IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received August 19, 2021, accepted September 23, 2021, date of publication September 29, 2021, date of current version October 7, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3116526

Introducing Image Classification Efficacies

GUOFAN SHAO[1], LINA TANG[2], AND HAO ZHANG[3]
[1]Department of Forestry and Natural Resources, Purdue University, West Lafayette, IN 47907, USA
[2]Key Laboratory of Urban Environment and Health, Institute of Urban Environment, Chinese Academy of Sciences, Xiamen 361021, China
[3]Department of Statistics, Purdue University, West Lafayette, IN 47907, USA

Corresponding author: Hao Zhang (zhanghao@purdue.edu)

ABSTRACT Accuracy assessment is essential in all image classification-related fields, ranging from molecular imaging to earth observation. However, existing accuracy metrics are too sensitive to class imbalance or lack explicit interpretations for assessing classification performance. Consequently, their scores may be misleading when they are applied to compare classification algorithms that address different image data sources. These limitations jeopardize the widespread application of deep learning classification methods for classifying different image types. We introduce the metrics of image classification efficacy from medicine and pharmacology to overcome the limitations of accuracy metrics. We include a baseline classification to derive the metrics of image classification efficacy and apply real-world and hypothetical examples to further examine their usefulness. Image classification efficacies can be applied at the map and class levels and for binary and multiclass classifications. The interpretability and comparability of image classification efficacies facilitate reliable classification method evaluation across data sources. We detail the procedures of classification efficacy assessment for image classification researchers and classifier users.

INDEX TERMS Accuracy, classification algorithms, classification assessment, image classification, machine learning, remote sensing.

I. INTRODUCTION

Machine learning, specifically deep learning, has been deployed in every field that involves image classification. Deep learning has transformed the way we classify images at any scale. At the micro scale, biomedical imaging can benefit from deep learning for a better understanding of irregular human body activities and early diagnosis of severe diseases [1]–[5]; at the macro scales, Earth surface characterization [6]–[8] and seven solid Earth geoscience [9] can be strengthened by applying deep learning. The main advantage of deep learning is that a well-trained neural network facilitates automated image classification and can be applied to many different image types. It is essential to assess the accuracy of classification outputs with a deep learning classification algorithm for its new applications [10], [11]. As deep-learning classification methods continue to diversify and advance, the rigorous assessment of neural networks becomes increasingly vital.

More than a dozen metrics have been invented for evaluating pattern recognition and computer vision [12], [13]. With or without modification, these metrics are extensively applied in image classification-related fields, from molecular imaging to earth observation. The existing accuracy metrics can be divided into three types:

- *Type I:* Accuracy metrics are directly derived from error matrices (also known as confusion matrices). These metrics for positive-negative binary classification include accuracy (or overall accuracy) at the map level and sensitivity, specificity, positive precision, and negative precision at the class level. Earth resource remote sensing often involves multiple classes and traditionally uses producer's accuracy (equivalent to sensitivity and specificity) and user's accuracy (equivalent to positive precision and negative precision) [14]. Although these accuracy metrics are interpretable, they are affected by the size distributions of classes and the values of these accuracy metrics are not as informative as to be expected [15]
- *Type II:* These accuracy metrics, which are the immediate derivatives of Type I accuracy metrics, typically include balanced accuracy (arithmetic mean of sensitivity and specificity) and F1 score (harmonic mean of positive precision and sensitivity). The balanced accuracy may reduce class imbalance effects but blurs accuracy interpretation whereas the F1 score may be interpretable

The associate editor coordinating the review of this manuscript and approving it for publication was Essam A. Rashed.

but is still affected by class imbalance [17]. Similar to machine learning applications, the F1 score has become increasingly popular. The mean of two accuracy values may prevent the lower accuracy value from alerting a potential flaw in classification. For example, binary classification output with a value of 0 for any single accuracy metric is useless.

- *Type III:* This type of metrics is rooted in Statistics and then introduced to image classifications to assess their performance. Such metrics include the Matthews correlation coefficient (MCC) and Cohen's Kappa coefficient (Kappa) [18]. Because they were developed in different contexts, these metrics are not interpretable for image classification accuracy assessment despite of their popularity. One common misinterpretation is that when the MCC or Kappa rate is equal to 0, the classification method is usually believed to be similar to random guessing. Remote sensing researchers suggest rejecting the use of Kappa for image classification accuracy assessment [19], [20]. Medical imaging researchers suggest that the MCC provides a more truthful and informative result than other metrics for binary classification assessment based on a series of studies [17], [21]–[23].

Among the three types of accuracy metrics, Type I metrics are the most commonly employed metrics in research. If a single accuracy value is reported in earth remote sensing, this value is highly likely the overall accuracy [15]. The rates of overall accuracy can be misleading. For example, the overall accuracy is 99.5% on average for all six binary, global burned area products [24], whereas the 16-class, global land-cover data have an overall accuracy of only 66.9% [25]. Based on the overall accuracy values, the global burned area classification seems much more successful than the global land-cover classification. Although class-level accuracy metrics are suggested to be more meaningful to the assessment of image classification performance or classification result accuracy, a one-size-fits-all assessment solution is not available [26], [27]. The values of class-level metrics are unequally sensitive to their proportions within the image extent. The same amount of error affects a large class relatively less than it affects a small class, and thus, the classification accuracy is more favorable to a large class than to a small class [15], [16]. As image classification is becoming more far-reaching in research and application, its assessment requires more generally dependable and informing measures.

