TITLE

Class imbalance on medical image classification: towards better evaluation practices for discrimination and calibration performance

AUTHORS

Candelaria Mosquera [a,b], Luciana Ferrer [c], Diego H Milone [d], Daniel Luna [a], Enzo Ferrante [d]

[a] Hospital Italiano de Buenos Aires, Buenos Aires, Argentina

[b] Universidad Tecnológica Nacional, Buenos Aires, Argentina

[c] Instituto de Ciencias de la Computación, UBA-CONICET, Argentina

[d] Institute for Signals, Systems and Computational Intelligence, sinc(i) CONICET-UNL, Santa Fe, Argentina.

* Corresponding author: E-mail: candelaria.mosquera@hospitalitaliano.org.ar. Address: Av. Juan B. Alberdi 447, Ciudad de Buenos Aires, Argentina. Phone: (+54) 11-5777-3200 (ext.: 2901).

ABSTRACT

Purpose:

This work aims to assess standard evaluation practices used by the research community for evaluating medical imaging classifiers, with a specific focus on the implications of class imbalance. The analysis is performed on chest x-rays as case-study and encompasses a comprehensive model performance definition, considering both discriminative capabilities and model calibration.

Materials and Methods:

We conduct a concise literature review to examine prevailing scientific practices used when evaluating x-ray classifiers. Then, we perform a systematic experiment on two major chest x-ray datasets to showcase a didactic example of the behavior of several performance metrics under different class ratios, and highlight how widely adopted metrics can conceal the performance in the minority class

Results:

Our literature study confirms that: (1) even when dealing with highly imbalanced datasets, the community tends to use metrics that are dominated by the majority class; and (2) it is still uncommon to include calibration studies for chest x-ray classifiers, albeit its importance in the context of healthcare. Moreover, our systematic experiments confirm that current evaluation practices may not reflect model performance in real clinical scenarios, and suggest complementary metrics to better reflect the performance of the system in such scenarios.

Conclusion:

Our analysis underscores the need for enhanced evaluation practices, particularly in the context of class-imbalanced chest x-ray classifiers. We recommend the inclusion of complementary metrics such as the AUC-PR, adjusted AUC-PR and balanced Brier score, to offer a more accurate depiction of system performance in real clinical scenarios, considering metrics which reflect both, discrimination and calibration performance.

CLINICAL RELEVANCE STATEMENT

This study underscores the critical need for refined evaluation metrics in medical imaging classifiers, emphasizing that prevalent metrics may mask poor performance in minority classes, potentially impacting clinical diagnoses and healthcare outcomes.

KEYWORDS

Deep learning, Computer-Assisted Diagnosis, X-Rays, Machine Learning, Prevalence

KEY POINTS

- We conduct a brief literature study to analyze common scientific practices in papers dealing with x-ray CAD systems based on AI.
- We highlight existing limitations in the reporting of evaluation metrics for x-ray CAD systems in highly imbalanced scenarios.
- We propose the adoption of alternative metrics and provide experimental evaluation on large scale datasets to support our recommendations.

MANUSCRIPT

1. Introduction

The application of machine learning to medical images has grown rapidly in the last years, showing high levels of performance [1]. However, traditional methods for evaluating performance seem insufficient to describe their impact on the real clinical diagnosis pathway [2]. Class imbalance is a distinctive characteristic of medical datasets, and the performance of classifiers should be assessed with metrics not dominated by the majority class [3].

The analysis of chest x-ray images is a typical imbalance scenario, where the low prevalence of radiological findings is evidenced in the large datasets [4] like ChestX-ray14 [5] and CheXpert [6]. Extensive work has been published on the use of chest x-ray datasets to train deep learning (DL) classifiers [4]. These studies use the area under the receiver operating characteristics curve (AUC-ROC) as the primary classification metric. As it does not depend on the class ratios of the test dataset, the comparison of models evaluated in datasets with different prevalence is simplified [7, 8]. However, ROC curves require special caution when used for imbalanced datasets [9-15]. AUC-ROC is insensitive to class imbalance, meaning it treats each class equally regardless of their prevalence in the dataset. In imbalanced datasets, AUC-ROC may yield high scores even when the model performs poorly on the minority class. To elucidate the implications for patient safety, consider this practical example. To validate the performance of a system for pneumonia detection in chest x-rays, the researchers use a set of 100 normal images and 100 images with pneumonia (balanced setting), and the system obtains 90 true positives, 90 true negatives, but 10 false positives and 10 false negatives. The precision, sensitivity and specificity are 0.90, and the AUC-ROC for this example is 0.90. Now, in the real setting, let us say the prevalence of pneumonia is actually 1 in 20 cases (5%). The researchers now evaluate the system in 200 images of this new setting (just 10 with pneumonia), and also obtain an AUC-ROC close to 0.90, because the AUC-ROC is insensitive to class imbalance. Then, they

might conclude that the system performance is maintained in the clinical routine. However, if looking in closer detail, in this new dataset the system (with the same proportion of errors) actually had 9 true positives and 18 false positives: even though the sensitivity and specificity are still high (0.90 and 0.91), the precision has decreased to 0.33. This means that the positive predictive value of the system is poor, so there would be 18 patients who might, for example, get unnecessary chest CT scans requested for pneumonia confirmation.

