Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2022.Doi Number

Assessing Inter-Annotator Agreement for Medical Image Segmentation

Feng Yang[1]*, Ghada Zamzmi[1], Sandeep Angara[1], Sivaramakrishnan Rajaraman[1], Senior Member, IEEE, André Aquilina[2], Zhiyun Xue[1], Stefan Jaeger[1], Emmanouil Papagiannakis[2], Sameer K Antani[1]*, Senior Member, IEEE

[1]National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA
[2]DYSIS Medical Ltd, Edinburgh, UK

*Corresponding author: Feng Yang, Sameer K Antani (e-mail: feng.yang2@nih.gov; sameer.antani@nih.gov).

This work was supported by the Intramural Research Program of the National Library of Medicine (NLM), National Institutes of Health (NIH) and the Lister Hill National Center for Biomedical Communications of the National Library of Medicine (NLM), National Institutes of Health.

ABSTRACT Artificial Intelligence (AI)-based medical computer vision algorithm training and evaluations depend on annotations and labeling. However, variability between expert annotators introduces noise in training data that can adversely impact the performance of AI algorithms. This study aims to assess, illustrate and interpret the inter-annotator agreement among multiple expert annotators when segmenting the same lesion(s)/abnormalities on medical images. We propose the use of three metrics for the qualitative and quantitative assessment of inter-annotator agreement: 1) use of a common agreement heatmap and a ranking agreement heatmap; 2) use of the extended Cohen's kappa and Fleiss' kappa coefficients for a quantitative evaluation and interpretation of inter-annotator reliability; and 3) use of the Simultaneous Truth and Performance Level Estimation (STAPLE) algorithm, as a parallel step, to generate ground truth for training AI models and compute Intersection over Union (IoU), sensitivity, and specificity to assess the inter-annotator reliability and variability. Experiments are performed on two datasets, namely cervical colposcopy images from 30 patients and chest X-ray images from 336 tuberculosis (TB) patients, to demonstrate the consistency of inter-annotator reliability assessment and the importance of combining different metrics to avoid bias assessment.

INDEX TERMS Reliability, agreement, inter-annotator, heatmap, STAPLE, Cohen's kappa, Fleiss' kappa

I. INTRODUCTION

In computational health research, it is typical to collect annotations from several expert annotators to capture the diversity of opinion, mitigate subjective biases, and compensate for factors such as the level of experience, expertise, or fatigue [1]–[4]. Therefore, it is necessary to (i) assess the extent of agreement between different annotators, called *inter-annotator reliability* or *inter-annotator agreement*, and (ii) develop an appropriate strategy for training reliable segmentation models that reflect the underlying agreement/disagreement between different annotators [5], [6].

A typical situation in which it may be necessary to assess inter-rater reliability is when multiple experts annotate images for the presence or severity of underlying disease(s). In this case, the judgments from different experts or annotators are with discrete categories, such as "presence" and "absence" of lesions, and qualitative assessments as lesions being "mild," "moderate," or "severe." These categories are mutually exclusive; therefore, each case falls into one of these categories. The kappa statistic [7], [8] is commonly used to evaluate inter-annotator reliability in categorical classification problems [9]–[11]. However, it is challenging to assess agreement when multiple experts draw boundaries around regions in images. To the best of our knowledge, there are no systematic studies assessing inter-annotator agreement on segmentation in medical images, which is crucial in evaluating the performance of region segmentation algorithms that may be used for the clinical assessment of various diseases.

Modern Artificial Intelligence (AI) techniques have accelerated the development of automated systems for detecting diseases in images and quantifying disease progression [1]. Simultaneous Truth and Performance Level Estimation (STAPLE) based methods [12], [13] and label fusion algorithms [14]–[16] are commonly used to obtain ground truth labels. Lampert *et al* [17] presented an in-depth study to quantify the effects of acquiring ground truth data from multiple annotators using different methods. They concluded that the STAPLE and maximizing a posteriori probability-based method algorithms find a reasonable balance between all annotations when the overall inter-rater segmentation variance is low.

The assessment of the inter-annotator agreement is essential for several reasons. First, ensuring that annotators' annotations are consistent is essential to designing and establishing stable AI algorithms. Second, an agreement is desirable for AI model training and evaluation since high levels of inter-annotator agreement help reduce noise and subjectivity. Third, it helps improve the validity of the annotations. Data with higher inter-annotator reliability indicates good quality and supports reproducible studies. The high inter-annotator agreement supports enhanced clinical decision-making and risk prediction.

In this paper, we systematically assess the inter-annotator reliability in two applications: lesion segmentation on cervical images and abnormality segmentation in Chest X-ray (CXR) images. Our work makes the following four main contributions:

- We propose two agreement heatmaps to visualize and quantify the inter-annotator reliability, including a common agreement heatmap and a ranking agreement heatmap. To the best of our knowledge, no prior works have studied similar agreement heatmaps, particularly the ranking agreement heatmap, toward evaluating the reliability of inter-annotator agreement.
- We extend kappa coefficients, particularly Fleiss' kappa coefficient, from categorical classification to pixel-wise segmentation, by generating and interpreting the new kappa tables for the image segmentation problem.
- We apply the STAPLE algorithm to generate the ground truth and compute Intersection over Union (IoU), sensitivity, and specificity to quantitatively evaluate the inter-annotator reliability and variability and compare the consistency of agreement using different metrics. The IoU, sensitivity, and specificity values facilitate the identification of raters with little or high agreement.
- We publish a new collection of annotations for the Shenzhen dataset.

The rest of this paper is organized as follows. In Section II, we describe the different metrics for qualitative and quantitative assessment of inter-annotator reliability. In Section III, we present the two datasets used in this study and demonstrate the evaluation results, followed by the Discussion and Conclusion in Section IV.

II. Related works

We follow the well-known PRISMA approach to search and select related works on inter-annotator reliability assessment in cervical cancer images and chest X-Rays (CXRs). We used different keywords to search in PubMed and Google Scholar. Examples of these keywords include "heatmap", "kappa", "multi-rater", "multiple annotations", "inter-rater", "inter-annotator", "cervical cancer images", "CXR", etc. Then, we read the abstract to confirm whether the papers or methods fit well. These criteria resulted in a total of 18 papers that were discussed below.

Heatmap. Heatmaps have been utilized in data analysis for over a century [18], and are recently widely used in radiomics and understanding of AI model predictions. In radiomics, large data tables are clustered and then color-coded to help identify patterns [19], [20]. In computer vision, an AI model may produce a heatmap that identifies the areas of the input image which contribute most to the model using class activation mappings [21]. In [22], the Class-selective Relevance Map (CRM) and Class Activation Map (CAM) are used to visualize the model prediction for ablative treatability classification in digitized cervix images. In [23], the weights from the last convolutional layers are used to generate a heatmap to emphasize the high-weight signals of coronavirus disease (COVID-19) in CXRs. In [24], a probability map is used to visualize segmentation/annotations of multiple organs and to reduce the use of tedious and prone-to-error manual annotations from CXRs. Despite the wide usage of heatmaps in radiomics and visualization of model predictions, current heatmaps are not suitable for multiple annotations. This is the first paper that proposes agreement heatmaps for inter-annotator reliability assessment.

Kappa statistics. In [25], [26], Cohen's kappa coefficient is used to evaluate inter-annotator agreement on the performance of visual inspection with acetic acid for precancerous lesion classification between two test providers. In [27], both Cohen's kappa and Fleiss' kappa coefficients are used to assess the inter-annotator and intra-annotator agreement for five or nine categories among twelve pathologists on 1790 cervical biopsy specimens from 850 patients. In [28], Cohen's kappa coefficient is used to evaluate the agreement between two radiologists on TB diagnosis. In [29], [30], Cohen's kappa coefficient is applied to assess the agreement between two annotators for CXR findings in COVID-19 and the diagnosis of pneumonia, respectively. In [31], weighted Cohen's kappa coefficient is calculated to assess the agreement between lung ultrasound and CXR for classification between normal, unilateral, or bilateral pulmonary infiltrates. Although kappa statistics have been widely used for categorical classification and certain works have applied Cohen's kappa in image segmentation, we are not aware of any previous work that explores and interprets Fleiss's kappa coefficient for medical image segmentation.

Staple consensus. STAPLE is a well-known expectation-maximization algorithm for multi-annotator segmentation evaluation that generates a ground truth segmentation map from the annotations of multiple experts while providing a performance measure associated with each individual annotation. In [32]–[34] the STAPLE algorithm is applied to cervigram images to generate a ground truth from multiple cervix segmentations, whereas in [32], [33] the authors also utilize the STAPLE algorithm to evaluate automatic cervix segmentation algorithms. In [35], [36], the authors applied the STAPLE method to build ground truth segmentations consensus from two bounding-box-based annotations for CXR lung abnormalities and CXR COVID-19 findings, respectively, as well as used IoU and Dice scores to compare the agreement between segmentation STAPLE consensus and segmentations from DL models.

