Computers in Biology and Medicine 134 (2021) 104497

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Comparison of metrics for the evaluation of medical segmentations using prostate MRI dataset

Ying-Hwey Nai [a,*,1], Bernice W. Teo [b], Nadya L. Tan [c], Sophie O'Doherty [a], Mary C. Stephenson [a,d], Yee Liang Thian [e], Edmund Chiong [f,g], Anthonin Reilhac [a,1]

[a] Clinical Imaging Research Centre, Yong Loo Lin School of Medicine, National University of Singapore, Singapore
[b] Nanyang Junior College, Singapore
[c] St. Joseph's Institution International, Singapore
[d] Centre for Translational MR Research, Yong Loo Lin School of Medicine, National University of Singapore, Singapore
[e] Department of Diagnostic Imaging, National University Hospital, Singapore
[f] Department of Surgery, Yong Loo Lin School of Medicine, National University of Singapore, Singapore
[g] Department of Urology, National University Hospital, Singapore

ARTICLE INFO

Keywords:
Prostate cancer
Medical image segmentation
Deep learning
Evaluation metrics
Rank evaluation

ABSTRACT

Nine previously proposed segmentation evaluation metrics, targeting medical relevance, accounting for holes, and added regions or differentiating over- and under-segmentation, were compared with 24 traditional metrics to identify those which better capture the requirements for clinical segmentation evaluation. Evaluation was first performed using 2D synthetic shapes to highlight features and pitfalls of the metrics with known ground truths (GTs) and machine segmentations (MSs). Clinical evaluation was then performed using publicly-available prostate images of 20 subjects with MSs generated by 3 different deep learning networks (DenseVNet, HighRes3DNet, and ScaleNet) and GTs drawn by 2 readers. The same readers also performed the 2D visual assessment of the MSs using a dual negative-positive grading of −5 to 5 to reflect over- and under-estimation. Nine metrics that correlated well with visual assessment were selected for further evaluation using 3 different network ranking methods - based on a single metric, normalizing the metric using 2 GTs, and ranking the network based on a metric then averaging, including leave-one-out evaluation. These metrics yielded consistent ranking with HighRes3DNet ranked first then DenseVNet and ScaleNet using all ranking methods. Relative volume difference yielded the best positivity-agreement and correlation with dual visual assessment, and thus is better for providing over- and under-estimation. Interclass Correlation yielded the strongest correlation with the absolute visual assessment (0–5). Symmetric-boundary dice consistently yielded good discrimination of the networks for all three ranking methods with relatively small variations within network. Good rank discrimination may be an additional metric feature required for better network performance evaluation.

1. Introduction

Medical image segmentation is an important task in the clinical workflow for image-guided treatment procedures, tracking disease progression by measuring tumor volume, disease staging and diagnosis, and image registration [1,2]. Manual segmentation is tedious and prone to inter-operator variability, and thus encouraged the development of automated or semi-automated methods using machine learning or deep learning. The performance of these methods is typically assessed using similarity or error metrics to compare the machine segmentation (MS) with a ground truth (GT), which is often manually drawn by trained human operators [2]. The number of available metrics, their variants, their relation to other metrics and more importantly the lack of understanding on their meaning make it difficult to interpret measures and identify those which are relevant for performance validation or comparison tasks. MS may be affected by anatomical variance and image quality such as artifacts, image inhomogeneity, noise, spatial resolution, volume averaging, contrast levels, and patient motion [2]. In addition,

* Corresponding author. Clinical Imaging Research Centre (CIRC), National University of Singapore, Centre for Translational Medicine (MD6), 14 Medical Drive, #B1-01, 117599, Singapore
E-mail addresses: mednyh@nus.edu.sg, yinghweynai@yahoo.com (Y.-H. Nai).
[1] The authors made an equal contribution to this manuscript.

https://doi.org/10.1016/j.compbiomed.2021.104497
Received 3 March 2021; Received in revised form 11 May 2021; Accepted 11 May 2021
Available online 15 May 2021

GT is also affected by factors such as individual skill levels, related clinical requirements, and subjective boundary uncertainty due to image quality [1] or individual preference (e.g. draw on, inside, or outside of the region boundary). GT also suffers from intra- and inter-reader variability and is sensitive to target size, where the smaller the target, the greater the variability. An alternative to using GT is to grade MS using human visual assessment (or manual grading). However, visual assessment is similarly exhausting, time-consuming, and subjective.

A major issue faced in segmentation methods comparison is the ranking of these methods, which is affected by the test data, the annotated GTs, choice of metrics, metric variant applied, and metric aggregation methods. Ranking can be performed with one or multiple metrics using metric-based (aggregate, then rank), or case-based ranking (rank then aggregate) [3]. For metric-based ranking, the metrics are either averaged and ranked accordingly or compared using significance tests [4,5] or selected metrics are normalized to a standard range using 2 GTs, averaged and ranked, to cater for inter-reader variability [6,7]. For case-based ranking, each case is ranked using the selected metrics and the final ranking is obtained by averaging the ranks for all cases [8]. There are many evaluation metrics measuring different and often complementary aspects of the segmentation and thus, they may not be comparable, and may yield contradicting scores and ranking [2,6].

The objective of this work is to compare various previously proposed segmentation evaluation metrics with traditional metrics to (1) determine if these metrics are better than traditional metrics and (2) to identify their relevance in clinical situations, which should correlate well with visual assessment of the segmentation. To simplify the comparison, we focused on comparing binary segmentation using (1) 2D synthetic shapes with known GTs and MSs, to identify features measured by the metric, and (2) actual clinical MRI data of whole prostate gland (WG) and their segmentation generated by three different deep learning networks. The comparison of MSs of WG was performed using GTs and visual assessment grading of 2 readers, and metrics that correlated well with visual assessment grading were identified. We attempted for the first time to use a dual negative-positive visual assessment grading system to evaluate metrics that could indicate over- and under-estimation, namely C-Factor (CF) [9] and relative volume difference (RVD). The changes in networks ranking with the selected metrics was then evaluated using 3 ranking methods, namely ranking based on the averages of selected metrics individually [4,5], normalizing the selected metrics using 2 GTs then averaging for ranking [7], and ranking each case using the selected metrics then averaging the rank [8].

2. Materials and methods

2.1. Evaluation metrics

Table 1 lists a total of 33 metrics that were evaluated in this work. We included most of the metrics that can be used for crisp (binary or hard) segmentation as evaluated by Taha and Hanbury [10] (refer for a comprehensive evaluation of 20 traditional metrics) and other commonly-employed metrics mentioned by other authors [2,6]. In brief, traditional evaluation metrics can be classified into the 6 following groups [10]: spatial-overlap, volume, pair-counting, information-theoretic, probabilistic, and spatial-distance, for both fuzzy and crisp segmentation. Most of the metrics can be derived using the four basic cardinalities of the confusion matrix, namely true positive (TP), true negative (TN), false positive (FP) and false negative (FN), which reflect the number of pixels in the MS that are classified correctly or incorrectly with respect to the GT as shown in Fig. 1. Spatial-overlap-based metrics measure the overlap of the four cardinalities of the confusion matrix. Volume-based metrics consider only the volume of the segmentation. Pair-counting-based metrics measure all possible object pairs in the four cardinalities of the confusion matrix. Information theoretical-based metrics focus on the entropy of the information. Probabilistic-based metrics describe the statistics of the voxels within the overlapping regions. Spatial-distance-based metrics measure the Euclidean distance between the boundaries of MS and GT.