Here we provide empirical evidence of the adoption of AUC-ROC as a central metric by conducting a literature analysis on chest x-rays classification. We then provide empirical evidence of AUC-ROC limitations on imbalanced scenarios through an experimental study on CheXpert and ChestX-ray14, and compare it with metrics based on the area under the precision-recall curve (AUC-PR), showing that these metrics better reflect the performance on both the majority and the minority class.

Another distinctive characteristic of the medical domain is the need for interpretable outputs to assist in clinical decision-making [16], which requires that models generate good posterior probabilities for the classes given the input. The quality of posteriors is affected by two aspects: their ability to discriminate the classes from each other and their calibration [22]. Well-calibrated posteriors are those that reflect the uncertainty of the system [17, 18, 19]. Model calibration is well studied for example in epidemiology literature [20, 21], but it is not commonly discussed in chest x-ray diagnosis studies. In particular, AUC metrics are immune to calibration problems, measuring only discrimination performance, failing to reflect the overall performance of the scores as posterior probabilities. Metrics such as the negative log likelihood or the Brier score, known as proper scoring rules [23], are sensitive to both discrimination and calibration performance [24] and properly reflect the quality of the posteriors. However, the impact of class imbalance should also be considered in calibration. If imbalance is large, traditional proper scoring rules might show good values because they are dominated by the majority class. Here

we discuss the use of alternative indicators to detect performance issues in low-prevalence pathologies.

2. Materials and methods

Our retrospective study was performed using publicly available data for which no institutional review board approval was required.

2.1 Literature analysis

We performed an illustrative literature review, including conference and journal research articles from two sources:

1. **PubMed**: we selected articles containing the expression “chest radiograph” or “chest x-ray” and the expressions “artificial intelligence”, “machine learning” or “deep learning” in the title, published between 2019 and 2022.

2. **International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)**: we evaluated the volumes “X-ray imaging” and “Computer-aided diagnosis” from the 2019 edition, “Machine Learning methodologies”, “Prediction and diagnosis”, “Machine Learning applications” and “Heart and lung imaging” from the 2020 edition, “Computer aided diagnosis”, “Outcome/Disease Prediction” and “Clinical Applications - Lung” from the 2021 edition, and “Computer aided diagnosis”, “Outcome and Disease Prediction” and “Heart and Lung Imaging” from the 2022 edition. We excluded articles that did not use chest x-rays as test set for a classification task.

We excluded reviews, editorial notes, author corrections, erratas or commentary publications, articles that do not inform the imbalance ratio of the test set or do not report classification results on chest x-ray images. We also excluded articles focusing on COVID-19 because it has been reported that many papers published during 2020 applying machine learning to detect and prognosticate for COVID19 using chest radiographs present poor-quality data, poor application

of machine learning methodology, poor reproducibility and introduce biases in study design [25]. We revised the articles and checked the following aspects:

- **The class ratios on the test set.** Following the criteria used in the community [25, 26], if the minority class represents less than 33% of images (imbalance ratio of 1:3), it was categorized as “Imbalanced”; otherwise as ”Balanced”.
- **The main metric for reporting classification performance.** This was determined by the following conditions: the primary outcome of the study as mentioned by the authors, else the first metric mentioned in the abstract, or the metric with most detailed reporting in figures or tables.
- Whether the study reports at least one of the following metrics, which more adequately reflect the performance of the system for the minority class: AUC-PR, positive predictive value (precision), negative predictive value, Matthews Correlation Coefficient or F-score.
- Whether model uncertainty or calibration were numerically reported or discussed in some way.

2.2 Experimental study

We performed a systematic experimental study to assess the behavior of various performance metrics across pathologies with different class ratios. Data processing and model training were performed with Pytorch[1], adapting the open framework by Cohen et al [28]. Our source code is fully available online.[2]

2.2.1 Data

[1] Torch 1.7.0 (https://pytorch.org/). Torchvision 0.8.1 (https://pytorch.org/vision/).
[2] https://github.com/cmosquer/imbalanceCXR

We used the ChestX-ray14 [5], from the National Institutes of Health, and the CheXpert [6] from Stanford University. ChestX-ray14 includes labels for 14 findings as positive or negative, while CheXpert is labeled for 13 findings as positive, negative, or uncertain. We used frontal chest x-rays and included only one image per patient. For CheXpert, we excluded images with positive "Support Devices" label. Class imbalance was measured considering the presence of pathology as the positive class (Figure 1).