Accuracy metrics are regularly utilized for accuracy assessment of image classification, although the values of some metric, such as the MCC and Kappa, do not strictly indicate the accuracies. As the name suggests, the rates of the MCC may indicate the correlation levels, whereas Kappa may indicate extent of agreement. When the accuracy metric values are compared between two image classifications, the classification efficacy is examined. Consequently, the word efficacy sometimes appeared as a verbal explanation for the effectiveness of image classification approaches in various fields [28]–[34]. Such use of efficacy makes sense only when different classifications use the same classification Scheme and address images with the same area Extent and the same data-acquisition Time (SET). In the medical fields, efficacy is a common term, and its values are computed by comparing the illness rates between sampled people with a treatment and sampled people without a treatment. Following the same concept, we generalize the evaluation of image classification methods with efficacy, which is quantified by referring to a standard baseline classification as a control to mitigate the class imbalance effects. The resulting image classification efficacy provides an alternative measure for assessing image classification. Next, we derive its equation, examine its robustness, and discuss its applicability.

II. METRICS OF IMAGE CLASSIFICATION EFFICACY

A. ERROR MATRIX AND TYPE I ACCURACY METRICS

An error matrix is a table that displays the number or percentage of cases correctly classified and those incorrectly classified (Table 1). Practically, only random samples are used to compose an error matrix. The reference values (also known as ground truthing) are assumed to be true and represent the actual population.

TABLE 1. General error matrix (also known as a confusion table).

Classification	Reference $j=1$	$j=2$	⋯	$j=J$	Classification total
$i=1$	n_{11}	n_{12}	⋯	n_{1J}	$n_{1\bullet}$
$i=2$	n_{21}	n_{22}	⋯	n_{2J}	$n_{2\bullet}$
⋯	⋯	⋯		⋯	⋯
$i=J$	n_{J1}	n_{J2}	⋯	n_{JJ}	$n_{J\bullet}$
Referene total	n_1	n_2	⋯	n_J	n

An error matrix resembles a contingency table in statistics. Hence, we follow the notations in a contingency table. The element n_{ij} represents the number of objects (or pixels) in class j that are classified to class i. The map-level (or overall) accuracy (A) is therefore $\sum_{j=1}^{n} n_{jj}/n$. The accuracy for each individual class is computed by using either the reference total or classification total. If the reference total is selected, the accuracy with respect to class j is $\mathrm{RA}_j = n_{jj}/n_j$(where, n_j is a simplified presentation of $n_{\bullet j}$, which is a commonly applied to represent a reference total). If the classificaiton total is applied, the accuracy for class j is $\mathrm{CA}_j = n_{jj}/n_{j\bullet}$.

Binary classification is conducted in many fields, and the two classes are commonly referred to as positive for class 1 and as negative for class 2 [12], [13]. In this case, researchers tend to use different terminologies:

- The true positive rate, which is also referred to as sensitivity in pharmacology and as recall in machine learning, is the percentage of positive objects that are classified correctly within the reference total of class positive. We prefer the term sensitivity to recall and denote it by $\mathrm{Se} = n_{11}/n_1$.
- The true negative rate, which is also referred to as specificity in pharmacology, is the percentage of negative

objects that are correctly classified within the reference total of class negative; it is denoted by $\mathrm{Sp} = n_{22}/n_{2}$.
- Positive precision is the number of correctly classified positive cases over the total number of positive cases given by the classifier; it is denoted by $\mathrm{Pp} = n_{11}/n_{1\cdot}$.
- Negative precision is the number of correctly classified negative cases over the total number of negative cases given by the classifier; it is denoted by $\mathrm{Np} = n_{22}/n_{2\cdot}$.

B. IMAGE CLASSIFICATION EFFICACY

In the medical field, the efficacy of a drug is defined by comparing the drug effects on the treatment group to those of a baseline group or the placebo group. Vaccine efficacy (VE) [35] is defined as

$$\mathrm{VE} = \frac{\mathrm{ARU} - \mathrm{ARV}}{\mathrm{ARU}} \tag{1}$$

where ARU is the attack rate in the unvaccinated population and ARV is the attack rate in the vaccinated population. The rates of ARU and ARV are usually determined with a double-blind randomized placebo-controlled trial with persons susceptible to disease.

This approach is perfectly transferable to quantify the effectiveness of image classification methods: a vaccine is equivalent to a classification method; an attack rate is comparable to classification error; the use of vaccine corresponds to the application of classification method; and a randomized placebo control is similar to a random classification as a baseline in image classification. Considering the overall accuracy A as an example, the map-level image classification efficacy (MICE) is expressed as

$$\mathrm{MICE} = \frac{(1 - A_0) - (1 - A)}{1 - A_0} = \frac{A - A_0}{1 - A_0} \tag{2}$$

where A_0 is the accuracy of a random classification as a baseline, which will be given explicitly here.

We now give an explicit formula for A_0 in a general setting of classifying n objects into J classes. Assume that n_j objects (or pixels) belong to class j so that $\sum_{j=1}^{J} n_j = n$. The random classification assigns a randomly chosen object to class j with probability n_j/n. The probability that an object in class j is correctly classified is n_j^2/n^2 (Appendix). Hence, the overall accuracy of the classification is

$$A_0 = \sum_{j=1}^{J} \left(\frac{n_j}{n}\right)^2 \tag{3}$$

We then have

$$\mathrm{MICE} = \frac{A - \sum_{j=1}^{n} \left(\frac{n_j}{n}\right)^2}{1 - \sum_{j=1}^{n} \left(\frac{n_j}{n}\right)^2} \tag{4}$$

Based on (4), the image classification efficacy is defined as the difference between the measure and the corresponding measure for random classification divided by one minus the random classification measure. The MICE value reaches its maximum of 1, if the classification is perfect (i.e., $A = 1$).

