Bolstering Performance Evaluation of Image Segmentation Models with Efficacy Metrics in the Absence of a Gold Standard

Lina Tang*, Jinyuan Shao*, Shiyan Pang, Yameng Wang, Aaron Maxwell, Xiangyun Hu, Zhi Gao, *Member, IEEE*, Ting Lan, Guofan Shao

***Abstract*—Image segmentation using deep learning has become overwhelmingly widespread. However, routine model testing methods can encounter evaluation inconsistencies or bias, largely due to how accuracy metrics respond to variations in class share distribution. Here we address the effects of class imbalance on model performance evaluation and demonstrates a refined approach that incorporates image classification efficacy (ICE) metrics within the context of semantic segmentation in remote sensing. This evaluation approach was applied in six segmentation experiments that involve multispectral and lidar data, single or multiple models tested with the same or different datasets, and binary and multiclass schemes. ICE metrics revealed unique aspects of model's segmentation capabilities compared to precision, recall, F-score, and overall accuracy. By mitigating the class imbalance effect, per-class efficacy enables precise class-level optimization of segmentation models, while whole-class efficacy facilitates evaluating a model's potential performance when adapted to new datasets. The suitability of the Kappa coefficient, ROC-AUC, and PR-AUC for model evaluation under class imbalance was discussed in comparison with ICE metrics. This efficacy-enhanced model evaluation protocol can be implemented for deep learning model training and testing. The routine use of this evaluation approach will strengthen the dependability and applicability of segmentation tools in various fields.**

***Index Terms*—Artificial intelligence, performance assessment, model testing, semantic segmentation, MICE, class imbalance.**

I. Introduction

IMAGE segmentation is a vital computer vision task with extensive applications in various fields, including medical imaging and earth observation [1]–[6]. With the fast development of numerous deep learning models, the integration of deep learning into image segmentation has evolved at such a rapid pace that it has fundamentally revolutionized the entire field of image segmentation over the past decade. However, the performance of deep learning models continues to be assessed using traditional accuracy metrics that are sensitive to class imbalance. This approach of segmentation evaluation has, to some extent, contributed to the reproducibility crisis in deep learning applications [7]–[10]. In the age of artificial intelligence, it is crucial to transform the evaluation of image segmentation model performance.

Semantic segmentation, also known as pixel classification in the field of computer vision, is a widely used image segmentation technique that assigns each pixel in an image to a specific class [11], [12]. In many application fields, semantic segmentation or pixel classification is also referred to as pixel-based image classification or simply image classification [13]–[18]. Alongside pixel-based image classification, there exists object-based image classification [19]. To evaluate these segmentation tasks, the results are compared with reference data, commonly known as ground truth, resulting in a confusion matrix or error matrix [20]. From this matrix, various accuracy metrics can be computed (Table I). While the primary accuracy metrics used in different academic fields are essentially the same, they may have different names [2], [21]–[26]. The most used per-class metrics include precision (and its synonyms), recall (and its synonyms), specificity, F-score (also known as F1 score and F-measure), Intersect over Union (IoU), and the Dice coefficient (Dice) (Table I). Although these metrics originated from binary segmentations, they have become popular in multiclass segmentations. On the other hand, whole-class metrics mainly include overall accuracy (OA or A), the kappa coefficient, and the means of class-level metrics, such as mean recall, mean F-score (mF), and mean IoU (mIoU) (Table I) [1], [11], [26], [27].

The overall accuracy of image segmentation is influenced by the distribution of class shares and the number of classes, and tends to exaggerate the model's performance on imbalanced datasets [28]–[31]. In other words, it is easier to achieve a high overall accuracy when segmenting image data with substantial disparities in class shares, such as global burned area mapping [32]. On the other hand, per-class accuracy often corresponds to the proportion of class shares [33]–[36]. This phenomenon, known as the "class imbalance effect" [2], [15], suggests that high accuracy values may not necessarily indicate desired quality of image segmentation. This is partic-

*Eequal contribution
Corresponding authors: Lina Tang and Jinyuan Shao
L. Tang is with the Key Laboratory of Urban Environment and Health, Chinese Academy of Sciences Institute of urban Environment, Xiamen, China
J. Shao is with the Department of Forestry and Natural Resources, Purdue University, West Lafayette, Indiana, USA
S. Pang is with the Faculty of Artificial Intelligence in Education, Central China Normal University, Wuhan, China
Y. Wang is with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China
A. Maxwell is with the Department of Geology and Geography, West Virginia University, Morgantown, WV, USA
X. Hu is with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China
Z. Gao is with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China
T. Lan is with the Key Laboratory of Urban Environment and Health, Chinese Academy of Sciences Institute of urban Environment, Xiamen, China
G. Shao is with the Department of Forestry and Natural Resources, Purdue University, West Lafayette, Indiana, USA

TABLE I
GENERAL TERMINOLOGY AND MATHEMATICAL EXPRESSIONS FOR SELECTED METRICS USED FOR EVALUATING BINARY OR MULTICLASS SEMANTIC SEGMENTATION.

Metric Name	Metric Computation	Metric Definition and/or Interpretation
Per-class Accuracy Metrics		
Commission error (Ce) (Prediction-total-based error)	$\mathrm{Ce}_j = \frac{n_{j+}-n_{jj}}{n_{j+}}$	False positive rate and false alarm rate in binary classification. Over-prediction error relative to class share proportion from classification.
Omission error (Oe) (Reference-total-based error)	$\mathrm{Oe}_j = \frac{n_{+j}-n_{jj}}{n_{+j}}$	False negative rate and miss rate in binary classification. Under-prediction error relative to class share proportion from reference.
Precision (P), positive predictive value for the positive class, negative predictive value for the negative class, or user's accuracy for any class	$\mathrm{P}_j = \frac{n_{jj}}{n_{j+}} = 1 - \mathrm{Ce}_j$	Positive predictive rate in binary classification. Per-class accuracy relative to class share proportion from classification.
Recall (R), sensitivity, hit rate, or true positive rate for the positive class, specificity, selectivity, or true negative rate for the negative class, or producer's accuracy for any class	$\mathrm{R}_j = \frac{n_{jj}}{n_{+j}} = 1 - Oe_j$	Hit rate, true positive rate in binary classification. Per-class accuracy relative class share proportion from reference. Some tolerance to the class imbalance effect.
F-score	$\text{F-score}_{\ j} = \frac{2\times(\mathrm{P}_j\times\mathrm{R}_j)}{\mathrm{P}_j+\mathrm{R}_j}$	Harmonious mean of precision and recall.
Dice value (Dice = F-score in binary classification)	$\text{Dice}_{\ j} = \frac{2\times n_{jj}}{n_{j+}+n_{+j}}$	Similarity between prediction and reference.
Intersect over Union (IoU)	$\mathrm{IoU}_j = \frac{n_{jj}}{n_{j+}+n_{+j}-n_{jj}}$	Accuracy of image segmentation or object detection.
Precision-based image classification efficacy (PE)	$\mathrm{PE}_j = \frac{\mathrm{P}_j-\frac{n_{+j}}{n}}{1-\frac{n_{+j}}{n}}$	Effectiveness of semantic segmentation on the top of random assignment. Model's performance after the class imbalance effect is lessened.
Recall-based image classification efficacy (RE)	$\mathrm{RE}_j = \frac{\mathrm{R}_j-\frac{n_{+j}}{n}}{1-\frac{n_{+j}}{n}}$	Effectiveness of semantic segmentation on the top of random assignment. Model's performance after the class imbalance effect is lessened.
Mean efficacy (Mean E)	$\text{Mean E}_j = \frac{\mathrm{PE}_j+\mathrm{RE}_j}{2}$	The average value of precision- and recall-based image classification efficacies
Whole-class Accuracy Metrics		
Overall accuracy (A)	$\mathrm{A} = \sum_{j=1}^{J}\frac{n_{jj}}{n}$	Percent pixels or objects correctly predicted for the entire data
Balanced accuracy (mean recall) (MA)	$\mathrm{ME} = \sum_{j=1}^{J}\frac{\mathrm{R}_j}{J}$	An accuracy measure. Generally insensitive to the class imbalance effect.
Mean F-score (mF)	$\mathrm{mF} = \sum_{j=1}^{J}\frac{\mathrm{F-\ score}_{\ j}}{J}$	The average value of F-scores.
Mean IoU (mIoU)	$\mathrm{mIoU} = \sum_{j=1}^{J}\frac{\mathrm{IoU}_j}{J}$	The average value of IoUs.
Kappa coefficient (KC)	$\mathrm{KC} = \frac{\mathrm{A}-\sum_{j=1}^{J}\frac{n_{j+}n_{+j}}{n^2}}{1-\sum_{j=1}^{J}\frac{n_{j+}n_{+j}}{n^2}}$	Agreement of two independent judges after removing chance agreement. Invalid and non-interpretable for segmentation evaluation.
Map-level image classification efficacy (MICE)	$\mathrm{MICE} = \frac{\mathrm{A}-\sum_{j=1}^{J}\left(\frac{n_{+j}}{n}\right)^2}{1-\sum_{j=1}^{J}\left(\frac{n_{+j}}{n}\right)^2}$	Effectiveness of semantic segmentation on the top of random assignment. Model's performance after the class imbalance effect is lessened.