As preprocessing we cropped non-squared images to a squared central crop, resized to 224x224 pixels, and scaled pixel values to [-1024,1024], as is standard practice for these datasets [29]. We performed data augmentation on the training set, according to the best results reported in Cohen et al [30]: random rotation of up to 45°, combined with scaling and translation of up to 15% of image size. 20% of images from each dataset were selected as the test set, and the remaining set was further split into 20% for tuning (validation) and 80% for training. The experiment was repeated five times for each dataset, using five different random seeds for the splitting.

2.2.2 Classifiers

We used a DenseNet-121 architecture [29] previously used in chest x-rays classification [29, 28, 32]. We used a final dense layer of 11 and 14 sigmoid units for CheXpert and ChestXRay14, respectively. We trained for 100 epochs with batch size of 64, Adam optimizer (weight decay of 1e-5) and binary cross entropy loss (with no class weight) [28]. The initial learning rate was 1e-3 and it was reduced by a factor of 10 every 40 epochs.

2.2.3 Performance metrics

We computed the average ROC and PR curves for the test set of both datasets across 5 runs. To calculate metrics that require a threshold, particularly recall (sensitivity), specificity and

precision (positive predictive value), we chose the operating point for each pathology as the threshold that maximized F1 score in the training set.

Since the datasets we used exhibit high class imbalance in several pathologies (e.g. in ChestX-ray 14 the percentage of patients with edema is just 0.24%), we also used the Adjusted AUC-PR introduced in Bugnon et al [33]:

$$\hat{AUC}_{PR} = 1 - \frac{\log(AUC_{PR})}{\log(AUC_{PR-RG})},$$

where the standard AUC-PR is normalized by the performance of a random-guess classifier, AUC-PR−RG. Since it is normalized considering the performance of a random-guess classifier, it eases comparison between pathologies with different imbalance ratios (similar to AUC-ROC), while taking into account the performance on the minority class (similar to AUC-PR).

We calculated the Brier score (BR) [24] for each binary classification task (i.e., each pathology class). This metric is defined as the average squared difference between the true class and the estimated class probability of each image in the test set

$$BR = \frac{1}{N}\sum_{n=1}^{N}(y_n - \hat{y_n})^2$$

where y_n is the binary class label for the *i*th image (0 or 1) and $\widehat{y}_n$ is its output on the corresponding sigmoid unit of the model.

As BR is a proper scoring rule, its optimal value corresponds to a perfect prediction: a system that is perfectly calibrated and perfectly discriminative will have BR=0. However, the BR could be misleading in imbalanced datasets: performance may be good on average over all samples, but poor for the minority class. To address this problem, Wallace et. al [34] propose a decomposition into scores calculated over positive (BR^+) and negative (BR^-) samples

$$BR^{+} = \frac{\sum_{y_i=1}(y_i - \hat{y}_i)^2}{N_{pos}},$$

$$BR^{-} = \frac{\sum_{y_i=0}(y_i - \hat{y}_i)^2}{N_{neg}},$$

where N_{pos} and N_{neg} are the number of positive and negative samples respectively. Therefore, the BR^{+} is actually the Brier score calculated only over the positive samples, while the BR^{-} is calculated only over the negative samples.

Here, an interesting alternative is the sum of both stratified BR, which can be used for evaluation or as objective function for optimization algorithms:

$$BR_{bal} = BR^{+} + BR^{-}$$

3. Results

3.1 Results for the literature analysis

PubMed and MICCAI searches resulted in 123 and 220 articles respectively. 17% of PubMed articles (n=21) were excluded because they were reviews, commentary publications or editorial notes; while 15% of PubMed articles (n=19) and 84% of MICCAI articles (n=185) were excluded because they did not evaluate a chest-x ray classification task. After exclusion, we evaluated 118 articles in total. A detailed table on this analysis is available as Supplementary Material. The primary study outcome is AUC-ROC in 91 works (77%), accuracy in 19 (16%), and other in 8 (7%). Moreover, 90% of studies do not report performance metrics that reflect the overall quality of the posteriors or their calibration, indicating that this aspect was not assessed. Regarding class ratios, 74% of articles use imbalanced test sets, suggesting this is a common scenario in chest x-rays analysis. Figure 2 compares performance evaluation aspects between articles with balanced and imbalanced test sets. 93% of studies with an imbalanced test set (81 of 87) use AUC-ROC or accuracy as main metric (Figure 2.a), and 46% (40 of 87) report only performance

metric that are dominated by the majority class (Figure 2.b), failing to assess model behavior in the minority class. Model calibration is considered in 16% of these works (14 of 87), while only one study with balance test set discusses it (Figure 2.c). Our analysis confirms our two initial hypothesis:

1. That even when dealing with highly imbalanced datasets, the community tends to focus on AUC-ROC and metrics dominated by the majority class; and
2. It is still uncommon to include metrics that assess the quality of the posteriors (considering the calibration aspect) in CAD papers, albeit its importance in the context of healthcare.