Nine metrics that were previously proposed by various authors to overcome the shortcomings of classical metrics, are briefly described here, while the details of their implementation can be found in supplementary material 1. Firstly, C-Factor (CF) can differentiate over- and under-segmentation and represents a trade-off between sensitivity and specificity [9]. Bidirectional local distance (BLD) was proposed to overcome the deficiency of minimum distance and normal perpendicular distance methods used in the traditional spatial-distance-based metrics [11]. Medical Similarity Index (MSI) was extended from BLD to include medical relevance by penalizing for over- or under-segmentations based on medical intent [1]. Objective Quality Metric (OQM) measures the quantity, area, external contour, and content of the similarity between GT and MS [12]. OQM takes into account human vision system (HVS) and measures four possible segmentation errors, namely: added region, added background, inside holes, and border holes. The last 5 metrics comes from a family of symmetric-boundary-overlap hybrid metrics that combine boundary information with traditional spatial-overlap-based metrics namely Dice Similarity Coefficient (DSC), Jaccard Index (JAC), True Positive Rate (TPR), True Negative Rate (TNR), and precision (PPV) to yield Symmetric Boundary Dice (SBD), Symmetric Boundary Jaccard (SBJ), Symmetric Boundary True Positive Fraction (SBTP), Symmetric Boundary True Negative Fraction (SBTN), Symmetric Boundary Precision (SBP) [2]. These 9 metrics were adapted accordingly and evaluation of all metrics was performed using an in-house program written in python.

Some evaluation metrics are mathematically-equivalent or inversely related, namely F1-Measure Score (FMS), Symmetric Volume Difference (SVD), True Positive Rate (Sensitivity or Recall, TPR), True Negative Rate (Specificity, TNR), Volumetric Overlap Error (VOE), False Discovery Rate (FDR) and Volumetric Distance (VD), against DSC, FNR, FPR, JAC, PPV and VS accordingly. Only one metric was retained for each situation as they show similar performance to the metric that they are related to (e.g. DSC = FMS = 1-SVD). In addition, we included metrics such as histogram intersection (HI), determined using MR image intensity, and normalized false discovery rate (nFDR), which normalizes FP over the GT rather than the rest of the image [10]. Out of these 33 metrics, only two metrics could reflect over- and under-estimation with plus-minus signs, namely RVD and one previously proposed metric, CF [9]. We finally added a new group named the “hybrid” group which is composed of metrics encompassing two of the six groups, mainly spatial-overlap and spatial-distance, or measures an entirely different aspect.

2.2. Evaluation of metrics with simple synthetic shapes

In order to better highlight features and pitfalls of each metric, evaluation of the metrics was first performed using simple synthetic shapes describing 6 different scenarios with known GTs and MSs, as shown in Fig. 2. In all scenarios but (e), the GT consisted of a circle with a radius of 50 mm in an image matrix of 200 × 200 (1 mm × 1 mm). In (a) and (b), the MS consisted of a circle, centered about the GT, but with a radius underestimated (a) or overestimated (b) by 10 mm. In (c), MS had the correct area and shape, but the center of mass (COM) was shifted to the left by 20 mm. In (d), the MS had the same area but was star-shaped. For (e), both GT and MS had radii halved of that in (a) of 25 and 20 mm. Lastly, (f) described the same situation in (a) but in an image matrix of 400 × 400 (1 mm × 1 mm).

Visually, we would expect (a) and (f) to yield the same metric value as the known relative differences were the same, regardless of the matrix size. Therefore, if the metric values were different for scenario (a) and (f), the metric is not suitable for comparison across journals or studies if the image matrix size is different. Similarly, if the metric yielded the same metric value for (a) and (e), it showed that the metric is able to

Table 1
List of 33 evaluation metrics with corresponding equations and references. The four cardinalities from the confusion matrix: true positive (TP), true negative (TN), false positive (FP), false negative (FN). n = total number of voxels, d(GT, MS) = Euclidean distance between corresponding paired points in GT and MS, S(GT) = Surface voxels of GT, $H_{GT}(i)$ = Histogram (i bin out of N total number of bins) of MR voxels within GT mask, $E_m(GT)$ is the marginal entropy of GT, $E_j(GT, MS)$ is the joint entropy of GT and MS, x_i refers to a single voxel in the image, p_i refers to a point on a surface, $E(S_{GT}, S_{MS}, x_i)$ is the local refinement error.