III. Inter-Annotator reliability assessment

A. Agreement heatmap

In this study, we propose two agreement heatmaps to assess the inter-annotator reliability that are different from those used in radiomics [19], [20] and for the understanding of AI model prediction [21]. We propose a common agreement heatmap that can be used for all kinds of annotations and a ranking agreement heatmap that can be used when a ranking order is also included in the annotation.

A common agreement heatmap is generated using the sum of annotation masks from multiple annotators. Assume $Mask\_i$ is a binary mask generated by the i^{th} annotator, in which non-zero values correspond to lesion areas. Then the common agreement heatmap can be computed using:

$$\text{Heatmap}(x,y) = \sum\nolimits_{i=1:N} \text{Mask\_i}(x,y), \qquad (1)$$

where N indicates the number of annotators, and (x,y) represents the coordinates. A higher pixel value in the heatmap indicates higher inter-annotator agreement. The maximum possible value in the heatmap is N.

If annotations include ranking numbers indicating the severity of lesions, then a ranking agreement heatmap can be generated using the average of ranking masks:

$$\text{Heatmap\_r}(x,y) = \frac{1}{\text{N}} \sum\nolimits_{i=1:5} \text{MaskR\_i}(x,y), \qquad (2)$$

where $MaskR\_i$ is a ranking mask generated by the i$^{\text{th}}$ annotator, and (x,y) represents the coordinates. The pixel value, which indicates the inter-annotator agreement, is called the heatmap agreement score in our study.

By taking the lesion annotations in cervical images as an example and assuming that the maximum number of annotated lesion areas is L, we use a non-linear function $y = [a^{x-b}]$ to represent the relationship between pixel values of lesion areas in a ranking mask and the ranking number, where y is an integer indicating the pixel value in the ranking mask, x indicates the ranking number, and a and b are constants that ensure the value of y will decrease in a non-linear manner and not go beyond zero. In this study, $y = [0.77^{x-13}]$ and $L = 10$. The two constants $a = 0.77$ and $b = 13$ are empirical values to ensure that pixel values from $x = 1$ to $x = L$ are different and decrease with a reducing speed and could be changed to other values that satisfy the two conditions. Since $x \in [1,10]$, pixel values in a ranking mask $y \in [2,23]$. When using (2) to generate the ranking agreement heatmap, the maximum value is 23, which indicates all the N annotators agree with a given lesion with rank = 1 (most severe). Fig. 1 shows an example of a ranking agreement heatmap on a cervical colposcopy image for precancer/cancer lesion annotation from five annotators. For the upper left area (in dark red), four annotators rated it as rank = 1 and one annotator rated it as rank = 2, and thus the pixel values in this area (in dark red) are calculated as (4×23+1×18)/5 = 22.

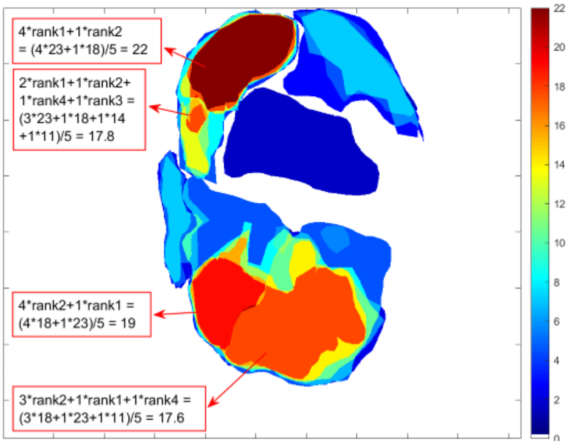


FIGURE 1. A ranking agreement heatmap generated from five annotations for cervical lesion segmentation.

The ranking agreement heatmap may provide more accurate guidance for further evaluation, such as biopsy sampling. In cervical cancer screening, colposcopy results are typically characterized by a colposcopic impression and the selection of the worst-appearing site for biopsy [37]. However, studies from screening and vaccination trials have suggested that colposcopic impression and biopsy sampling are poorly reproducible [38] and fail to detect 30% to 50% of prevalent high-grade squamous intraepithelial lesions (HSILs) [39], [40]. These data also suggest that taking more biopsies increases the detection of HSILs. A ranking agreement heatmap may help identify the most suspicious lesion area(s), i.e., the highest scoring area(s).

B. Extending Kappa Coefficient from Classification to Segmentation

Although the kappa coefficient [7], [41] has been widely used to assess inter-annotator agreement, most previous works explored it with classification tasks such as cervical cancer classification [25]–[27] and disease/survival prediction in CXR [29]–[31], among others. Our literature review reveals that the Fleiss' Kappa coefficient has never been explored with medical image segmentation. In this section, we first describe the calculation of the general kappa coefficient and then explain how we extend it to image segmentation.

Kappa statistic is a quantitative metric initially proposed to measure the extent of agreement between multiple annotators when classifying a given group of subjects into several categories. In medicine, kappa statistic is used to determine the agreement between categorical ratings made by two or more annotators and agreement between categorical ratings made by the same annotator on two or more occasions, representing *inter-annotator reliability* and *intra-annotator reliability*, respectively. It is calculated as the observed agreement beyond chance divided by the maximum agreement beyond chance:

$$k = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e} \quad , \tag{3}$$

where $\bar{P}$ is the proportion of observed agreements and $\bar{P}_e$ is the proportion of agreements expected by chance. Cohen's kappa coefficient [7] and Fleiss' kappa coefficient [41] are the most widely used kappa statistics. They are discussed below:

Cohen's kappa coefficient [7]**:** Cohen's kappa coefficient is commonly used to measure the level of agreement between two annotators, who each classify *M* subjects into *K* mutually exclusive categories. $\bar{P}$ and $\bar{P}_e$ in (3) can be calculated as shown in (4) and (5).

$$\bar{P} = \frac{Number\ in\ agreement}{Total} \quad , \tag{4}$$

$$\bar{P}_e = \frac{1}{M^2}\sum_{k=1}^{K} n_{k1} n_{k2} \ , \tag{5}$$

where n_{ki} is the number of times that annotator *i* predicted category *k*. Taking the simplest case of two-category classification rated by two annotators for *M* patients as an example, the agreement matrix (also called confusion matrix when measuring between ground truth and predicted results) is shown in Table I. *A+D* is the number of patients for whom two annotators agree, and *B+C* is the number of patients for whom they disagree. Then, $\bar{P}$ and $\bar{P}_e$ can be calculated by the following formula.

$$\bar{P} = \frac{(A+D)}{A+B+C+D} \quad , \tag{6}$$

$$\bar{P}_e = \frac{(A+B)(A+C)}{(A+B+C+D)^2} + \frac{(C+D)(B+D)}{(A+B+C+D)^2} \quad . \tag{7}$$

In this work, we extend Cohen's kappa from classification to segmentation as follows:

- First, a similar agreement matrix as in Table I is established, which now accounts for each image instead of all patients. Fig. 2 illustrates the counts for Cohen's kappa agreement matrix in image segmentation. A, B, C, and D indicate the number of pixels included in both segmentations, only in the segmentation from Annotator 2, only in the segmentation from Annotator 1, and outside of both segmentations, respectively. In segmentation, categories 1 and 2 are segmentation and background, respectively.
- Then, the final Cohen's kappa coefficient on all patients is calculated as the average of Cohen's kappa coefficients of all images.

TABLE I.
AGREEMENT MATRIX OF AGREEMENT AND DISAGREEMENT FROM TWO ANNOTATORS. A+D IS THE NUMBER OF PATIENTS FOR WHOM TWO ANNOTATORS AGREE, AND B+C IS THE NUMBER OF PATIENTS FOR WHOM THEY DISAGREE.