3.2 Results for the experimental analysis

We present results measured in the held-out test set as an average of five runs of training for each dataset. Figure 3 shows discrimination metrics for all pathologies, sorted by decreasing positive class ratio. The three AUC measures have similar values in pathologies whose class positive ratio is higher than 30%, meaning that when there are no significant imbalances, all these metrics are good indicators of discriminative performance. However, starting at the *lung lesion* class in CheXpert (positive class ratio 22.41%) the PR-based measures drop noticeably. Interestingly, pathologies with similar high values of AUC-ROC show a large difference in AUC-PR measures when the class imbalance is very different (for example, atelectasis and pneumonia in CheXpert, or effusion and hernia in ChestX-ray14). This suggests that evaluating the AUC-PR curve is especially important to understand model behavior in imbalanced datasets. If only the AUC-ROC was considered, we would conclude that the model is equally good for most tasks, while the actual behavior is significantly different. For example, the AUC-ROC values for effusion and edema on the ChestX-ray14 dataset are close (0.89 and 0.91, respectively), suggesting that discrimination performance for these pathologies was similar. However, their AUC-PR values differ greatly (0.39 vs 0.05), indicating that the system has a poor usability for edema detection. The Adjusted AUC-PR also correctly indicates the difference

(approx 0.7 for effusion vs approx 0.4 for edema), but without falling too low as the AUC-PR and also not overestimating the discriminative performance as the AUC-ROC. Importantly, a high AUC-ROC value might not imply an acceptable performance in highly imbalanced classification tasks.

We observe that specificity is high across pathologies: in spite of a large number of false positives, the much larger number of true negatives dominates the numerator of its formula. On the other hand, precision tends to decrease along with the positive class ratio since it is sensitive to false positives. The same specificity value might not imply a comparable performance in imbalanced datasets; instead, precision becomes more informative in these scenarios. Another important fact is observed in the class Nodule on ChestX-ray14 (positive class ratio 5.40%), where the Adjusted AUC-PR begins to separate from the AUC-PR. In that point a significant increase in specificity is observed, at the cost of another big drop in recall (and the same precision). Here it can be seen how the Adjusted AUC-PR is a better indicator of the balance between recall and precision, as well as being less sensitive to very low positive class ratios, since it is normalized by the performance of a random-guess classifier.

Figure 4 shows ROC and PR curves for the pathologies with highest and lowest positive class ratio in each dataset. We can see how lung opacity and pneumonia detection have similar ROC plots, but their PR plots are strikingly different. Regarding ChestX-Ray14, the detection of edema presents a high mean AUC-ROC value, however the AUC-PR is very low. For strongly imbalanced classes, there is a greater variation of performance across split seeds, showing that the dependence on specific positive samples is stronger as imbalance increases.

Given fixed score distributions for each class, decreasing the positive class ratio does not change the AUC-ROC but it does affect the AUC-PR: for the same recall level the precision will get worse. The AUC-ROC of a random classifier would be 0.5 regardless of the positive class ratio in the test set, while its AUC-PR would vary along with this ratio [10]. In imbalanced

scenarios, a significant portion of the ROC curve could correspond to operating points with a large number of false positives (i.e., points with low precision). As a consequence, the AUC-ROC is highly dominated by the performance on operating points that would not be acceptable in an actual application, making it an inadequate metric for such scenarios. On the other hand, in the PR curve, those points with a large number of false positives correspond to low values of precision and a lower height of the curve, reducing the value of the AUC-PR.

Figure 5 shows the average BR for each class. Unlike AUC metrics, which measure only discrimination power, the BR is affected also by model calibration, representing a comprehensive assessment of performance. However, if only the standard BR was considered, we would conclude that the model has a similar performance across pathologies. Instead, we see that BR^{+} is higher than BR^{-} for pathologies with low positive class ratio. This shows that performance is consistently worse for the minority class than the majority one across pathologies. A central finding of this work is showing that a problem in the classification of the minority class might only be noted when using BR_{bal}, the balanced version of Brier. Using the standard BR, this performance problem would go unnoticed. The BR_{bal} can be used to measure overall performance in imbalanced data: a high score implies that the posterior probabilities are poor and, hence, not useful for interpretation, which could be caused by poor calibration, poor discrimination, or a combination of both—either in the minority or the majority class, or both.

4. Discussion

The central objective of this work was to study how performance is affected by the positive class ratios of chest pathologies in the context of chest x-ray classification. We observed that AUC-ROC alone can be misleading: the fact that the ROC curve includes multiple operating points with low precision values makes it less informative for imbalanced scenarios than the PR curve.