Metric Group	Abbr.	Name	Equations	Brief Description	Reference
Spatial-Overlap	DSC	Dice Similarity Coefficient	$DSC = \frac{2\lvert GT \cap MS\rvert}{\lvert GT\rvert + \lvert MS\rvert} = \frac{2TP}{2TP + FP + FN} = 1 - SVD = FMS$	Amount of overlap over total number of pixels in GT and MS	[2,9,10,13]
	JAC	Jaccard Index	$JAC = \frac{\lvert GT \cap MS\rvert}{\lvert GT \cup MS\rvert} = \frac{TP}{TP + FP + FN} = 1 - VOE$	Amount of overlap over divided by their union	[2,10,13]
	FPR	False Positive Rate/ Fallout	$FPR = \frac{FP}{FP + TN} = 1 - TNR$	Falsely segmented pixels over total negative pixels	[2,10,13]
	FNR	False Negative Rate/ Miss Rate	$FNR = \frac{FN}{FN + TP} = 1 - TPR$	Falsely segmented pixels over total pixels in GT	[2,10,13]
	PPV	Positive Predictive Value/Precision	$PPV = \frac{\lvert GT \cap MS\rvert}{\lvert MS\rvert} = \frac{TP}{TP + FP} = 1 - FDR$	Amount of overlap over with respect to MS	[2,10,13]
	nFPR	Normalized FPR	$nFPR = \frac{FP}{TP + FN}$	Normalizing falsely segmented pixels over GT rather than rest of image	[14]
	HI	Histogram intersection	$HI = \sum_{i=1}^{N} \min(H_{GT}(i), H_{MS}(i))$	Similarity of gray-scale probability distributions of MS and GT	[15]
	CF	C-Factor[a, β]	$CF = \begin{cases} d, & p \geq q \ \wedge p > 1 - q \\ -d, & p < q \ \wedge p > 1 - q \\ \textit{undefined}, & p \leq 1 - q \end{cases}$ $d = \frac{2p(1-q)}{p + (1-q)} + \frac{2(1-p)q}{(1-p) + q}$, $p = \frac{TP}{TP + FN}$, $q = \frac{TN}{TN + FP}$	Discrepancy measure with trade-off between FP and TP Reflects over- and under-estimation	[9]
Error (Spatial Overlap)	GCE	Global Consistency Error	$GCD = \frac{1}{n}\min\{\sum_{i}^{n} E(S_{GT}, S_{MS}, x_i), \sum_{i}^{n} E(S_{MS}, S_{GT}, x_i)\} = \frac{1}{n}\min\left\{\frac{FN(FN + 2TP)}{TP + FN} + \frac{FP(FP + 2TN)}{TN + FP}, \frac{FP(FP + 2TP)}{TP + FP} + \frac{FN(FN + 2TN)}{TN + FN}\right\}$	MS error averaged over all voxels	[10]
Probabilistic	ICC	Interclass Correlation	$ICC = \frac{MS_b - MS_w}{MS_b + MS_w}$, $MS_b = \frac{2}{n-1}\sum_{x}(m(x) - \mu)^2$, $MS_w = \frac{1}{n}\sum_{x}(GT(x) - m(x))^2 + (MS(x) - m(x))^2$, $m(x) = (GT(x) + MS(x))/2$, $\mu = (Mean(GT) + Mean(MS))/2$	Measure of conformity or correlations between pairs of observations that may not have an order	[10]
	KAP	Cohen Kappa Coefficient	$KAP = \frac{f_a - f_c}{n - f_c}$, $f_a = TP + TN$, $f_c = \frac{(TN + FN)(TN + FP) + (FP + TP)(FN + TP)}{n}$	Measure of agreement between GT and MS	[10]
	AUC	Area under Receiver Operator Characteristics	$AUC = 1 - \frac{FPR + FNR}{2}$	Area under the simple trapezoid	[10]
	ACC	Accuracy	$ACC = \frac{TP + TN}{n}$	Amount of TP and TN over the total image volume	[13,16]
	INF	Bookmaker Informedness	$INF = TPR + TNR - 1$	Probability that a prediction (MS) is informed in relation to the condition (GT)	[13]
	MK	Markedness	$MK = PPV + NPV - 1$	Probability that a condition (GT) is marked by the predictor (MS)	[13]
Volume	RVD	Relative Volume Difference[β]	$RVD = \frac{GT - MS}{GT}$	Difference in volume between GT and MS with respect to GT	[2,6]
	VS	Volumetric Similarity	$VS = 1 - \frac{\lvert FN - FP\rvert}{2TP + FP + FN} = 1 - VD$	Measure of similarity with consideration to the volumes of GT and MS	[10]
Information-Theoretic	MI	Mutual Information	$MI = E_m(GT) + E_m(MS) - E_j(GT, MS)$	Measures amount of info one variable has about the other	[10]
	VOI	Variation of Information	$VOI = E_m(GT) + E_m(MS) - 2MI(GT, MS)$	measures amount of info lost (or gained) when changing from one variable	[10]
Pair-Counting	PRI[b]	Probabilistic Rand Index	$PRI = \frac{a + b}{a + b + c + d}$	Similarity measure between clusterings	[10]
	ARI[b]	Adjusted Rand Index	$ARI = \frac{2(ad - bc)}{c^2 + b^2 + 2ad + (a + d)(c + b)}$	PRI with correction for chance	[10]
Spatial-Distance	AHD	Average Hausdorff distance/Max. Symmetric Surface Distance	$AHD = \max(d(S_{gt}, S(MS)), d(S_{ms}, S(GT)))$	Hausdorff distance averaged over all points	[6]
	ASD				[2,6]

(*continued on next page*)

Table 1 (*continued*)

Metric Group	Abbr.	Name	Equations	Brief Description	Reference
		Average Symmetric Surface Distance	$ASD = \frac{1}{\lvert S(GT)\rvert + \lvert S(MS)\rvert}\left(\sum_{S_{gt}\in S(GT)} d(S_{gt},\ S(MS)) + \sum_{S_{ms}\in S(MS)} d(S_{ms},\ S(GT))\right)$	Average of all the distances from points on the boundary of MS to the boundary of the GT, and vice versa	
	MedSD	Median Symmetric Surface Distance	$MedSD = median(d(S_{gt},\ S(MS)), d(S_{ms},\ S(GT)))$	Median of all the distances from points on the boundary of MS to the boundary of the GT, and vice versa	[17]
	RMSD	Root Mean Square Symmetric Surface Distance	$RMSD = \sqrt{\frac{1}{\lvert S(GT)\rvert + \lvert S(MS)\rvert}}\cdot\sqrt{\left(\sum_{S_{gt}\in S(GT)} d^2(S_{gt},\ S(MS)) + \sum_{S_{ms}\in S(MS)} d^2(S_{ms},\ S(GT))\right)}$	Root Mean Square of all the distances from points on the boundary of MS to the boundary of the GT, and vice versa	[2,6]
	BLD	Bidirectional Local Distance[a]	$BLD = max(FMinD, BMaxD)$	Max. distance among the min. distances found at point on the GT surface	[11]
Hybrid	MSI	Medical Similarity Index[a]	$MSI\left(T,R\right) = \frac{\sum_{i=1}^{n} MCF(LDP(T_i,R)), il, ol)}{n}$	Uses medical consideration function to penalize for over- and under-segmentation	[1]
	OQM	Objective Quality Metric[a]	$OQM = \frac{\propto S_q + S_{o-r} + \lambda S_c}{1 + \propto + \lambda}$	Measures the quantity, area, external contour, and content of the similarity between GT and MS	[12]
	SBD	Symmetric Boundary Dice[a]	$SBD(G,\ M) = \frac{\sum_{x\varepsilon\partial G} DSC(N_x) + \sum_{y\varepsilon\partial M} DSC(N_y)}{\lvert\partial G\rvert + \lvert\partial M\rvert}$	Hybrid metric combining boundary info with DSC	[2]
	SBJ	Symmetric Boundary Jaccard[a]	$SBJ(G,\ M) = \frac{\sum_{x\varepsilon\partial G} JAC(N_x) + \sum_{y\varepsilon\partial M} JAC(N_y)}{\lvert\partial G\rvert + \lvert\partial M\rvert}$	Hybrid metric combining boundary info with JAC	[2]
	SBTP	Symmetric Boundary True Positive Fraction[a]	$SBTP(G,\ M) = \frac{\sum_{x\varepsilon\partial G} TPR(N_x) + \sum_{y\varepsilon\partial M} TPR(N_y)}{\lvert\partial G\rvert + \lvert\partial M\rvert}$	Hybrid metric combining boundary info with True Positive Fraction	[2]
	SBTN	Symmetric Boundary True Negative Fraction[a]	$SBTN(G,\ M) = \frac{\sum_{x\varepsilon\partial G} TNR(N_x) + \sum_{y\varepsilon\partial M} TNR(N_y)}{\lvert\partial G\rvert + \lvert\partial M\rvert}$	Hybrid metric combining boundary info with True Negative Fraction	[2]
	SBP	Symmetric Boundary Precision[a]	$SBP(G,\ M) = \frac{\sum_{x\varepsilon\partial G} PPV(N_x) + \sum_{y\varepsilon\partial M} PPV(N_y)}{\lvert\partial G\rvert + \lvert\partial M\rvert}$	Hybrid metric combining boundary info with Precision	[2]

[a] Refer to supplementary material 1 for details. [β] Note that CF and RVD are metrics that can reflect over- and under-estimation with plus-minus signs.

[b] $a = \frac{1}{2}[TP(TP-1) + FP(FP-1) + TN(TN-1) + FN(FN-1)]$, $b = \frac{1}{2}[(TP+FN)^2 + (TN+FP)^2 - (TP^2 + TN^2 + FP^2 + FN^2)]$, $c = \frac{1}{2}[(TP+FP)^2 + (TN+FN)^2 - (TP^2 + TN^2 + FP^2 + FN^2)]$, $d = \frac{n(n-1)}{2} - (a+b+c)$.