		Annotator 1	
		Category 1	Category 2
Annotator 2	Category 1	A	B
	Category 2	C	D

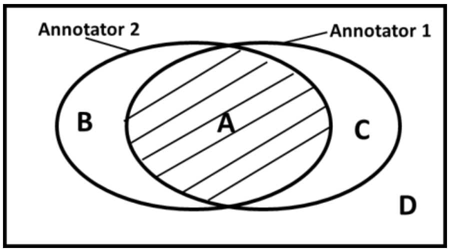


FIGURE 2. **Annotation results from two annotators, with A, B, C, and D indicating the number of pixels included in both annotated areas, only in the annotated area of Annotator 2, only in the annotated area of Annotator 1, and in both background areas, respectively.**

Fleiss' kappa coefficient [41]**:** Fleiss' kappa coefficient is proposed to measure the reliability of agreement between more than two annotators when classifying several subjects into different categories. Let *M* be the total number of subjects, let *N* be the number of annotators per subject, and let *K* be the number of categories into which classifications are made. The subjects are indexed by *i = 1, ... M* and the categories are indexed by *j = 1, ... K*. Let n_{ij} represent the number of annotators who classify the i^{th} subject to the j^{th} category. The Fleiss' kappa coefficient can be calculated in four steps. First, the proportion p_j of all classifications to the j^{th} category is calculated using the following equation:

$$p_j = \frac{1}{MN}\sum_{i=1}^{M} n_{ij} \quad . \tag{8}$$

Second, the agreement rate P_i among the *N* annotators for the i^{th} subject can be indexed by the proportion of agreeing pairs out of all the $N(N-1)$ possible pairs of classifications [41]:

$$\begin{aligned} P_i &= \frac{1}{N(N-1)}\sum_{j=1}^{K} n_{ij}(n_{ij}-1) \\ &= \frac{1}{N(N-1)}\left[\sum_{j=1}^{K} n_{ij}^2 - N\right] . \end{aligned} \tag{9}$$

Third, $\bar{P}$ and $\bar{P}_e$ that indicate the mean of P_i and the squared total value of p_j, respectively, are computed using the following equations:

$$\bar{P} = \frac{1}{M}\sum_{i=1}^{M} P_i = \frac{1}{MN(N-1)}\left(\sum_{i=1}^{M}\sum_{j=1}^{K} n_{ij}^2 - MN\right), \quad (10)$$

$$\bar{P}_e = \sum_{j=1}^{K} p_j^2 \,, \quad (11)$$

Finally, the Fleiss' kappa coefficient is calculated using (3).

Table II shows an example of the application of the Fleiss kappa coefficient, where 15 annotators classify ten patients into five categories. We denote the number of annotators who classify the i^{th} subject to the j^{th} category as $\boldsymbol{n_{ij}}$. The row of Subject 1 shows that all fifteen annotators agree that the first subject belongs to Category 5. The sum of each row consistently equals the number of annotators.

TABLE II.
COUNTS OF AGREEMENT BETWEEN FIFTEEN ANNOTATORS ON THE FIVE-CATEGORY CLASSIFICATION OF TEN SUBJECTS. $\boldsymbol{n_{ij}}$ IS THE NUMBER OF ANNOTATORS WHO CLASSIFY THE ITH SUBJECT TO THE JTH CATEGORY.

$\boldsymbol{n_{ij}}$		Category				
		1	**2**	**3**	**4**	**5**
Subject	**1**	0	0	0	0	15
	2	0	2	6	4	3
	3	0	0	3	5	7
	4	0	3	9	3	0
	5	1	2	8	1	3
	6	7	7	0	0	1
	7	3	2	7	3	0
	8	2	6	3	2	2
	9	7	5	2	1	0
	10	0	2	2	3	8

TABLE III.
COUNTS OF AGREEMENT BETWEEN FIFTEEN ANNOTATORS ON BINARY SEGMENTATION OF ONE IMAGE. $\boldsymbol{n_{ij}}$ IS THE NUMBER OF ANNOTATORS WHO ANNOTATE THE *I*TH SUBJECT TO THE *J*TH CATEGORY. THE TOTAL NUMBER OF ANNOTATORS IS 15.

$\boldsymbol{n_{ij}}$		Category	
		ROI	**Background**
Pixel	**1**	15	0
	2	13	2
	3	15	0
	4	12	3
	5	13	2
	6	8	7
	7	13	2
	8	10	5
	…	…	…
	n	13	2

We extend Fleiss's kappa coefficient from classification to image segmentation as follows:

- First, a Fleiss' kappa table is established for each image. As shown in Table III, each row indicates a pixel, and each column represents a category (ROI and background. We denote the number of annotators who annotate the i^{th} pixel to the j^{th} category as $\boldsymbol{n_{ij}}$, corresponding to an element in Table III. Again, the sum of each row consistently equals the number of annotators.
- Then, p_j and P_i are calculated using (8) and (9), and $\bar{P}$ and $\bar{P}_e$ are computed using (10) and (11).
- The Fleiss' kappa coefficient for one image is then calculated using (3). The final Fleiss' kappa coefficient for all patients is calculated as the average of Fleiss' kappa coefficients of all images.

The extension of kappa coefficients to segmentation is important for evaluating the inter-annotator and intra-annotator reliability of multiple segmentations, and for assessing automatic segmentation models. Table IV lists the difference between extended and general kappa coefficients.

TABLE IV.
CHARACTERISTICS FOR KAPPA COEFFICIENTS BEFORE AND AFTER EXTENSION.

	Multiple annotators	**Classification**	**Segmentation**
Fleiss's kappa	Yes	Yes	No
Cohen's kappa	only for two	Yes	No
Extended Fleiss' kappa	Yes	Yes	Yes
Extended Cohen's kappa	only for two	Yes	Yes

Interpretation for the kappa coefficient: The kappa coefficient ranges from -1 (worst) to +1 (best). The higher the coefficient, the better the inter-annotator reliability. Landis and Koch [42] proposed the following interpretation for two-annotator two-class classification: values ≤ 0 as no agreement, 0.01–0.20 as slight agreement, 0.21–0.40 as fair agreement, 0.41– 0.60 as moderate agreement, 0.61–0.80 as substantial agreement, and 0.81–1.00 as almost perfect agreement. Similar formulations have been proposed [43]–[45], but with slightly different descriptors. Sim and Wright [8] pointed out that the effects of prevalence and bias on the kappa coefficient should be considered when judging the reliability of agreement and that the number of categories would affect the kappa value.

C. STAPLE-based Consensus [12], [46]

A common question arises when we have segmentations from multiple annotators: what is the ground truth we can use to train an AI-based model? STAPLE algorithm [12], which is a widely used approach to aggregate multiple annotations, which uses expectation-maximization (EM) to find sensitivity and specificity values that maximize the data likelihood. In this study, we use the STAPLE algorithm to characterize the inter-annotator variability through annotators' sensitivity and specificity estimation and to generate a consensus reference segmentation. Further, sensitivity, specificity, and IoU scores are calculated to evaluate the performance of each annotator.

Consider an image of N voxels. Let $\mathbf{q} = (q_1, q_2, \ldots, q_M)^T$ be a column vector of M elements, with each element a sensitivity parameter characterizing one of M segmentations, and $\mathbf{r} = (r_1, r_2, \ldots, r_M)^T$ be a column vector of M elements, with each element a specificity parameter characterizing the performance of one of M segmentations. Let $\mathbf{D}$ denote an $N \times M$ matrix that describes binary decisions made for each segmentation at each image voxel. Let $\mathbf{T}$ denote an indicator vector containing N elements representing hidden true binary segmentation where for each voxel, the structure of interest is recorded as present (1) or absent (0). The complete data can be written as ($\mathbf{D}$, $\mathbf{T}$) and the probability mass function as $f(\mathbf{D}, \mathbf{T}|\mathbf{q}, \mathbf{r})$. The goal of the STAPLE algorithm is to estimate the performance level of the annotators characterized by ($\mathbf{q}$, $\mathbf{r}$) using the EM algorithm, which maximizes the data log-likelihood function

$$(\mathbf{q}', \mathbf{r}') = argmax_{\mathbf{q},\mathbf{r}} \ln f(\mathbf{D}, \mathbf{T}|\mathbf{q}, \mathbf{r}). \tag{12}$$

We use the STAPLE-based consensus as the ground truth and measure its agreement with the segmentation from each annotator using sensitivity, specificity, and IoU scores. The sensitivity and specificity are defined as,

$$\text{Sensitivity} = \frac{TP}{FN + TP}\ , \tag{13}$$

$$\text{Specificity} = \frac{TN}{FP + TN}\ , \tag{14}$$

where TP, FP, TN, and FN indicate true positives, false positives, true negatives, and false negatives, respectively. The IoU metric, also named the Jaccard Index, is defined as a ratio between the area of overlap and the area of union between two segmentations:

$$\text{IoU} = \frac{\text{Area of overlap}}{\text{Area of union}}\ . \tag{15}$$

IV. Experimental results

A. Data

We use two datasets to assess inter-annotator reliability: (i) cervical colposcopy images and (ii) the Shenzhen TB CXR dataset [47], [48].

