

SSEGEP: Small Segment Emphasized Performance Evaluation Metric for Medical Image Segmentation

Ammu Raju and Neelam Sinha

International Institute of Information Technology, Bangalore, India

Abstract

Automatic image segmentation is a critical component of medical image analysis, and hence quantifying segmentation performance is crucial. Challenges in medical image segmentation are mainly due to spatial variations of regions to be segmented and an imbalance in distribution of the classes. Commonly used metrics treat all detected pixels indiscriminately. However, pixels in smaller segments must be treated differently from pixels in larger segments, as smaller ones aid in early treatment of associated disease and are also easier to miss. To address this, we propose a novel evaluation metric for segmentation performance, emphasizing smaller segments, by assigning higher weightage to smaller segment pixels. "Small SEGment Emphasized Performance" (SSEGEP) metric, based on weighted false positives, was proposed and evaluated using synthetic images and 4 publicly available clinical datasets of eye (fundus imaging, $n = 33$), breast (mammogram, $n = 108$), liver (CT, $n = 131$), and pancreas cancer (CT, $n = 107$), where n refers to the number of images from the dataset that were used in the study. Mean opinion score (MOS) was calculated from the scores (5-scale scores) of 33-fundus-image segmentation as assigned by 15 researchers (2–5 years of experience in image analysis). Statistical analysis was performed for the other datasets to quantify the relevance of the proposed approach. Across 33 fundus images, where the largest exudate is 1.41% and the smallest is 0.0002% of the image, the proposed metric is 30% closer to MOS, as compared to Dice Similarity Coefficient (DSC). Statistical significance testing resulted in a p value of order 10^{-18} that shows the significance of SSEGEP for hepatic tumors compared to DSC. The proposed metric is found to perform better for the images having multiple segments of a single class.

Keywords: medical image segmentation, evaluation metric, segmentation evaluation, multiple segments, small lesion emphasized
DOI: 10.31526/JMLFS.2022.229

1. INTRODUCTION

Automated disease diagnosis requires automated segmentation of region of interest (ROI), on which subsequent analysis depends. Hence, it is important to evaluate segmentation performance. Segmentation is done in different ways depending on the image modality and characteristics of the region to be segmented. Some common methods include histogram analysis, region segmentation, and edge detection. In a typical CAD (*Computer-Aided Detection and Diagnosis*) system, image segmentation is usually followed by feature extraction and classification. One successful model used in medical image segmentation is U-net [1]. Since subsequent diagnosis depends on segmentation, it is important to utilize robust segmentation performance evaluation metrics.

General Image Segmentation and Medical Image Segmentation

Evaluation of medical image segmentation should not be done in a similar way to the general segmentation scenarios. Commonly used metrics include *accuracy*, *sensitivity*, *specificity*, Dice Similarity Coefficient (DSC, also known as F1-score), Intersection Over Union (IOU, also known as Jaccard Index), and boundary distance scores which are applicable to general segmentation scenarios. Information about the overlap area between ground truth and segmented image is used to evaluate the quality of general image segmentation method. However, medical image segmentation evaluation should also consider the diagnosis aspect, like what is more important in disease diagnosis. Some examples of medical segmentation scenarios include the detection and monitoring of tumor progress, which is important in surgical planning [2, 3, 4], and change patterns in white matter that assist in early diagnosis and treatment of brain diseases [5]. There are also medical scenarios where the size/volume of the segments conveys more information on disease grading/diagnosis. Metrics such as DSC, IOU, and volumetric overlap error are used to evaluate segmentation based on size/volume of region. In certain cases, the contour of the segment is of interest, which requires the segmentation algorithm to provide exact boundary delimitation. Boundary distance-based scores such as Hausdorff distance are used in such a scenario. However, in other cases, like in radiotherapy, the location of the segment may be of interest.

Segmentation Challenges in Medical Images

The region to be segmented in medical images is not always distinct and localized. There may be regions with variations in location, size, shape, and brightness. Preprocessing techniques usually take care of brightness and color. Other challenges include the following.

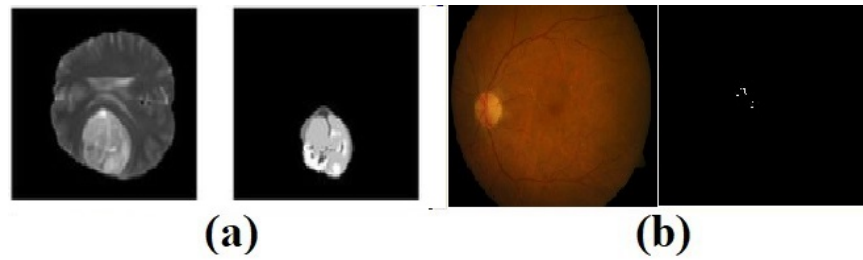


FIGURE 1: (a) MRI scan of glioma (images taken from BraTS dataset [6, 7, 8]), (b) fundus images (images taken from DIARETDB1 dataset [9]). In each of the image pairs, the image on the left is the original, while the image on the right is the segmentation ground truth hand-marked by experts. The fraction of pixels in the ROI, in each of the cases, is very small. The number of TN pixels is large in all these cases compared to the TP pixels.

- (1) Imbalance in data distribution: when the size of the segment is significantly smaller than the background, the metrics based on the calculation of true negatives (TN) are not suitable. The number of TN pixels is always larger in medical images as shown in Figure 1. *Accuracy*, *sensitivity*, *specificity*, and positive predictive value (PPV) are not suitable for pixel-level segmentation evaluation since TN pixels, which constitute the major portion of the image, are treated in a similar way to other pixels. Pixel *accuracy* and *specificity* are usually high for medical images since TN are more. Quality of segmentation cannot be measured only with *sensitivity* as it does not consider the false positives (FP). As PPV considers only the positive pixels, it cannot be used to assess the overall segmentation quality. Therefore, such classical metrics are not efficient in segmentation evaluation. DSC and IOU can be used in this case to overcome the above-stated drawback as they consider the overlap between ground truth and prediction.
- (2) Variation in size: another challenge faced in medical segmentation is due to the size variation of the regions to be segmented. The region may appear clustered and not localized at a point. Here, the challenge of the segmentation algorithm is to extract the tiniest segment. Smaller segments usually indicate the presence of associated disease at an earlier stage, and hence these small-sized segments are required to be emphasized more compared to larger ones. Commonly-used metrics like *sensitivity*, *specificity*, DSC, and IOU do not discriminate the segments based on the area. Therefore, it is imperative to design a metric that highlights the detection of smaller segments. Such a scenario can be seen in fundus images with exudates of various sizes as shown in Figure 2(a). A small protruding exudate may aid in the early detection of diabetic retinopathy. Another scenario is calcifications on mammograms as seen in Figure 2(b) which help to identify breast cancer.
- (3) Multiclass segmentation with class imbalance: multiclass segmentation arises when there are segments of different classes in a single image, and class imbalance occurs when there is a relative size difference in segments of each class (label). For example, the image in Figure 3(b) consists of pancreas tissue and a tumor (which occupies a smaller space compared to the pancreas) within it. The loss function of the deep learning network (for segmentation) favors the large-sized pancreas more than the tumor. However, the loss function should be efficient in extracting both segments. A commonly used loss function for medical segmentation is Dice loss, which is derived based on the evaluation metric DSC. However, here, generalized Dice loss [10] which works for multilabel cases should be used as it is not a simple binary segmentation. Another medical scenario is multiclass segmentation with more than one segment of various sizes in each category of class. Liver and tumor segmentation (Figure 3(a)) illustrate this situation with liver tissue and multiple hepatic tumors as the targets. In such a scenario, *generalized Dice score* is not sufficient to quantify the segmentation performance, as there may be segments of different sizes in each category of the label, and it is required to penalize the loss incurred by each pixel based on the area of the segment it belongs to.