Note: Symbol n is the total sample size, n_{jj} is the number or percent of sample points correctly classified as class j, and J is the total number of classes. The subscript symbol $j+$ is the prediction total and $+j$ is the reference total.

ularly relevant when a purpose class has fake high accuracy while a non-purpose class exhibits fake low accuracy, or vice versa. As a result, it is impossible to directly compare the performance among models when the test data consists of varying class share distributions [12]. Furthermore, these class share-induced changes in accuracy values can cause erroneous conclusions regarding the reproducibility of a particular deep learning model [7]–[10].

Along with accuracy metrics, the kappa coefficient is also a prominent measure employed in accuracy assessment in remote sensing. Originally developed to evaluate the agreement between two independent judges while accounting for chance agreement in social science [37], the kappa coefficient has become one of the most controversial metrics in accuracy assessment [38]–[40]. Three reasons make the kappa coefficient inappropriate for accuracy assessment. Firstly, its assumption of noncorrectness for both judgments is inconsistent with the accuracy assessment paradigm in image segmentation, where reference data are presumed to be correct. Secondly, the chance agreement component ($n_{j+}n_{+j}$; Table I) depends on the assumption of independence between judgments, which is not invalid in deep-learning-based image segmentation, where training and reference data often originate from the same population. Finally, the kappa coefficient is not an accuracy

metric.

The desired segmentation power of a model is characterized by its ability to consistently achieve high accuracy in separating classes, particularly confused classes, regardless of fluctuations in class share distribution. However, due to the class imbalance effect, conventional accuracy metrics can introduce bias in evaluating model performance and may not reflect the real segmentation power of the model. The straightforward connection between the real and apparent segmentation powers of a model is expressed as follows:

$$P_r = P_a - B \tag{1}$$

where, P_r is the model's real segmentation power, P_a is model's apparent segmentation power, and B is the bias in model performance caused by class share variations.

The conceptual model of Eq (1) is well represented by using image classification efficacy (ICE) metrics, which lessen the class imbalance effect by considering class share-proportional random probability as a general baseline (Table I) [41]. If ICE $= 1$, the segmentation is perfect; if ICE < 0, the segmentation is worse than random assignment and is therefore considered ineffective. The per-class accuracy of the baseline is equivalent to the class share proportion in percentage terms, meaning that the larger a class, the greater its accuracy value. The overall accuracy of the baseline increases with the skewness of the class size distribution and decreases with the number of classes. For instance, the overall accuracy for binary baseline segmentation with a class share ratio of $0.75 : 0.25$ is 0.625 $(0.75^2 + 0.25^2)$, surpassing that for a class share ratio of $0.5 : 0.5$ $(0.5^2 + 0.5^2 = 0.500)$. These two accuracy values are both greater than the overall accuracy for a four-class baseline segmentation with an equal share $(4 \times 0.25^2 = 0.250)$. Due to this computational mechanism, ICE mitigates the bias in segmentation performance caused by varying class share distributions and class counts. To facilitate the applications of ICE, Shao et al. divided its values into eight scales [41]. According to this scaling, for example, Zheng et al. 's image segmentation has reached the extraordinary level (map-level ICE or MICE = 0.77), confirming the great performance of their model as indicated by high OA [42].

The ICE metrics are potentially applicable across various academic fields involving image segmentation, but their explicit application within this context has not been sufficiently elucidated. One primary question is, why is there a need for efficacy metrics when accuracy metrics already exist? Using six segmentation experiments in remote sensing, this paper aims to systematically clarify the limitations of segmentation evaluation using accuracy metrics, explicitly interpret ICE metrics across diverse segmentation scenarios, and enlighten efficacy-reinforced evaluation of image segmentation models. This contribution paves the way for implementing the transformed practices for assessing the real segmentation power of deep learning models in various fields.

II. Segmentation Experiments

Each image segmentation task is unique and there is no one-size-fits-all evaluation method for different segmentation tasks. We used six segmentation experiments to demonstrate and explain why and how an efficacy-reenforced evaluation approach is implemented with model testing. In six segmentation experiments, multispectral and lidar data were used as input data, single or multiple models were tested with the same or different datasets, and binary and multiclass schemes were both considered. Depending on the specific nature of image segmentation, the experimental results were evaluated with both accuracy and efficacy metrics. Class imbalance was quantified with class share percentage for binary segmentation and the Coefficient of Variance (CV) of class share proportions for multiclass segmentation.

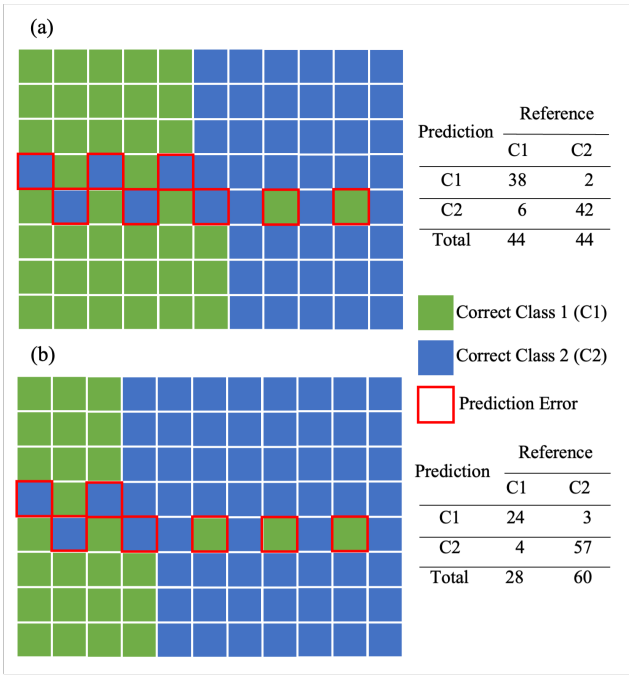


Fig. 1. Comparison of two image segmentations, assuming that the true positive rates remain the same for the two segmentations. (a) two classes have an even share. (b) two classes have uneven class shares.

A. Experiment I: A hypothetical case (balance vs imbalance)

This experiment demonstrates how moderate changes in class share proportions affect per-class and whole-class accuracy and efficacy. It is assumed that the true positive rates remain approximately the same, one image segmentation involves an even distribution ($50 : 50$; Fig. 1a), while the other involves an uneven distribution between two classes ($32 : 68$; Fig. 1b).

B. Experiment II: Binary segmentation (building vs nonbuilding), single model, single data source

Experiment II was to differentiate building pixels from nonbuilding pixel (background). The dataset was Massachusetts Buildings Dataset (MBD) [43], one of the most used building segmentation datasets in computer vision community. MBD consisted of 151 images and each of them sized 1,500 × 1,500

pixels covering an area of 2.25 square kilometers (pixel size was 1 m). Buildings occupied a small portion of the landscape and this segmentation exercise involved a class imbalance problem. We randomly divided the dataset into training set, validation set, and test set with 106, 15, and 30 images, respectively. Subsequently, the testing datasets were evenly split into three subsets, and each subset maintained a class proportion ratio of approximately 4 : 1 (uneven), 9 : 1 (more uneven), and 19 : 1 (most uneven), respectively. We trained a vanilla U-Net using training and validation sets with binary cross entropy loss, and evaluated model performance against each of the three subsets of the test data.

C. Experiment III:Binary segmentation (change vs no change), two models, single data source

Pang et al. developed a deep-learning model called "prior semantic information-guided change detection" (PSI-CD), showcasing its excellence through various accuracy metrics [44]. We used the same change detection dataset of the WHU Building dataset (WHU-CD) http://gpcv.whu.edu.cn/data/building_dataset) for this experimentation. The WHU-CD dataset consisted of bitemporal images, covering an area of 20.5 km^2, collected in Christchurch, New Zealand, in 2012 and 2016, respectively. This dataset includes 2,386 pairs of 512 × 512 images with a resolution of 0.3 meters. Each image pair has a corresponding change label. The total number of image pairs was randomly divided into two subsets: 1,559 for training and 827 for testing. Class-share ratios between change and no change were computed for each image pair. Subsequently, the 827 test images were nearly equally divided into three subsets based on their class-share ratios. The average class share proportions were 15:1 (No change to Change) for subset 1 (uneven), 36:1 for subset 2 (more uneven), and 122:1 for subset 3 (most uneven). The FC-EF model is an extended U-Net architecture, comprising four max-pooling and upsampling layers. The PSI-CD is a Siamese convolutional neural network structure, and consists of three modules. A semantic segmentation module is a pre-trained network, a change analysis module calculates the change features from the two-period semantic features, and a decoder module is used to output the final change detection patch.