The cervical image dataset includes 510 images from 30 patients acquired using the dynamic spectral imaging (DSI) colposcope (DYSIS by DYSIS Medical, Edinburgh, UK) with a digital video camera. For each patient, a series of 17 images were acquired with an interval of seven to ten seconds, from before to after the application of acetic acid, while visualizing the cervix using the DYSIS digital colposcope [49]. After completing the data acquisition, biopsies were taken according to colposcopic impression. Histology is considered the gold standard for this dataset. The biopsy sampling position was not reported. According to the biopsy results, the dataset includes six CIN3 (Cervical Intra-epithelial Neoplasia 3) patients, four CIN2 patients, 11 CIN1 patients, six negative patients, and three patients with unknown biopsy results. Five experienced colposcopists annotated lesion areas on the image at the 56th second (considered to be a typical indication of aceto-whiteness) via an interactive video/image annotation tool CVAT [50], with a ranking number ordering lesions according to perceived severity. Rank=1 indicates the most severe lesion. CIN1 is not considered as precancer since it usually regresses without treatment. CIN2 or CIN3 is not cancerous but may become cancer and spread to nearby normal tissue if not treated.

The Shenzhen TB dataset [47], [48] includes 326 CXRs from non-TB patients and 336 CXRs from TB patients. Two radiologists annotated the 336 TB patients for 19 abnormalities, including pleural effusion, apical thickening, single nodule (non-calcified), pleural thickening (non-apical), calcified nodule, small infiltrate (non-linear), cavity, linear density, severe infiltrate (consolidation), thickening of the interlobar fissure, clustered nodule (2mm-5mm apart), moderate infiltrate (non-linear), adenopathy, calcification (other than nodule and lymph node), calcified lymph node, miliary TB, retraction, other, and unknown [48]. The cases in the dataset have been confirmed by culture, and that typical TB appearance in imaging combined with a positive response to anti-TB medication was a criterion for confirming TB. The CXRs vary in dimensions but are approximately 3000×3000 pixels. The Shenzhen TB dataset and annotations from one radiologist are publicly available[1]. We will make the collection of annotations from the second radiologist published along with the paper.

B. Pre-processing

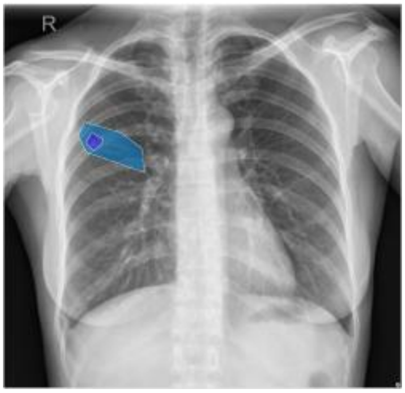


FIGURE 3. **An example of overlapping abnormalities in a CXR image annotated by the same annotator.**

In the Shenzhen CXR dataset, different abnormalities annotated by the same annotator for a given image may overlap. As shown in Fig. 3, the dark blue area is annotated as a "calcified nodule," and the light blue area is annotated as a "clustered nodule." The dark blue area belongs to both categories. This study does not consider such multi-label

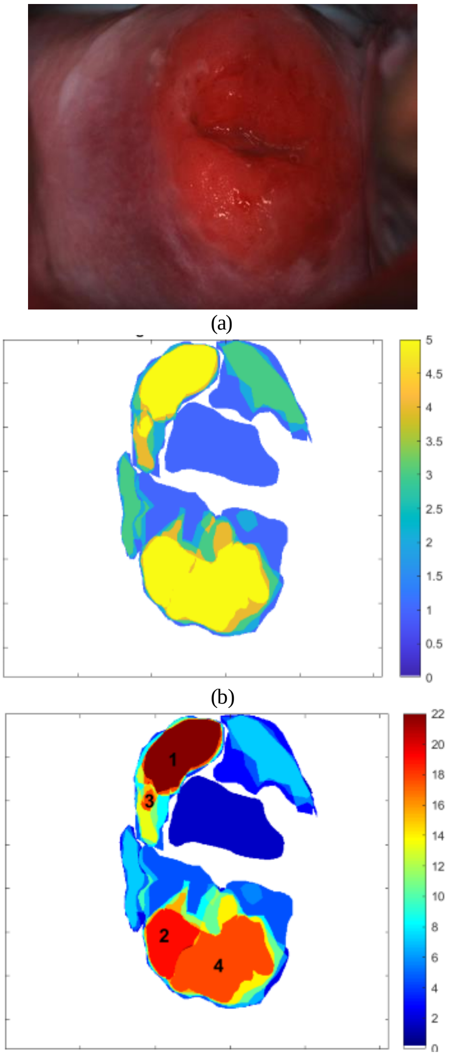
[1] https://data.lhncbc.nlm.nih.gov/public/Tuberculosis-Chest-X-ray-Datasets/Shenzhen-Hospital-CXR-Set/index.html

problems because they may cause overcounted agreement in heatmaps. To give each pixel a unique label, we excluded the small area from the larger one. In addition, we do not differentiate abnormality types in agreement heatmaps, kappa analysis, and STAPLE consensus. That is, all 19 abnormalities are considered an abnormality union, and the Kappa coefficient calculation and STAPLE consensus are obtained based on binary segmentation, i.e., foreground TB-consistent pixels and background pixels.

C. Results

Qualitative and quantitative analyses are performed to measure the inter-annotator reliability in terms of the agreement heatmap(s), kappa statistics, and STAPLE consensus along with sensitivity, specificity, and IoU score.

1) COLPOSCOPY IMAGES


(a)

(b)

(c)

FIGURE 4. **Dynamic cervical images and corresponding agreement heatmaps from five annotators. (a) Dynamic colposcopy images. (b) The common agreement heatmap obtained from five segmentations. The areas in yellow (upper left and lower part) had the highest value of 5. (c) The ranking agreement heatmap computed from the five ranking annotations. The top four severe lesion areas correspond to the four areas with the top four high values, marked as lesions 1, 2, 3, and 4.**

Fig. 4 shows an example of cervical lesion segmentations from five annotators. The histology for this case was CIN3. Fig.4(a) shows the dynamic images in colposcopy based on which the five annotators determined lesion areas by comparing the aceto-whitening changes. Fig. 4(b) and Fig. 4(c) show the common agreement heatmap and the ranking agreement heatmap of the annotations, respectively. We observe in the common agreement heatmap that the highest agreement between the five annotators was achieved in the upper left (in yellow) and bottom areas (in yellow). In contrast, in the ranking agreement heatmap, there is better discrimination, and the top four severe lesion areas correspond to different values (colored in dark red to orange). In Table V we present the overall agreement among five annotators in cervical colposcopy images using Fleiss' kappa coefficient. We observe that the five annotators achieve a fair agreement on the 30 patients and achieve a moderate agreement when excluding patients with histology label Unknown and CIN2, whose diagnosis and distinction from CIN1 and CIN3 is a well-recognized problem in clinics.

TABLE V.
RELIABILITY OF AGREEMENT BETWEEN FIVE ANNOTATORS FOR THE CERVICAL IMAGES FROM 30 PATIENTS USING FLEISS KAPPA ANALYSIS.

Number of patients	Fleiss' kappa	Agreement level	Patients
30	0.3740±0.2305	Fair	All
20	0.4122±0.2083	Moderate	Excludes patients with the label "CIN2" or "Unknown."

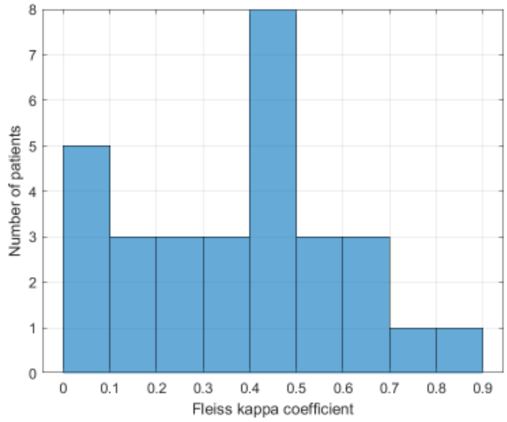


FIGURE 5. **Fleiss' kappa coefficient vs patient number.**

Fig. 5 shows the distribution of Fleiss' kappa coefficients for the 30 patients. There is slight agreement (kappa coefficient range 0-0.20) for eight patients, among whom two are unknown type in histology, four are Negative and CIN1 and two are CIN2. There is fair agreement (kappa coefficient range 0.21-0.4) for six patients, among whom four are Negative and CIN1, one is CIN2, and one is CIN3. There is moderate agreement (kappa coefficient range 0.41-0.6) for 11 patients, including one negative case, seven CIN1 cases, two CIN2 cases, and one Unknown case. There is substantial agreement (kappa coefficient range 0.61-0.8) for four patients, corresponding to one Negative case, one Unknown case, and two CIN3 cases. There is almost perfect agreement (kappa coefficient range 0.81-1.0) for a CIN3 patient.

