Advances in Artificial Intelligence and Machine Learning; Short Communication 2 (4) 471-476 Received 20-10-22; Accepted 2-11-22; Published 8-11-22

Strengthening Machine Learning Reproducibility for Image Classification

Guofan Shao shao@purdue.edu
Department of Forestry and Natural Resources,
Purdue University, West Lafayette, Indiana, USA.

Hao Zhang zhanghao@purdue.edu
Department of Statistics,
Purdue University, West Lafayette, Indiana, USA.

Jinyuan Shao jyshao@purdue.edu
Department of Forestry and Natural Resources,
Purdue University, West Lafayette, Indiana, USA.

Keith Woeste keith.woeste@usda.gov
Hardwood Tree Improvement and Regeneration Center,
US Department of Agriculture Forest Service.
Department of Forestry and Natural Resources, Purdue University, West Lafayette, Indiana, USA.

Lina Tang lntang@iue.ac.cn
Key Laboratory of Urban Environment and Health,
Institute of Urban Environment,
Chinese Academy of Sciences, Xiamen, China.

Corresponding Authors: Guofan Shao and Lina Tang

Abstract

Machine learning (ML) reproducibility needs to be informed with reliable evaluation measures. However, routine image classification is evaluated using metrics that are highly sensitive to class prevalence. Consequently, the reproducibility of ML models remains unclear due to class imbalance-induced noise. We suggest regularly using class imbalance-resistant evaluation metrics, including balanced accuracy, area under precision-recall curve, and image classification efficacy, for the evaluation of the reproducibility of ML models. Each of these evaluation metrics is conceptually consistent with and logically complements the others, and their joint use can help explain different aspects of classification performance at the whole-class level and individual class level. These metrics can be used for the validation, testing, and/or transfer of ML classifiers. Comprehensive analysis using these metrics as a routine approach strengthens the reproducibility of ML models.

Keywords: Deep learning, classification, Discriminative power, Class imbalance, Efficacy.

Citation: Guofan Shao, et al. Strengthening Machine Learning Reproducibility for Image Classification. Advances in Artificial Intelligence and Machine Learning. 2022;2(4):32.

1. CHALLENGES

Image classification (also known as semantic segmentation) is implemented in a variety of fields, ranging from molecular imaging to earth observation. Machine learning (ML) algorithm makes automated image classification of new datasets possible, especially when the new data to be analyzed is collected using the same types of sensors and if it covers the same types of objects as those used to train the original model [1]. If an ML model is reproducible, it can be used by an independent group to obtain 'the same result' using their own datasets [2]. The reproducibility crisis of ML models remains, however, a crucial concern [3-5]. As to image classification, because no two images are identical due to data variation (e.g., pixel and radiometric resolution), object variation (e.g., contrast and color), and variation in scale and extent trained ML model does not produce identical results when new image data are classified. Instead, ML reproducibility is a function of the variation in the accuracy of image classification [6-8]. Specifically, when a trained model is used for classifying two or more image datasets, its reproducibility level is expressed by the changes in classification accuracy: the smaller the accuracy variation, the greater a model's reproducibility [6]. This argument is valid if accuracy information perfectly reflects the classifier's discriminative power irrespective of data and object variation.

There are dozens of metrics available for classification evaluation. The most used metrics include precision, recall, F score (also known as F-measure), intersection over union (IoU), and/or accuracy [1,4,7,8] (FIGURE 1). The use of one or more evaluation metrics is essential, but the practice is not consistent because (1) there is no standard for the choice of accuracy metrics within a field, (2) the terminology and choice of evaluation metrics vary among fields, and (3) most evaluation metrics are sensitive to class sizes or their distribution, a phenomenon known as the class imbalance effect [9]. In brief, at the whole-class level, overall accuracy (or simply accuracy) is likely high when class distributions are uneven; at the class level, accuracy tends to increase with class size [10]. The overestimated accuracy as the result of the class imbalance effect contributes to the so-called 'AI chasm' in digital medicine [11,12], meaning that high accuracy does not ensure clinically meaningful outcomes. At the class level, as indicated by a hypothetical example in FIGURE 2, the value of F score or IoU for a majority class (> 90%) can be 2-3 times of that for a minority class (< 10%) when map-level accuracy remains the same. Such a case is not the worst when compared with many real-world AI applications in classifying imbalanced data [13].