D. Experiment IV: Multiclass segmentation (land use types), multiple models, single data source

Using Landsat TM data acquired in September 2007, level-I land cover maps were developed for the midwest of the USA. Covering an area of 1,323 km^2, the landscape predominantly featured Agriculture (64%), followed by Forest (19%), Urban areas (11%), and Water (6%). A group of graduate students used conventional supervised and unsupervised image classification algorithms with Erdas Imagine to generate 23 land use and land cover maps.The purpose of this exercise was to compare accuracy variations among land use land cover maps generated with the same dataset. Each map was assessed using with the same 4,800 randomly sampled points labeled by referring to two-meter-resolution RGB orthophotos acquired by the US National Agriculture Imagery Program (NAIP) in 2007. The ten best classifications were selected for this comparative analysis based on overall accuracy and MICE.

E. Experiment V:Multiclass segmentation (land use types), single model, single data source

Experiment V involved multi-modal joint segmentation tasks using the N3C-California dataset and the IKD-Net framework [36]. N3C-California is a comprehensive, annotated dataset that includes over 10,000 LiDAR and imagery patches. IKD-Net is an innovative and efficient architecture designed to extract features directly from raw multi-modal data rather than from their simplified derivatives. Its end-to-end, disentangled dual-stream backbone ensures the integrity of information across heterogeneous modalities. The testing dataset, drawn from the N3C-California dataset, comprised 1,080 image patches, each sized at 512 × 512 pixels, and included four classes: Ground, Tree, Building, and Other. Class Ground emerged as the dominant class, Urban was the codominant class, and Other was the smallest class. The CV was computed for each labeled image patch across the four classes. Subsequently, all testing image patches were ranked based on their CV values and then evenly divided into three subsets representing high (CV = 1.29), middle (CV = 0.83), and low (CV = 0.71) unevenness. Following the methodology outlined by Wang et al. [36], we executed the IKD-Net on each subset of testing-image patches integrated with Lidar data.

F. Experiment VI: Multiclass segmentation (forest types), single model, data from different areas

Maxwell et al. predicted forest community types, total aboveground live biomass (AGLBM), and species-specific AGLBM for the states of Michigan, Oregon, and West Virginia, USA [45]. For the task of forest type mapping, the input data included the Landsat multispectral time series and the 10 m spatial resolution National Elevation Dataset (NED). Random forest was used to differentiate forest community types. The number of forest types ranged from seven to nine across the three states. These forest landscapes vary in terms of forest characteristics, terrain, management practices, and disturbance histories, all of which affect the model's performance. Notably, distinguishing between broadleaved tree species, especially in West Virginia, proved more challenging than distinguishing between coniferous and broadleaved trees in Michigan and Oregon. The CV based on plot counts by forest types revealed differences in class share unevenness among the three states: 0.73 for Michigan, 1.20 for Oregon, and 1.89 for West Virginia.

III. Results

The model performance from the six segmentation examples or experiments is expressed with precision, recall, F-score, precision-based efficacy, recall-based efficacy, mean efficacy, overall accuracy, and MICE (Table II). In the subsequent four subsections, we provide a summary of the general trends and key takeaways.

In binary segmentation, per-class efficacy is not regularly affected by class share proportion. If FP equals

TABLE II
SELECTED RESULTS OF SIX SEGMENTATION EXPERIMENTS.

Segmentation Experiment	Scenario	Class Name	Class Share (%)	Precision	Recall	F-score	Precision-based Efficacy	Recall-based Efficacy	Mean Efficacy	Overall Accuracy	MICE
I	A: Even Shares	Green	50.0	0.950	0.864	0.905	0.900	0.727	0.814	0.909	0.818
		Blue	50.0	0.875	0.955	0.913	0.750	0.909	0.830		
	B: Uneven Shares	Green	31.8	0.889	0.857	0.873	0.837	0.790	0.814	0.920	0.817
		Blue	68.2	0.934	0.950	0.942	0.794	0.843	0.818		
II	A: Uneven	Building	20.5	0.645	0.815	0.720	0.554	0.767	0.661	0.870	0.602
		Non-Building	79.5	0.949	0.884	0.916	0.750	0.437	0.593		
	B: More Uneven	Building	10.8	0.602	0.816	0.693	0.554	0.794	0.674	0.922	0.594
		Non-Building	89.2	0.977	0.935	0.955	0.785	0.395	0.590		
	C: Most Uneven	Building	5.2	0.573	0.891	0.697	0.550	0.885	0.717	0.960	0.593
		Non-Building	94.8	0.994	0.964	0.979	0.881	0.301	0.591		
III	A: FC-EF Uneven	Change	6.3	0.878	0.763	0.816	0.87	0.747	0.804	0.979	0.817
		No-Change	93.7	0.984	0.993	0.988	0.749	0.887	0.812		
	B: FC-EF More Uneven	Change	2.7	0.838	0.759	0.797	0.833	0.753	0.791	0.989	0.801
		No-Change	97.3	0.993	0.996	0.994	0.753	0.849	0.798		
	C: FC-EF Most Uneven	Change	0.8	0.348	0.254	0.294	0.342	0.248	0.288	0.990	0.384
		No-Change	99.2	0.994	0.996	0.995	0.249	0.520	0.337		
	D: PSI-CD Uneven	Change	6.3	0.951	0.884	0.916	0.948	0.876	0.912	0.990	0.914
		No-Change	93.7	0.992	0.997	0.994	0.877	0.952	0.915		
	E: PSI-CD More Uneven	Change	2.7	0.945	0.875	0.909	0.943	0.872	0.908	0.995	0.910
		No-Change	97.3	0.997	0.999	0.998	0.872	0.948	0.910		
	F: PSC-CD Most Uneven	Change	0.8	0.941	0.885	0.912	0.940	0.885	0.913	0.996	0.914
		No-Change	99.2	0.997	0.943	0.969	0.885	0.943	0.914		
IV		Water	6.2	0.909	0.915	0.912	0.903	0.910	0.906	0.868	0.757
		Urban Area	10.9	0.739	0.646	0.689	0.707	0.602	0.655		
		Forest	19.3	0.858	0.724	0.785	0.824	0.658	0.741		
		Agriculture	63.6	0.888	0.943	0.915	0.691	0.844	0.768		
V	A: CV = 0.71	Other	1.4	0.302	0.830	0.443	0.293	0.828	0.560	0.930	0.893
		Tree	22.3	0.923	0.864	0.893	0.901	0.825	0.863		
		Building	33.7	0.974	0.972	0.973	0.961	0.958	0.960		
		Ground	42.6	0.954	0.935	0.944	0.920	0.886	0.903		
	B: CV = 0.83	Other	1.4	0.331	0.856	0.477	0.321	0.854	0.588	0.941	0.905
		Tree	18.3	0.930	0.860	0.894	0.915	0.828	0.872		
		Building	29.6	0.978	0.979	0.979	0.969	0.971	0.970		
		Ground	50.7	0.967	0.950	0.958	0.933	0.899	0.916		
	C: CV = 1.29	Other	1.9	0.314	0.819	0.454	0.301	0.815	0.558	0.943	0.871
		Building	10.9	0.967	0.968	0.967	0.963	0.964	0.963		
		Tree	14.6	0.945	0.880	0.911	0.935	0.859	0.897		
		Ground	72.6	0.984	0.955	0.969	0.941	0.838	0.889		
	D: Entire Dataset	Other	1.6	0.316	0.833	0.458	0.305	0.830	0.568	0.938	0.896
		Tree	18.3	0.932	0.867	0.898	0.916	0.837	0.877		
		Building	24.4	0.974	0.974	0.974	0.966	0.966	0.966		
		Ground	55.7	0.971	0.949	0.960	0.935	0.884	0.910		
VI	A: Michigan CV = 0.73	Six classes	—	—	—	—	—	—	—	0.631	0.535
	B: Oregon CV = 1.20	Nine classes	—	—	—	—	—	—	—	0.698	0.598
	C: West Virginia CV = 1.89	Eight classes	—	—	—	—	—	—	—	0.785	0.557

FN, precision equals recall in binary segmentation; thus, the major class has a greater precision (or recall) value than the minor class. When FP does not equal FN, precision and recall also differ, but the F-score of the major class exceeds the F-score of the minor class (Figs. 2a–2c). Conversely, the average value of precision- and recall-based efficacy of the major class does not necessarily differ from that of the minor class (Figs. 2d–2f), reflecting different interpretations of segmentation quality. The similar ICE values produced by Experiments I and III (Fig. 2) indicate that the two classes have similar segmentation qualities despite the major class being more accurate as indicated by F-score. Sometimes ICE values display an opposite trend to accuracy values (Experiment II in Fig. 2), which explains why the major class has worse segmentation qualities than the minor class.