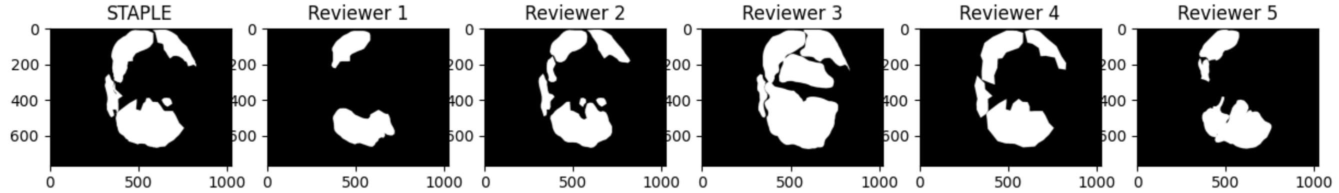


FIGURE 6. **STAPLE-based consensus and binary masks from the five annotators on a cervical image. The STAPLE image indicates the STAPLE-based consensus, and Reviewer 1 to Reviewer 5 indicates the binary masks generated from annotations of five annotators.**

The STAPLE-based consensus is shown in Fig. 6, which is a weighted voting result of annotations from the five reviewers. Sensitivity and specificity are calculated to compare the agreement between each annotator with STAPLE-based consensus, and we find that reviewers 2, 3, and 4 consistently achieve higher sensitivity than reviewers 1 and 5 among the 30 patients, whereas reviewer 3 consistently has the lowest specificity among 30 patients. Fig. 7 shows the inter-annotator agreements in each histology class. The Fleiss' kappa coefficient, IoU score, and normalized ranking agreement heatmap value are shown in blue, orange, and gray, respectively. The three different metrics consistently show that the agreement of multiple annotators increases from the Negative category to the CIN3 category, except for CIN2.

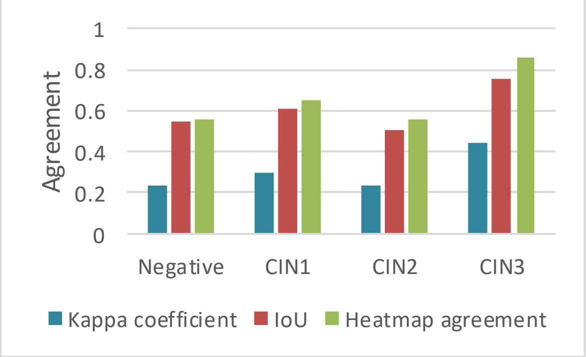
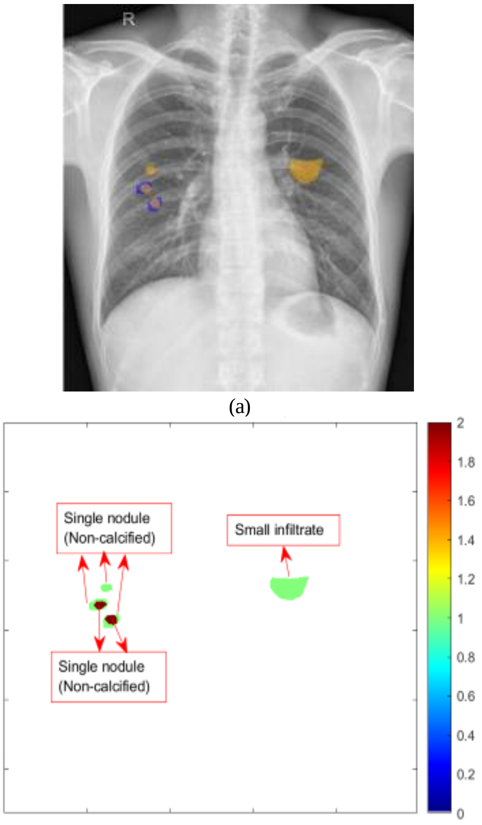


FIGURE 7. **Agreement vs categories in histology for the cervical dataset. Y-axis indicates the agreement score by different metrics, while the x-axis indicates the different categories in histology. The mean Fleiss' kappa coefficient, mean IoU score and mean value of normalized ranking agreement heatmap value are shown in blue, orange, and gray, respectively. IOU is calculated by comparing the ground truth by STAPLE with each segmentation of the five annotators. The heatmap agreement score in each category is computed using the following steps: 1) the ranking heatmap is normalized using the equation Y = (X-Min)/(Max-Min); 2) the mean heatmap score in the given category is calculated by averaging the normalized heatmap values in this category.**

2) *SHENZHEN TB CXR DATASET*

Fig. 8 shows an example of abnormality annotations from two annotators in a TB CXR image. Fig. 8 (a) illustrates annotations from both annotators overlapping with the original CXR. Blue areas are abnormalities annotated by the first annotator, while the second annotator annotates yellow areas. Fig. 8 (b) is a common agreement heatmap that shows the agreement between the two annotators. From the figure, we can see that they achieve an agreement on the nodule areas in dark red.



(a)

(b)

FIGURE 8. **Agreement heatmap of abnormality annotations from two experienced annotators on a chest X-ray image of a TB patient. (a) Annotated abnormalities from two reviewers overlapped with the original CXR. The two blue areas are annotated by Annotator 1, and the yellow areas are annotated by Annotator 2. (b) Common agreement heatmap from two annotators who achieve the highest agreement at the two dark red areas in the right lung.**

Fig. 9 shows the distribution of Cohen's kappa coefficients for 336 CXRs, and Fig.10 illustrates examples at different agreement levels. Among the 336 CXRs, 17 cases for which both annotators annotated abnormalities but in non-overlapping different positions, corresponding to

Cohen's kappa coefficients of less than zero and indicating that there is no agreement between the two annotators, as shown in the first bar in Fig. 9. An example of such situation is shown in Fig. 10 (a). The first annotator did not find any abnormalities in six of the 336 CXRs, and the second annotator did not find any abnormalities in 12 CXRs. For the 18 patients, there is no overlapping ROI between annotation masks from the two annotators, and thus Cohen's kappa coefficients are zero, which is shown in the second bar in Fig. 9. The second bar in Fig. 9 also includes 15 patients for whom annotated masks from the two annotators overlap only in a very small area. Corresponding examples are shown in Fig.10 (b) and (c). Table VI lists the overall agreement between the two annotators using Cohen's kappa coefficient. The two annotators achieve a moderate agreement on the 336 CXRs.

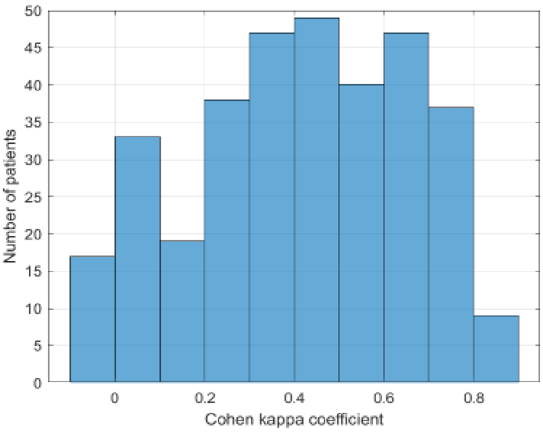


FIGURE 9. Histogram of Cohen's kappa coefficients for 336 CXRs in the Shenzhen dataset. Kappa coefficients less than zero indicate no agreement between two annotators, e.g., for patients for whom two annotators annotated in different locations, shown in Fig.10(a). Kappa coefficients of zero correspond to patients for whom one of the two annotators does not find any abnormality in the CXR, as shown in Fig. 10(b).

TABLE VI.
RELIABILITY OF AGREEMENT BETWEEN TWO ANNOTATORS FOR CHEST X-RAY TB PATIENTS USING COHEN'S KAPPA ANALYSIS.

No of patients	Cohen's kappa	Agreement level	Patients
336	0.4145±0.2427	Moderate	All included
318	0.4380±0.2279	Moderate	Excludes patients in Condition 1
301	0.4630±0.2078	Moderate	Excludes patients in Condition 1 or Condition 2

Note: Condition 1 – one annotator finds no abnormality; Condition 2 – there is no overlap between the abnormality annotations from two annotators.