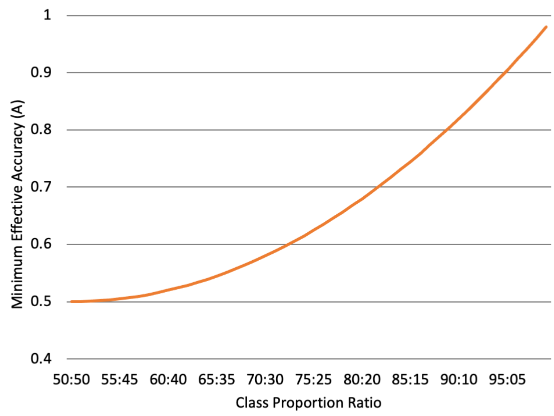


FIGURE 1. Changes in classification accuracy with binary class proportion or size ratios when MICE = 0 (equation 4).

MICE = 0 when $A = A_0$. If MICE is < 0, the classification result should be disregarded because it is even worse than the output of the random classification. When the MICE rate is between 0 and 1, misclassification is reduced compared with the random classification. We define the classification accuracy as the minimum effective accuracy, which increases with the ratio of class proportions for binary classification, when the MICE = 0 (Fig. 1).

Often it is worthwhile or necessary to examine how good the classification results are for a particular class or classes. Based on the baseline probabilities (Appendix), we obtain class-level image classification efficacy as

$$\mathrm{RE}_j = \frac{\mathrm{RA}_j - \frac{n_j}{n}}{1 - \frac{n_j}{n}} \tag{5}$$

and

$$\mathrm{CE}_j = \frac{\mathrm{CA}_j - \frac{n_j}{n}}{1 - \frac{n_j}{n}} \tag{6}$$

where RE_j is the reference-total-based image classification efficacy for class j and CE_j is the classification-total-based image classification efficacy for class j.

For binary classifications, we refer to the terms in pharmacology and machine learning to name the following class-specific, image classification efficacies as sensitivity efficacy (SeE), specificity efficacy (SpE), positive precision efficacy (PpE), and negative precision efficacy (NpE). Each of these efficacies provides the assessment of classification from a different perspective in a way similar to sensitivity, specificity, positive precision, and negative precision.

III. METRIC PERFORMANCE AND INTERPRETATION

A. BINARY CLASSIFICATION

We use seven classifications for image data with class-size ratios near 9:1 to explain the unique usefulness of image classification efficacies (Tables 2 and 3). The first three cases show that the MCC and Kappa can be quite sensitive for a slight change in the classification result for a minor class.

TABLE 2. Results of six image classifications with positive and negative classes.

Case	True Positive	False Negative	False Positive	True Negative	Class-Size Ratio
1	90	1	9	0	91:9
2	90	1	8	1	91:9
3	90	1	7	2	91:9
4	70	20	2	8	90:10
5	73	17	1	9	90:10
6	75	16	0	9	91:9
7	85	5	5	5	90:10

TABLE 3. Comparison of map-level derivatives of seven error matrices (table 2).

Case	A	MICE	MCC	Kappa
1	0.90	0.39	-0.03	-0.02
2	0.91	0.45	0.20	0.15
3	0.92	0.51	0.35	0.30
4	0.78	-0.22	0.39	0.32
5	0.82	0.00	0.49	0.42
6	0.84	0.02	0.54	0.46
7	0.90	0.44	0.44	0.44

TABLE 4. Comparison of class-level derivatives of seven error matrices (table 2).

Case	Sn	Sp	Pp	Np	SeE	SpE	PpE	NpE
1	0.99	0.00	0.91	0.00	0.88	-0.10	-0.01	-0.10
2	0.99	0.11	0.92	0.50	0.88	0.02	0.09	0.45
3	0.99	0.22	0.93	0.67	0.88	0.15	0.20	0.63
4	0.78	0.80	0.97	0.29	-1.22	0.78	0.72	0.21
5	0.81	0.90	0.99	0.35	-0.89	0.89	0.86	0.27
6	0.82	1.00	1.00	0.36	-0.95	1.00	1.00	0.30
7	0.94	0.50	0.94	0.50	0.44	0.44	0.44	0.44

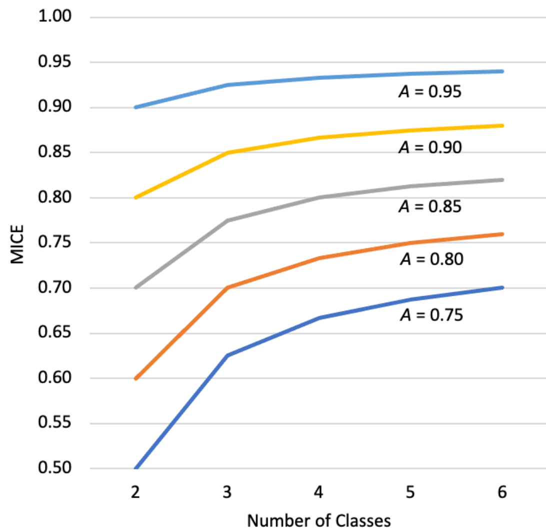


FIGURE 2. Responses of MICE (%) to the number of balanced classes with the same overall accuracy rates.

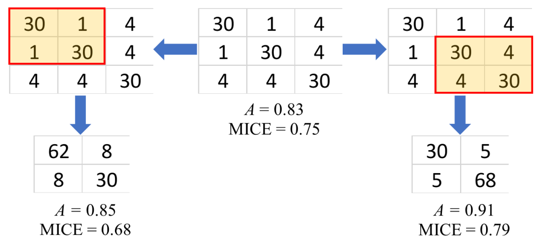


FIGURE 3. Error matrices explaining the effectiveness of class aggregation from three to two classes in terms of image classification accuracy and efficacy.