In multiclass segmentation, per-class efficacy fine-tunes accuracy distribution patterns. In multiclass segmentation, the level of per-class accuracy is affected by both class share proportion and misassignment errors in multiple classes; thus, the relationship between per-class accuracy and class share proportion in multiclass segmentation is not as strong as in binary segmentation. Nevertheless, a dominant class can still have relatively high accuracy (Figs. 3b and 2e) but may not have high efficacy (Figs. 3c and 3f). When a small class has high accuracy (Fig. 3b), it must have high efficacy (Fig. 3c). The different patterns of accuracy (Figs. 3b and 3c) and ICE (Figs. 3e and 3f) indicate that ICE can reduce the class imbalance effect at the class level. For example, both the largest and smallest classes have high accuracy (F-score of Agriculture = 0.915 and F-score of Water = 0.912), but

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2024.3446950

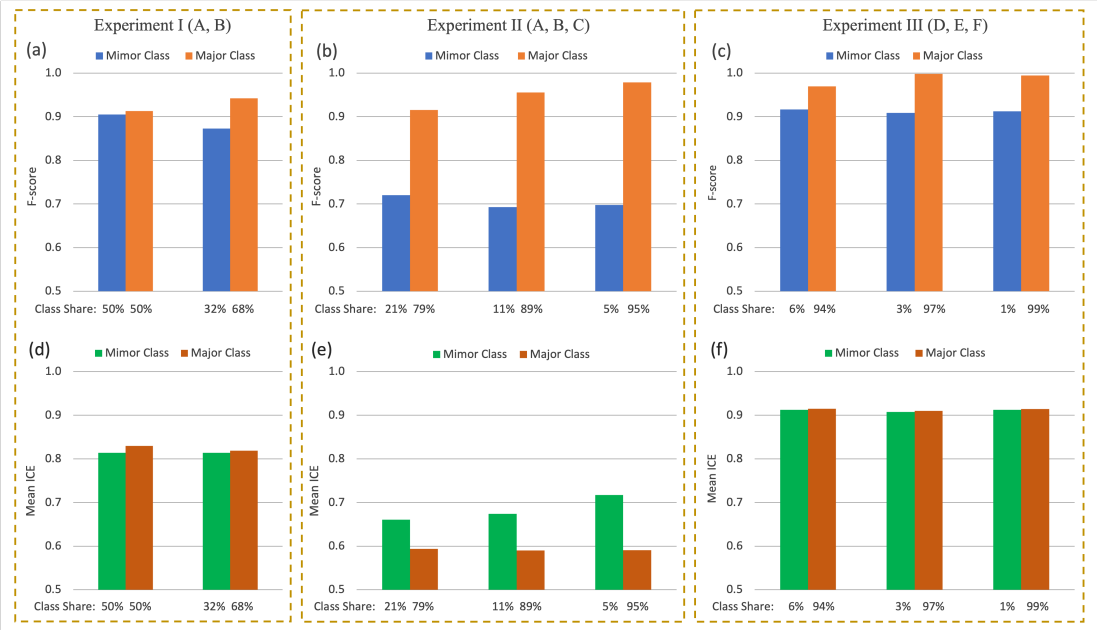


Fig. 2. Changes in F-score and Mean ICE with class share ratios in three segmentation experiments (Table II). (a)–(c) F-score. (d)–(f) Mean Efficacy (Table I).

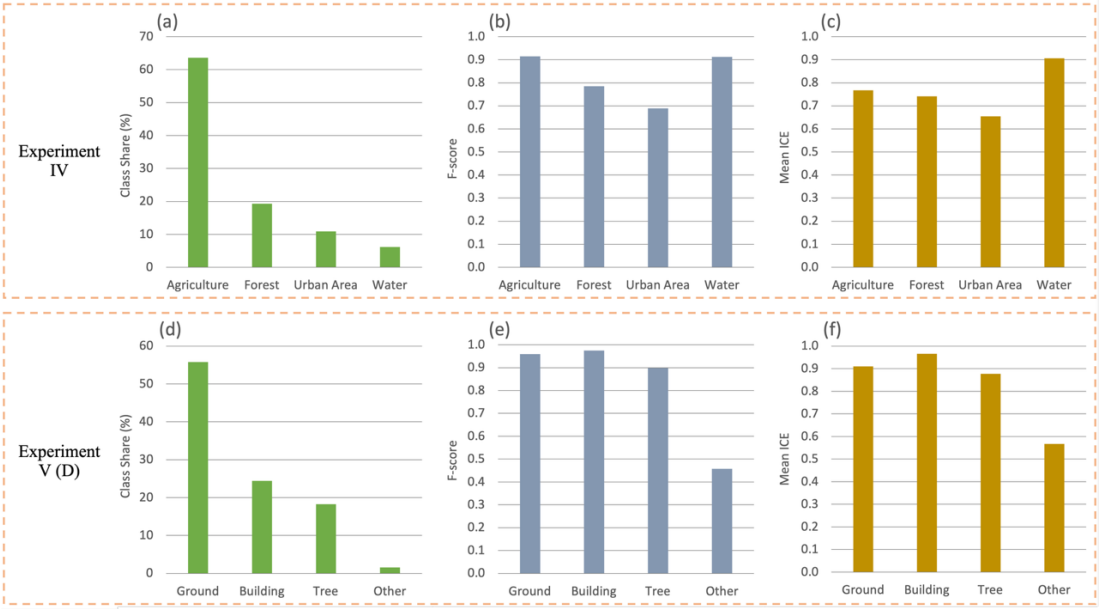


Fig. 3. Changes in per-class F-score and Mean ICE with class share distribution in two segmentation experiments with multiple classes (Table II). (a) & (d) Class share distribution. (b) & (e) F-score. (c) & (f) Mean ICE (Table I).

Agriculture has a much lower ICE value (0.768) than Water (0.906; Figs. 3a and 3c and Table II). When a class is severely rare (e.g., Other in Fig. 3d), its accuracy can be rather low (Fig. 3e), but its efficacy may not be as low relative to that of other classes (Fig. 3f). The efficacy histograms suggest that Water and Building are segmented most effectively among the four classes in Experiment IV, and Urban Area and Other are segmented least effectively among the four classes in Experiment V.

Class share distribution does not typically affect whole-class efficacy. Overall accuracy is proportionate to the unevenness of class share distribution, particularly when image segmentation is at an ordinary accuracy level (Figs. 4a and 4d). This trend becomes less notable when image segmentation is highly accurate (Figs. 4b and 4c). Conversely, the distribution pattern of overall accuracy differs for MICE, being independent of class share unevenness (Fig. 4). The stable MICE values occur in binary segmentations (Figs. 4a and 4b), demonstrating its insensitivity to class share distribution. In multiclass segmentation, the variations in MICE values show different compositions of confused multiple classes between reference datasets (Figs. 4c and 4d and Table II).

Efficacy amplifies the signal for the performance of segmentation models. Let p_1 be the share proportion of class

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2024.3446950

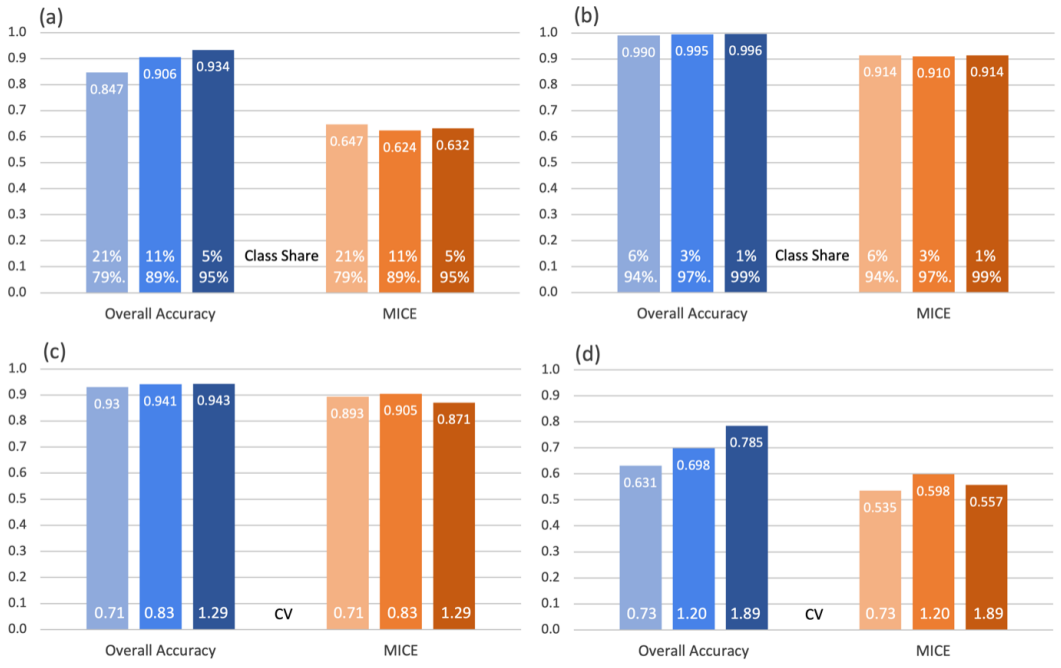


Fig. 4. Changes in overall accuracy and MICE with class share distributions (Table II). (a) & (b) Binary segmentations. (c) & (d) multiclass segmentations. CV stands for Coefficient of Variance for class share proportions.