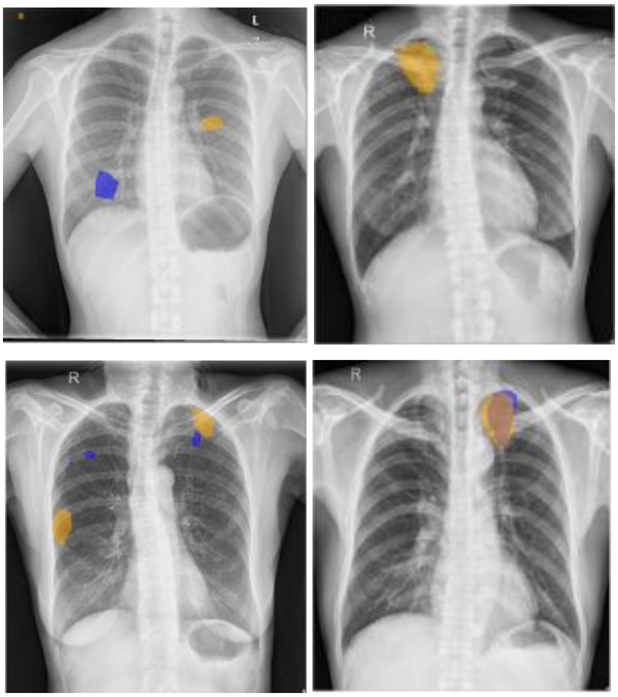
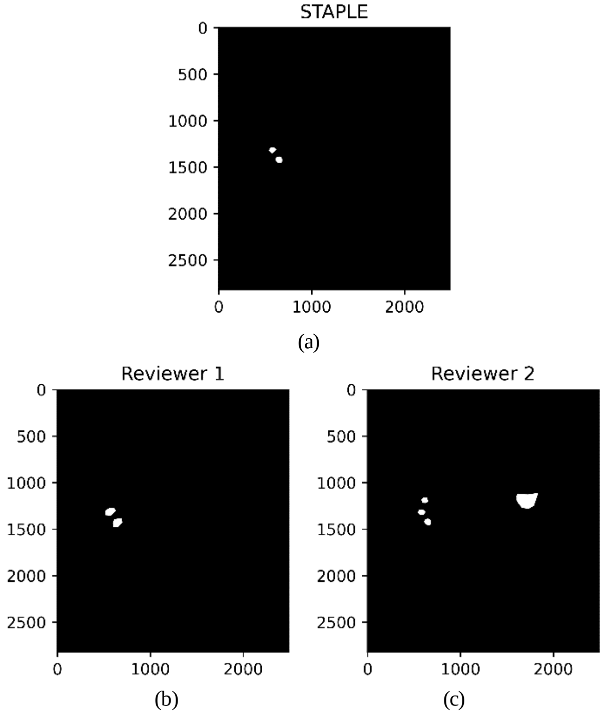


FIGURE 10. Annotations at different levels of agreement. (a) No agreement. Kappa = -0.0032; (b) No agreement. Kappa = 0; (c) Slight agreement, Kappa = 0.0100; (d) Almost perfect agreement, Kappa = 0.8096. Blue and yellow areas correspond to abnormalities annotated by Annotator 1 and Annotator 2, respectively.



(a)

(b)

(c)

FIGURE 11. STAPLE-based consensus and binary masks from two annotators for the patient shown in Fig.8. (a) STAPLE-based consensus. (b) Abnormality annotations from Annotator 1. (c) Abnormality annotations from Annotator 2.

The STAPLE-based consensus and annotations from two annotators for the CXR example in Fig.8 are illustrated in

Fig. 11. To achieve fast convergence, STAPLE was only performed on the 301 CXRs for which the kappa score is larger than zero. The histogram of IoU scores for both annotators for 301 CXRs (with STAPLE consensus as the ground truth) is shown in Fig.12, in which patients are excluded if no abnormality is found by one annotator or no overlap is found between the abnormality annotations between two annotators. The mean IoU score for 301 CXRs across two annotators is 0.5407.

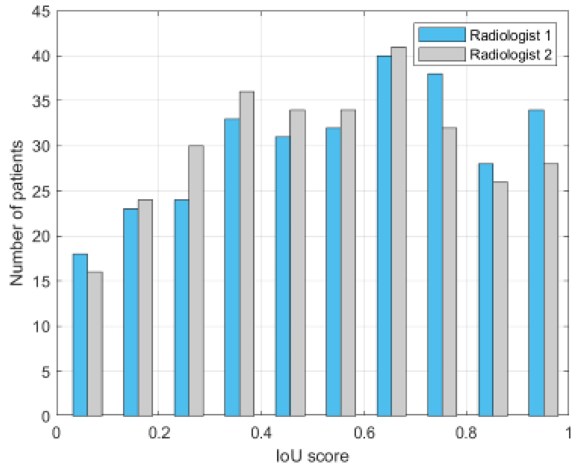


FIGURE 12. Histogram of IoU scores for 301 CXRs in the Shenzhen dataset. Patients are excluded if no abnormality is found by one annotator, or if no overlap is found between the abnormality annotations between two annotators.

V. Discussion and Conclusion

Evaluation of ground truth annotations and inter-annotator agreement from multiple segmentations in image segmentation is challenging. Human annotators can easily achieve high agreement on the presence or absence of an object in a certain image region but are more likely to generate different contours when they are asked to annotate the outline of the same object in the same image. It would be more challenging to achieve an agreement when the annotations are for many objects with different sizes and different types (the situation in our CXR dataset).

It is important to evaluate the inter-annotator agreement before using the annotations to train AI models since the model performance is affected by the inter-annotator agreement. It is generally assumed that human inter-annotator agreement defines the upper limit on our ability to measure automated performance [51].

In colposcopy, agreement heatmaps with ranking can highlight the most suspicious cervical lesions, which could be used to guide biopsy sampling. The diagnosis of CIN2 and its distinction from CIN1 and CIN3 is a well-recognized problem since CIN2 is not biologically homogeneous, with some aligning with CIN1 (intermediate pattern) and some with CIN3 (transforming pattern) [46]. Therefore, it is reasonable to observe low agreement in CIN2, which is consistent with agreement heatmap, kappa analysis, and STAPLE-based IoU score. The small number of samples may also contribute to this issue.

Extending kappa coefficients from classification to segmentation makes it possible to quantitatively evaluate and interpret the inter-annotator reliability of multiple segmentation annotations. To our knowledge, this is the first paper to extend and interpret the Fleiss' kappa table for medical image segmentation.

STAPLE is a proven algorithm dealing with multiple annotations for obtaining the ground truth and analyzing the inter-annotator reliability and variability. Although it has been widely used, STAPLE has some limitations: 1) it fails to converge when multiple annotations have no overlap, which may occur in lesion/abnormality annotations; 2) Since STAPLE relies on majority voting, it could tend to underestimate the edges of the structure that are traced; 3) If there is a high degree of variability between the annotators, it requires a greater number of annotators to obtain a meaningful or acceptable consensus.

Evaluating the inter-annotator agreement with different metrics is important to avoid bias when using a single metric. For example, the agreement heatmap from the dynamic cervical image dataset showed that the five annotators achieved high agreement for the most suspicious lesions, whereas the kappa coefficient showed that the inter-annotator agreement is fair, and adjustments (such as changing the number of annotators, limit the maximum number of lesions per image, etc.) could be made to improve agreement and generalization before training models. We also should note that the size of annotations could generate potential biases for the metrics. For example, it is easy to achieve more than 50% overlap between two annotators when segmenting objects that occupy a major part of an image, e.g., the lungs in a CXR or CT for 3D. However, this may not be even qualitatively considered an acceptable inter-annotator agreement. In this case, increasing the acceptable kappa threshold (e.g., kappa>0.8) could be considered as one approach. This study focused on lesions that tend to occupy a smaller part of the image. Challenges related to unifying segmentation disagreements between larger objects will be explored in future work.

The proposed metrics in this study have several clinical and practical implications for both health organizations and individuals. In clinical practice, the proposed metrics can be applied to a small subset to assess the robustness of annotations before applying it to a large dataset; this can allow clinicians to have a better understanding of the annotations and fine-tune annotation instructions. Further, organizations and individuals can use the proposed metrics to assess the quality of annotations in existing datasets before releasing these datasets or using them for model development. From a machine learning standpoint, the proposed metrics could be used to build robust annotations, leading to robust training and development for AI models. Further, the proposed metrics can be used to generate clean labels or remove noisy labels, which can enhance the performance of AI models and lead to faithful diagnosis/prediction.

To conclude, this paper systematically assessed the inter-annotator reliability for medical image segmentation using different metrics, including agreement heatmaps, extended kappa statistics, and STAPLE analysis. Experimental results on dynamic cervical images showed that an overall fair agreement is achieved between the five annotators, and the highest agreement is achieved on CIN3, which is the most severe category in histology. Experiments on TB CXR images showed an overall moderate agreement between the two annotators. We consider the annotations with a good agreement due to the uncertainty of the number and size of abnormalities when performing annotations. There are 19 abnormalities among the 336 CXRs, while each image may include zero to 19 abnormalities, and the size of the abnormality can be a very small area, such as a single nodule, or a large area, such as moderate infiltrate or miliary. When the annotated area is very small, background agreement D will be much larger than foreground agreement A in (6) and (7), which therefore results in a very small Cohen's kappa coefficient using (3). The reliability measurements on the agreement of multiple annotators are consistent with different metrics, including agreement heatmaps, kappa coefficient, and IoU score, which proves the validity of annotations as well as our proposed metrics. Finally, it is important to note that although the proposed agreement heatmaps and the extended kappa coefficients were evaluated on two medical imaging datasets, such metrics can be easily applied to any other data with multiple segmentation annotations.