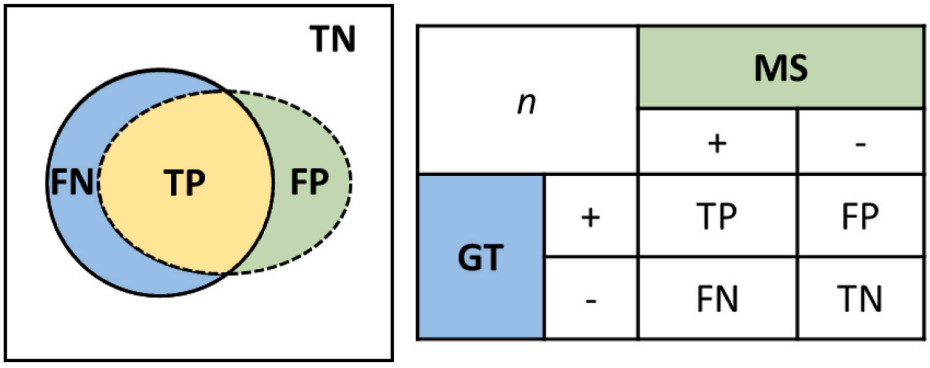

Fig. 1. Four basic cardinalities of the confusion matrix. The pixels in the overlap region of GT (Full Line) and MS (Dashed Line) are classified correctly as positive (TP, Yellow) and the pixels in the background are classified correctly as negative (TN, White). Pixels in MS that are incorrectly classified as positive (FP, Green) or not classified (FN, Blue) with respect to the GT in the entire image (box) with a total of n pixels.

reflect the relative differences proportionally as the GT and MS in (e) had radii halved of that in (a). (b) and (a) should preferably show the same or close values as the known differences were the same even though MS was bigger in (b) and smaller in (a). In (c) and (d), the MS had the same area as the GT, but one was shifted by 20 mm to the left, while (d) was centered about the GT but with star-shape. Thus, we would expect (d) to have higher value than (c). Metrics that changed due to changes in target size (e) or image size (f) were identified as they affect comparison even though the similarity or differences between MS and GT were the same as that in (a).

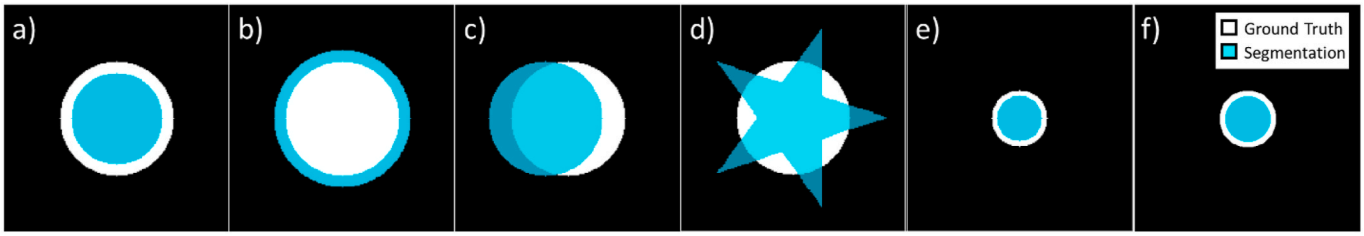

Fig. 2. Evaluation of metrics using simple synthetic shapes with GT in white and MS in light blue.

2.3. Evaluation of metrics and network ranking with clinical image data

2.3.1. Clinical image data

T2-weighted magnetic resonance (MR) images were extracted from the PROSTATEx Challenge dataset (https://prostatex.grand-challenge.org/), with 120, 20 and 20 subject data for training, validation and evaluation. The images were acquired on the Siemens 3T MR scanners (either MAGNETOM Trio or Skyra) without endorectal coil and covered the entire prostate region with a voxel size of $0.5 \times 0.5 \times 3.6$ mm^3. The images were corrected for field inhomogeneity and their intensity linearly scaled to within 0–1000. The images were then cropped and resliced into a matrix size of 192×192 x 46 with a voxel size of $0.5 \times 0.5 \times 2$ mm^3, covering the entire prostate [18].

2.3.2. Deep learning-based machine segmentation (MS)

MSs were generated for all 20 evaluation subjects using three deep learning networks - DenseVNet (DVN) [19], HighRes3DNet (HRN) [20], and ScaleNet (SN) [21]. DVN is a monomodal network where feature maps are first computed using a stride convolution followed by a cascade of dense feature stacks and stride convolutions to generate activation maps, then a convolution unit to reduce the number of features before bilinear upsampling back to image size [19]. HighRes3DNet relies on dilated convolutions and residual connections to create an end-to-end mapping from image volume to voxel-level dense segmentation [20]. ScaleNet is a multimodal network with HighRes3DNet as its backend, while the frontend merges the data from the backend to the frontend independently of the number of input modalities [21]. The GT masks for training and validation were drawn by a research fellow with 3 years of experience in segmenting prostates. Details of the network training and optimization can be found in our previous work [18].

2.3.3. Inter reader ground truth (GT)

Prostate masks covering the WG were first manually drawn by two students trained in segmenting the prostates on T2w images using the Medical Imaging Interaction Toolkit (MITK) software (https://www.mitk.org) [18]. The masks were subsequently corrected independently by a research fellow with 3 years of experience in segmenting prostates (reader 1), and by an experienced medical physicist with over 10 years' experience of delineating regions for radiotherapy (reader 2), leading to two sets of GTs.

2.3.4. Visual assessment of machine segmentation (MS)

The quality of the MSs generated by the networks was assessed visually by the same research fellow and medical physicist using an eleven-grade discrete scale (-5, $+5$) to express 5 levels of severity of under- (negative) and of over-segmentation (positive), with 0 representing perfect segmentation (Fig. 3). This dual grading was used to investigate the correlation of the visual assessment with metrics that quantify over- and under-estimation, while its absolute value was used for correlation with other metrics. Both manual segmentation and visual assessment were implemented slice-by-slice in the transverse direction. Pearson correlation (R), and Cohen's Kappa (κ) were used to evaluate the agreement between the readers' grading. Quadratic weighting was applied to κ to chance-correct the distance distribution of the readers' classification, due to the use of dual negative-positive rating with a center grade of 0 [22].