The +/− sign of the MCC or Kappa can be flipped, or the value can be doubled, whereas MICE values remain relatively stable, when the number of true negatives increase by only 1. Cases 4–6 show that the MCC and Kappa values are considered to be fair or moderate, but the MICE values indicate that the classification is worse than or similar to random classification. This finding suggests that the MICE exhibits different behaviors from the MCC and Kappa.

The overall accuracy of the seven classifications ranged from 0.78 to 0.92 (Table 3). Such levels of classification accuracy sound reasonable for real-world image classifications but could be misleading because of the class imbalance effects. For example, in Case 5, the overall accuracy is 0.82, whereas the efficacy shows that it performs just as the random classification, and therefore, has a poor performance. When A is less than 0.82, the MICE will have a negative value, indicating that the classification method is worse than the random classification. The accuracy rates of the subject classification and baseline classification experience the class imbalance effects and the computation of the MICE assists in mitigating the class imbalance effects.

Cases 4–6 suggest the importance of the efficacy values at both the map level and class level. Despite the fair or moderate values of the MCC and Kappa in these three cases, the MICE values indicate that the overall accuracy is not acceptable, and the sensitivity efficacy (SeE) provides an explanation. A single negative value of these efficacies is sufficient for rendering the classification unacceptable. Cases 2 and 3 are not unacceptable, according to the MICE, MCC, and Kappa, but far from satisfactory because two (SpE and PpE) of the four class-level efficacy values are rather low.

When a classification has symmetrical errors (i.e., false positive = false negative), Se > Sp because $n_1 > n_2$ (Case 7) (Table 4). This finding explains the problem that the rates of recall and selectivity tend to be related to class size. In contrast, SeE and SpE do not have such a problem because SeE is always equal to SpE when false positive equals false negative, which is mathematically provable.

Because the minimum effective classification accuracy is related to class size ratios (Fig. 1), it is important to use image classification efficacy to evaluate the performance of image classification. For example, the minimum effective accuracies are 0.58 and 0.82 when the class-size ratios are 70:30 and 90:10, respectively. Therefore, an accurate rate of 0.80 is pretty good for binary classification with a class-size ratio of 70:30 but fails for binary classification with a class-size

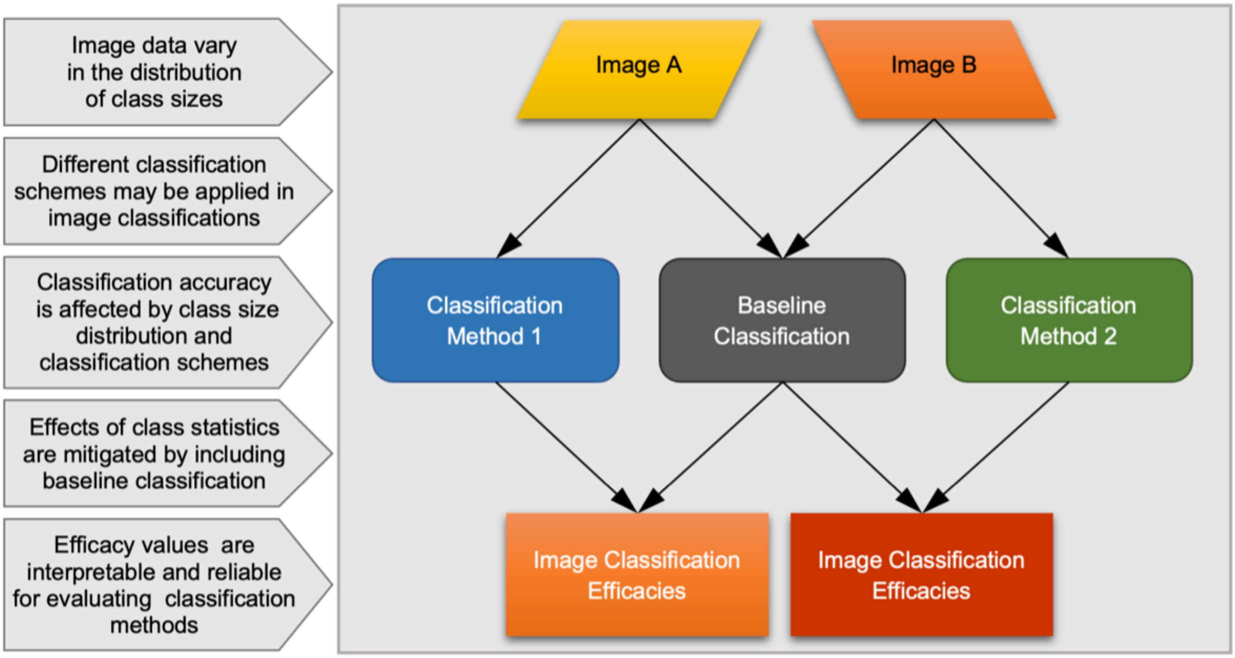


FIGURE 4. Diagram explaining the approach with image classification efficacy to evaluate the performance of image classification methods that involve different images and/or classification schemes.

ratio 90:10. The proportion of burned area is 0.37% within the global mapping extent and the overall accuracy is 99.5% on average among six global burned area products [24]. In this case, the average MICE value is only 0.29, indicating that global burned area classifications are more similar to a random classification than to a perfect classification.