In addition, machine learning classifiers normally underestimate the probability of a rare class [14,15]. Problems that arise when there are differences in class proportion between lab-controlled and real-world classifications confuse and complicate the judgement of ML reproducibility for image classification. Because the performance of a weak classifier can be obscured by a high accuracy value (the reverse is true, too), research into the class imbalance problem in AI-based classification has been extensive [6,10,13-18]. In contrast, the problem of the class imbalance effect on classification evaluation has received scant attention [7,10,19].

		Reference (e.g., Manually Labelled)			
		Positive (P)	Negative (N)	Class-Level Accuracy	Class-Level ICE
Classification	Positive (P)	True Positive (TP)	False Positive (FP)	P. Precision (Pp) $= TP/(TP+FP)$	$PpE = \frac{Pp - n_p/n}{1 - n_p/n}$
	Negative (N)	False Negative (FN)	True Negative (TN)	N. Precision (Np) $= TN/(TN+FN)$	$NpE = \frac{Np - n_n/n}{1 - n_n/n}$
	Actual Class Proportion	$n_p/n = (TP+FN)/n$	$n_n/n = (FP+TN)/n$	F score $= 2 \times (Pp \times Se)/(Pp + Se)$ Intersection over Union $= TP/(n_p + FP)$ Accuracy $(A) = (TP+TN)/n$ Balanced Accuracy$= (Se+Sp)/2$ Map-Level ICE (MICE) $= \left[A - \left(\frac{n_p}{n}\right)^2 - \left(\frac{n_n}{n}\right)^2\right] \Big/ \left[1 - \left(\frac{n_p}{n}\right)^2 - \left(\frac{n_n}{n}\right)^2\right]$	
	Class-Level Accuracy	Sensitivity (Se) (Recall) $= TP/(TP+FN)$	Specificity (Sp) $= TN/(TN+FP)$		
	Class-Level ICE	$SeE = \frac{Se - n_p/n}{1 - n_p/n}$	$SpE = \frac{Sp - n_n/n}{1 - n_n/n}$		

Figure 1: Confusion matrix of binary classification and the formulas of selected evaluation metrics, including image classification efficacy (ICE).

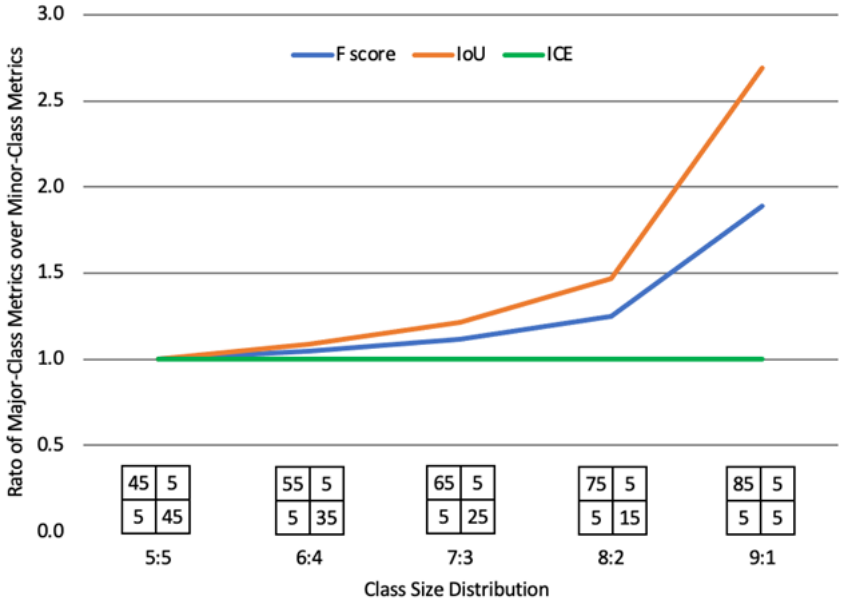

Figure 2: Comparison of class-level evaluation metrics: their ratios of major class over minor when map-level accuracy remains 90% and classification errors are symmetric.