All metrics were generated using the 2 sets of GTs in 2D to match the slice-by-slice visual assessment grading done by the same reader who

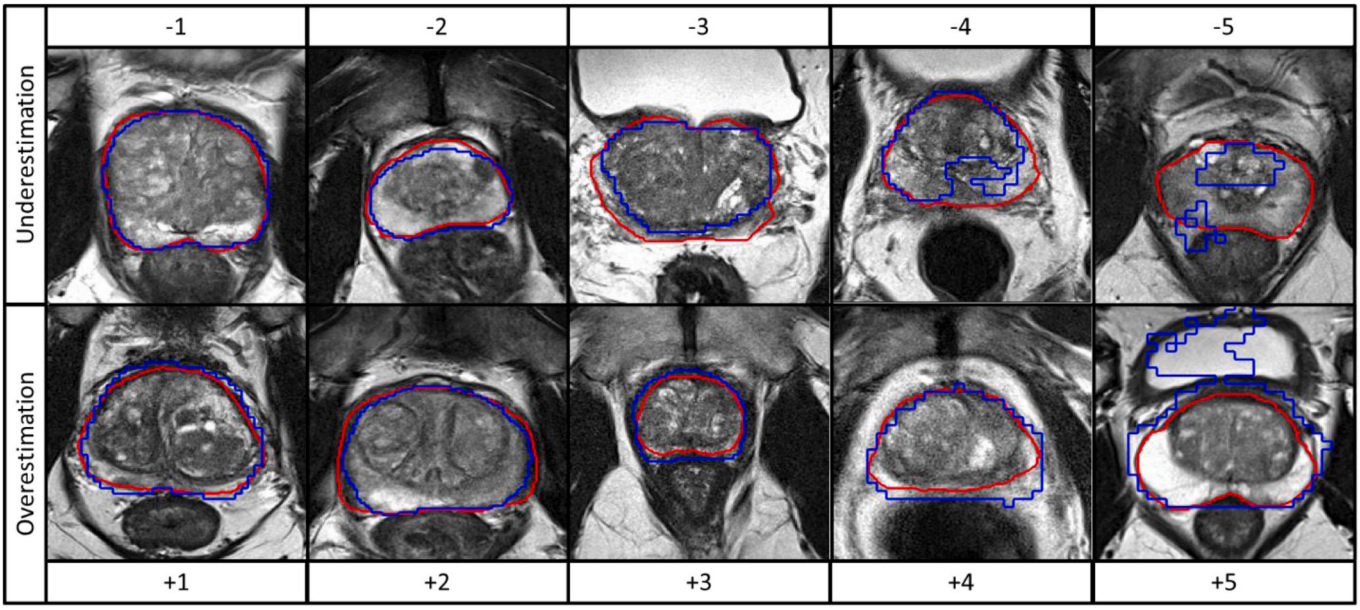

Fig. 3. Dual negative-positive grading applied for visual assessment, with the MS (blue) and GT (red). The images depicted were given the same grade by both readers. The GTs shown here were drawn by the less experienced reader 1.

drew the GTs. Only slices where both readers have delineated prostate gland were selected for visual assessment and metrics evaluation. This resulted in the assessment of a total of 1197 slices (399 slices x 3 networks) for each set of GTs. Spearman rank-order correlation coefficient (ρ) was measured to determine the strength and direction of the monotonic association between the visual grading and the evaluation metrics. All metrics were correlated with the absolute visual assessment grading of 0–5, except for two metrics (RVD and CF), which could reflect over- and under-estimation, were correlated with the dual negative-positive visual assessment grading of −5 to 5. One metric from each category with the best correlation with visual assessment was then identified for network ranking evaluation.

2.3.5. Network ranking

The ranking of the 3 networks (DVN, HRN, and SN) was carried out using 3 methods: (1) averaging a single selected metric and then rank the networks based on the averaged metric value [4,5], (2) normalizing the selected metrics using 2 GTs and then averaging the values for ranking, and (3) ranking each case using the selected metrics and then averaging the ranking [8]. The ranking methods 1 and 3 were carried out with each set of GTs to determine the amount of variation in ranking caused by the readers. For ranking method 2, the metric was first generated with the MSs (MET_{MS}), as well as, the GTs of reader 1 (MET_{GT1}), using the GTs drawn by the more experienced reader (reader 2) as the reference. The relative values for the selected metrics were then calculated by dividing the metrics of MSs over that of GTs of reader 1 (MET_{MS}/MET_{GT1}). For metrics with values that lie outside the range of 0 and 1 (e.g. spatial-distance-based metrics), the rounded minimum and maximum values of all segmentations were used to normalize the relative values. As the choice of the metrics, metric aggregation methods, annotators establishing the GTs and, test subjects used affect the ranking or final evaluation of networks, leave-one-out evaluation was further carried out to determine the changes in values and ranking of the networks. Ideally, good evaluation metrics will show small variation in values within each network, but greater differences across networks, which will ensure consistency in network ranking [3]. As such, leave-one-out evaluation was also performed by removing all slices of one subject each time for all 20 subjects.

3. Results

3.1. Evaluation with fixed shapes

Table 2 shows the 33 metrics generated using simple synthetic shapes which describes 6 different scenarios with known GTs and MSs. All of the selected spatial-overlap-based, volume-based, spatial-distance-based, and hybrid metrics were not affected by matrix size as in (a) vs. (f). For probabilistic-based metrics, only ICC, KAP, ACC and MARK were affected. The error of spatial-overlap-based metrics, namely GCE, information theoretic-based, and pair-counting-based metrics were all affected by matrix size in (a) vs. (f). PRI showed no difference from (a) to (d) and hence shows poor suitability for segmentation evaluation. Volume-based metrics, VS and RVD, showed no difference in (c) and (d), between circle and star shapes, as their areas are the same. Out of 33 metrics, only RVD and CF reflected over- and under-estimation of the MSs with plus-minus-signs correctly for (a), (b), (e) and (f). MSI showed higher values for shifted circle (c), followed by star-shaped (d) then slightly smaller or larger volumes of (a) and (b), with the same value for both (a) and (b). OQM showed higher values for (b), followed by (a), (c) then (d) accordingly, which is more similar to the human

Table 2
Evaluation results of 33 metrics generated for 6 different scenarios with GT in white and MS in light blue: (a) MS is smaller than GT by 10 mm radius, (b) MS is larger by 10 mm radius, (c) MS is the same size but shifted to left by 20 mm, (d) MS is star-shaped with the same area as GT, (e) MS and GT are half the size of that in (a), and (f) MS and GT are the same as that of (a) but in an image matrix twice that of (a).

Metric Group	Parameter	(a)	(b)	(c)	(d)	(e)	(f)
Spatial-Overlap	DSC	0.781	0.82	0.748	0.759	0.781	0.781
	JAC	0.641	0.695	0.597	0.612	0.641	0.641
	FPR	0.000	0.107	0.062	0.059	0.000	0.000
	FNR	0.359	0.000	0.252	0.241	0.359	0.359
	PPV	1.000	0.695	0.748	0.759	1.000	1.000
	nFPR	0.000	0.439	0.252	0.241	0.000	0.000
	HI	0.641	1.000	0.748	0.759	0.641	0.641
	CF	−0.529	0.193	−0.512	−0.492	−0.528	−0.529
Error[a]	GCE	0.116	0.146	0.182	0.175	0.029	0.029
Probabilistic	ICC	0.739	0.763	0.686	0.701	0.772	0.772
	KAP	0.741	0.766	0.686	0.701	0.773	0.981
	AUC	0.820	0.946	0.843	0.850	0.820	0.820
	ACC	0.930	0.914	0.901	0.906	0.982	0.982
	INF	0.641	0.893	0.686	0.701	0.641	0.641
	MK	0.919	0.695	0.686	0.701	0.982	0.982
Volume	RVD	−0.359	0.439	0.000	0.000	−0.359	−0.359
	VS	0.781	0.820	1.000	1.000	0.781	0.781
Information-Theoretic	MI	0.361	0.464	0.286	0.299	0.155	0.155
	VOI	0.185	0.395	0.428	0.415	0.046	0.046
Pair-Counting	PRI	0.685	0.685	0.685	0.685	0.907	0.068
	ARI	0.669	0.662	0.587	0.604	0.758	0.984
Spatial-Distance[b]	AHD	10	10	20	33	5	10
	ASD	10	10	12	11	5	10
	MedSD	9	9	14	9	4	9
	RMSD	216	238	325	385	74	216
	BLD	10	10	20	22	5	10
Mixed/Hybrid	MSI	0.5	0.5	1	0.603	0.527	0.5
	OQM	0.707	0.735	0.685	0.684	0.703	0.707
	SBD	0.289	0.296	0.351	0.321	0.354	0.289
	SBJ	0.216	0.222	0.272	0.240	0.253	0.216
	SBTP	0.216	0.455	0.398	0.339	0.253	0.216
	SBTN	0.555	0.278	0.496	0.508	0.910	0.555
	SBP	0.445	0.222	0.416	0.406	0.867	0.445