B. MULTICLASS CLASSIFICATION

It is not surprising that binary classification usually has greater overall accuracy than multiclass classification [24], [25], [36]. This phenomenon is the classification scheme effect, which makes overall accuracy incomparable between two classifications that involve different numbers of classes [15]. With the same overall accuracy, the MICE values increase with respect to the number of map classes (Fig. 2). Such an increase in the MICE with the number of map classes makes sense as it reflects the notion that it is more difficult to classify more classes than to classify fewer classes. This result explains another advantage of MICE to overall accuracy. The same MICE value (0.70) is obtained when overall accuracy = 0.85 for two classes; when overall accuracy = 80% for three classes; and when overall accuracy = 75% for six classes (Fig. 2). These three classification methods have the same effectiveness although the overall accuracy values are different. For the global land-cover classification [25], the MICE = 0.63, although its overall accuracy is only 0.67. Such a relatively high MICE value suggests that the global land-cover classification is more effective than the global burned area classifications (MICE = 0.29 on average) despite their almost perfect overall accuracy (A = 99.5% on average) [24].

C. EFFICACY RESPONSES TO CLASS AGGREGATION

Image classification is often performed by following a hierarchical classification system [36], [37], which allows lower-level classes to be aggregated into higher-level classes. Such aggregation usually ensures that overall accuracy cannot be reduced except when only combining classes without misclassification errors between them. If the misclassification error is relatively small, the overall accuracy can increase, but the MICE values may decrease, suggesting that such an aggregation does not improve the classification effectiveness (Fig. 3 left). When combining classes with substantial errors between them, the overall accuracy and MICE values can increase (Fig. 3 right). This kind of effective aggregation is assumed to be the case when aggregation follows a hierarchical classification system. For example, the overall accuracies of the 2011 US National Land Cover Database (NLCD) at Classification Level II and Classification Level I were 82% and 88%, respectively [37]. The corresponding MICE values are 80% and 85%, respectively, confirming that class aggregation from Level II to Level I of the NLCD is effective.

IV. USE OF IMAGE CLASSIFICATION EFFICACY

The advantage of image classification efficacy is that it can mitigate the effects of class imbalance and classification schemes on classification assessment and thus, emphasize the true effectiveness of classification methods (Fig. 4). Therefore, image classification efficacy can function as a general metric for comparing different image classification methods with different class proportions. With rapid advancements in image classification techniques, periodic reviews are

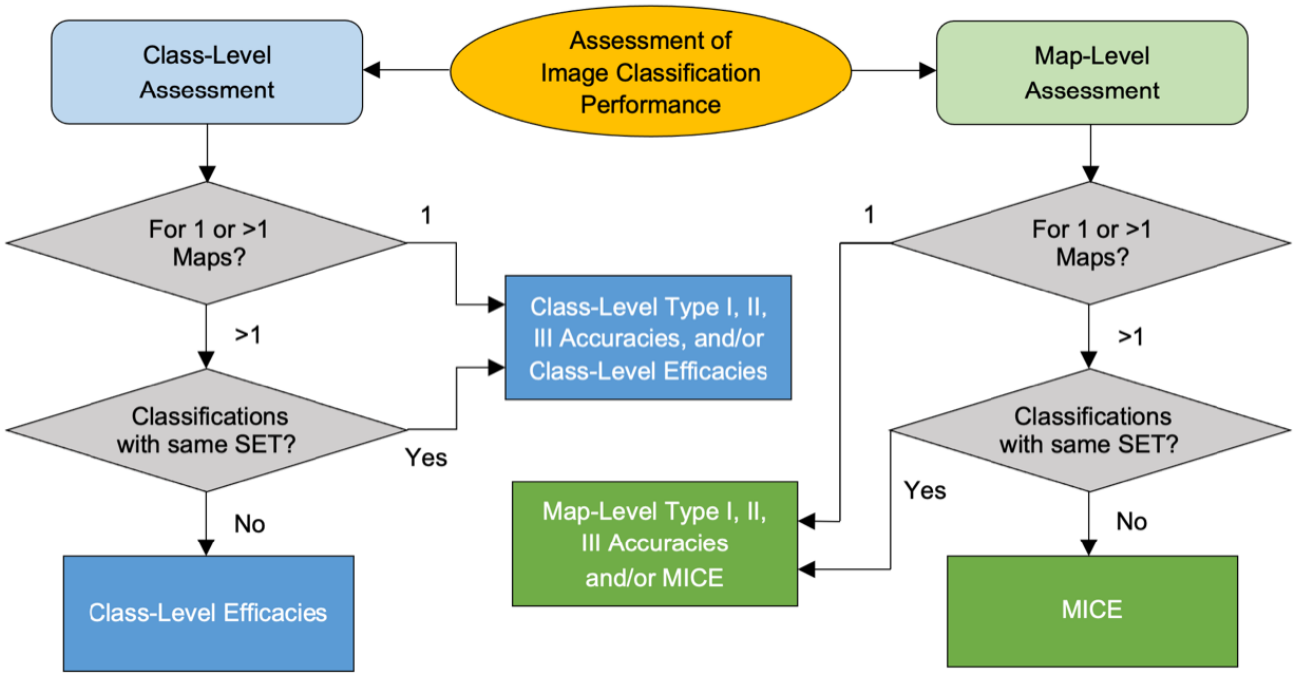


FIGURE 5. Flowchart of classification assessment with image classification accuracy and efficacy metrics. SET stands for classification Scheme, area extent, and data-acquisition time.

becoming increasingly important [2]–[4], [6], [8], [38], [39]. These reviews inevitably involve classification methods that have experimented with different data sources. Comprehensive reviews on image classification techniques can be strengthened by using image classification efficacies.