2. A THEORETICAL PERSPECTIVE

Mathematically, ML reproducibility for image classification can be expressed as:

$$s_t^2 = s_d^2 + s_c^2 \tag{1}$$

where s_t^2 is a total evaluation metric variance as the indicator of ML reproducibility, s_d^2 is the variance of the metric of a classifier's discriminative power, where variance is due to changes in data and object features, and s_c^2 is the variance of the metric of class prevalence, where variance is due to changes in image scale, extent, and/or location.

Equation 1 implies that a decrease in the variance of the metrics for discriminative power and class prevalence will result in an increase in ML reproducibility for image classification. The term s_d^2 provides information about ML reproducibility, whereas s_c^2 includes the noise caused by the class imbalance effect. Both terms are embedded in evaluation metrics that are sensitive to the class imbalance effect. As explained above, the class imbalance effect is so great that the term s_c^2 can strongly influence the magnitude of s_t^2. Because the contribution of s_c^2 is inseparable from s_d^2, a practically feasible and effective approach to minimize s_c^2 is to regularly employ class imbalance-resistant metrics for classification evaluation.

3. RECOMMENDATIONS

For image classification evaluation, balanced accuracy (the arithmetic mean of sensitivity and specificity), area under precision-recall curve (AUPRC), and image classification efficacy (ICE) are the most effective metrics to reduce the class imbalance effect [19]. Balanced accuracy is a whole-level metric. AUPRC is used for evaluating positive-class classification performance of ML models under a range of decision probability thresholds. For this reason, precision-recall curve cannot be plotted with a single confusion matrix. The metrics of ICE are map-level ICE (MICE) and class-level ICE; both can be derived from individual confusion matrices (FIGURE 1) (TABLE 1).

Given a confusion matrix with skewed class distribution: TP, FP, FN, and TN, the class proportions are $n_p = TP + FN$ and $n_n = FP + TN$ (FIGURE 1). Assuming the false positive rate and false negative rate remain unchanged, by dividing each element in the positive column with n_p and dividing each element in the native column with n_n, we can obtain an equal class proportion: $n'_p = n'_n = 1$. From the new confusion matrix with even class distribution, the accuracy is $A' = (TP' + TN') / \left(n'_p + n'_n\right) = \left(TP/n_p + TN/n_n\right)/2 =$ (sensitivity + specificity)$/2$. This reasoning explains why balanced accuracy is resistant to the class imbalanced effect.

The precision-recall curve is often recommended to reduce the class imbalance effect [20,21], because AUPRC can incorporate the random classification baseline, defined as n_p/n. The principle of ICE is rooted in the methodology of medical efficacy, an operational metric in the field of medicine. ICE and AUPRC include the same random classification baseline [19], so their evaluations are consistent in principle. The two metrics are complementary because they can be implemented at different stages of ML model's development. One advantage of ICE metrics (over AUPRC) is that they are derived from a single confusion matrix in the same way as other simple metrics,

and ICE metrics can be implemented at the whole-class and individual-class levels (FIGURE 1) (TABLE 1). As illustrated in FIGURE 2, class-level ICE values can remain unchanged with the changes in class proportions. AUPRC and ICE are easily interpretable; if their value is equal to or below zero, the discriminative power of the ML model is not better than random guess. ICE values are lower than corresponding accuracy values because ICE values deduct the contribution from random factors that are proportional to class prevalence [19]; in this way, ICE metrics can reduce the class imbalance effect on image classification performance. Therefore, using these evaluation metrics together as a standardized approach strengthens the evaluation of discriminative power and classification reproducibility of ML models. Routine reporting of these metrics will improve consistent communication on the reliability of ML models, especially between ML model developers and users.

Table 1: Three evaluation metrics suitable for evaluating ML-based image classification.