[a] Error of Spatial-overlap group.
[b] Units in mm (rounded to the whole number).

perspective. On the other hand, CF showed higher absolute value for (a), followed by (c) and (d), with (b) yielding unexpectedly low value. The metrics of symmetric-boundary family yielded comparatively smaller values, with higher values obtained for (c), (d), (b) then (a) for SBD and SBJ.

3.2. Agreement of visual assessment between readers

Relatively good agreements between readers were obtained for both absolute (0–5) and all (−5 to 5) grading with $R \geq 0.68$ and $\kappa \geq 0.660$, with higher agreements observed for SN, followed by DVN then HRN (Table 3). The more experienced reader, Reader 2, yielded slightly higher mean absolute grading, but smaller deviations across all networks. The differences in mean absolute grading between the two readers were the smallest in SN, followed by DVN and HRN, with HRN yielding a much greater difference. The overall plus-minus sign agreement in grading between the readers is about 85.5% with the highest agreement obtained for SN of 95% (Table 3).

3.3. Evaluation of GTs between readers

The average 20 WG volumes generated by readers 1 and 2 were 64.1 ± 45.3 cm^3 [21.2, 199.2] and 59.6 ± 41.1 cm^3 [20.9, 176.4] (Averaged over the 2 sets of GTs: 61.8 ± 43.1, [21.1, 187.8] cm^3). The 33 metrics were also generated for the GTs drawn by reader 1, using the GTs drawn by reader 2 as reference: DSC (0.907 ± 0.082), JAC (0.839 ± 0.120), FPR (0.020 ± 0.019), FNR (0.039 ± 0.065), PPV (0.872 ± 0.123), nFPR (0.177 ± 0.257), HI (0.978 ± 0.060), absolute CF (aCF, 0.107 ± 0.090), GCE (0.039 ± 0.025), ICC (0.894 ± 0.086), KAP (0.894 ± 0.086), AUC (0.971 ± 0.031), ACC (0.978 ± 0.015), INF (0.941 ± 0.061), MK (0.866 ± 0.120), absolute RVD (aRVD, 0.173 ± 0.257), VS (0.928 ± 0.084), MI (0.456 ± 0.178), VOI (0.136 ± 0.071), PRI (0.764 ± 0.108), ARI (0.868 ± 0.093), AHD (4.791 ± 3.090), ASD (1.505 ± 1.094), MedSD (1.197 ± 0.993), RMSD (42 ± 29), BLD (4.518 ± 2.862), MSI (0.835 ± 0.107), OQM (0.836 ± 0.079), SBD (0.680 ± 0.139), SBJ (0.578 ± 0.139), SBTP (0.807 ± 0.125), SBTN (0.699 ± 0.155) and SBP (0.666 ± 0.176).

3.4. Spearman rank-order correlation matrix

Fig. 4 shows the Spearman rank-order correlation matrix of the 33 evaluation metrics generated for all MSs for all networks and for both readers, whereby the GTs were drawn by the same reader for visual assessment, and visual assessment (VA) grading. Refer to Supplementary Figs. 1–8 for Spearman rank-order correlation matrix for the individual networks and all networks for each respective reader. Stronger correlation (larger $|\rho|$ values) was observed between the VA grading and metrics generated using the GTs drawn by the more experienced reader 2. Between the two metrics that reflect over- and under-estimation, only RVD yielded $|\rho| \geq 0.7$ for reader 1 and $|\rho| \geq 0.8$ for reader 2, consistently across all networks, while CF, yielded $|\rho| \geq 0.6$ for DVN and HRN, but not SN for both readers. For plus-minus sign agreement between the metrics with VA agreement, CF yielded 66.8% and 74.3%, while RVD yielded 84.7% and 91.1% for readers 1 and 2 respectively. For all other metrics, ASD, AHD, KAP, ICC, ARI, DSC, JAC, OQM, SBJ, and SBD consistently yielded $|\rho| \geq 0.7$ with VA, except for HRN network when evaluated by the less experienced reader 1.

Between the two metrics that can reflect over- and under-estimation, RVD yielded the highest $|\rho|$ values consistently with VA and with better plus-minus-sign agreement than that between the readers, while CF yielded only moderate $|\rho|$ with VA and poor plus-minus sign agreement. Probabilistic-based metrics, namely ICC and KAP have consistently ranked the top two metrics across the three networks and two readers, with ICC consistently ranked first. For spatial-overlap and spatial-distance-based metrics, DSC and ASD were consistently ranked higher than their related metrics. ARI was ranked the best pair-counting-based metrics and showed good correlation with VA, with $|\rho| \geq 0.7$. Information-theoretic-based metrics showed no consistent trend, with both VOI and MI doing much poorer than other metrics. For volume-based metrics, RVD yielded better correlation than VS with VA, even though RVD was correlated with the full VA grading of −5 to 5, while VS was correlated with VA grading of 0–5. Among the previously proposed hybrid or mixed metrics, OQM, SBJ, and SBD consistently yielded $|\rho| \geq 0.7$. From the same symmetric-boundary family, SBD consistently yielded slightly higher $|\rho|$ than SBJ, hence only SBD was selected for further evaluation. As such, 7 evaluation metrics, namely DSC, ICC, RVD, ARI, ASD, OQM, and SBD were selected for ranking evaluation, with VA grading.

3.5. Performance ranking of the three networks

For ranking method 1, the ranking of the three networks was based on each metric individually (Top of Table 4). The network rankings were consistent across all metrics and similar to that of VA, with HRN ranked first followed by DVN and SN for both readers. The differences in metric values between readers were largest with ASD for infinite-ranged metrics and SBD and ARI for fixed-ranged (0–1) metrics. ASD showed the greatest difference across networks due to the large range of values, followed by RVD. Ranking method 2 requires normalization of the selected metrics using 2 GTs, before averaging across all cases. Thus, VA could not be evaluated in this case. Most metrics showed small differences across networks except for RVD due to the great differences in volume at the apex and base of the WG (Middle of Table 4). However, the value differences were greater than that from ranking method 1. Ranking was also consistent for all selected metrics with HRN ranked first then DVN and SN. Ranking method 3 requires the ranking for each case and then averaging the ranks of all cases for each metric (Bottom of Table 4). The ranking was consistent across all metrics and was similar to VA, with small differences in ranking values between the two readers across the three networks. All three ranking methods yielded the same rankings with all 7 metrics with similar ranking to that of VA.