1, then the MICE of binary segmentation is computed as (Table I):

$$\text{MICE} = \frac{A-p_1^2-(1-p_1)^2}{1-p_1^2-(1-p_1)^2} = \frac{A}{2p_1(1-p_1)} - \frac{p_1^2+(1-p_1)^2}{2p_1(1-p_1)} \qquad (2)$$

Thus, the slope with respect to A of MICE is inversely proportional to $2p_1(1-p_1)$. The minimum slope is 2 under an even share distribution ($p_1 = 0.5$). The more skewed the share distribution, the greater the slope (Fig. 5a), which implies that when class shares are extremely uneven, a small difference in overall accuracy can lead to a large difference in MICE; thus, MICE can enhance the expression of segmentation quality. For example, in Experiment III, two deep learning models, PSI-CD and FC-EF, were tested against the same reference data with a class ratio of $99.2 : 0.8$, which resulted in similarly high overall accuracy (0.996 and 0.990) but substantially different MICE values (0.914 and 0.384, Table II), confirming that the PSI-CD model is much more effective than the FC-EF model for image segmentation (Fig. 5b).

IV. Discussion

The outcomes of the six experiments primarily focus on the comparison between accuracy and efficacy metrics as a means of assessing image segmentations. Further interpretation of the results is needed to fully elucidate the significance of the transformed practice of segmentation evaluation with ICE.

Accuracy metrics lead to inconsistent evaluations of model performance between training and application. The reproducibility of deep learning models means their consistent performance for image segmentation [9]. In remote sensing, image segmentation is aimed to generate maps. Traditionally, one trained image segmentation model normally results in a single immediate map product. Contemporary segmentation using deep learning makes it possible to transfer a trained model to a new geographic location where new maps are made, and consistent performance of deep learning models is crucial. Because the test data used for evaluating a trained a deep learning model by the deep learning engineer may be different from the data from a user in terms of class share distribution, accuracy values may not be repeatable (Fig. 6). This type of variation in accuracy is inevitable if a deep learning model is trained with lab-controlled benchmark data but it is applied with real-world data [52]. For example, if the class of interest is a dominant class for the test data with training but it is a rare class with application, its F-score would be lowered in application (Fig. 6a), signifying the reproducibility problem. The way overall accuracy is influenced by class share distribution is different from F-score. Because class share distribution affects accuracy values, a confusion matrix must be a representative of the real-world population [26]. Perhaps this ought to be a gold standard for accuracy assessment in remote sensing. In this regard, geographically stratified partitioning techniques are not recommended if the study area is not uniform or class proportions change across a landscape. Traditionally, stratified random sample points are located by using the map product under assessment to generate a confusion matrix. When multiple maps for the same geographical area need to be assessed, such stratified random sampling is not an efficient option because class proportions may vary among these maps. It is therefore easier if reference information is collected with simple random sampling for the entire mapping area though the initial investment may be high. This practice is similar to that benchmark data are segmentate by using different models.

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2024.3446950

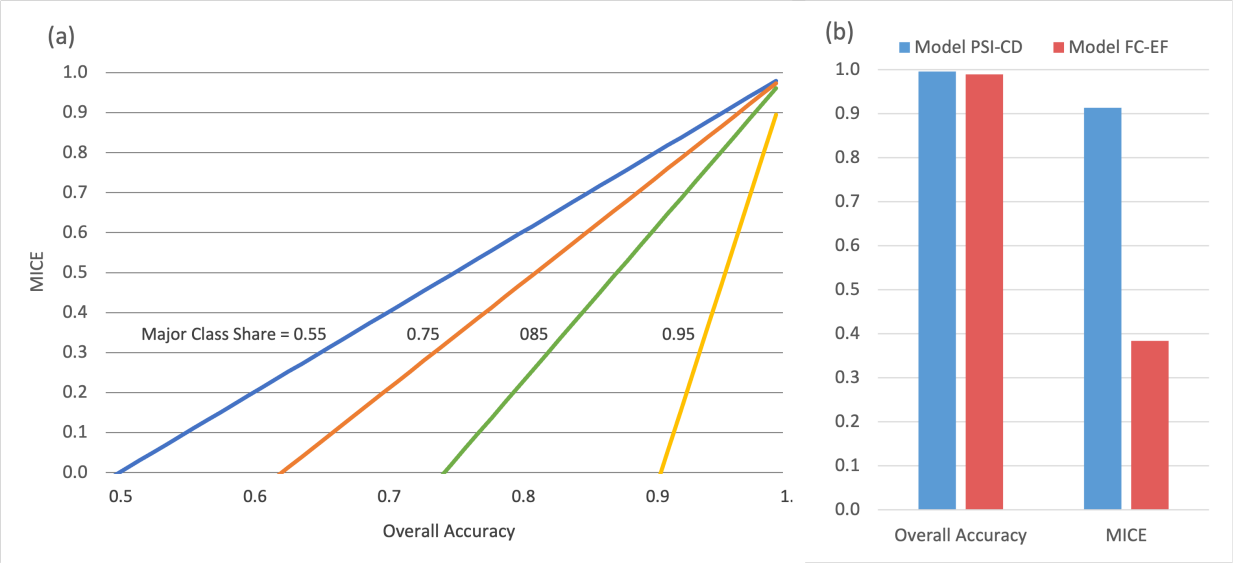


Fig. 5. MICE-overall accuracy relationship. (a) Changes in slopes of their linear relationship based on Eq. 2. (b) Results of Segmentation Experiment III C and F with class share proportions of 99.2 and 0.8 (Table II).

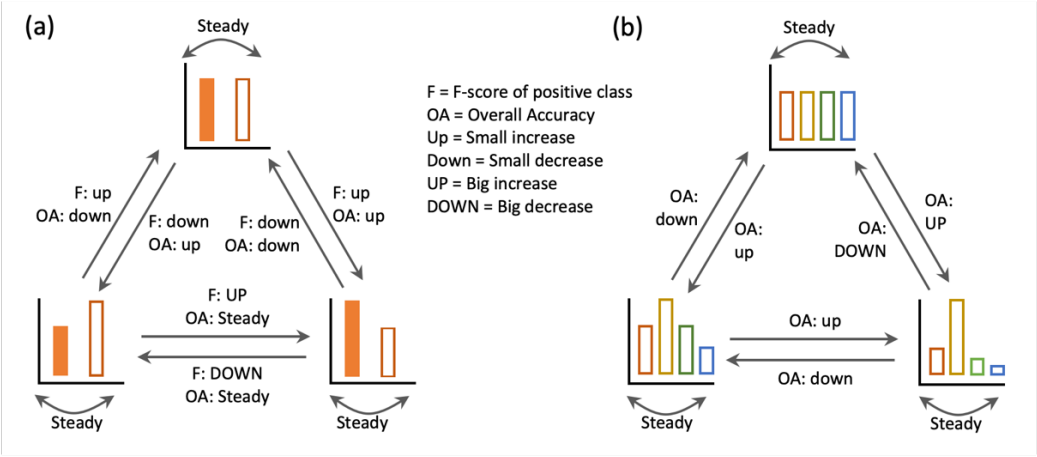


Fig. 6. Illustration model's inconsistent behaviors indicated by varying accuracy between model training and application due to changes in class share proportions of testing data. (a) Binary-class segmentation. (b) Multi-class segmentation.

Because deep learning models are usually trained on large amount reference data, even a fraction (e.g., 10%) of the reference data would be sufficient for generating a population confusion matrix for a map under evaluation.

Jointly using accuracy metrics cannot resolve the class imbalance effect. If the overall quality is intended to capture every facet of map quality, it necessitates the incorporation of multiple accuracy metrics [53]. Typical binary classification results in a confusion matrix involving a positive class and a negative class [20]. In cases where either class can be considered the positive or interesting class, it is possible to compute precision and recall for each class pair, as is done in multiclass scenarios. In the hypothetical example above (Experiment I; see Fig. 1), the true positive rate of each class remained relatively constant despite changes in class proportions, resulting in unchanged recall values for each class between the two segmentations. This scenario supports the view that recall is theoretically insensitive to class imbalance [54]. However, because precision is sensitive to class imbalance, F-score is still proportionate to class share proportion. Using multiple accuracy metrics is essential because different metrics express different aspects of segmentation quality. Nevertheless, it is important to note that many accuracy metrics are correlated [26], and therefore using multiple metrics may only partially avoid the class imbalance effect. For example, researchers often use mean F-score and mean IoU, which are effective in many cases [5], [46]–[51]. Plotting 182 pairs of these two metrics from these publications reveals an almost one-to-one correlation ($R^2 = 0.99$; Fig. 7a). Although the correlation between mF and overall accuracy is not as strong ($R^2 = 0.84$; Fig. 7b), it is worth noting that overall accuracy and mF are significantly related. Since overall accuracy suffers from the class imbalance effect, mF and mIoU also influenced by the class imbalance effect.