Importance of Small-Sized Segments in Medical Images

- (1) Exudates are patches of fatty deposits on the eye formed due to leaky blood vessels. They appear as yellow-white structures in color fundus images with variable sizes, shapes, and contrast. As exudates are among the common early clinical signs of diabetic retinopathy (DR) [11], their detection would be critical in diagnosing the health status of the eye. The challenges faced in exudate segmentation include drastic variations in size (small to large as an optic disc) and shape and severe imbalanced distribution of classes (more number of nonexudate pixels). The size of diabetic hard exudates changes with the progression of disease [12]; hence, the identification of tiny exudates is very important as it helps in the early treatment. Therefore, the performance evaluation should be in such a way that it emphasizes the smaller exudates more. As the commonly used metrics treat all the TP pixels in a similar way, they are not able to quantify the small exudate segmentation. Qing Liu et al. [13] have used *sensitivity* and PPV for exudate segmentation evaluation. Sreeparna et al. [14] used *sensitivity*, *specificity*, *accuracy*, and area under curve of Receiver Operating Characteristic (ROC) as metrics to evaluate the detection of hard exudates using mean shift and normalized cut method.
- (2) Another medical segmentation scenario is cancer detection. Volume scans using CT for anatomies such as abdomen are used to detect tumors [15]. These tumors show diversity in location, size, and the way they are scattered across the anatomy, leading to several challenges in their precise segmentation. Other examples in cancer detection include using MRI to detect brain [16] and spinal cord [17] tumor. Pancreatic tumor is smaller in size compared to the large pancreatic tissue, and a common algorithm for multiclass segmentation (pancreatic tissue and tumor) is a challenge faced in segmentation. The loss

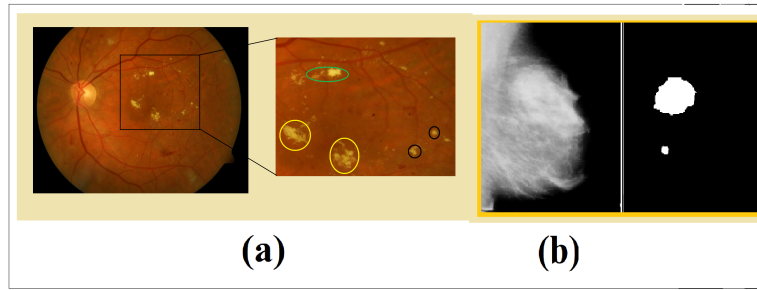


FIGURE 2: Variation in size of regions to be segmented. (a) Fundus image with exudates of various sizes circled in distinct colors. Largest exudates circled in yellow, medium-sized ones in green, and smaller ones in black. (b) Mammogram (left) and ground truth with calcifications (right) of varying sizes. Calcifications are shown in white color.

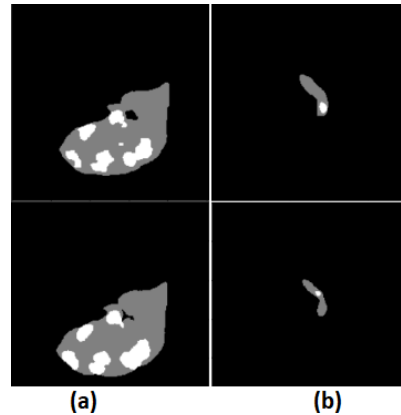


FIGURE 3: Ground truth images and segmentation showing two different classes (tissue and tumor). Top row shows the ground truth slice and bottom row shows corresponding segmentation. (a) Liver (gray color) with tumors of varying sizes in white color. (b) Pancreas (gray color) and tumor (white color).

function used in training deep networks should overcome this imbalance due to size differences. Most recent works use cross-entropy and *generalized Dice score* [10] in this scenario. Numerous lesions of varying sizes may appear in liver tissue called hepatic lesions, which can be malignant. Studies [18] have revealed the prevalence of small hepatic lesions found at CT in patients with cancer. These small lesions represent metastases in 11.6% of patients as identified by the research study. The size of the lesions also changes during the course of cancer, and patients with small hepatic metastases also had associated extra hepatic metastases.

- (3) Calcifications are the earliest signs of breast cancer. Information about the size, density, and distribution of breast microcalcifications can give an idea about the benign or malignant nature of cancer. It is found that decreasing calcifications is not a benign finding on mammogram [19]. Spontaneous decreases in calcifications are not necessarily indicative of better clinical outcomes. It has been demonstrated that a decrease in or complete resolution of breast calcifications is most concerning when it is associated with an extra breast mass, architectural distortion, or increased density [20, 21]. Small clusters of punctate or granular calcifications may represent high-grade DCIS (ductal carcinoma in situ), where an aggressive clinical approach is recommended [22]. If the calcifications are larger in size and well defined in shape, they are less suspicious. However, if they appear smaller in size or appear in varying shapes, they are cancerous [23]. Hence, it is important to detect smaller calcifications in mammograms.

The medical scenarios explained above state that the detection of the smallest lesion is very critical in further medical analysis. The smallest exudate in Figure 2(a) occupies only 0.0002% of the image; the smallest tumor in Figure 3(a) and (b) occupies 0.0015% and 0.00004% of the respective images. A small protruding lesion may indicate the start of a disease. In most cases, one of the factors which are used to grade disease severity is the size/volume of the lesion or region of interest. As explained before, the size of the hard exudates increases during the course of diabetic retinopathy [12], and lesions become bigger with stages of cancer like hepatic cancer [18] and breast cancer. Hence, the detection of a tiny lesion helps in the early treatment of associated disease which can thereby avoid unnecessary biopsies. Furthermore, the detection of a small liver tumor may point to the presence of a tumor in another region [18]. These studies enunciate the idea that a performance evaluation metric that emphasizes these tiny lesions is imperative.

DSC is commonly used to evaluate various segmentation methods [1, 24, 25, 26, 27]. Carole Sudre et al. [10] have used *generalized Dice score* for segmentation assessment for unbalanced tasks. Nur et al. [28] used the saliency method based on region to detect exudates in retinal images of Diabetic Retinopathy. They have calculated *accuracy*, *sensitivity*, and *specificity* for the evaluation of the proposed method. Fang Lu et al. [29] used volumetric overlap error (VOE), relative volume difference (RVD), average sym-

metric surface distance (ASD), root mean square symmetric surface distance (RMSD), and maximum symmetric surface distance (MSD) to evaluate the deep learning algorithm with graph cut refinement to automatically segment liver in CT scans. Varduhi Yeghiazaryan et al. [30] have introduced a family of metrics to quantify the difference or similarity of segments by considering their average overlap in fixed-size neighborhoods of points on the boundaries of those segments.

To quantify the segmentation performance of images with varying sizes of segments (exudates or tumors), we require a metric that emphasizes the smaller-sized segments. As seen before, DSC assigns the same weightage to all the TP pixels irrespective of whether they are part of a large/small region, which seems inappropriate. The small segments should be penalized more while calculating the segmentation loss since the loss of a few pixels in a small segment is more severe than that in a large segment. The existing metrics to the best of our knowledge do not discriminate between the detection of larger and smaller segment pixels. DSC and IOU metrics reflect the human understanding when the dataset contains images with only large or only small segments with fewer false positives. For a clinical evaluation where screening has to be done, image-level evaluation using *sensitivity* and *specificity* is enough. However, to compare different segmentation methods or to optimize a single method, metrics that consider overlap between ground truth and segmented images are required. As medical datasets usually consist of segments of varying sizes, evaluation with DSC and IOU is not sufficient as they treat all segments in the same way. The proposed metric, which incorporates the area of segments and which penalizes the loss of smaller segments would be more effective in comparing the segmentation approaches. This paper is organized as follows: Section 2 discusses the commonly used evaluation metrics, Section 3 presents the proposed approach in detail, and Section 4 demonstrates the experimental results and analysis with the datasets used, followed by conclusion in Section 6.

2. EXISTING EVALUATION METRICS

Different evaluation metrics are used in practice to compare the effectiveness of various segmentation algorithms. The evaluation metric to be used depends on the problem and anatomy of the segmented region. To compare various medical segmentation methods, we need evaluation at the pixel level for the entire image. The metric definitions of the most popular evaluation methods are given in Table 1.