The comparison of AUC-ROC and AUC-PR has been studied in prior work [3, 35]. Ozenne et al [11] showed that AUC-PR had a higher correlation with precision across simulations with various imbalance ratios, and concluded that it reflected better the discriminant ability of a biomarker in rare diseases. Sahiner et al [12] compared the statistical power of these two metrics, and indicated that AUC-PR can offer a statistical advantage when the positive class ratio is low, while in a balanced scenario AUC-ROC has slightly higher power. The fact that the PR curve is prevalence-dependent has practical implications: it hinders the comparison across datasets with different class ratios, and requires a test set with a prevalence that matches that of the true population to make valid statistical inferences. As a practical example, consider the comparison of the performance for atelectasis detection in the ChestX-ray14 dataset vs. the CheXpert dataset in our study. Both present a good AUC-ROC of approximately 0.80: however, the positive class ratio of atelectasis is very different in the two settings (47% for CheXpert vs. 5% for ChestX-ray14). Therefore, the AUC-PR value varies greatly across these settings: around 0.80 for CheXpert and 0.20 for ChestX-ray14. To determine which classifier presents a better performance, we would need to consider the prevalence of the target clinical setting. Although these issues can be solved with an adequate experimental design —by reporting the test class ratios and by applying properly-designed correction methods— AUC-ROC can hide many errors when the dataset is unbalanced, as is often the case in practice. Another appealing characteristic of the ROC and PR plots is their visual interpretability, as they provide an overview of performance across a wide range of operating points. However, [10] showed that, in the context of imbalanced datasets, ROC plots can be deceptive, due to an intuitive but mistaken interpretation of specificity. Therefore, AUC-PR should be reported alongside AUC-ROC to provide a comprehensive assessment of the discrimination power of a classifier. Moreover, we also propose to adopt the Adjusted AUC-PR introduced in Bugnon et al [33], as an alternative indicator which turns out to be less sensitive to very low positive class ratios, due to the random-guess normalizing factor. Although these metrics and aspects might be well-known in

the machine learning community, our literature analysis reveals that it is not usually addressed in x-ray image classification studies. The literature search was not exhaustive; however, we consider that the selected sample of articles is representative of the current common practices in the field of chest x-ray automated diagnosis.

All discussed AUC indices are ranking metrics, as they depend on the probability that a randomly chosen positive instance is ranked higher than a randomly chosen negative instance. For this reason, they do not consider the problem of threshold optimization. This can be a limitation in clinical practice, where obtaining a final binary classification is critical to assist in decision-making. The selection of an optimal operating point is a key aspect of model performance [34]. Metrics that are calculated for a fixed operating point—such as the confusion matrix—can grade the quality of the decisions that are based on this threshold. They should be reported to describe the model's actual utility for a classification task, and class imbalance as well as prevalence shifts should be considered when interpreting them [36]. Recall and specificity are metrics that are not affected by the imbalance ratio: when the negative class predominates strongly over the positive class, high values of these metrics can be hiding a great number of false positives. On the contrary, precision reveals this problem when it occurs. Although this has been previously reported [7, 10] and it was further confirmed in our experimental results, we observed that several works still fail to report precision (positive predictive value) when evaluating chest x-rays classifiers.

Although the calibration aspect is critical for the successful translation of DL models to clinical practice [16], our literature review showed that most articles do not address this issue. Proper scoring rules like BR, can be used to consider this aspect, as they quantify performance including both discrimination and calibration [37]. BR is affected by large class imbalance and dominated by the performance in the majority class, as was confirmed by our experimental study. To address this problem, the BR_{bal} is a useful metric in imbalanced datasets, which reflects performance in both the minority and the majority class.

We hope that this work will contribute to improve evaluation practices for CAD systems by improving the comprehension of model behaviour in real scenarios, and ultimately help to translate medical AI into the clinical setting.

FIGURES

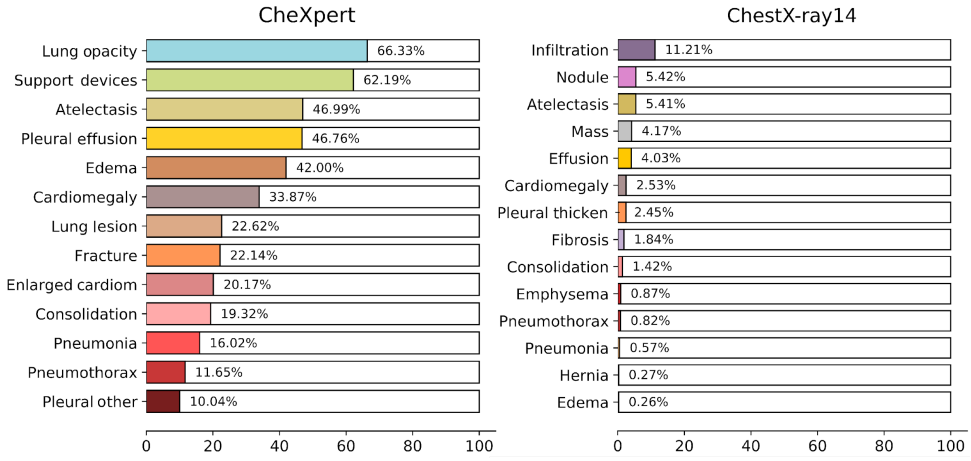

Figure 1: Positive class ratio for the radiological findings annotated in two major chest x-ray datasets. The colored area represents the proportion of positive observations for each finding. These values were computed using only one image per patient and only frontal views. Uncertain and empty labels were excluded.