References

[1] W. Ji *et al.*, "Learning Calibrated Medical Image Segmentation via Multi-rater Agreement Modeling," 2021. doi: 10.1109/CVPR46437.2021.01216.

[2] M. Schaekermann, G. Beaton, M. Habib, L. I. M. Andrew, K. Larson, and L. A. W. Edith, "Understanding expert disagreement in medical data analysis through structured adjudication," *Proceedings of the ACM on Human-Computer Interaction*. 2019. doi: 10.1145/3359178.

[3] L. Joskowicz, D. Cohen, N. Caplan, and J. Sosna, "Inter-observer variability of manual contour delineation of structures in CT," *Eur. Radiol.*, 2019, doi: 10.1007/s00330-018-5695-5.

[4] H. Fu *et al.*, "A retrospective comparison of deep learning to manual annotations for optic disc and optic cup segmentation in fundus photographs," *Transl. Vis. Sci. Technol.*, 2020, doi: 10.1167/tvst.9.2.33.

[5] M. L. McHugh, "Interrater reliability: The kappa statistic," *Biochem. Medica*, 2012, doi: 10.11613/bm.2012.031.

[6] S. Möller, B. Debrabant, U. Halekoh, A. K. Petersen, and O. Gerke, "An extension of the bland–altman plot for analyzing the agreement of more than two raters," *Diagnostics*, 2021, doi: 10.3390/diagnostics11010054.

[7] J. Cohen, "A Coefficient of Agreement for Nominal Scales," vol. XX, no. 1, pp. 37–46, 1960.

[8] J. Sim and C. C. Wright, "The kappa statistic in reliability studies: Use, interpretation, and sample size requirements," *Phys. Ther.*, vol. 85, no. 3, pp. 257–268, 2005, doi: 10.1093/ptj/85.3.257.

[9] A. C. Speciale *et al.*, "Observer variability in assessing lumbar spinal stenosis severity on magnetic resonance imaging and its relation to cross-sectional spinal canal area," *Spine (Phila. Pa. 1976).*, 2002, doi: 10.1097/00007632-200205150-00014.

[10] C. R. Burton, "Research in Health Care: Concepts, Designs and Methods," *J. Adv. Nurs.*, 2002, doi: 10.1046/j.1365-2648.2002.02288_7.x.

[11] J. Belur, L. Tompson, A. Thornton, and M. Simon, "Interrater Reliability in Systematic Review Methodology: Exploring Variation in Coder Decision-Making," *Sociol. Methods Res.*, 2021, doi: 10.1177/0049124118799372.

[12] S. K. Warfield, K. H. Zou, and W. M. Wells, "Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation," *IEEE Trans. Med. Imaging*, vol. 23, no. 7, pp. 903–921, 2004, doi: 10.1109/TMI.2004.828354.

[13] X. Liu, A. Montillo, E. T. Tan, and J. F. Schenck, "iSTAPLE: improved label fusion for segmentation by combining STAPLE with image intensity," 2013. doi: 10.1117/12.2006447.

[14] Q. Liu, Q. Dou, L. Yu, and P. A. Heng, "MS-Net: Multi-Site Network for Improving Prostate Segmentation with Heterogeneous MRI Data," *IEEE Trans. Med. Imaging*, 2020, doi: 10.1109/TMI.2020.2974574.

[15] G. Chen *et al.*, "Automatic Pathological Lung Segmentation in Low-Dose CT Image Using Eigenspace Sparse Shape Composition," *IEEE Trans. Med. Imaging*, 2019, doi: 10.1109/TMI.2018.2890510.

[16] S. Yu *et al.*, "Difficulty-aware glaucoma classification with multi-rater consensus modeling," 2020. doi: 10.1007/978-3-030-59710-8_72.

[17] T. A. Lampert, A. Stumpf, and P. Gançarski, "An Empirical Study Into Annotator Agreement, Ground Truth Estimation, and Algorithm Evaluation," *IEEE Trans. Image Process.*, vol. 25, no. 6, pp. 2557–2572, 2016, doi: 10.1109/TIP.2016.2544703.

[18] L. Wilkinson and M. Friendly, "History of the Cluster Heat Map,"*Am. Stat.*, 2009.

[19] R. J. Gillies, P. E. Kinahan, and H. Hricak, "Radiomics: Images are more than pictures, they are data," *Radiology*, 2016, doi: 10.1148/radiol.2015151169.

[20] M. T. Freedman and T. Osicka, "Heat Maps. An Aid for Data Analysis and Understanding of ROC CAD Experiments," *Acad. Radiol.*, 2008, doi: 10.1016/j.acra.2007.07.010.

[21] P. Rajpurkar *et al.*, "Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists," *PLoS Med.*, 2018, doi: 10.1371/journal.pmed.1002686.

[22] P. Guo *et al.*, "Network visualization and pyramidal feature comparison for ablative treatability classification using digitized cervix images," *J. Clin. Med.*, 2021, doi: 10.3390/jcm10050953.

[23] W. Kusakunniran *et al.*, "COVID-19 detection and heatmap generation in chest x-ray images," *J. Med. Imaging*, 2021, doi: 10.1117/1.jmi.8.s1.014001.

[24] D. Pal, P. B. Reddy, and S. Roy, "Attention UW-Net: A fully connected model for automatic segmentation and annotation of chest X-ray," *Comput. Biol. Med.*, vol. 150, p. 106083, 2022, doi: https://doi.org/10.1016/j.compbiomed.2022.106083.

[25] A. O. Raifu *et al.*, "Determinants of cervical cancer screening accuracy for visual inspection with acetic acid (VIA) and lugol's iodine (VILI) performed by nurse and physician," *PLoS One*, 2017, doi: 10.1371/journal.pone.0170631.

[26] B. Sherigar, A. Dalal, G. Durdi, Y. Pujar, and H. Dhumale, "Cervical cancer screening by visual inspection with acetic acid-interobserver variability between nurse and physician," *Asian Pacific J. Cancer Prev.*, 2010.

[27] A. Malpica *et al.*, "Kappa statistics to measure interrater and intrarater agreement for 1790 cervical biopsy specimens among twelve pathologists: Qualitative histopathologic analysis and methodologic issues," 2005. doi: 10.1016/j.ygyno.2005.07.040.

[28] S. Jaeger *et al.*, "Automatic tuberculosis screening using chest radiographs," *IEEE Trans. Med. Imaging*, 2014, doi: 10.1109/TMI.2013.2284099.

[29] M. Balbi *et al.*, "Chest X-ray for predicting mortality and the need for ventilatory support in COVID-19 patients presenting to the emergency department," *Eur. Radiol.*, 2021, doi: 10.1007/s00330-020-07270-1.

[30] M. Hassen *et al.*, "Radiologic Diagnosis and Hospitalization among Children with Severe Community Acquired Pneumonia: A Prospective Cohort Study," *Biomed Res. Int.*, 2019, doi: 10.1155/2019/6202405.

[31] M. M. González *et al.*, "Comparison of lung ultrasound versus chest x-ray for detection of pulmonary infiltrates in covid-19," *Diagnostics*, 2021, doi: 10.3390/diagnostics11020373.

[32] S. Lotenberg, S. Gordon, R. Long, S. Antani, J. Jeronimo, and H. Greenspan, "Automatic evaluation of uterine cervix segmentations," 2007. doi: 10.1117/12.708889.

[33] S. Gordon, S. Lotenberg, R. Long, S. Antani, J. Jeronimo, and H. Greenspan, "Evaluation of uterine cervix segmentations using ground truth from multiple experts," *Comput. Med. Imaging Graph.*, 2009, doi: 10.1016/j.compmedimag.2008.12.002.

[34] Z. Xue, L. R. Long, S. Antani, L. Neve, Y. Zhu, and G. R. Thoma, "A unified set of analysis tools for uterine cervix image segmentation," *Comput. Med. Imaging Graph.*, 2010, doi: 10.1016/j.compmedimag.2010.04.002.

[35] S. Rajaraman, L. R. Folio, J. Dimperio, P. O. Alderson, and S. K. Antani, "Improved semantic segmentation of tuberculosis—consistent findings in chest x-rays using augmented training of modality-specific u-net models with weak localizations," *Diagnostics*, 2021, doi: 10.3390/diagnostics11040616.

[36] S. Rajaraman, S. Somapudi, P. O. Alderson, L. R. Folio, and S. K. Antani, "Analyzing inter-reader variability affecting deep ensemble learning for COVID-19 detection in chest radiographs," *PLoS One*, 2020, doi: 10.1371/journal.pone.0242301.