The metrics of image classification efficacy are particularly useful for comparing classification methods and thus, their relative differences are more important than their absolute values. This does not mean that the efficacy scores should not have a target. As previously discussed, a negative value of image classification efficacy means that the classification is unacceptable, which is the bottom line. The question is how high is high enough? It is understandable if an image classification analyst considers accuracy target. For example, Anderson [40] proposed an accuracy target of 85% for land use land cover classification with satellite remote sensing data. Referring to a binary classification for a class size ratio of 75:25, which is the median of 50:50 and 100:0 ratios, the MICE equals 60%, corresponding to an overall accuracy of 85%. Therefore, we can subjectively set the target of the image classification efficacy scores to 60%. We then divided the positive efficacy values into six levels: 0–0.19 indicates slight progress, 0.20–0.39 denotes moderate progress, 0.40–0.59 represents barely satisfactory, 0.60–0.74 indicates satisfactory, 0.75–0.89 denotes extraordinary, and 0.90–0.99 represents almost perfect. By using this scale, for example, the efficacies of US NLCD datasets [37] are extraordinary at classification levels I and II; the global land-cover classification [25] is satisfactory; and the six global burned area products [24] show moderate progress on average.

The introduction of image classification efficacy does not mean complicating existing classification assessment practices. The misuse of existing classification accuracy metrics can be avoided by employing image classification efficacy. To better conduct image classification efficacy assessment, we summarize the assessment procedures under different circumstances and for different purposes (Fig. 5).

If classification methods that need to be compared are executed with the same images and classification scheme, their comparative assessment can be made directly with Type I accuracy metrics. Otherwise, it will become risky to conduct conventional accuracy assessments. In this case, the MICE and class-level efficacy metrics should be utilized.

V. CONCLUSION

The derivation of image classification efficacies has followed the broadly understandable vaccine efficacy. Image classification efficacy means the effectiveness of image classification relative to random assignment. The metrics of image classification efficacy are applicable to binary and multiclass classification, and suitable for both class-level and map-level efficacy assessments. More importantly, the values of image classification efficacy mitigate the effects of class proportions and classification schemes, and thus, are useful for comparing classification methods that are tested with different images. The introduction of image classification efficacy meets the critical need to rectify the strategy for the assessment of image classification performance as image classification methods are becoming more diversified. The metrics of image classification efficacy can be employed to assess image classifications in all the relevant fields, ranging from molecular imaging to earth observation remote sensing. In any case, researchers are encouraged to provide image data, training data, and reference data when they report their classification progress so that image classification efficacies can be computed when needed.

APPENDIX

In this appendix, we provide proofs for (3), (5) and (6).

A. PROOF OF (3)

By definition, A_0 is the probability that a randomly chosen object is classified correctly. The addition rule of probability implies that

$$A_0 = \sum_{j=1}^{J} P$$

where,

$$\begin{aligned} &P\,(\text{object classified correctly and belonging to class } j)\\ &\quad = P\,(\text{object classified correctly} \mid \text{belonging to class } j)\\ &\qquad \times P\,(\text{object belonging to class } j) \end{aligned}$$

Because there are n_j objects in class j and the classification is random, the two probabilities in the right hand side of the last equation are both n_j/n. Therefore,

$$A_0 = \sum_{j=1}^{J} \left(\frac{n_j}{n}\right)^2.$$

B. PROOFS OF (5) AND (6)

For a random classification, the probability that it classifies correctly an object in class j is clearly n_j/n. This serves as the baseline probability. By definition of efficacy, RE_j and CE_j are given explicitly by (5) and (6), respectively.

REFERENCES

[1] S. L. Goldenberg, G. Nir, and S. E. Salcudean, "A new era: Artificial intelligence and machine learning in prostate cancer," *Nature Rev. Urol.*, vol. 16, no. 7, pp. 391–403, Jul. 2019.

[2] A. Nunes *et al.*, "Using structural MRI to identify bipolar disorders–13 site machine learning study in 3020 individuals from the ENIGMA bipolar disorders working group," *Mol. Psychiatry*, vol. 25, no. 9, pp. 2130–2143, 2020.

[3] J. Schlemper, O. Oktay, M. Schaap, M. Heinrich, B. Kainz, B. Glocker, and D. Rueckert, "Attention gated networks: Learning to leverage salient regions in medical images," *Med. Image Anal.*, vol. 53, pp. 197–207, Apr. 2019.

[4] X. Liu, L. Faes, A. U. Kale, S. K. Wagner, D. J. Fu, A. Bruynseels, T. Mahendiran, G. Moraes, M. Shamdas, C. Kern, J. R. Ledsam, M. K. Schmid, K. Balaskas, E. J. Topol, L. M. Bachmann, P. A. Keane, and A. K. Denniston, "A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: A systematic review and meta-analysis," *Lancet Digit. Health*, vol. 1, no. 6, pp. e271–e297, Oct. 2019.

[5] L. Li, Y. Chen, Z. Shen, X. Zhang, J. Sang, Y. Ding, X. Yang, J. Li, M. Chen, C. Jin, C. Chen, and C. Yu, "Convolutional neural network for the diagnosis of early gastric cancer based on magnifying narrow band imaging," *Gastric Cancer*, vol. 23, no. 1, pp. 126–132, Jan. 2020.

[6] G. Grekousis, "Artificial neural networks and deep learning in urban geography: A systematic review and meta-analysis," *Comput., Environ. Urban Syst.*, vol. 74, pp. 244–256, Mar. 2019.