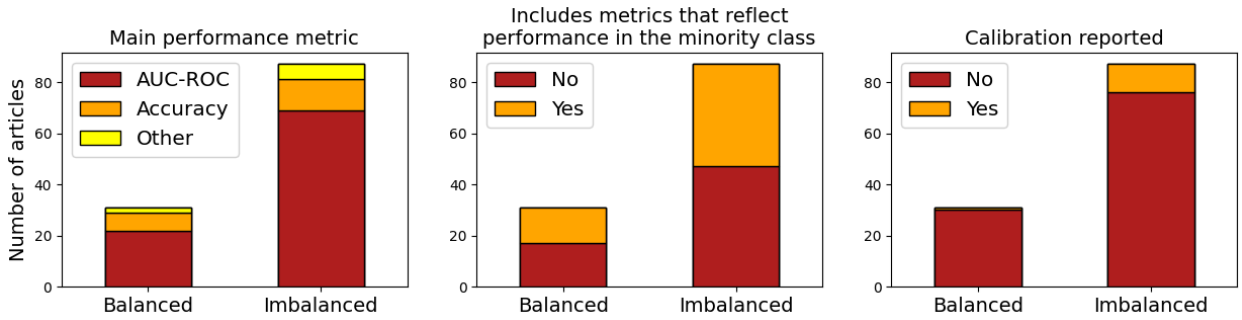

Figure 2: Results from the literature analysis on chest x-ray classification. Comparison between articles with balanced and imbalanced test sets for three aspects of performance assessment.

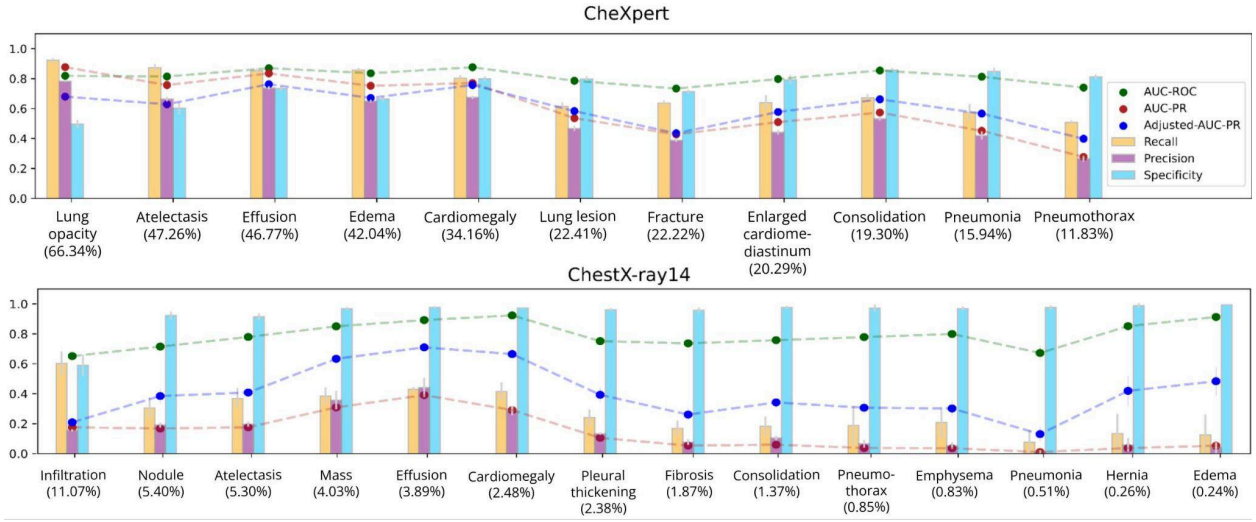

Figure 3: Discrimination metrics as imbalance increases. Values in parentheses indicate the average positive class ratio in the test set.