Metric	Formula
Accuracy (Acc)	$Acc = \frac{ I - G \cup S + G \cap S }{ I }$
True positive rate (TPR)/Sensitivity	$TPR = \frac{ G \cap S }{ G }$
True negative rate (TNR)/Specificity	$TNR = \frac{ I - S \cup G }{ I - G }$
False positive rate (FPR)	$FPR = \frac{ S - G \cap S }{ I - G }$
False negative rate (FNR)	$FNR = \frac{ S \cup G - S }{ G }$
Intersection over union (IOU)	$IOU = \frac{ G \cap S }{ G \cup S }$
Dice similarity coefficient (DSC)	$DSC = \frac{2 G \cap S }{ S + G }$
Generalized Dice score (GD)	$GD = 2 \frac{\sum_{i=1}^l \frac{1}{ G_i } G_i \cap S_i }{\sum_{i=1}^l \frac{1}{ G_i } (S_i + G_i)}$

TABLE 1: Most popular evaluation metrics and their definitions are given. “| |” represents the cardinality. G and S constitute the set of nonzero pixels in ground truth and prediction, respectively. $|I| = TP + FP + FN + TN$, “ l ” is the label length in multiclass segmentation.

An imbalance in data distribution is noticed in medical image segmentation where the number of TN pixels is higher compared to other pixels, which makes classical metrics like *accuracy* and *specificity* inappropriate for performance evaluation. Overlap metrics like DSC and IOU take into account the area of overlap between the ground truth and segmented image. Both these metrics measure the same aspects, and hence using them together as evaluation metrics is not required. Boundary distance-based scores calculate the distance between the boundaries of segments in ground truth and prediction. They are used when boundary delineation of segmentation is important. They do not take into account the size information of regions. It is not recommended that Hausdorff distance be directly used in medical image segmentation as it is sensitive to outliers [31] and time-consuming.

If images in the dataset consist of segments of various sizes, the segmentation algorithm should be able to detect most of the pixels in the smaller segments, even if it misses a few larger segment pixels. Thus, to assess the quality of prediction, the evaluation metric should be able to penalize more for smaller segment pixels than for larger segment pixels. This aspect is not considered in simple overlap and boundary distance-based metrics. For the evaluation of multiclass segmentation performance as in pancreatic tumor, generalized *Dice* score [10] which gives a weightage of the inverse of the total area for each label is used. However, if

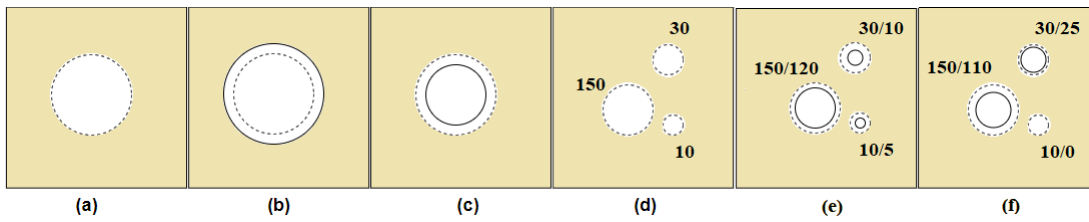


FIGURE 4: Synthetic images to illustrate the limitations of evaluation metrics: (1) boundary distance-based score and (2) Dice Similarity Coefficient. (b) and (c) depict two different segmentation scenarios of over- and under-segmentation compared to the ground truth by the same differences in radius, leading to the same boundary distance-based scores. (e) and (f) are two segmentation scenarios where the same number of pixels was missing from the ground truth: (e) $TP = 120 + 10 + 5 = 135$, $FP = 0$, $FN = 55$; (f) $TP = 110 + 25 + 0 = 135$, $FP = 0$, $FN = 55$, leading to the same DSC. Ground truth is shown using dashed lines, and predictions are shown as solid lines. The numbers shown in the figure indicate the number of pixels that make up the area of the nearest segment. The numerator in the fractions denotes the number of pixels in a ground truth segment, while the denominator denotes the number of pixels in a predicted segment.

there are many segments in a single class (many hepatic lesions), DSC is not sufficient for performance evaluation as it cannot discriminate between the segments in a single class. Few synthetic images which represent different binary segmentation scenarios are constructed as shown in Figure 4. The distance between the boundaries of ground truth and prediction for the images in (b) and (c) is the same, even though two different cases are depicted, which shows the failure of the boundary-based metric. Since the same number of pixels is lost in (e) and (f), DSC will be the same for both, even though a full small segment is lost as seen in (f). This drawback is overcome by the proposed metric as it weighs the TP pixels based on the area of the segment it is derived from. Higher weightage is given to the smaller segment pixels and lower weightage to the larger segment pixels.

3. METHODOLOGY

3.1. Proposed Metric

We have proposed an evaluation metric to quantify the performance of various segmentation scenarios, especially the cases where small segments in the prediction need to be emphasized more. We have illustrated the analysis in medical scan images which contain segments of various sizes. The proposed metric, SSEGEP, penalizes the loss of smaller segments more in comparison to the larger ones. Each pixel in the ground truth image is assigned a weightage which is calculated based on the area of the segment it belongs to. The weightage of a segment is the inverse of its area, which results in high-weighted small-sized segments and low-weighted bigger segments. Accordingly, while calculating SSEGEP, instead of a simple count of true positive pixels, the weighted sum of TP pixels is considered. FP pixels are also weighted, which is given as the inverse of TP count. If the task given is to quantify the combined segmentation of organ and tumor, two different weights are given to the FP pixels: one for tumor case, which is given as the inverse of TP count of tumor mask, and the other for organ case, which is given as the inverse of TP count of organ mask.

3.2. Implementation and Explanation

The general block diagram for 2D images is shown in Figure 5. Ground truth and segmented images are given as inputs. Segments of labels 1 and 2 are given in different colors. The segmentation evaluation methodology is comprised of assigning weightage to segments in ground truth image, finding true positive pixels in each segment, finding and assigning weightage to the false positive segments, and finally calculating the proposed metric. Initially, from the ground truth image, distinct segments belonging to each label are found. Here, in the block diagram, there are two different labels, and every distinct segment of these two labels is found. Similarly, the distinct segments of each label are found from the segmented image. This is done by finding the contours of the input image. Then, a weightage is allotted to these ground truth segments which is given as the inverse of the area of that segment as shown in Figure 5. Then, the number of pixels in TP segments (overlapping region of segments between ground truth and prediction) is calculated, which is given by a_i as shown in the block diagram. XOR operation is performed separately for each label between the ground truth and segmented image to find the false positive images for each label (here, two FP images are obtained as there are two labels), and the separate count of these pixels is recorded. Weightage of inverse TP count is given to those FP pixels as shown in the block diagram. Therefore, when the number of FP is very much higher than that of TP, the SSEGEP value becomes small, indicating poor segmentation, and when the number of FP is small compared to TP, SSEGEP value becomes high, indicating good segmentation (provided that the number of FN is less). The final SSEGEP value is calculated from this set of metrics which quantifies the quality of the prediction.

