 353, 279, 85, 424, 137, 258, 367, 10, 454, 251, 275, 140, 208, 438, 356, 79, 52, 444, 90, 252, 111, 296, 325, 293, 312, 321, 387, 358, 41, 397, 54, 159, 369, 71, 180, 301, 74, 363, 24, 86, 49, 253, 270, 390, 329, 333, 348, 238, 233, 6, 124, 448, 359, 236, 243, 335, 210, 58, 317, 284, 21, 192, 12, 379, 274, 202, 415, 377, 373, 271, 82, 193, 376, 242, 183, 88, 276, 322, 395, 327, 239, 338, 319, 149, 288, 408, 136, 128, 46, 163, 132, 28, 207, 411, 302, 412, 227, 101, 221, 145, 328, 277, 234, 73, 60, 418, 92, 20, 198, 103, 27, 410, 151, 174, 29, 287, 11, 304, 347, 161, 285, 437, 341, 366, 5, 225, 334, 423, 297, 47, 324, 147, 371, 38, 323, 53, 184, 344, 378, 162, 394, 414, 106, 110, 426, 346, 130, 264, 432, 179, 440, 191, 94, 189, 362, 316, 220, 175, 421, 39, 188, 19, 401, 389, 419, 309, 224, 109, 95, 197, 97, 31, 244, 135, 351, 133, 352, 402, 63, 405, 23, 81, 305, 396, 330, 266, 303, 294, 43, 62, 364, 295, 168, 181, 165, 114, 89, 66, 250, 248, 278, 13, 430, 25, 416, 34, 398, 199, 84, 56, 381, 76, 35, 269, 291, 404, 228, 185, 241, 453, 211, 160, 15, 209, 173, 420, 326, 286, 64, 273, 9, 299, 382, 139, 442, 96, 372, 201, 30, 446, 386, 403, 127, 196, 69, 155, 75, 360, 231, 235, 310, 267, 255, 345, 349, 171, 126, 190, 452, 342, 57, 182, 59, 455, 439, 123, 383, 112, 59, 357, 55, 142, 214, 203, 374, 314, 170, 87, 187, 300, 355, 262, 7, 260, 268, 280, 1, 44, 298, 265, 120, 249, 91, 422, 436, 18, 108, 113, 176, 283, 245, 433, 158, 218, 122, 313, 172, 380, 365, 154, 385, 290, 388, 247, 216, 67, 98, 368, 177, 164, 289, 320, 32, 186, 400, 16, 237, 42, 431, 157, 375, 272, 406, 259, 315, 117, 77, 152, 306, 434, 72, 166, 392, 449, 230, 246, 26, 445, 232, 146, 118, 107, 61, 129, 50, 143, 427, 144, 370, 311, 336, 447, 105, 263, 256, 215, 125, 167, 240, 226, 195, 48, 194, 200, 3, 213, 354, 340, 45, 141, 102, 332, 51, 443, 33, 68].

Those papers that reference this claim without citation include [78, 212, 428, 116, 8, 223, 407, 4, 217, 261, 70, 169, 343, 153, 219, 206, 204, 2, 350, 156, 134, 22, 115, 257, 353, 279, 85, 424, 258, 367, 10, 454, 251, 275, 140, 356, 79, 52, 444, 90, 252, 111, 296, 293, 321, 358, 41, 54, 159, 369, 71, 180, 301, 74, 363, 24, 86, 253, 270, 390, 329, 333, 348, 233, 6, 124, 448, 359, 236, 243, 335, 210, 284, 21, 12, 379, 274, 202, 415, 377, 373, 271, 82, 376, 242, 183, 88, 322, 395, 239, 338, 149, 288, 408, 136, 128, 46, 163, 28, 412, 101, 328, 73, 60, 20, 103, 151, 29, 287, 11, 304, 161, 285, 341, 5, 423, 324, 371, 53, 184, 344, 162, 414, 110, 346, 264, 179, 191, 94, 189, 362, 316, 175, 188, 19, 389, 419, 109, 95, 351, 133, 23, 396, 303, 43, 364, 168, 165, 114, 250, 248, 430, 25, 416, 34, 398, 199, 56, 76, 35, 269, 291, 404, 228, 241, 453, 211, 160, 15, 420, 286, 9, 299, 139, 442, 96, 372, 30, 196, 69, 75, 360, 231, 235, 267, 171, 190, 342, 57, 439, 383, 112, 59, 357, 214, 314, 87, 187, 300, 262, 7, 260, 280, 44, 120, 249, 91, 18, 108, 113, 176, 283, 245, 122, 172, 380, 154, 385, 290, 216, 67, 164, 289, 186, 400, 16, 42, 431, 157, 375, 77, 306, 166, 449, 230, 232, 146, 118, 107, 61, 370, 311, 336, 447, 105, 263, 256, 125, 240, 3, 354, 340, 45, 141, 102, 332, 443]