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2024.3446950

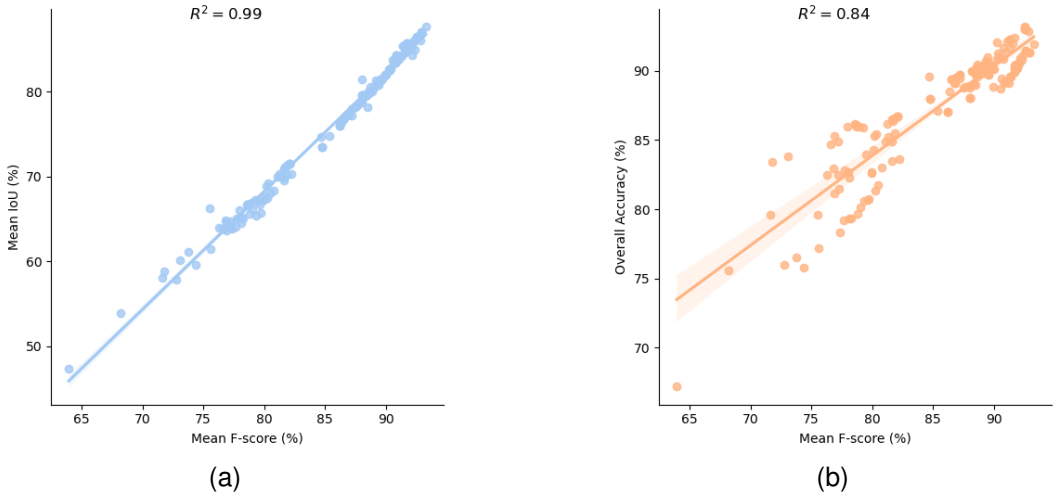


Fig. 7. Scatter plots between mean F-score and mean IoU and between mean F-score and overall accuracy (N = 182) [5], [46]–[51]. (a) Scatter plot between mean F-score and mean IoU. (b) between mean F-score and overall accuracy.

Efficacy metrics target model's real segmentation power. The performance of a model with moderate accuracy and less pronounced class imbalances is more sensitive to class distribution when assessed using accuracy metrics, whereas per-class and whole-class efficacy values remain relatively stable (Figs. 2 and 4). In such scenarios, ICE values with slight variations are consistent indicators of model performance. On the contrary, in highly skewed distributions, the overall accuracy may approach 1, and slight fluctuations in overall accuracy can lead to significant changes in MICE (Fig. 5). Under such circumstances, MICE serves as an amplifying indicator of model performance. The stable and fluctuating responses of ICE metrics to varying class share distributions and accuracy values are both crucial expressions of a model's real segmentation power. More specifically, per-class efficacy enhances the detection of classes for the precisely fine-tuning of segmentation models, and whole-class efficacy amplifies the signal of model performance for the overall comparison of segmentation models.

When confronted with an extremely rare class in image segmentation, additional efforts are often made to enhance its segmentation [3], [27]–[29], [31], [34], [55], [56]. This class imbalance problem in semantic segmentation sometimes can be challenging for deep learning engineers, especially when the rare class holds significance for accurate segmentation. The use of accuracy metrics sensitive to class share distribution can exacerbate this issue [3], [32]. When evaluating segmentation using ICE metrics, a small class may not necessarily be the least effective. As depicted in Fig. 2e, ICE metrics prove especially valuable when dealing with highly uneven class share distributions.In this regard, ICE metrics should be incorporated into the optimization process of a deep learning model. By using ICE metrics in the training stage, model's inconsistent behaviors from training to application can be relieved, and the segmentation strategies of individual classes, rare or common, can be determined without many influences of class imbalance. Deciding whether to prioritize the improvement of rare class segmentation depends on the ICE values of individual classes.

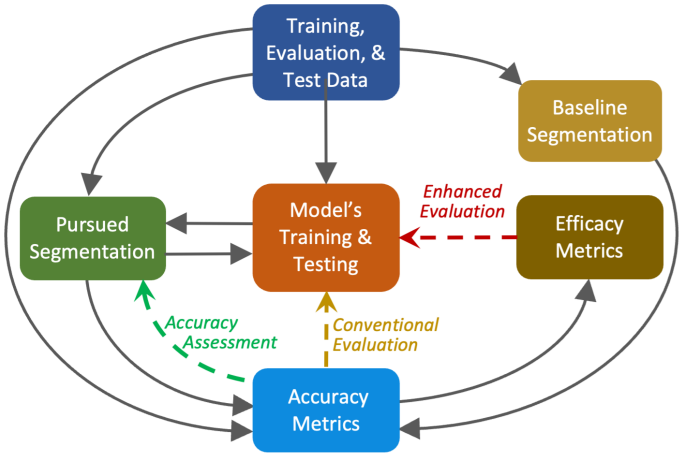


Fig. 8. Graphic comparison of accuracy assessment and model performance evaluation, including conventional and enhanced evaluation approaches.

The traditional accuracy assessment of segmentation outcomes and contemporary evaluations of deep learning model performance are two separate but related tasks (Fig. 8). The use of class imbalance-sensitive accuracy metrics to quantify model performance introduces uncertainty regarding the model's real segmentation power. Therefore, reliance on accuracy metrics alone is insufficient for evaluations of model performance. In contrast, ICE metrics are designed to determine the effectiveness of image segmentation, and are both interpretable and resilient to the class imbalance effect. Efficacy metrics can be used in the same way as accuracy metrics in model's validation during training. The incorporation of ICE metrics into model's validation can improve feedback regarding the optimization and fine-tune of deep learning models. Considering that foundation models have begun to gain popularity in many fields involving image segmentation and related tasks [6], [57], the integrated use of accuracy and efficacy metrics is becoming essential to ensure

model's reliable applications.

ROC-AUC and PR-AUC vs ICE. Deep learning models are often evaluated along a range of decision probability thresholds during its training. The overall performance of a model can be shown by the area under the receiver operating characteristic curve (ROC-AUC) for a positive class in binary segmentation. In remote sensing, an ROC-AUC is a plot between false positive rate (FPR) as x and true positive rate (TPR) (also called recall) as y at different decision thresholds (Fig. 9a) [22]. The greater an ROC-AUC value, the stronger mode's performance. The total ROC-AUC consists of two parts divided by class size-independent random assignment line when TPR = FPR. The area below the 'base line' is a constant (0.5) and thus, the performance of a model is characterized by the area above the baseline assuming TPR $>$ FPR. At a given decision threshold, the segmentation power of a model is proportional to the difference between TPR and FPR. Despite TPR and FPR are individually insensitive to class imbalance [54], ROC-AUC's tolerance to the class imbalance effect has not reached a broad consensus. Perhaps this is because the baseline does not consider class imbalance. For multi-class segmentation, which is common in remote sensing, ROC-AUC needs to be adjusted [28].

The area under precision-recall curve (PR-AUC) is also used to evaluate the performance of deep learning models (Fig. 9b). PR-AUC is computed by including or excluding the area below the class proportion baseline (i.e., n_j/n) [25], [26], [28]. Past studies show its varied degrees of sensitivity to class imbalance. Precision is sensitive to class imbalance but recall is not; the effective area of PR-AUC, after deducting the area below n_j/n baseline, can reduce the class imbalance effect (Fig. 9b). The effective area of PR-AUC is theoretically consistent with ICE metrics because both consider the random classification baseline defined as n_j/n. Therefore, PR-AUC and ICE are supposed to be mutually supportive.

ROC-AUC and PR-AUC are used for per-class evaluations while ICE metrics are useful for per-class and whole-class evaluations; ROC-AUC and PR-AUC are computed with multiple confusion matrices obtained from a model under training while ICE metrics are derived from a single matrix that can be obtained from a single execution of a trained model.

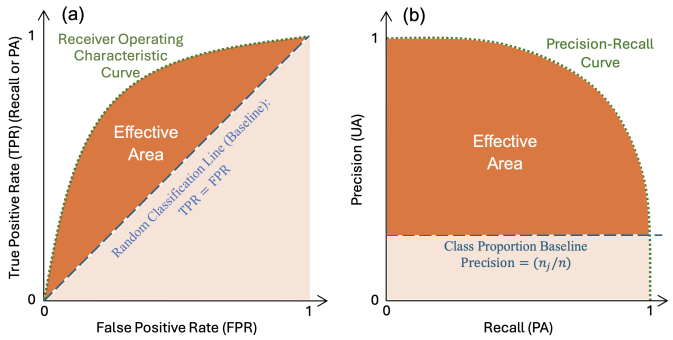


Fig. 9. Illustration of the area under a curve as a metric for the evaluation of model performance for a set of decision thresholds. (a) Area under the receiver operating characteristic curve (ROC-AUC). (b) Area under precision-recall curve (PR-AUC). Each AUC contains an effective area, critical for the evaluation of model performance.