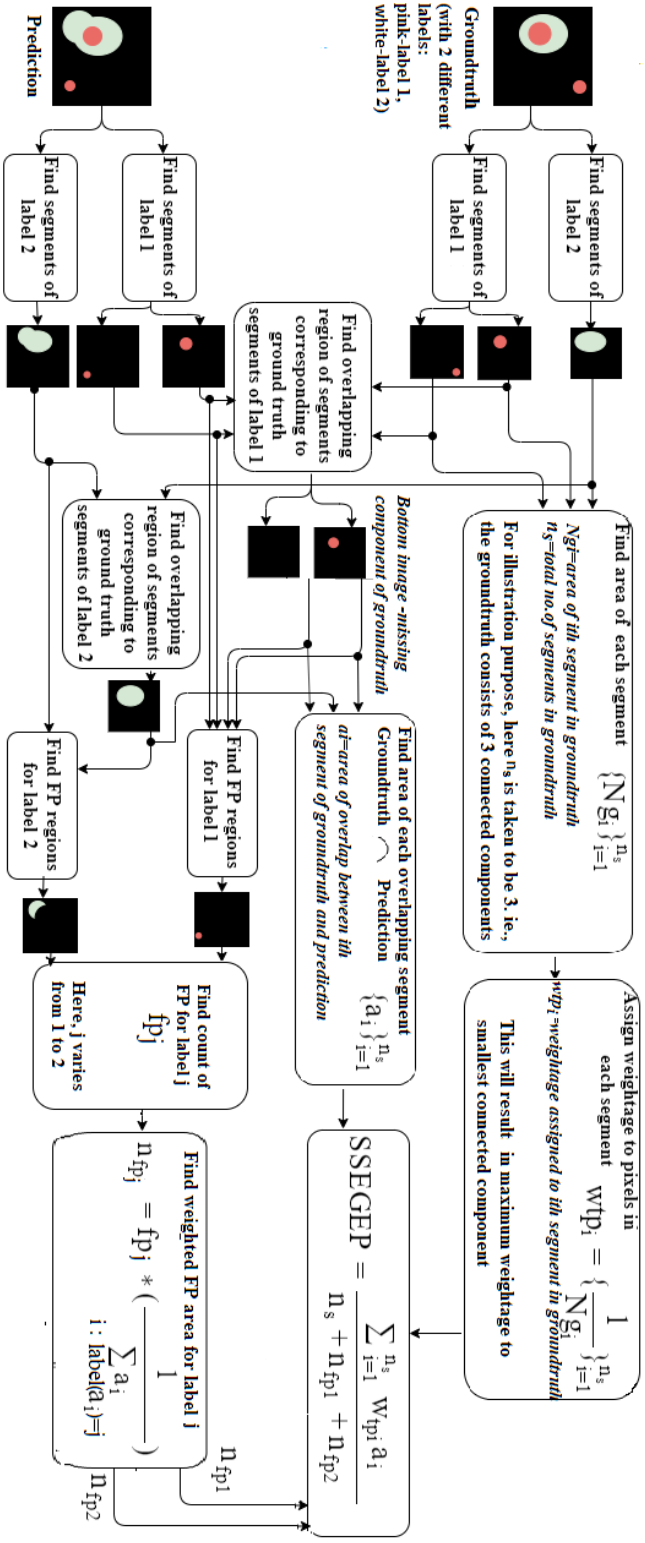


FIGURE 5: General outline of proposed evaluation approach. Weightage is assigned to TP pixels and FP pixels after finding the segments from the ground truth image to calculate the proposed metric. FN pixels are implicitly taken care of while assigning weightage to TP pixels since a higher number of FN pixels lead to small value of a_i and less value of numerator in SSEGEP. Here, n_s is the total number of segments (including label 1 and 2).

The proposed metric is given by the following equation:

$$\text{SSEGEP} = \frac{\sum_{i=1}^{n_s} w_{tpi} a_i}{n_s + \sum_j n_{fpj}}, \quad (1)$$

where w_{tpi} and a_i represent the weightage and TP count of the i th segments, respectively, and n_s represents the total number of segments in the ground truth image. Here, i is used to indicate segment and j to indicate label

$$n_{fpj} = fpj * \left(\frac{1}{\sum_{\text{label}(a_i)=j} a_i} \right), \quad (2)$$

where fpj represents the total FP count for segments of label j .

4. EXPERIMENTAL DETAILS AND MATERIALS USED

The experiments were performed on real scan images of three different modalities: fundus imaging, mammogram, and CT, to analyze and compare the proposed metric in different scenarios with commonly-used metrics.

4.1. Datasets Used

We used four different publicly available datasets to evaluate the proposed metric:

- (1) DIARETDB1, containing fundus images (imaging modality = fundus imaging, anatomy = eye),
- (2) CBIS-DDSM, containing mammogram images (imaging modality = mammogram, anatomy = breast),
- (3) MSD challenge, containing liver images (imaging modality = CT, anatomy = liver),
- (4) MSD challenge, containing pancreas images (imaging modality = CT, anatomy = pancreas).

DIARETDB1 database (<https://www.it.lut.fi/project/imageret/diaretdb1/>) [9] consists of 89 fundus images, of which 48 had signs of hard exudates as marked by experts. Therefore, these 48 cases were selected for the study, of which 15 cases with diverse features were manually selected for training, and the 33 remaining cases were used for testing. The segmented images were generated using the multi-space clustering approach [32] (refer to Supplementary Material for network and training information). The original images of 1500 pixels \times 1152 pixels were downsized to 800 pixels \times 615 pixels, to enable the running of the segmentation algorithm. The largest exudate present is 1.41% and the smallest is 0.0002% of the processed image. A few sample images in the database with exudates are shown in Figure 6.

CBIS-DDSM [33] (available at <https://wiki.cancerimagingarchive.net/display/Public/CBIS-DDSM>) database is considered, which is a standardized version of DDSM database. It contains 753 cases of calcifications and 891 cases of mass. Out of 753 cases of calcifications, there are 414 benign and 339 malignant cases. 264 images were selected from malignant calcification cases, of which 156 images were used for training, and the 108 remaining images were used for evaluation of the proposed metric. The mammogram images are in DICOM format, and updated ROI segmentation is provided for each lesion. U-net based segmentation approach [1] is used to obtain the segmentation results for calcifications. The images were cropped to remove the majority-black pixels in the background for U-net segmentation (refer to Supplementary Material for network and training information). The first column in Figure 11 shows sample mammogram images with calcifications.

Medical Segmentation Decathlon liver dataset (source: IRCAD, Hopitiaux Universitaires) [34] (available at <http://medicaldecathlon.com/results.html>) contains 201 3D volumes with liver and tumor as targets for segmentation, of which 131 images are grouped into the training set and 70 images grouped into the testing set. The dataset is rich in terms of the challenges and variability of the images contained in it. Since ground truth labels are not available for images from the testing set, the evaluation of the proposed metric is performed using the ground truth and segmentation results of 131 images from the training set. The segmentation results for those 131 images were obtained using the nnU-Net framework proposed by Fabian Isensee et al. [35] (refer to Supplementary Material for network and training information of nnU-Net framework). The ground truth images of the training set are multi-labeled, where 0, 1, and 2 are the respective labels for background, liver, and tumor. Few sample slices with tumors are shown in Figure 7. Tumors appear in diverse sizes as seen in the figure.

Medical Segmentation Decathlon pancreas dataset (source: Memorial Sloan Kettering Cancer Center) [34] contains 420 3D volumes of CT images with segmentation targets being pancreas and tumor, of which 282 images are grouped into the training set and 139 images grouped into the testing set. The dataset contains cases with good variability. Since ground truth labels are not available for images from the testing set, the evaluation of the proposed metric is performed using the ground truth and segmentation results of 107 images taken from the training set. The segmentation results for those 107 images are obtained using the nnU-Net framework proposed by Fabian Isensee et al. [35] (refer to Supplementary Material for network and training information of nnU-Net framework). Here, only one tumor is present in a pancreas segment. In some slices, the pancreas tissue and tumor are of similar size, whereas in a few slices, the tumor is small compared to the tissue, leading to class imbalance.

4.2. Experimental Details

Segmentation evaluation was done on the following.

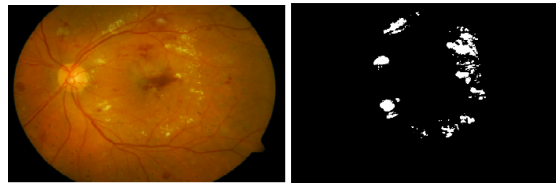


FIGURE 6: Representative image (from DIARETDB1). Left: original image; right: ground truth image.

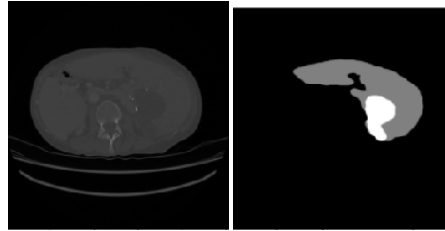


FIGURE 7: Representative images (from Liver_tumor dataset). Left: original image slice; right: ground truth image slice.

(a) Images with multiple segments from the same class: The images may contain segments of various sizes in which smaller segments are crucial in disease diagnosis. For example, the liver contains multiple tumors, and the fundus image consists of exudates and mammograms with calcifications of various sizes.

(b) Images with multiple segments from two different classes: Usually a single algorithm may not be efficient in extracting segments of different labels in multi-class segmentation. In scenarios where a common algorithm is required to perform multi-class segmentation, where the smaller segments are more important in disease diagnosis, it is imperative to use an evaluation metric that discriminates the segments in such a way as to weigh smaller segments more. As the proposed metric SSEGEP assigns size-based weight to the segments in the input image, the smallest segment gets the highest importance and the largest one gets the least weightage.