[7] L. Ma, Y. Liu, X. Zhang, Y. Ye, G. Yin, and B. A. Johnson, "Deep learning in remote sensing applications: A meta-analysis and review," *ISPRS J. Photogramm. Remote Sens.*, vol. 152, pp. 166–177, Jun. 2019.

[8] T. Kattenborn, J. Leitloff, F. Schiefer, and S. Hinz, "Review on convolutional neural networks (CNN) in vegetation remote sensing," *ISPRS J. Photogramm. Remote Sens.*, vol. 173, pp. 24–49, Mar. 2021.

[9] K. J. Bergen, P. A. Johnson, M. V. de Hoop, and G. C. Beroza, "Machine learning for data-driven discovery in solid Earth geoscience," *Science*, vol. 363, no. 6433, Mar. 2019, Art. no. eaau0323.

[10] A. E. Maxwell, T. A. Warner, and L. A. Guillén, "Accuracy assessment in convolutional neural network-based deep learning remote sensing studies—Part 1: Literature review," *Remote Sens.*, vol. 13, no. 13, p. 2450, 2021.

[11] A. E. Maxwell, T. A. Warner, and L. A. Guillé, "Accuracy assessment in convolutional neural network-based deep learning remote sensing studies—Part 2: Recommendations and best practices," *Remote Sens.*, vol. 13, no. 13, p. 2591, 2021.

[12] P. Baldi, S. Brunak, Y. Chauvin, C. A. F. Andersen, and H. Nielsen, "Assessing the accuracy of prediction algorithms for classification: An overview," *Bioinformatics*, vol. 16, no. 5, pp. 412–424, May 2000.

[13] D. M. W. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation," 2020, *arXiv:2010.16061*. [Online]. Available: http://arxiv.org/abs/2010.16061

[14] R. G. Congalton and K. Green, *Assessing the Accuracy of Remotely Sensed Data: Principles and Practices*, 3rd ed. Boca Raton, FL, USA: CRC Press, 2019, pp. 69–76.

[15] G. Shao, L. Tang, and J. Liao, "Overselling overall map accuracy misinforms about research reliability," *Landscape Ecology*, vol. 34, no. 11, pp. 2487–2492, Nov. 2019.

[16] S. Stehman and J. Wickham, "A guide for evaluating and reporting map data quality: Affirming Shao et al. 'Overselling overall map accuracy misinforms about research reliability,'" *Landscape Ecol.*, vol. 35, pp. 1263–1267, May 2020.

[17] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, pp. 1–13, 2020.

[18] R. A. Fisher, *Statistical Methods for Research Workers*, 13th ed. New York, NY, USA: Hafner, 1958, pp. 150–183.

[19] R. G. Pontius, Jr., and M. Millones, "Death to Kappa: Birth of quantity disagreement and allocation disagreement for accuracy assessment," *Int. J. Remote Sens.*, vol. 32, no. 15, pp. 4407–4429, Aug. 2011.

[20] G. M. Foody, "Explaining the unsuitability of the Kappa coefficient in the assessment and comparison of the accuracy of thematic maps obtained by image classification," *Remote Sens. Environ.*, vol. 239, Mar. 2020, Art. no. 111630.

[21] D. Chicco, N. Tötsch, and G. Jurman, "The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation," *BioData Mining*, vol. 14, no. 1, p. 13, 2021.

[22] D. Chicco, V. Starovoitov, and G. Jurman, "The benefits of the Matthews correlation coefficient (MCC) over the diagnostic odds ratio (DOR) in binary classification assessment," *IEEE Access*, vol. 9, pp. 47112–47124, 2021.

[23] D. Chicco, M. J. Warrens, and G. Jurman, "The Matthews correlation coefficient (MCC) is more informative than Cohen's Kappa and Brier score in binary classification assessment," *IEEE Access*, vol. 9, pp. 78368–78381, 2021.

[24] M. Padilla, S. V. Stehman, R. Ramo, D. Corti, S. Hantson, P. Oliva, I. Alonso-Canas, A. V. Bradley, K. Tansey, B. Mota, J. M. Pereira, and E. Chuvieco, "Comparing the accuracies of remote sensing global burned area products using stratified random sampling and estimation," *Remote Sens. Environ.*, vol. 160, pp. 114–121, Apr. 2015.

[25] J. Scepan, "Thematic validation of high-resolution global land-cover data sets," *Photogramm. Eng. Remote Sens.*, vol. 65, pp. 1051–1060, Sep. 1999.

[26] G. M. Foody, "Harshness in image classification accuracy assessment," *Int. J. Remote Sens.*, vol. 29, pp. 3137–3158, May 2008.

[27] S. V. Stehman and G. M. Foody, "Key issues in rigorous accuracy assessment of land cover products," *Remote Sens. Environ.*, vol. 231, Sep. 2019, Art. no. 111199.

[28] J. Han, D. Zhang, G. Cheng, L. Guo, and J. Ren, "Object detection in optical remote sensing images based on weakly supervised learning and high-level feature learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 6, pp. 3325–3337, Jun. 2015.

[29] O. Okwuashi and C. E. Ndehedehe, "Deep support vector machine for hyperspectral image classification," *Pattern Recognit.*, vol. 103, Jul. 2020, Art. no. 107298.

[30] J.-U. Hou, S. W. Park, S. M. Park, D. H. Park, C. H. Park, and S. Min, "Efficacy of an artificial neural network algorithm based on thick-slab MRCP images for the automated diagnosis of common bile duct stones," *J. Gastroenterol. Hepatol.*, Jun. 2021.

[31] T. J. Lark, I. H. Schelly, and H. K. Gibbs, "Accuracy, bias, and improvements in mapping crops and cropland across the united states using the USDA cropland data layer," *Remote Sens.*, vol. 13, no. 5, p. 968, Mar. 2021.