Those that do so while citing only other papers that themselves never reference or argue Claim 1 include [428, 219, 204, 350, 156, 134, 22, 454, 251, 275, 356, 79, 52, 444, 90, 358, 41, 159, 369, 74, 86, 270, 333, 6, 124, 359, 243, 210, 284, 12, 379, 415, 377, 373, 395, 239, 136, 128, 328, 73, 20, 103, 29, 11, 304, 285, 5, 162, 414, 264, 191, 94, 189, 175, 419, 133, 303, 364, 165, 250, 25, 416, 398, 35, 228, 241, 453, 211, 160, 15, 286, 139, 96, 372, 30, 69, 75, 360, 231, 235, 171, 112, 59, 357, 314, 262, 7, 44, 120, 18, 113, 176, 283, 122, 172, 154, 385, 290, 216, 289, 431, 157, 375, 230, 146, 118, 107, 61, 311, 336, 3, 45, 141, 332]

All papers identified, manual screening results, and extracted quotes will be made available upon publication.

Claim	References	Commentary
Precision-recall curves or other associated metrics <i>may</i> more appropriately reflect deployment objectives than the receiver operating characteristic.	[78, 212, 428, 279, 85, 137, 6]	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	[212, 204, 2, 258, 275, 356, 79, 90, 325, 159, 180, 301, 329, 395]	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. 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 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	[212, 428, 257, 79]	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. 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.
AUPRC depends on prevalence, which is a desirable property	[282]	This claim is underspecified, and un-true. AUROC always means the same thing, probabilistically, and that meaning independent from class imbalance.
AUPRC better captures differentiating a positive sample with high score from a “hard” negative sample (“hard” meaning one also with high score)	[193]	This is unfounded; both AUROC and AUPRC are weighted expectations over the model’s false positive rate—AUPRC cares more about samples in regions of low firing rate, not explicitly about positive or minority samples.
AUROC is otherwise “optimistic” in low-prevalence settings	[78, 4, 417, 261, 70, 134, 425, 251, 363, 238, 6, 448, 236, 335, 58, 317, 12, 202, 183, 149]	This claim is unfounded; the AUROC clearly depends on the model’s recall. Besides, if recall is the measure of interest, then that should be measured explicitly.
AUPRC focuses more on the positive (minority) class	[393, 417, 361, 281, 22, 115, 85, 10, 275, 356, 296, 293, 82, 193, 239, 319, 288]	
AUROC can not appropriately detect models with poor recall	[282]	

Table 2: Various arguments and our responses to them present on a subset of papers for this claim in the literature.

I.5 Code Availability

All code pertaining to the literature review search can be found in the following GitHub repository: <https://github.com/Lassehhansen/ArxivMLClaimSearch>

Claim	References	Valid?	Commentary
Precision-recall curves or other associated metrics <i>may</i> more appropriately reflect deployment objectives than the receiver operating characteristic. AUPRC does not depend on the number of true negatives, so will be less optimistic than the AUROC	[78, 212, 339, 435, 40, 292, 331, 451, 222, 428] [212, 121, 79]	✓	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. 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	[212, 435, 121, 254, 331, 451, 222, 428, 79]	✓	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	[339, 121, 435]		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)	[331]	✓	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 3: Various arguments and our responses to them present on a subset of papers for this claim in the literature.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: The claims made in the abstract reflect this paper's contributions and scope.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: We discuss limitations and future work in Section 6

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory Assumptions and Proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [\[Yes\]](#)

Justification: The requisite assumptions and associated proofs for all theorems are provided in full technical detail in the appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental Result Reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [\[Yes\]](#)

Justification: We describe the experiments completely and fully release our code. All datasets used are either synthetic and reproducible in the code itself or publicly available.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: As stated above, all data used is either synthetic and fully reproducible or publicly available.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental Setting/Details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [NA]

Justification: While we do have some model training results to assess metrics (and for such results all training and test details are fully described in this paper and the full set of parameters and code is publicly released), this study is not actually a modelling study, but rather a study of machine learning metrics, so our main contribution is not a modeling result that is directly dependent on released test/split/hyperparameter details.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment Statistical Significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We perform appropriate statistical significance tests and report them in this work.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments Compute Resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Our synthetic experiments can be replicated in a colab notebook with the provided Jupyter notebook file, and the real data experiments are adequately described with released code.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code Of Ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines?>

Answer: [Yes]

Justification: We conform with the code of ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader Impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Our work is about correcting a major misunderstanding in the ML community regarding a widely used evaluation metric, and the potential fairness implications of this misunderstanding. In that sense, our entire work is clearly about the potential societal impacts of this misunderstanding, and how it should be corrected. We also clearly discuss the limitations of our work in Section 6.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: We do not release new data or models in this work.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [NA]

Justification: Only public datasets, appropriately cited, are used in this work.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: No new assets are released in this work.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: No crowdsourcing experiments were done in this work, nor was any research with human subjects done. The manual annotation of reviewed papers in this work was solely performed by authors of this work.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: No human subjects research was performed in this work.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.