V. Conclusions

The primary motivation behind evaluating segmentation models' performance is to ensure their reliability for their applications. A fundamental requirement is that these models demonstrate consistent prediction accuracy when implemented with new data. Because accuracy metrics are largely sensitive to class share variations, they struggle with the characterization of consistent performance. Consequently, solely relying on accuracy to assess segmentation outcomes may not reflect model proficiency. In contrast, the metrics of image classification efficacy consider class imbalances and offer unique insights crucial for the evaluation of model performance. By utilizing both per-class and whole-class efficacy metrics, we can effectively compare segmentation qualities and fine-tune models with precision. Integrating efficacy metrics into model training and testing enhances models' reliability for their operational use. It is hopeful that this efficacy-reinforced approach will contribute to a higher standard for evaluating deep learning models that are used to address real-world challenges.

The way this enhanced evaluation approach is implemented is shown in the six samples of this paper. Although these examples are from the field of remote sensing, the evaluation technique is applicable in image segmentation in other fields, such as medical imaging. Consequently, the enhanced information on model performance can help narrow the chasm from model evaluation to clinical impact [10], [58].

The six segmentation experiments were limited to semantic segmentation. Scene segmentation and object-based segmentation are not considered in this study. These examples are from remote sensing but segmentation cases from other fields need to be investigated with ICE metrics. This study did not consider the impacts of ICE metrics on the estimation of application variables, such as class area and change rates. The segmentation experiments in this study did not incorporate ICE metrics into model's training. In particular, we did not conduct segmentation experiments with repeated random data partitioning.

Acknowledgments

This study was supported by the National Key Research and Development Program of China (no. 2022YFF1301303).

References

[1] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 7, pp. 3523–3542, 2021.

[2] D. Müller, I. Soto-Rey, and F. Kramer, "Towards a guideline for evaluation metrics in medical image segmentation," *BMC Res. Notes*, vol. 15, no. 1, p. 210, 2022.

[3] M. Yeung, E. Sala, C.-B. Schönlieb, and L. Rundo, "Unified focal loss: Generalising dice and cross entropy-based losses to handle class imbalanced medical image segmentation," *Comput. Med. Imag. Grap.*, vol. 95, p. 102026, 2022.

[4] B. Janga, G. P. Asamani, Z. Sun, and N. Cristea, "A review of practical ai for remote sensing in earth sciences," *Remote Sens.*, vol. 15, no. 16, p. 4112, 2023.

[5] Y. Mo, H. Li, X. Xiao, H. Zhao, X. Liu, and J. Zhan, "Swin-conv-dspp and global local transformer for remote sensing image semantic segmentation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 5284–5296, 2023.

[6] J. Ma, Y. He, F. Li, L. Han, C. You, and B. Wang, "Segment anything in medical images," *Nat. Commun.*, vol. 15, no. 1, p. 654, 2024.

[7] F. Renard, S. Guedria, N. D. Palma, and N. Vuillerme, "Variability and reproducibility in deep learning for medical image segmentation," *Sci. Rep.*, vol. 10, no. 1, p. 13724, 2020.

[8] B. J. Heil, M. M. Hoffman, F. Markowetz, S.-I. Lee, C. S. Greene, and S. C. Hicks, "Reproducibility standards for machine learning in the life sciences," *Nat. Methods*, vol. 18, no. 10, pp. 1132–1135, 2021.

[9] R. F. Laine, I. Arganda-Carreras, R. Henriques, and G. Jacquemet, "Avoiding a replication crisis in deep-learning-based bioimage analysis," *Nat. Methods*, vol. 18, no. 10, pp. 1136–1144, 2021.

[10] G. Shao, H. Zhang, J. Shao, K. Woeste, and L. Tang, "Strengthening machine learning reproducibility for image classification," pp. 471–476, 2022.

[11] Y. Mo, Y. Wu, X. Yang, F. Liu, and Y. Liao, "Review the state-of-the-art technologies of semantic segmentation based on deep learning," *Neurocomputing (Amst)*, vol. 493, pp. 626–646, 2022.

[12] J. Lv, Q. Shen, M. Lv, Y. Li, L. Shi, and P. Zhang, "Deep learning-based semantic segmentation of remote sensing images: a review," *Front. Ecol. Evol.*, vol. 11, p. 1201125, 2023.

[13] Q. Weng, Z. Mao, J. Lin, and W. Guo, "Land-use classification via extreme learning classifier based on deep convolutional features," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 5, pp. 704–708, 2017.

[14] Y. Li, H. Zhang, X. Xue, Y. Jiang, and Q. Shen, "Deep learning for remote sensing image classification: A survey," *Wiley Interdiscip. Rev.: Data Min. Knowl. Discovery*, vol. 8, no. 6, p. e1264, 2018.

[15] A. Luque, A. Carrasco, A. Martín, and A. de Las Heras, "The impact of class imbalance in classification performance metrics based on the binary confusion matrix," *Pattern Recogn.*, vol. 91, pp. 216–231, 2019.

[16] C. A. Ramezan, T. A. Warner, and A. E. Maxwell, "Evaluation of sampling and cross-validation tuning strategies for regional-scale machine learning classification," *Remote Sens.*, vol. 11, no. 2, p. 185, 2019.

[17] B. Manifold, S. Men, R. Hu, and D. Fu, "A versatile deep learning architecture for classification and label-free prediction of hyperspectral images," *Nat. Mach. Intell.*, vol. 3, no. 4, pp. 306–315, 2021.

[18] X. Wu, D. Hong, and J. Chanussot, "Convolutional neural networks for multimodal remote sensing data classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–10, 2021.

[19] L. Ma, M. Li, X. Ma, L. Cheng, P. Du, and Y. Liu, "A review of supervised object-based land-cover image classification," *ISPRS J. Photogramm. Remote Sens.*, vol. 130, pp. 277–293, 2017.

[20] K. M. Ting, *Confusion matrix*. Springer US, 2011, pp. 209–209.

[21] C. Liu, P. Frazier, and L. Kumar, "Comparative assessment of the measures of thematic classification accuracy," *Remote Sens. Environ.*, vol. 107, no. 4, pp. 606–616, 2007.

[22] D. Powers, "Evaluation: From precision, recall and f-measure to roc, informedness, markedness correlation," *J. Mach. Learn Tech.*, vol. 2, no. 1, pp. 37–63, 2011.

[23] D. Hand and P. Christen, "A note on using the f-measure for evaluating record linkage algorithms," *Stat. Comput.*, vol. 28, pp. 539–547, 2018.

[24] R. G. Congalton and K. Green, *Assessing the accuracy of remotely sensed data: principles and practices*. CRC press, 2019.

[25] R. Padilla, S. L. Netto, and E. A. Da Silva, "A survey on performance metrics for object-detection algorithms," in *2020 international conference on systems, signals and image processing (IWSSIP)*. IEEE, 2020, pp. 237–242.

[26] 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.

[27] M. Saini and S. Susan, "Tackling class imbalance in computer vision: a contemporary review," *Artif. Intell. Rev.*, vol. 56, no. Suppl 1, pp. 1279–1335, 2023.

[28] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural Networks*, vol. 106, pp. 249–259, 2018.

[29] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," *J. Big Data*, vol. 6, no. 1, pp. 1–54, 2019.

[30] G. Shao, L. Tang, and J. Liao, "Overselling overall map accuracy misinforms about research reliability," *Landscape Ecol.*, vol. 34, pp. 2487–2492, 2019.

[31] S. Susan and A. Kumar, "The balancing trick: Optimized sampling of imbalanced datasets—a brief survey of the recent state of the art," *Eng. Rep.*, vol. 3, no. 4, p. e12298, 2021.

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

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

[34] M. H. Hesamian, W. Jia, X. He, and P. Kennedy, "Deep learning techniques for medical image segmentation: achievements and challenges," *J. Digital Imaging*, vol. 32, pp. 582–596, 2019.

[35] K. Karra, C. Kontgis, Z. Statman-Weil, J. C. Mazzariello, M. Mathis, and S. P. Brumby, "Global land use/land cover with sentinel 2 and deep learning," in *2021 IEEE international geoscience and remote sensing symposium IGARSS*. IEEE, 2021, pp. 4704–4707.

[36] Y. Wang, Y. Wan, Y. Zhang, B. Zhang, and Z. Gao, "Imbalance knowledge-driven multi-modal network for land-cover semantic segmentation using aerial images and lidar point clouds," *ISPRS J. Photogramm. Remote Sens.*, vol. 202, pp. 385–404, 2023.

[37] J. Cohen, "A coefficient of agreement for nominal scales," vol. 20, no. 1, pp. 37–46, 1960.

[38] 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, 2011.

[39] P. Olofsson, G. M. Foody, M. Herold, S. V. Stehman, C. E. Woodcock, and M. A. Wulder, "Good practices for estimating area and assessing accuracy of land change," *Remote Sens. Environ.*, vol. 148, pp. 42–57, 2014.

[40] 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, p. 111630, 2020.

[41] G. Shao, L. Tang, and H. Zhang, "Introducing image classification efficacies," *IEEE Access*, vol. 9, pp. 134 809–134 816, 2021.

[42] Y. Zheng, L. Tang, and H. Wang, "An improved approach for monitoring urban built-up areas by combining npp-viirs nighttime light, ndvi, ndwi, and ndbi," *J. Clean. Prod.*, vol. 328, p. 129488, 2021.