[32] K. Jia, S. Li, Y. Wen, T. Liu, and D. Tao, "Orthogonal deep neural networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 4, pp. 1352–1368, Apr. 2021.

[33] A. Repici, M. Badalamenti, R. Maselli, L. Correale, F. Radaelli, E. Rondonotti, E. Ferrara, M. Spadaccini, A. Alkandari, A. Fugazza, A. Anderloni, P. A. Galtieri, G. Pellegatta, S. Carrara, M. Di Leo, V. Craviotto, L. Lamonaca, R. Lorenzetti, and C. Hassan, "Efficacy of real-time computer-aided detection of colorectal neoplasia in a randomized trial," *Gastroenterology*, vol. 159, no. 2, pp. 512–520, 2020.

[34] M. S. Seo, J. Lee, J. Park, D. Kim, and D.-G. Choi, "Sequential feature filtering classifier," *IEEE Access*, vol. 9, pp. 97068–97078, 2021.

[35] W. A. Orenstein, R. H. Bernier, T. J. Dondero, A. R. Hinman, J. S. Marks, K. J. Bart, and B. Sirotkin, "Field evaluation of vaccine efficacy," *Bull. World Health Org.*, vol. 63, no. 6, p. 1055, 1985.

[36] K. Fenske, H. Feilhauer, M. Förster, M. Stellmes, and B. Waske, "Hierarchical classification with subsequent aggregation of heathland habitats using an intra-annual rapideye time-series," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 87, May 2020, Art. no. 102036.

[37] J. Wickham, S. V. Stehman, L. Gass, J. A. Dewitz, D. G. Sorenson, B. J. Granneman, R. V. Poss, and L. A. Baer, "Thematic accuracy assessment of the 2011 national land cover database (NLCD)," *Remote Sens. Environ.*, vol. 191, pp. 328–341, Mar. 2017.

[38] M. E. Paoletti, J. M. Haut, J. Plaza, and A. Plaza, "Deep learning classifiers for hyperspectral imaging: A review," *ISPRS J. Photogramm. Remote Sens.*, vol. 158, pp. 279–317, Dec. 2019.

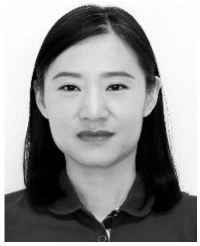
[39] G. Cheng, X. Xie, J. Han, L. Guo, and G.-S. Xia, "Remote sensing image scene classification meets deep learning: Challenges, methods, benchmarks, and opportunities," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, no. 99, pp. 3735–3756, Jun. 2020.

[40] J. R. Anderson, "A land use and land cover classification system for use with remote sensor data," U.S. Geological Survey Professional, Washington, DC, USA, Paper 964, 1976.

GUOFAN SHAO received the Ph.D. degree in computer modeling of forest dynamics from the Chinese Academy of Sciences, in 1989. He also received the certificates of geographic information systems from the Environmental Systems Research Institute (ESRI), Germany, in 1988, and postdoctoral education from the University of Virginia, USA, between 1991 and 1993.

From 1994 to 1996, he was a Research Associate with Virginia Long-Term Ecological Research. Since 1997, he has been an Assistant Professor, an Associate Professor, and a Full Professor with Purdue University. He regularly teaches remote sensing courses at the undergraduate and graduate levels. He is the author of 154 peer-reviewed journal articles and six books.

Prof. Shao was a recipient of the Excellent Associate Editor Award from Springer's *Journal of Forestry Research*, from 2016 to 2020.



LINA TANG received the Ph.D. degree in ecology from the Chinese Academy of Sciences, in 2006. She also received postdoctoral education from Purdue University, USA, between 2007 and 2008.

From 2009 to 2011, she was an Associate Professor with the Institute of Urban Environment, Chinese Academy of Sciences. Since 2009, she has been a Full Professor with the Institute of Urban Environment. She is the author of 130 peer-reviewed journal articles and five books. Her research interests include ecological remote sensing applications and urban ecological patterns, processes, and effects.

Prof. Tang's two papers were awarded, in 2020, as Influential Academic Papers in *Acta Ecologica Sinica*, from 2010 to 2019. She has served as an Associate Editor for *International Journal of Sustainable Development & World Ecology*, an Editorial Board Member for *Acta Ecologica Sinica*, and the Vice Chairman of the Ecological Management Committee of the Chinese Ecological Society.

HAO ZHANG was born in Anhui, China, in 1966. He received the B.S. degree in mathematics and the M.S. degree in statistics from Peking University, China, in 1986 and 1989, respectively, and the Ph.D. degree in statistics from Michigan State University, East Lansing, MI, USA, in 1995.

From 1995 to 1997, he was an Assistant Professor with Marquette University, Milwaukee, WI, USA. From 1997 to 2007, he was an Assistant Professor and an Associate Professor with Washington State University, Pullman, WA, USA. Since 2007, he has been a Professor of statistics with Purdue University, West Lafayette, IN, USA. He worked as the Head of the Department of Statistics, from 2015 to 2020. His research interests include methodological development and theoretical studies of statistical models for massive spatio-temporal data.

Dr. Zhang is a fellow of the American Statistical Association and an Elected Member of the International Statistical Institute. He is an Associate Editor of the *Journal of the American Statistical Association* and the journal *Environmetrics*. He has served as an Associate Editor for the journals *Statistics & Probability Letters* and *Statistica Sinica*. He was a Section Editor of *Encyclopedia of Environmetrics* (2012, Revised Edition).