[37] D. Saslow *et al.*, "American Cancer Society, American Society for Colposcopy and Cervical Pathology, and American Society for Clinical Pathology screening guidelines for the prevention and early detection of cervical cancer," *Am. J. Clin. Pathol.*, 2012, doi: 10.1309/AJCPTGD94EVRSJCG.

[38] L. S. Massad, J. Jeronimo, and M. Schiffman, “Interobserver agreement in the assessment of components of colposcopic grading,” *Obstet. Gynecol.*, 2008, doi: 10.1097/AOG.0b013e31816baed1.

[39] U. Indraccolo, E. Santi, P. Iannone, C. Borghi, and P. Greco, “Number of colposcopic cervical biopsies and diagnosis of cervical intraepithelial neoplasia: A prospective study,” *Eur. J. Gynaecol. Oncol.*, 2021, doi: 10.31083/j.ejgo4204100.

[40] M. H. Stoler *et al.*, “The accuracy of colposcopic biopsy: Analyses from the placebo arm of the Gardasil clinical trials,” *Int. J. Cancer*, 2011, doi: 10.1002/ijc.25470.

[41] J. L. Fleiss, “Measuring nominal scale agreement among many raters,” *Psychol. Bull.*, 1971, doi: 10.1037/h0031619.

[42] J. R. Landis and G. G. Koch, “The Measurement of Observer Agreement for Categorical Data,” *Biometrics*, 1977, doi: 10.2307/2529310.

[43] B. S. Everitt and J. L. Fleiss, “Statistical Methods for Rates and Proportions.,” *Biometrics*, 1981, doi: 10.2307/2530193.

[44] J. Ludbrook, “PRACTICAL STATISTICS FOR MEDICAL RESEARCH,” *Australian and New Zealand Journal of Surgery*. 1991. doi: 10.1111/j.1445-2197.1991.tb00019.x.

[45] P. E. Shrout, “Measurement reliability and agreement in psychiatry,” *Statistical Methods in Medical Research*. 1998. doi: 10.1191/096228098672090967.

[46] R. Van Baars *et al.*, “Investigating diagnostic problems of CIN1 and CIN2 associated with high-risk HPV by combining the novel molecular biomarker PanHPVE4 With P16INK4a,” *Am. J. Surg. Pathol.*, 2015, doi: 10.1097/PAS.0000000000000498.

[47] S. Jaeger, S. Candemir, S. Antani, Y.-X. J. Wáng, P.-X. Lu, and G. Thoma, “Two public chest X-ray datasets for computer-aided screening of pulmonary diseases.,” *Quant. Imaging Med. Surg.*, 2014, doi: 10.3978/j.issn.2223-4292.2014.11.20.

[48] F. Yang *et al.*, “Annotations of Lung Abnormalities in the Shenzhen Chest X-ray Dataset for Computer-Aided Screening of Pulmonary Diseases,” *Data*, vol. 7, no. 7, 2022, doi: 10.3390/data7070095.

[49] J. A. Louwers *et al.*, “Dynamic spectral imaging colposcopy: Higher sensitivity for detection of premalignant cervical lesions,” *BJOG An Int. J. Obstet. Gynaecol.*, 2011, doi: 10.1111/j.1471-0528.2010.02806.x.

[50] CVAT.ai Corporation, “Computer Vision Annotation Tool (CVAT),” 2022. https://github.com/opencv/cvat.

[51] J. Resnik, Philip; Lin, “Inter-annotator agreement and upper bounds,” in *The Handbook of Computational Linguistics and Natural Language Processing*, S. Clark, C.F. Alexander; Lappin, Ed. Wiley-Blackwel, 2010, pp. 271–295. doi: 10.1002/9781444324044.

Dr. **Feng Yang,** PhD, joined the Lister Hill National Center for Biomedical Communications (LHNCBC), National Library of Medicine (NLM) in October 2017. She is currently a Research Fellow at NLM of NIH. Dr. Yang had been working as a Principal Investigator, Associate Professor at Beijing Jiaotong University in China from 2012 to 2019. Dr. Yang received her PhD degree from the National Institute of Applied Science (INSA Lyon) in France in 2011, and her BS and MS degrees from Northwestern Polytechnical University in China in 2005 and 2007, respectively. Her current research interests include machine learning and artificial-intelligence-based biomedical image processing and analysis, and cardiac image processing. She has so far published more than 70 research papers, including 30 journal articles, 1 book chapter, and 40 conference proceedings. She has been the special session chairman of IEEE ICSP from 2012 to 2022.

Dr. **Ghada Zamzmi**, PhD, joined the Communication Engineering Branch at Lister Hill National Center for Biomedical Communications (LHNCBC) in February 2019. She received her PhD degree from University of South Florida in 2018. She focuses on using computational sciences and engineering techniques toward advancing the healthcare of vulnerable populations (e.g., infants and minority groups). Ghada research interests include medical imaging, affective and cognitive computing, human behavior analysis, and healthcare application. She has >12 journal papers published in top tier journals (e.g., Transactions on Affective Computing), >15 conference publications and two patents. She served in the program committee in several top conferences in Computer Vision including CVPR, NeurIPS, and MICCAI. She chaired several academic workshops and events in her area of interests, led and participated in several mentoring programs. Ghada received different prestigious awards such as MIT Innovator under 35 and IEEE Computational Life Sciences best PhD Thesis Award (2019). She’s selected as the North America Ambassador for the international organization Women in AI.

Sandeep Angara is a Research Scientist at the National Library of Medicine, National Institute of Health. His research interest includes deep learning, machine learning, and medical imaging. He received his Master's in Electrical Engineering from The University of Texas at Tyler. He has nine years of experience building computer vision applications using deep learning and machine learning. He has many published papers in various conferences and journals.

Dr**. Sivaramakrishnan Rajaraman**, PhD, is working as a research scientist at the National Library of Medicine (NLM), the National Institutes of Health (NIH). Dr. Rajaraman received his Ph.D. in Information and Communication Engineering from Anna University, India. He is a versatile researcher with expertise in machine learning, data science, biomedical image analysis, and computer vision. He has more than 15 years of experience in academia where he taught core and allied subjects in biomedical engineering. He has authored several national and international journal and conference publications in his area of expertise. Dr. Rajaraman is an Editorial Board member of the journals PLOS ONE and MDPI Electronics. He is reviewing manuscripts for more than 75 journals including those published by Nature, Lancet, IEEE, MDPI, Elsevier, and other leading conferences including CVPR, EMBS, CBMS, and MICCAI. Dr. Rajaraman is a Life Member of the Society of Photo-optical Instrumentation Engineers (SPIE), a regular member of the Institute of Electrical and Electronics Engineers (IEEE), IEEE Engineering in Medicine & Biology Society (EMBS), and the Biomedical Engineering Society (BMES).

Dr**. Zhiyun Xue**, PhD, is a Staff Scientist at the Lister Hill National Center for Biomedical Communications (LHC) at the National Library of Medicine (NLM). She obtained her Ph.D. from Lehigh University USA, M.S. and B.S. from Tsinghua University, China. Dr. Xue has been working at LHC since 2006 on a number of medical imaging informatics projects. By applying her knowledge and expertise in the fields of machine learning, image processing, and computer vision to analyze biomedical images in different modalities, Dr. Xue puts her R&D efforts in those projects with the goals of advancing the research in biomedical informatics and data science, assisting clinicians at the point-of-care, improving the health of the people, and addressing the needs of the underserved population.

Dr**. Stefan Jaeger** is a staff scientist at the Lister Hill National Center for Biomedical Communications at the United States National Library of Medicine (NLM), which is part of the National Institutes of Health (NIH). He received his diploma in computer science from the University of Kaiserslautern and his Ph.D. from the University of Freiburg, Germany, also in computer science. Dr. Jaeger has held research positions at the Chinese Academy of Sciences, the University of Maryland, the University of Karlsruhe, and Daimler, among others. At NLM, he supervises machine learning and data science research for diagnosing infectious diseases and conducts research into image informatics and artificial intelligence for clinical care and education. He has over a hundred publications in these areas, several of which received best paper nominations, including two patents.

Dr. **Sameer Antani** is currently a Tenure Track Investigator, in Computational Health Research Branch, the Lister Hill National Center for Biomedical Communications (LHNCBC), National Library of Medicine (NLM). His research interests include machine learning and artificial intelligence, medical image and signal processing, and information retrieval for solving challenging problems and advancing health. Dr. Antani is a senior member of the International Society of Photonics and Optics (SPIE), the Institute of Electrical and Electronics Engineers (IEEE). He serves as the vice chair for computational medicine on the IEEE Computer Society's Technical Committee on Computational Life Sciences (TCCLS) and the IEEE Life Sciences Technical Community (LSTC).