3.6. Changes in network ranking with leave-one-out evaluation

Fig. 5 shows the changes in average values, relative values, and ranks of the 7 selected metrics using ranking methods 1 to 3 for both readers during the leave-one-out evaluation. Note that for VA, RVD, and ASD

Table 3
Mean absolute grading of each reader, plus-minus sign agreement between readers and the agreement of visual assessment grading of the 2 readers for all segmentations and all three individual networks (DVN: DenseVNet, HRN: HighRes3DNet, SN: ScaleNet) using Cohen's Kappa (κ), with quadratic weighting, and Pearson's Correlation (R) for absolute grading (0–5) or all grading (−5 to 5).

Networks	All		DVN		HRN		SN	
Readers	Reader 1	Reader 2	Reader 1	Reader 2	Reader 1	Reader 2	Reader 1	Reader 2
Mean Absolute Grading (stdev)	2.742 (1.080)	2.848 (1.000)	2.734 (0.882)	2.739 (0.889)	2.226 (1.084)	2.561 (0.908)	3.226 (1.005)	3.243 (1.000)
Plus-Minus Sign Agreement (%)	85.5		83.5		77.9		95.0	
Agreement	Absolute	All	Absolute	All	Absolute	All	Absolute	All
κ	0.777	0.753	0.747	0.663	0.660	0.647	0.863	0.840
R	0.783	0.764	0.747	0.681	0.708	0.680	0.863	0.842

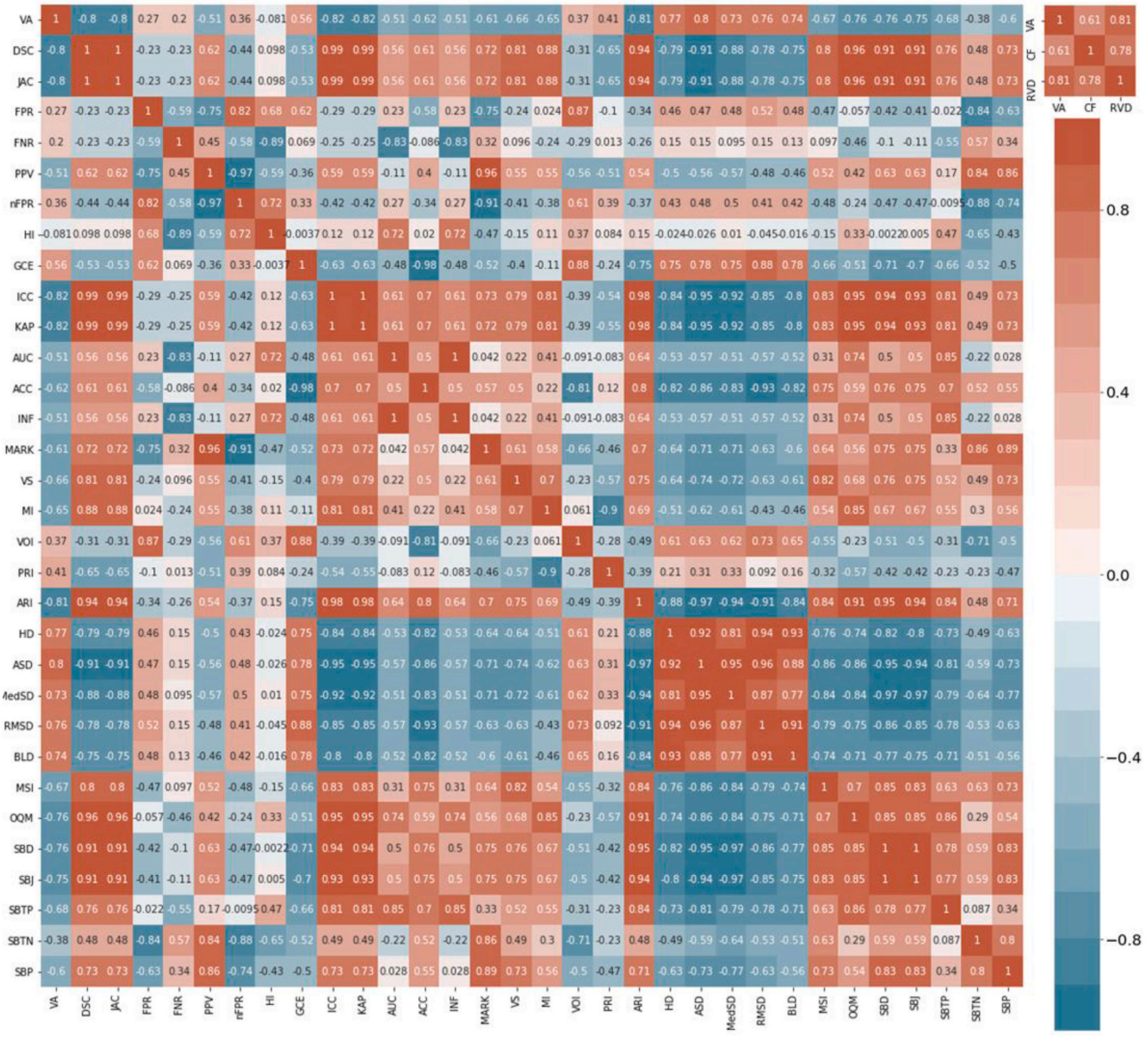

Fig. 4. Spearman rank-order correlation matrix of the 33 evaluation metrics, generated for all MSs for all networks and for both readers, and visual assessment (VA) grading. The 2 metrics that could reflect over- and under-estimation were correlated with the VA grading of −5 to 5 as shown on the top right-hand corner, while the rest were correlated with VA grading of 0–5.

using ranking methods 1, the smaller the values, the better the network. Ranking method 1 generally yielded small variation within each network, but the variation across the network may overlap due to the small differences in values. ARI, SBD, RVD, and ASD showed clearer separation of the 3 networks, though the variation within each network is larger for RVD and ASD (Fig. 5a–c). Greater separation between the networks was obtained for DSC, ICC, ARI, and SBD with the more experienced reader 2 (Fig. 5b) than reader 1 (Fig. 5a). Greater discrimination between HRN and DVN in visual assessment grading was observed for reader 1 than reader 2 (Fig. 5c). For ranking method 2, all metrics yielded clear separation except for OQM and RVD with the greatest discrimination across networks obtained with SBD while achieving small variation within each network (Fig. 5d). For ranking method 3, all metrics showed clear separation across the network, and for both readers, though the differences in average rank values were smaller for RVD and OQM (Fig. 5e–f). The discrete scale of VA led to many overlaps, thus yielding lower ranking for VA than other metrics.

4. Discussion

In this study, we compared 9 proposed segmentation evaluation metrics with 24 traditional metrics to identify features or metrics that better capture what is required for the clinical evaluation of segmentations. Evaluation was carried out first using 2D synthetic shapes to identify the features measured by the previously proposed metrics with known GTs and MSs. Clinical evaluation was then performed using publicly-available prostate images of 20 subjects with MSs generated by 3 different networks (DVN, HRN, and SN). Comparison of MSs was carried out using GTs drawn by 2 readers who also performed the 2D visual assessment. Metrics that correlated well with visual assessment grading were identified and selected for further evaluation using 3 different network ranking methods.