Further, the relevance of the proposed metric, SSEGEP is analyzed with two different approaches.

4.2.1. Mean Opinion Score

The proposed metric is compared with *accuracy*, *DSC*, and *IOU*, and its relevance is tested with mean opinion score calculated based on the scores assigned by 15 subjects to the segmentation results obtained for 33 fundus images. 15 subjective scores were sought from researchers with varying levels of experience (2 to 5 years) in image analysis. The scores given were 0, 0.25, 0.5, 0.75, 1, where 0 and 1 indicated poor and best segmentation, respectively. Then, the average of the absolute difference between the mean opinion score and the value of the evaluation metric is found, which indicates the deviation of that metric from the mean opinion score.

4.2.2. Statistical Analysis

As we have a greater number of images in liver, pancreas, and mammogram datasets, the statistical significance of SSEGEP was measured separately for each dataset by hypothesis testing. To assess the statistical significance of the proposed metric, Welch's *t*-test was performed since Levene's test for homogeneity of variances indicated unequal variances between groups and sample sizes were unequal. The null hypothesis is given as follows: "The two populations have equal mean."

The two populations that we considered here are a set of images from 2 different categories. The first category consists of values of evaluation metric obtained for "good segmentation" cases, and the second category consists of values of evaluation metric obtained for "poor segmentation" cases. The threshold for TPR and FPR which are used to decide the samples in two groups were selected based on the condition: if TPR is greater than 80% and FPR is less than 20%, the corresponding segmented image belongs to "good segmentation" category. If TPR is less than 40% and FPR is more than 50%, the image belongs to "bad segmentation" category. Extreme cases of good and poor segmentation categories are selected for better understanding. Hypothesis testing was performed with a confidence level of 99.9999. We wanted to check if the evaluation metrics obtained for both categories were overlapping or if the evaluation metrics were able to differentiate between the two categories. The statistical significance testing was performed separately for *DSC* and the proposed metric. Then we compared the *p*-values obtained with both the metrics to find the metric resulting in a smaller *p*-value. The smaller *p*-value with the proposed metric indicated that it has a better discriminating capability between good and bad segmentation. The statistical tests were performed only to compare *DSC* and proposed metric as *DSC* is mostly used in medical segmentation scenarios.

All metric values *DSC*, *accuracy*, *IOU*, and *SSEGEP* lie in the range of 0 to 1. *Accuracy*, *DSC*, *IOU*, and *SSEGEP* are computed for fundus images, mammograms, and liver images involving single-class segmentation, and *generalized Dice score* and *SSEGEP* are computed for multi-class segmentation (liver and pancreas). Python 3.6 is used for the implementation.

5. RESULTS AND DISCUSSION

The effectiveness of our metric in the evaluation of segmentation performance is demonstrated in this section. Results obtained on applying the proposed metric to each of the datasets have been reported. Comparisons with existing widely used metrics have also been presented. Besides, hypothesis testing in order to establish the statistical significance of the proposed metric is also carried out.

5.1. Experiments with Synthetic Images

Few synthetic images with multiple segments of varying sizes which represent different segmentation scenarios are constructed as shown in Figure 8 for illustrating the proof of concept. Consider Figure 8(a) as the ground truth image. Figure 8(b)–(d) need to be considered as predictions made with different segmentation algorithms. From visual inspection, it is obvious that Figure 8(d) is a better prediction than others since all the segments including the smallest one are segmented out. Figure 8(b) consists of FP and FN pixels due to which the segmentation is worst compared with other segmentation results. However, in Figure 8(c), the smallest segment is completely missing, and the loss of that smallest segment should be penalized most. The values of commonly-used metrics wrongly indicate that the result of the second segmentation in Figure 8(c) is better than that of the third segmentation in Figure 8(d) (it is because the number of correctly detected exudate pixels with the second algorithm is greater than that of the third algorithm). However, a small exudate is missing in Figure 8(c), which means the corresponding segmentation algorithm is not efficient in detecting smaller exudates. The proposed metric can discriminate these cases as it weighs the TP pixels based on the area of the segment it is derived from. SSEGEP for Figure 8(d) is higher than that for Figure 8(c) which indicates that the quality of prediction is better with algorithm 3 than with algorithm 2.

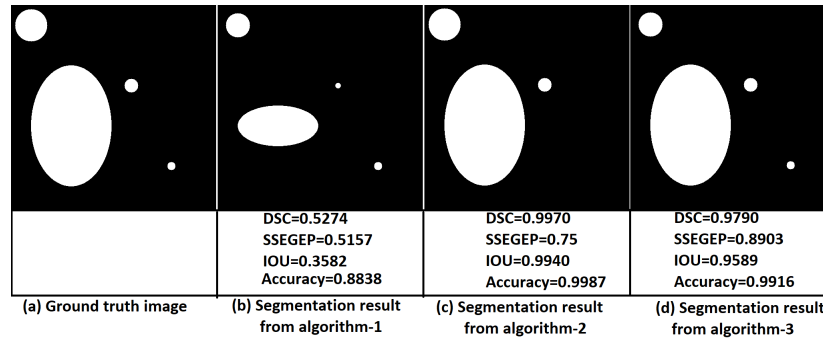


FIGURE 8: A sample synthetic ground truth image (a) and predictions (b)–(d) with different segmentation algorithms to illustrate the proof of concept. Calculated metric values are also given.

5.2. Experiments with Real Scan Images

5.2.1. Experiments with Retina Images: Multiple Segments of Same Type

Figure 9 shows a few fundus images with their ground truth and segmentation. The values of the proposed metric and commonly-used metrics are compared with the average MOS. Accuracy is not sufficient to evaluate segmentation performance for imbalanced data (here, the number of TN pixels is large). DSC can be used when there is no much variation in exudate size (Figure 9(a), (b), and (e)). However, DSC fails to quantify segmentation performance when there are exudates of varying sizes (Figure 9(d)). In such a case, SSEGEP outperforms other metrics as it is able to discriminate different segments according to the count of the pixels in them.

Evaluation metric	Spearman's rank correlation	Average deviation w.r.t. MOS
Accuracy	0.76	0.15
DSC	0.85	0.11
IOU	0.83	0.11
SSEGEP (proposed)	0.89	0.08

TABLE 2: Correlation of evaluation metrics with the mean MOS for the DIARETDB1 database is shown in the second column. Average deviation of metrics from the MOS is given in column 3. MOS is calculated from scores assigned by 15 subjects, and comparison is done on a scale of 0 to 1.

The average of the difference between mean opinion score and value of the evaluation metric (accuracy, DSC, IOU, and SSEGEP) along with correlation of the metrics with mean MOS is tabulated in Table 2. The deviation value is least with the proposed metric, which indicates its higher correlation with human interpretability.

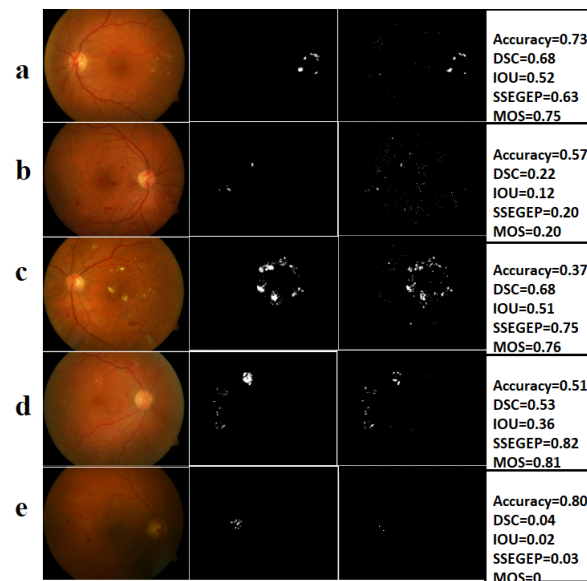


FIGURE 9: First column represents the original fundus images. Corresponding ground truths and predictions (using multi-space clustering) are shown in the second and third column, respectively. Values of evaluation metrics (accuracy, DSC, IOU, and SSEGEP) and MOS are given in the last column.