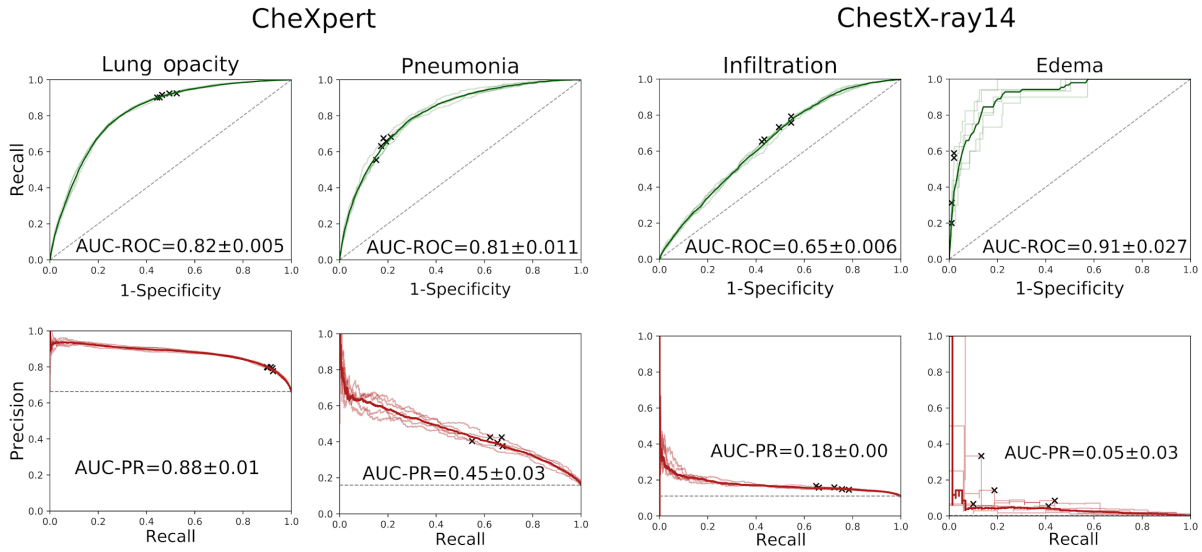

Figure 4: ROC curve (top) and PR curve (bottom) for the pathologies with highest and lowest imbalance ratio. Black crosses indicate the operating point corresponding to maximum F1 score for each run. Bold line represents the mean across runs.

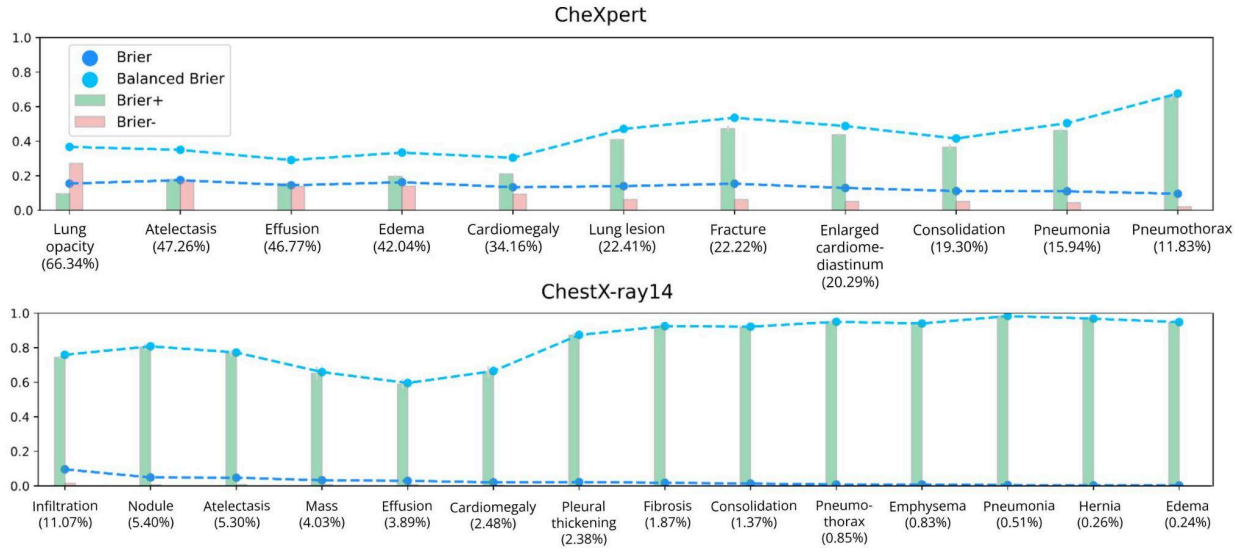

Figure 5: Brier, BR+ (Brier over positive samples), BR- (Brier over negative samples) and Balanced Brier (sum of BR+ and BR-)t, sorted by decreasing positive class ratio. Values in parentheses indicate the average positive class ratio for each class.

REFERENCES

1. Yu KH, Beam AL, Kohane IS (2018) Artificial intelligence in healthcare. Nature biomedical engineering, 2(10):719–731.
2. Beam AL, Manrai AK, Ghassemi M (2020) Challenges to the reproducibility of machine learning models in health care. Jama, 323(4):305–306.
3. Luque A, Carrasco A, Martín A, de las Heras A (2019) The impact of class imbalance in classification performance metrics based on the binary confusion matrix. Pattern Recognition, 91:216–231.
4. Çallı E, Sogancioglu E, van Ginneken B, van Leeuwen KG, Murphy K (2021) Deep learning for chest x-ray analysis: A survey. Medical Image Analysis, page 102125.
5. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM (2017) Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2097–2106.
6. Irvin J, Rajpurkar P, Ko M, Yu Y, Ciurea-Ilcus S, Chute C, Marklund H, Haghgoo B, Ball R, Shpanskaya K, et al. (2019) Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, pages 590–597.
7. Erickson BJ, Kitamura F (2021) Magician's corner: 9. performance metrics for machine learning models.
8. de Hond AA, Steyerberg EW, van Calster B (2022) Interpreting area under the receiver operating characteristic curve. The Lancet Digital Health, 4(12):e853–e855.
9. López V, Fernández A, García S, Palade V, Herrera F (2013) An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. Information Sciences, 250(11):113–141.