Suitability	Balanced Accuracy	Area under Precision-Recall Curve	Image Classification Efficacy
Whole-class or individual-class?	Whole Classes	Class Level	Both
One or more confusion matrices required?	One	Multiple	One
What is interpretability?	Accuracy Alternative	Model Performance	Classification Efficacy
Is baseline classification incorporated?	No	Should be	Yes
Can be used as a loss function?	No	No	Yes
Can be used for validation?	Yes	Yes	Yes
Can be used for testing and/or transfer?	Yes	No	Yes
Can explain positive class error level?	No	Yes	Yes
Can explain negative class error level?	No	No	Yes
What is computation consideration?	None	Imposing Baseline	Population Matrix

References

[1] Moen E, Bannon D, Kudo T, Graf W, Covert M,et al. Deep Learning for Cellular Image Analysis. Nat. Methods.2019;16:1233-1246.

[2] https://www.acm.org/publications/policies/artifact-review-and-badging-current

[3] Hutson M. Artificial Intelligence Faces Reproducibility Crisis. Science.2018;359:725-726.

[4] Laine R, Arganda-Carreras I., Henriques R, Jacquemet G. Avoiding a Replication Crisis in Deep-Learning-Based Bioimage Analysis. Nat. Methods.2021;18:1136-1144.

[5] https://www.nature.com/articles/s43588-021-00152-6

[6] Bizzego A, Bussola N, Chierici M, Maggio V, Francescatto M, et al. Evaluating Reproducibility of AI Algorithms in Digital Pathology With DAPPER. PLoS Comput. Biol..2019;15:e1006269

[7] Coveney P, Groen D, Hoekstra A. Reliability and Reproducibility in Computational Science: Implementing Validation, Verification and Uncertainty Quantification in Silico. Philos. Trans. A Math. Phys. Eng. Sci..2021;379:20200409.

[8] Heil B, Hoffman M, Markowetz F, Lee S, Greene C, et al. Reproducibility Standards for Machine Learning in the Life Sciences. Nat. Methods.2021;18:1132-1135.

[9] Shao G, Tang L, Liao J. Overselling Overall Map Accuracy Misinforms About Research Reliability. Landsc. Ecol..2019;34:2487-2492.

[10] Ohsaki M, Wang P, Matsuda K, Katagiri S, Watanabe H, et al. Confusion-Matrix-Based Kernel Logistic Regression for Imbalanced Data Classification. IEEE Trans. Knowl. Data Eng.. 29, 1806-1819 (2017,9)

[11] Topol E. High-Performance Medicine: The Convergence of Human and Artificial Intelligence. Nat. Med..2019;25:44-56.

[12] Marwaha J, Kvedar J. Crossing the Chasm From Model Performance to Clinical Impact: The Need to Improve Implementation and Evaluation of AI. NPJ Digit Med.2022;5:25.

[13] Johnson J, Khoshgoftaar T. Survey on Deep Learning With Class Imbalance. Journal Of Big Data.2019;6:27.

[14] Buda M, Maki A, Mazurowski M. A Systematic Study of the Class Imbalance Problem in Convolutional Neural Networks. Neural Netw..2018;106:249-259.

[15] Megahed F, Chen Y, Megahed A, Ong Y, Altman N, et al. The Class Imbalance Problem. Nat. Methods.2021;18:1270-1272.

[16] https://link.springer.com/referenceworkentry/10.1007/978-0-387-30164-8_110/

[17] Visa S, Ralescu A. Issues in Mining Imbalanced Data Sets-a Review Paper. Proceedings Of The Sixteen Midwest Artificial Intelligence And Cognitive Science Conference.2005;67-73.

[18] Ali A, Shamsuddin SM,Ralescu AL. Classification With Class Imbalance Problem: A Review. Int. J. Advance Soft Compu. Appl.2015;7:176-204.

[19] Shao G, Tang L, Zhang H. Introducing Image Classification Efficacies. IEEE Access.2021;9:134809-134816.

[20] Sofaer H, Hoeting J, Jarnevich C. The Area Under the Precision-Recall Curve as a Performance Metric for Rare Binary Events. Methods Ecol. Evol..2019;10:565-577.

[21] Ashtiani F, Geers A, Aflatouni F. An On-Chip Photonic Deep Neural Network for Image Classification. Nature.2022;606:501-506.