5.2.2. Experiments with Mammogram Images: Multiple Segments of Same Type

The bar plots of SSEGEP values and DSC for a few sample mammogram images are shown in Figure 10. DSC and SSEGEP values for images of good segmentation category are almost in the same range. This implies that both metrics can reflect good segmentation. However, the DSC values for many images under the poor segmentation category are higher than those of SSEGEP. Even though the segmentation result is poor, DSC is high. However, the proposed metric can reflect the poor quality of segmentation since those values are lower.

Figure 11 shows sample mammogram images with their ground truth and segmentation masks. The values of the proposed metric and commonly used metrics are also shown. As seen for fundus images, sensitivity and accuracy are not sufficient to quantify segmentation performance. Panels (a) to (e) consist of calcifications of varying sizes. Smaller calcifications indicate higher severity of the disease. Even though a smaller segment is lost in mammogram segmentation for those panels, commonly-used metrics fail to reflect it. For example, DSC is 0.91 for mammograms in panel (a) even though the segmentation algorithm failed to detect the smaller segment. However, SSEGEP is 0.43, which indicates poor segmentation in terms of medical diagnosis. Panels (f) and (g) consist of calcifications of similar sizes; hence, the values of the proposed metric, SSEGEP, are comparable to those of commonly-used metrics. Panel (h) comprises a single large segment, and hence all metric values are of a similar range.

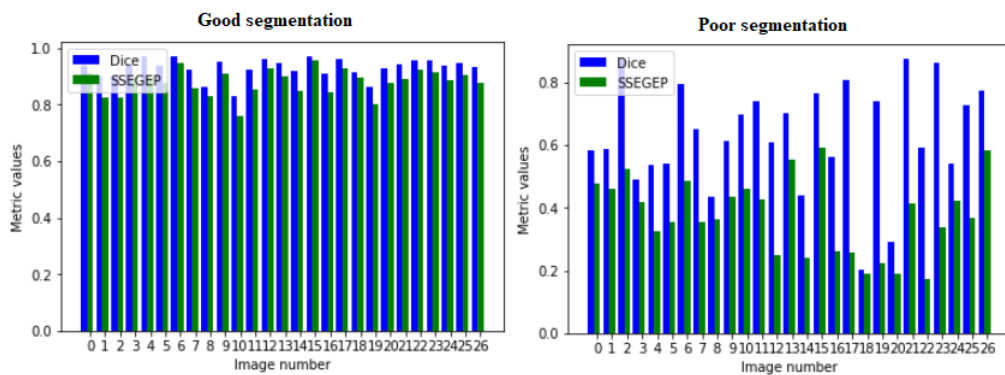


FIGURE 10: Bar plots for SSEGEP and DSC values for “good” and “bad” segmentation categories for 27 mammogram images (as mentioned in Section 4.2.2). Metric values are comparable for images in good segmentation group. However, DSC is greater than SSEGEP for images under poor segmentation category.

























a				Accuracy=0.96 DSC=0.91,IJU=0.84 SSEGEP=0.43
b				Accuracy=0.95 DSC=0.68,IJU=0.52 SSEGEP=0.39
c				Accuracy=0.95 DSC=0.76,IJU=0.61 SSEGEP=0.59
d				Accuracy=0.99 DSC=0.49,IJU=0.32 SSEGEP=0.29
e				Accuracy=0.97 DSC=0.81,IJU=0.68 SSEGEP=0.40
f				Accuracy=0.98 DSC=0.89,IJU=0.81 SSEGEP=0.81
g				Accuracy=0.99 DSC=0.54,IJU=0.37 SSEGEP=0.42
h				Accuracy=0.94 DSC=0.94,IJU=0.88 SSEGEP=0.88

FIGURE 11: First column represents the original mammogram images. Corresponding ground truths and predictions (using U-net) are shown in the second and third columns, respectively. Values of evaluation metrics (accuracy, DSC, IOU, and SSEGEP) are given in the last column.

The results for the statistical significance test are tabulated in Table 3. The p -value for SSEGEP is less than that of DSC, which indicates the better discriminating capability of the proposed metric.

5.2.3. Experiments with Liver Images

Multiple Segments (Tumors) of the Same Type. Figures 12, 13, and 14 show the evaluation metrics obtained for a few sample cases of (1) images with all small tumors, (2) images with small and large tumors, and (3) images with similar-sized (medium-sized) tumors. Here, performance evaluation is conducted only for tumors. *Accuracy* for all the cases is high (0.98, 0.99) as the number of TN pixels is high compared to other pixels. As seen in Figure 12, all the tumors appear small. SSEGEP and IOU values are varying in a similar way. DSC is slightly higher than IOU as the former gives higher weightage to TP pixels. Since tumors are smaller in size, even though few pixels are missed in segmentation, they should be highly penalized, which is reflected by the proposed metric.

Figure 13 consists of tumors of varying sizes. There are tiny tumors and large tumors present in a single tissue. Therefore, the loss of a small tumor should be highly penalized. However, still, the values of other metrics are closer to 0.90. However, SSEGEP value is in the range of 0.70–0.75, thereby reflecting the level of quality of segmentation.

Figure 14 comprises scan images with similar- and moderately sized tumors. The SSEGEP values are comparable with other commonly used metrics because of similar-sized segments. There is a FP present in the third case, which resulted in a smaller SSEGEP value compared to the other two cases. Since weightage of FP is decided by the inverse of TP pixels and number of TP pixels is less in this case, SSEGEP value is not too low.

Hence, we can conclude that if there are tumors of varying sizes in the ground truth and if the segmentation algorithm fails to extract the smaller tumors, it should be severely penalized (as seen in Figure 13). As it can be seen in Figure 13, even though other metric values are at a higher side indicating best segmentation, the proposed metric is less than that since smaller tumors are lost in segmentation. Commonly used metrics fail to discriminate between the varying-sized distinct tumors. However, the proposed metric assigns different weightage to tumors based on their size, emphasizing smaller ones.

Multi-Class Case with Many Segments of Different Types (Liver and Tumor) with Class Imbalance. Figure 15 shows sample multi-class segmentation and the corresponding evaluation metrics. Here, SSEGEP is compared with the generalized multi-class *Dice* value, which is the commonly used performance metric for unbalanced multi-class segmentation. Even though generalized multi-class *Dice* can discriminate the two different labels, it fails to differentiate the multi-sized tumors present. As seen in Figure 15, the liver occupies more space compared to the tumors. Even though small tumors are not identified in each case, DSC is the same as it gives the same weightage to every tumor irrespective of its size. However, SSEGEP value is not high unlike DSC, since the proposed metric emphasizes the smallest tumor more by assigning a higher weightage to it, thereby penalizing the loss of smaller segments. The results for hypothesis testing are tabulated in Table 3. The test says that the inputs come from different distributions for both the metrics, with the p -value for SSEGEP being less than that of DSC. This implies that both the metrics (SSEGEP and DSC) can distinguish between good and poor segmentation, with a better discriminating capability for SSEGEP.

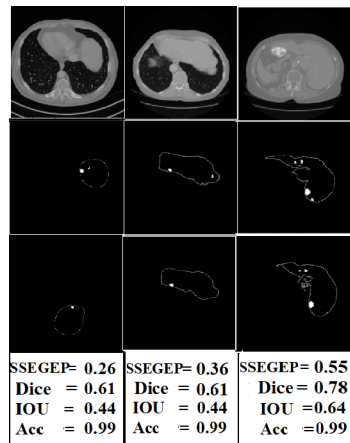


FIGURE 12: Performance evaluation of tumor segmentation (using nnU-Net) of liver (all tumors are smaller in size) is illustrated in the figure. The first row represents the original image slices. The second row represents the ground truth images. Corresponding segmentation is shown in the third row. The values of evaluation metrics (only for tumor), SSEGEP, DSC, IOU, and accuracy, are given in the last row.

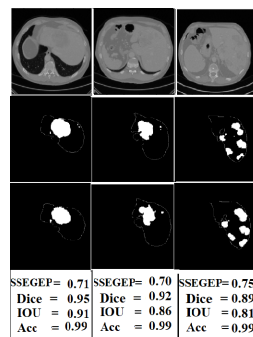


FIGURE 13: Performance evaluation of tumor segmentation (using nnU-Net) of liver (consisting of larger and smaller tumors) is illustrated in the figure. The first row represents the original image slices. The second row represents the ground truth images. Corresponding segmentation is shown in the third row. The values of evaluation metrics (only for tumor), SSEGEP, DSC, IOU, and accuracy, are given in the last row.