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

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

12. Sahiner B, Chen W, Pezeshk A, Petrick N (2017) 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 Medical Imaging 2017: Image Perception, Observer Performance, and Technology Assessment, volume 10136, page 101360G. International Society for Optics and Photonics.

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

14. Davis J, Goadrich M (2006) The relationship between precision-recall and roc curves. In Proceedings of the 23rd international conference on Machine learning, pages 233–240.

15. Varoquaux G, Colliot O (2023) Evaluating machine learning models and their diagnostic value. In Olivier Colliot, editor, Machine Learning for Brain Disorders. Springer.

16. Kompa B, Snoek J, Beam AL (2021) Second opinion needed: communicating uncertainty in medical machine learning. NPJ Digital Medicine, 4(1):1–6.

17. Ovadia Y, Fertig E, Ren J, Nado Z, Sculley D, Nowozin S, Dillon JV, Lakshminarayanan B, Snoek J (2019) Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. arXiv preprint arXiv:1906.02530.

18. Dawid AP (1982) The well-calibrated bayesian. Journal of the American Statistical Association, 77(379):605–610.

19. Mukhoti J, Kulharia V, Sanyal A, Golodetz S, Torr PHS, Dokania PK (2020) Calibrating deep neural networks using focal loss. arXiv preprint arXiv:2002.09437.

20. Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW (2016) A calibration hierarchy for risk models was defined: from utopia to empirical data. Journal of clinical epidemiology, 74:167–176.

21. Collins GS, Moons KGM (2019) Reporting of artificial intelligence prediction models. The Lancet, 393(10181):1577–1579.

22. Blattenberger G, Lad F (1985) Separating the brier score into calibration and refinement components: A graphical exposition. The American Statistician, 39(1):26–32.

23. Gneiting T, Raftery AE (2007) Strictly proper scoring rules, prediction, and estimation. Journal of the American statistical Association, 102(477):359–378.

24. Brier GW (1950) Verification of forecasts expressed in terms of probability. Monthly weather review, 78(1):1–3.

25. Batista, G. E., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD explorations newsletter, 6(1), 20-29.

26. Google Machine Learning Foundational Courses. Imbalanced data. Published by Google Developers. https://developers.google.com/machine-learning/data-prep/construct/sampling-splitting/imbalanced-data. Accessed March 1, 2024.

27. Roberts M, Driggs D, Thorpe M, Gilbey J, Yeung M, Ursprung S, Aviles-Rivero AI, Etmann C, McCague C, Beer L, et al. (2021) Common pitfalls and recommendations for using machine learning to detect and prognosticate for covid-19 using chest radiographs and ct scans. Nature Machine Intelligence, 3(3):199–217.

28. Cohen JP, Hashir M, Brooks R, Bertrand H (2020) On the limits of cross-domain generalization in automated x-ray prediction. In Medical Imaging with Deep Learning.

29. Rajpurkar P, Irvin J, Zhu K, Yang B, Mehta H, Duan T, Ding D, Bagul A, Langlotz C, Shpanskaya K, et al. (2017) Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. arXiv preprint arXiv:1711.05225.

30. Cohen JP, Bertin P, Frappier V (2019) Chester: A web delivered locally computed chest x-ray disease prediction system. arXiv preprint arXiv:1901.11210.

31. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 4700–4708.

32. Larrazabal AJ, Nieto N, Peterson V, Milone DH, Ferrante E (2020) Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. Proceedings of the National Academy of Sciences, 117(23):12592–12594.

33. Bugnon LA, Yones C, Milone DH, Stegmayer G (2019) Deep neural architectures for highly imbalanced data in bioinformatics. IEEE Transactions on Neural Networks and Learning Systems.

34. Wallace BC, Dahabreh IJ (2014) Improving class probability estimates for imbalanced data. Knowledge and information systems, 41(1):33–52.

35. García V, Sánchez JS, Mollineda RA (2012) On the effectiveness of preprocessing methods when dealing with different levels of class imbalance. Knowledge-Based Systems, 25(1):13–21.

36. Godau P, Kalinowski P, Christodoulou E, Reinke A, Tizabi M, Ferrer L, Jäger P, Maier-Hein L (2023) Deployment of image analysis algorithms under prevalence shifts. arXiv preprint arXiv:2303.12540.

37. Ramos D, Franco-Pedroso J, Lozano-Diez A, Gonzalez-Rodriguez J (2018) Deconstructing cross-entropy for probabilistic binary classifiers. Entropy, 20(3):208.