[43] V. Mnih, *Machine learning for aerial image labeling*. University of Toronto (Canada), 2013.

[44] S. Pang, X. Li, J. Chen, Z. Zuo, and X. Hu, "Prior semantic information guided change detection method for bi-temporal high-resolution remote sensing images," *Remote Sens.*, vol. 15, no. 6, p. 1655, 2023.

[45] A. E. Maxwell, B. T. Wilson, J. J. Holgerson, and M. S. Bester, "Comparing harmonic regression and glad phenology metrics for estimation of forest community types and aboveground live biomass within forest inventory and analysis plots," *Int. J. Appl. Earth Obs.*, vol. 122, p. 103435, 2023.

[46] L. Chen, X. Dou, J. Peng, W. Li, B. Sun, and H. Li, "Efcnet: ensemble full convolutional network for semantic segmentation of high-resolution remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2021.

[47] L. Gao, H. Liu, M. Yang, L. Chen, Y. Wan, Z. Xiao, and Y. Qian, "Stransfuse: Fusing swin transformer and convolutional neural network for remote sensing image semantic segmentation," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 14, pp. 10 990–11 003, 2021.

[48] R. Li, S. Zheng, C. Zhang, C. Duan, J. Su, L. Wang, and P. M. Atkinson, "Multiattention network for semantic segmentation of fine-resolution remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–13, 2021.

[49] X. Meng, Y. Yang, L. Wang, T. Wang, R. Li, and C. Zhang, "Class-guided swin transformer for semantic segmentation of remote sensing imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[50] Y. Xu, S. Zhou, and Y. Huang, "Transformer-based model with dynamic attention pyramid head for semantic segmentation of vhr remote sensing imagery," *Entropy-switz.*, vol. 24, no. 11, p. 1619, 2022.

[51] X. Meng, L. Zhu, Y. Han, and H. Zhang, "We need to communicate: Communicating attention network for semantic segmentation of high-resolution remote sensing images," *Remote Sens.*, vol. 15, no. 14, p. 3619, 2023.

[52] J. Thiyagalingam, M. Shankar, G. Fox, and T. Hey, "Scientific machine learning benchmarks," *Nat. Rev. Phys.*, vol. 4, no. 6, pp. 413–420, 2022.

[53] 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, 2020.

[54] A. Tharwat, "Classification assessment methods," *Appl. Comput. Inf.*, vol. 17, no. 1, pp. 168–192, 2020.
[55] F. M. Megahed, Y.-J. Chen, A. Megahed, Y. Ong, N. Altman, and M. Krzywinski, "The class imbalance problem," *Nat. Methods*, vol. 18, no. 11, pp. 1270–7, 2021.
[56] Q. Du Nguyen and H.-T. Thai, "Crack segmentation of imbalanced data: The role of loss functions," *Eng. Struct.*, vol. 297, p. 116988, 2023.
[57] Z. Yan, J. Li, X. Li, R. Zhou, W. Zhang, Y. Feng, W. Diao, K. Fu, and X. Sun, "Ringmo-sam: A foundation model for segment anything in multimodal remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–16, 2023.
[58] H. R. Sofaer, J. A. Hoeting, and C. S. Jarnevich, "The area under the precision-recall curve as a performance metric for rare binary events," *Methods Ecol. Evol.*, vol. 10, no. 4, pp. 565–577, 2019.

Lina Tang Lina Tang received the B.S. and M.S. degrees in ecology from Northeast Normal University, Changchun, China, in 2000 and 2003, respectively, and the Ph.D. degree in ecology from the Institute of Applied Ecology, Chinese Academy of Sciences, Shenyang, China, in 2006.

She is currently a Professor with Institute of Urban Environment (IUE), Chinese Academy of Sciences, and a director of Urban Ecological Environment Planning and Management Research Center, IUE. Her research interests include urban ecology, human-nature interactions, urban remote sensing, and landscape ecology.

Jinyuan Shao received the B.S. in information engineering from Huaqiao University, Xiamen, China, and the M.S. in Ecology from the University of Chinese Academy of Sciences, Beijing, China, in 2018. He is currently a PhD candidate at Purdue University.

His research interests are remote sensing image segmentation and LiDAR point cloud analysis and visualization by using computer vision and deep learning methods.

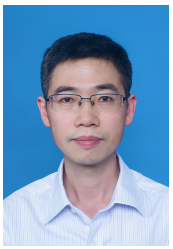
Shiyan Pang received the B.S. degree in Science and Technology of Surveying and Mapping from Wuhan University, Wuhan, China, in 2009, the M.S. and Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2012 and 2015, respectively. From 2016 to 2018, she was a Post Doctor with the Collaborative Innovation Center for Geospatial Technology and the School of Resource and Environmental Sciences, Wuhan University, Wuhan, China. She is an Associate Professor with the Faculty of Artificial Intelligence in Education, Central China Normal University, Wuhan, China. Her research interests include machine learning, deep learning, and semantic segmentation and change detection of remotely sensed data.

Yameng Wang received the B.S. degree in remote sensing science and technology in 2018 from Wuhan University, Wuhan, China, where she is currently working toward the Ph.D. degree in photogrammetry and remote sensing with Wuhan University, Wuhan, China.

Her research interests include remote sensing image processing and machine learning.

Aaron Maxwell received his BS in Chemistry, Biology, and Environmental Science from Alderson Broaddus College. He completed his MS and PhD in Geology at West Virginia University.

He is currently an Assistant Professor in the Department of Geology and Geography at West Virginia University with a research focus in remote sensing and spatial predictive mapping and modeling.



Xiangyun Hu received the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2001. From 2002 to 2005, he was a Post-Doctoral Research Fellow with the Department of Earth and Space Science and Engineering, Lassonde School of Engineering, York University, Toronto, ON, Canada. He has developed a feature extraction technology SmartDigitizer acquired by PCI Geomatics, Leica Geosystems, and Microsoft. From 2005 to 2010, he was a Senior Software Engineer with ERDAS Inc., Atlanta, GA, USA.

He is currently a Professor and the Head of the Department of photogrammetry with the School of Remote Sensing and Information Engineering, Wuhan University. He is also an Adjunct Professor with Hubei Luojia Laboratory, Wuhan. Recently, he has been leading a team developing an open-source deep learning framework—LuojiaNET. He has authored more than 60 papers in journals and conferences in intelligent feature extraction of remotely sensed data.

Zhi Gao (Member, IEEE) received the B.Eng. and Ph.D. degrees from Wuhan University, Wuhan, China, in 2002 and 2007, respectively. In 2008, he joined the Interactive and Digital Media Institute, National University of Singapore (NUS), Singapore, as a Research Fellow (A) and a Project Manager. In 2014, he joined the Temasek Laboratories, NUS (TL@NUS), as a Research Scientist (A) and a Principal Investigator.

He is currently working as a Full Professor with the School of Remote Sensing and Information Engineering, Wuhan University. He has published more than 70 research papers in top journals and conferences, such as *International Journal of Computer Vision* (IJCV), IEEE TRANSACTIONS ON PATTERN ANALYSIS ON AND MACHINE INTELLIGENCE (TPAMI), IEEE TRANSACTIONS ON INDUSTRIAL ELECTRONICS (TIE), IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING (TGRS), IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS (TITS), *ISPRS Journal of Photogrammetry and Remote Sensing (JPRS)*, *Neurocomputing*, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY (TCSVT), the Conference on Computer Vision and Pattern Recognition (CVPR), the European Conference on Computer Vision (ECCV), the Asian Conference on Computer Vision (ACCV), and the British Machine Vision Conference (BMVC). Since 2019, he has been supported by the Distinguished Professor Program of Hubei Province and the National Young Talent Program, China. His research interests include computer vision, machine learning, and remote sensing and their applications. In particular, he has a strong interest in vision for intelligent systems and intelligent system-based vision. Dr. Gao serves as an Associate Editor for *Unmanned Systems* journal.

Ting Lan received the M.S. degree in cartography and GIS from Fujian Normal University, Fuzhou, China, in 2018 and the Ph.D. degree in environment economy and environment management at the Institute of Urban Environment, Chinese Academy of Sciences, Xiamen, China, in 2022.

She is currently an assistant researcher of the Institute of Urban Environment, Chinese Academy of Sciences. Her research interests include environment effects of urban form and sustainable urban development.

Guofan Shao received the B.S. and M.S. in forestry from Northeastern Forestry University, Harbin, China, in 1982 and 1985, respectively, the Ph.D. degree in Ecology from the Chinese Academy of Sciences, in 1989. He received GIS certification from the Environmental Systems Research Institute (ESRI), Germany, in 1988, and postdoctoral education in computer modeling from the University of Virginia, USA, between 1991 and 1993.

He is currently a full Professor with Purdue University. He teaches remote sensing courses at the both undergraduate and graduate levels. He is among top 2% of world's most-cited scientists. His research interests include land use land cover mapping with remotely sensed imagery and accuracy assessment.