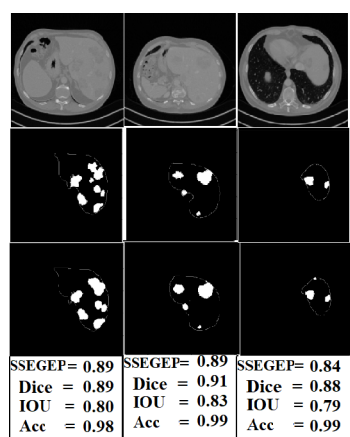


FIGURE 14: Performance evaluation of tumor segmentation (using nnU-Net) of liver (similar- and moderately sized tumors) is illustrated in the figure. The first row represents the original image slices. The second row represents the ground truth images. Corresponding segmentation is shown in the third row. The values of evaluation metrics (only for tumor), SSEGEP, DSC, IOU, and accuracy, are given in the last row.

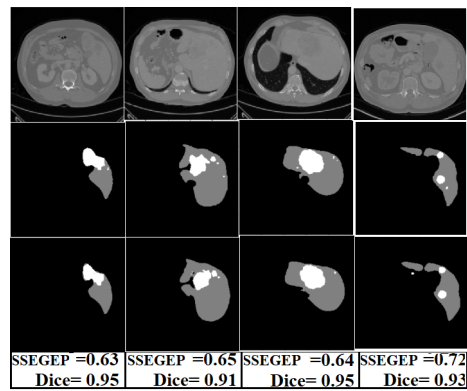


FIGURE 15: Combined segmentation (using nnU-net) of liver and tumor with the evaluation metrics is given. Gray colored region in the image refers to the liver, and white color refers to the tumor. The first row represents the original image slices. The second row represents the ground truth images. Corresponding segmentation is shown in the third row. The values of evaluation metrics, SSEGEP and DSC, are given in the last row.

5.2.4. Experiments with Pancreas Images: Multi-Class Case with One Segment Each of Different Type (Pancreas and Tumor)

Here, only one tumor is present in a pancreas segment. Therefore, the evaluation is done for the combined pancreas and tumor image as explained for the liver and tumor cases. Sample images with generalized multi-class *Dice* and SSEGEP values are shown in Figure 16. The tumor is not detected for ground truth in the first and last column in Figure 16. For all cases, DSC and SSEGEP values vary similarly, since each class (pancreas and tumor) contains only one segment (one tissue and one tumor) unlike combined liver and tumor cases (in liver and tumor segmentation, there were multiple tumors in the liver tissue). Hence, for multi-label segmentation evaluation where there is only one segment present in a class (or label), either generalized *Dice* or SSEGEP can be used.

The results for hypothesis testing are tabulated in Table 3. The *p*-value is the same for both metrics, indicating the applicability of both metrics in that case. This is because there is only one segment corresponding to each label in an image slice and both SSEGEP and generalized *Dice* act similarly for that scenario, whereas in liver tumor segmentation there are tumors of various sizes, and the size variation within a label is not reflected by DSC value.

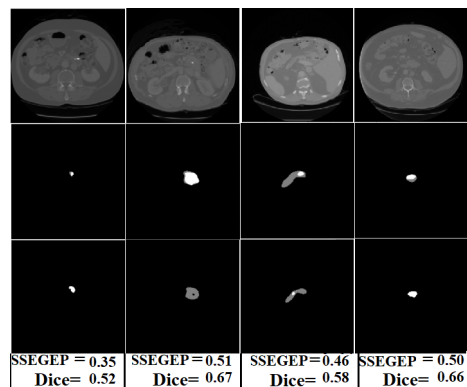


FIGURE 16: Combined segmentation (using nnU-net) of pancreas and tumor with the evaluation metrics is given. Gray colored region in the image refers to the pancreas, and white color refers to the tumor. The first row represents the original image slices. The second row represents the ground truth images. Corresponding segmentation is shown in the third row. The values of evaluation metrics, SSEGEP and DSC, are given in the last row.

Segmentation scenarios	<i>p</i> value (DSC)	<i>p</i> value (SSEGEP)
Calcification	1.8e-19	3.5e-35
Tumor (liver)	7.6e-12	1.2e-18
Combined liver tumor	2.9e-17	3.6e-19
Combined pancreas tumor	3.3e-19	3.3e-19

TABLE 3: Hypothesis testing with $\alpha = 0.00001$.

5.3. Analysis of Impact of Parameter Changes on SSEGEP

The various parameters used in the study are as follows.

- (1) Number of connected components in the ground truth: if the number of connected components in the ground truth, as an extreme case, happens to be only 1, the proposed metric will not result in any significantly different way compared to DSC, since it simply involves the calculation of total overlap area like DSC.
- (2) The relative sizes of each of the connected components in the ground truth: if the sizes of each of the connected components in the ground truth are all comparable, the advantage of assigning them different weights will not lead to many distinctions across them. However, if the sizes of the connected components in the ground truth are very diverse, the significance of assigning differential weights to each of them becomes very perceptible.
- (3) Impact of weightage on TP: weightage assigned to TP depends on the size of the corresponding connected component in the ground truth. If the connected component is of smaller size, higher weightage is assigned to that TP pixel, and vice versa.
- (4) Impact of weightage on FP: weightage assigned to FP depends on the size of total TP pixels. When the number of FP pixels is less, the segmentation performance is better, and when FP is more, segmentation performance is poor. To weight the FP pixels, inverse of TP pixels is considered. Hence, if the FP number is higher than TP number, a higher weightage is assigned to FP, thereby leading to smaller value of SSEGEP.

Study Limitations

The above experiments say that the proposed metric gives better performance than other metrics for the evaluation of the segmentation performance of medical images with segments of varying sizes. For images with segments of comparable size, SSEGEP performs in a comparable way to other overlap-based metrics (DSC, IOU). If segmentation is done at random or by chance, the metric value should ideally be zero. However, this is not the case with SSEGEP. It will be near zero. We have computed the metric only for 2D segmentation as of now and need to extend it for 3D segmentation as well. Statistical analysis was performed between good and bad segmentation classified by thresholding TPR and FPR only. It would be better to consider the number of segments and size of the ROIs as well, but it is difficult for cases with multiple segments and diverse sizes. Moreover, we have not performed visual assessments to measure MOS for liver, mammogram, and pancreas datasets.

6. CONCLUSION

Segmentation of smaller lesions that appear in medical image scans is important as they assist in the early treatment of associated diseases. The evaluation metric used should be able to quantify segmentation performance emphasizing these smaller lesions. The proposed metric assigns weightage to each segment in the image based on its area in such a way as to emphasize the smaller segments more. The applicability of our metric across different imaging techniques is shown by testing on four public datasets of different anatomies, and its significance is proved with mean opinion score and hypothesis testing. It is observed that the proposed metric is 30% closer to MOS, as compared to DSC for fundus images, and statistical significance testing resulted in a p -value of the order 10^{-18} with SSEGEP for liver tumor compared to DSC. The experimental results show that the proposed metric outperforms other metrics with respect to human interpretability and understanding for those images with multiple segments of various sizes.

CONFLICTS OF INTEREST

The authors declare that there are no conflicts of interest regarding the publication of this paper.

FUNDING

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

ACKNOWLEDGMENTS

We would like to thank Fabian Isensee for providing the segmentation results of pancreas and liver images that are obtained using nnU-Net [35]. We also acknowledge Sanjeev Dubey for providing detailed analysis in GitHub about the segmentation algorithm [32] used for fundus images.

References

- [1] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.
- [2] Kelly Zou, Simon Warfield, Aditya Bharatha, Clare Tempany, Michael Kaus, Steven Haker, William Wells, and Ron Kikinis. Statistical validation of image segmentation quality based on a spatial overlap index. *Academic radiology*, 11:178–89, 2004.
- [3] KH Zou, WM Wells, R Kikinis, and SK Warfield. Three validation metrics for automated probabilistic image segmentation of brain tumours. *Stat Med*, 23:1259–82, 2004.
- [4] Andrew Worth, Makris Nikos, Verne Caviness, and David Kennedy. Neuroanatomical segmentation in mri: Technological objectives. *International Journal of Pattern Recognition and Artificial Intelligence*, 11:1161–87, 11 1997.

- [5] SK Warfield, CF Westin, CRG Guttman, MS Albert, FA Jolesz, and R Kikinis. Fractional segmentation of white matter. In *International Conference on Energy Efficient Technologies for Sustainability (ICEETS)*, pages 109–117, 1999.
- [6] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993–2024, 2014.
- [7] Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S Kirby, John B Freymann, Keyvan Farahani, and Christos Davatzikos. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Scientific data*, 4(1):1–13, 2017.
- [8] Spyridon Bakas, Mauricio Reyes, Andras Jakab, Stefan Bauer, Markus Rempfler, Alessandro Crimi, Russell Takeshi Shinohara, Christoph Berger, Sung Min Ha, Martin Rozycki, et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. *arXiv preprint arXiv:1811.02629*, 2018.
- [9] RVJPH Kälviäinen and H Uusitalo. Diaretdb1 diabetic retinopathy database and evaluation protocol. In *Medical Image Understanding and Analysis*, volume 2007, page 61. Citeseer, 2007.
- [10] Carole Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and Manuel Jorge Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 240–248, 09 2017.
- [11] R Klein, B.E.K. Klein, S E Moss, M D Davis, and David Demets. The wisconsin epidemiologic study of diabetic retinopathy. vii. diabetic nonproliferative retinal lesions. *Ophthalmology*, 94:1389–400, 12 1987.
- [12] Prisca Loganadane, Bernard Delbosc, and Maher Saleh. Short-term progression of diabetic hard exudates monitored with high-resolution camera. *Ophthalmic Research*, pages 1–7, 11 2018.
- [13] Qing Liu, Beiji Zou, Jie Chen, Wei Ke, Kejuan Yue, Zailiang Chen, and Guoying Zhao. A location-to-segmentation strategy for automatic exudate segmentation in colour retinal fundus images. *Computerized Medical Imaging and Graphics*, 55, 09 2016.
- [14] Sreeparna Banerjee and Diptoneel Kayal. Detection of hard exudates using mean shift and normalized cut method. *Biocybernetics and Biomedical Engineering*, 36, 10 2016.
- [15] Anju Sahdev. Ct in ovarian cancer staging: how to review and report with emphasis on abdominal and pelvic disease for surgical planning. *Cancer Imaging*, 16(1):1–9, 2016.
- [16] Mahmoud Khaleel Abd-Ellah, Ali Ismail Awad, Ashraf AM Khalaf, and Hesham FA Hamed. A review on brain tumor diagnosis from mri images: Practical implications, key achievements, and lessons learned. *Magnetic resonance imaging*, 61:300–318, 2019.
- [17] Mert Ciftdemir, Murat Kaya, Esref Selcuk, and Erol Yalniz. Tumors of the spine. *World journal of orthopedics*, 7(2):109, 2016.
- [18] Lawrence Schwartz, Eric Gandras, Sandra Colangelo, Matthew Ercolani, and David Panicek. Prevalence and importance of small hepatic lesions found at ct in patients with cancer. *Radiology*, 210:71–4, 02 1999.
- [19] Quan Nguyen, Nga Nguyen, Linden Dixon, Flavia Monetto, and Angelica Robinson. Spontaneously disappearing calcifications in the breast: A rare instance where a decrease in size on mammogram is not good. *Cureus*, 12, 06 2020.
- [20] Benoit Mesurrolle, Fawaz Halwani, Vincent Pelsler, Jean Gagnon, Ellen Kao, and Francine Tremblay. Spontaneous resolving breast microcalcifications associated with breast carcinoma. *The breast journal*, 11:478–9, 11 2005.
- [21] H. Seymour, J. Cooke, and R. Given-Wilson. The significance of spontaneous resolution of breast calcification. *The British journal of radiology*, 72 853:3–8, 1999.
- [22] A Evans, K Clements, A Maxwell, H Bishop, A Hanby, G Lawrence, et al. Lesion size is a major determinant of the mammographic features of ductal carcinoma in situ: findings from the sloane project. *Breast Cancer Research*, 65:181–4, 2010.
- [23] Breastcancer.org. *Understanding Breast Calcifications*, 2018 (accessed October 13, 2020).
- [24] Ona Wu, Stefan Winzeck, Anne-Katrin Giese, Brandon Hancock, Mark Etherton, Mark Bouts, Kathleen Donahue, Markus Schirmer, Robert Irie, Steven Mocking, Elissa McIntosh, Raquel Bezerra, Konstantinos Kamnitsas, Petrea Frid, Johan Wasselius, John Cole, Huichun Xu, Lukas Holmegaard, Jordi Jimenez-Conde, and Natalia Rost. Big data approaches to phenotyping acute ischemic stroke using automated lesion segmentation of multi-center magnetic resonance imaging data. *Stroke*, 50, 06 2019.
- [25] Gabriele Piantadosi, Mario Sansone, Roberta Fusco, and Carlo Sansone. Multi-planar 3d breast segmentation in mri via deep convolutional neural networks. *Artificial Intelligence in Medicine*, page 101781, 12 2019.
- [26] Piotr Chudzik, Somshubra Majumdar, Francesco Caliva, Bashir Al-Diri, and Andrew Hunter. Exudate segmentation using fully convolutional neural networks and inception modules. In *Medical Imaging 2018: Image Processing*, volume 10574, pages 785 – 792. SPIE, 2018.
- [27] Sylvain Gouttard, Martin Styner, Marcel Prastawa, Joseph Piven, and Guido Gerig. Assessment of reliability of multi-site neuroimaging via traveling phantom study. In Dimitris Metaxas, Leon Axel, Gabor Fichtinger, and Gábor Székely, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2008*, pages 263–270, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg.
- [28] N Nur and Handayani Tjandrasa. Exudate segmentation in retinal images of diabetic retinopathy using saliency method based on region. *Journal of Physics: Conference Series*, 1108:012110, 11 2018.
- [29] Fang Lu, Fa Wu, Peijun Hu, Zhiyi Peng, and Dexing Kong. Automatic 3d liver location and segmentation via convolutional neural network and graph cut. *International Journal of Computer Assisted Radiology and Surgery*, 12, 09 2016.
- [30] Varduhi Yeghiazaryan and Irina Voiculescu. Family of boundary overlap metrics for the evaluation of medical image segmentation. *Journal of Medical Imaging*, 5:1, 02 2018.
- [31] Abdel Aziz Taha and Allan Hanbury. Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool. *BMC medical imaging*, 15:29, 2015.
- [32] Sanjeev Dubey and Utkarsh Mittal. Exudate detection in fundus images: Multispace clustering approach. In *Third International Conference, ICICCT 2018*, pages 109–117, 01 2019.
- [33] Rebecca Sawyer Lee, Francisco Gimenez, Assaf Hoogi, Kanae Kawai Miyake, Mia Gorovoy, and Daniel L Rubin. A curated mammography data set for use in computer-aided detection and diagnosis research. *Scientific data*, 4(1):1–9, 2017.
- [34] Amber L. Simpson, Michela Antonelli, Spyridon Bakas, Michel Bilello, Keyvan Farahani, Bram van Ginneken, Annette Kopp-Schneider, Bennett A. Landman, Geert J. S. Litjens, Bjoern H. Menze, Olaf Ronneberger, Ronald M. Summers, Patrick Bilic, Patrick Ferdinand Christ, Richard K. G. Do, Marc Gollub, Jennifer Golia-Pernicka, Stephan Heckers, William R. Jarnagin, Maureen McHugo, Sandy Napel, Eugene Vorontsov, Lena Maier-Hein, and M. Jorge Cardoso. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *CoRR*, abs/1902.09063, 2019.
- [35] Fabian Isensee, Jens Petersen, Simon A. A. Kohl, Paul F. Jäger, and Klaus H. Maier-Hein. nnu-net: Breaking the spell on successful medical image segmentation. *CoRR*, abs/1904.08128, 2019.