4.1. Visual assessment

We attempted for the first time to do a dual negative-positive visual assessment grading to evaluate metrics that could indicate over- and

Table 4
The average values (top), average relative values (middle) and average ranking (bottom) of the 7 selected metrics with visual assessment grading (VA) across the three networks (DVN, HRN, and SN) for the readers, and the respective ranking of the networks using the three ranking methods. The absolute values of RVD were used.

Metrics	Network	VA	DSC	ICC	RVD[a]	ARI	ASD[b]	OQM	SBD
Ranking Method 1									
Reader 1	DVN	2.734	0.836	0.815	0.200	0.778	2.392	0.742	0.530
	HRN	2.226	0.856	0.837	0.176	0.803	2.012	0.759	0.552
	SN	3.266	0.800	0.767	0.348	0.716	3.550	0.732	0.463
Reader 2	DVN	2.739	0.820	0.798	0.267	0.760	2.599	0.739	0.504
	HRN	2.561	0.845	0.826	0.219	0.791	2.134	0.759	0.531
	SN	3.243	0.771	0.735	0.476	0.681	4.046	0.718	0.425
Ranking	1st	HRN	HRN	HRN	HRN	HRN	HRN	HRN	HRN
	2nd	DVN	DVN	DVN	DVN	DVN	DVN	DVN	DVN
	3rd	SN	SN	SN	SN	SN	SN	SN	SN
Ranking Method 2									
Relative Values	DVN	-	0.919	0.865	5.003	0.850	0.900	0.900	0.771
	HRN	-	0.948	0.897	4.565	0.996	0.919	0.925	1.175
	SN	-	0.856	0.788	11.888	0.791	0.832	0.869	0.642
Ranking	1st	–	HRN	HRN	HRN	HRN	HRN	HRN	HRN
	2nd	–	DVN	DVN	DVN	DVN	DVN	DVN	DVN
	3rd	–	SN	SN	SN	SN	SN	SN	SN
Ranking Method 3									
Reader 1	DVN	1.6	1.9	1.9	1.9	1.9	1.9	2.1	1.8
	HRN	1.2	1.5	1.5	1.7	1.5	1.5	1.6	1.6
	SN	2.1	2.5	2.5	2.4	2.6	2.5	2.3	2.5
Reader 2	DVN	1.4	1.8	1.8	1.8	1.8	1.9	1.9	1.8
	HRN	1.2	1.5	1.5	1.7	1.5	1.5	1.5	1.5
	SN	1.9	2.7	2.7	2.6	2.7	2.7	2.5	2.6
Ranking	1st	HRN	HRN	HRN	HRN	HRN	HRN	HRN	HRN
	2nd	DVN	DVN	DVN	DVN	DVN	DVN	DVN	DVN
	3rd	SN	SN	SN	SN	SN	SN	SN	SN

[a] Smaller values reflect a better network for ranking methods 1 and 2.
[b] Normalized by the maximum value of 21 mm for ranking method 2.

under-estimation. The positivity-agreement between the readers was 85.5%, while CF yielded 66.8% and 74.3%, and RVD yielded 84.7% and 91.1% for readers 1 and 2 respectively. The results indicated that RVD is a better indicator of over- or under-estimation than CF and that the readers generally looked at the overall MS volume in evaluating over- and under-estimation of the segmentations. This is further supported by the stronger correlation obtained with RVD, compared to CF (balance between sensitivity and specificity) and FPR and FNR (inversely correlated with specificity and sensitivity), with VA using Spearman rank-order correlation matrix (Fig. 4). Despite the use of dual negative-positive grading and a larger scale of -5 to 5, we obtained higher agreements of $R = 0.783$ and 0.764 for absolute and all grading than Taha and Hanbury [5], with $R = 0.62$, with 2 readers. Greater difference was observed in the grading between the two readers for HRN, which was ranked 1st by all selected metrics and ranking methods. HRN generated MSs that segmented closer to the prostate border compared to DVN and SN. The greater discrepancy indicated the difficulty in assessing the target boundary and the limitation of visual assessment in discriminating MSs at the target boundary. Third-order polynomial regressions of RVD and DSC with VA are shown in Supplementary Fig. 9.

4.2. Fixed shape evaluation and visual assessment correlation

All information-theoretic (MI and VOI), pair-counting (PRI and ARI), and error of spatial-overlap-based (GCE), and some probabilistic-based metrics, (ICC, KAP, AUC, and MARK) metrics were affected by matrix size (Table 2). Therefore, these metrics are not suitable for comparison across images of different matrix sizes. Moreover, information-theoretic-based metrics yielded poor correlation with VA (Fig. 4) and hence are not suitable for clinical segmentation evaluation. PRI also yielded a poor correlation with VA (Fig. 4), with no change in values in the evaluation of the synthetic shape (Table 2), thus it is not a good metric. Conversely, ARI showed a strong correlation with VA across the three networks and for both readers (Fig. 4). ICC and KAP were consistently ranked as the top 2 metrics with the strongest correlation with VA (Fig. 4) and good synthetic shape outcomes (Table 2). Volume-based metrics (VS and RVD) are not affected by matrix size, but could not differentiate other features apart from volume (Table 2). RVD thus may only be suitable in providing over- and under-estimation as supported by better plus-minus-sign agreement with visual assessment, compared to CF, FPR, and FNR, and had the best correlation with VA (Fig. 4). All spatial-distance-based metrics were not affected by matrix size (Table 2). HD and BLD only yielded the expected values for (a) to (c), and ASD, RMSD, and BLD yield values with ranking more similar to visual expectations for the fixed shapes evaluation with (a) and (b) performing better than (c) and (d). Particularly, ASD yielded a high correlation with VA with $|\rho| \geq 0.7$ (Fig. 4). For spatial-overlap-based metrics, only DSC and JAC yielded values similar to expectations for the fixed shapes evaluation (Table 2) and yielded a high correlation ($\rho \geq 0.7$) with VA (Fig. 4). HI and nFDR did not perform better than the other commonly-applied metrics for segmentation evaluation and yielded only average correlation with VA ($|\rho| \geq 0.6$).

All hybrid metrics were not affected by matrix size (Table 2), with OQM only yielding values more similar to visual expectations for the fixed shapes evaluation, with (a) $\approx$ (b), (c) $\approx$ (d) and (a) and (b) $>$ (c) and (d). This is expected as OQM was developed with HVS incorporated. OQM also yielded a high correlation with VA of $\rho \geq 0.8$ for all networks by the more experienced reader 2 and $\rho \geq 0.65$ for reader 1. However, OQM did not yield a higher correlation than DSC or JAC with VA. This may be due to the removal of texture information from the formulation or requirement of more fine-tuning to the 3 weighting parameters applied in the formulation (refer to supplementary material). SBD and SBJ showed unexpected fixed shapes results with (c) and (d) $>$ (a) and (b), but achieved (a) $\approx$ (b), (c) $\approx$ (d). They yielded smaller values and slightly poorer correlation with VA than the original DSC and JAC, with $\rho \geq 0.7$ for all networks and readers except for HRN when graded by the less experienced reader 1.

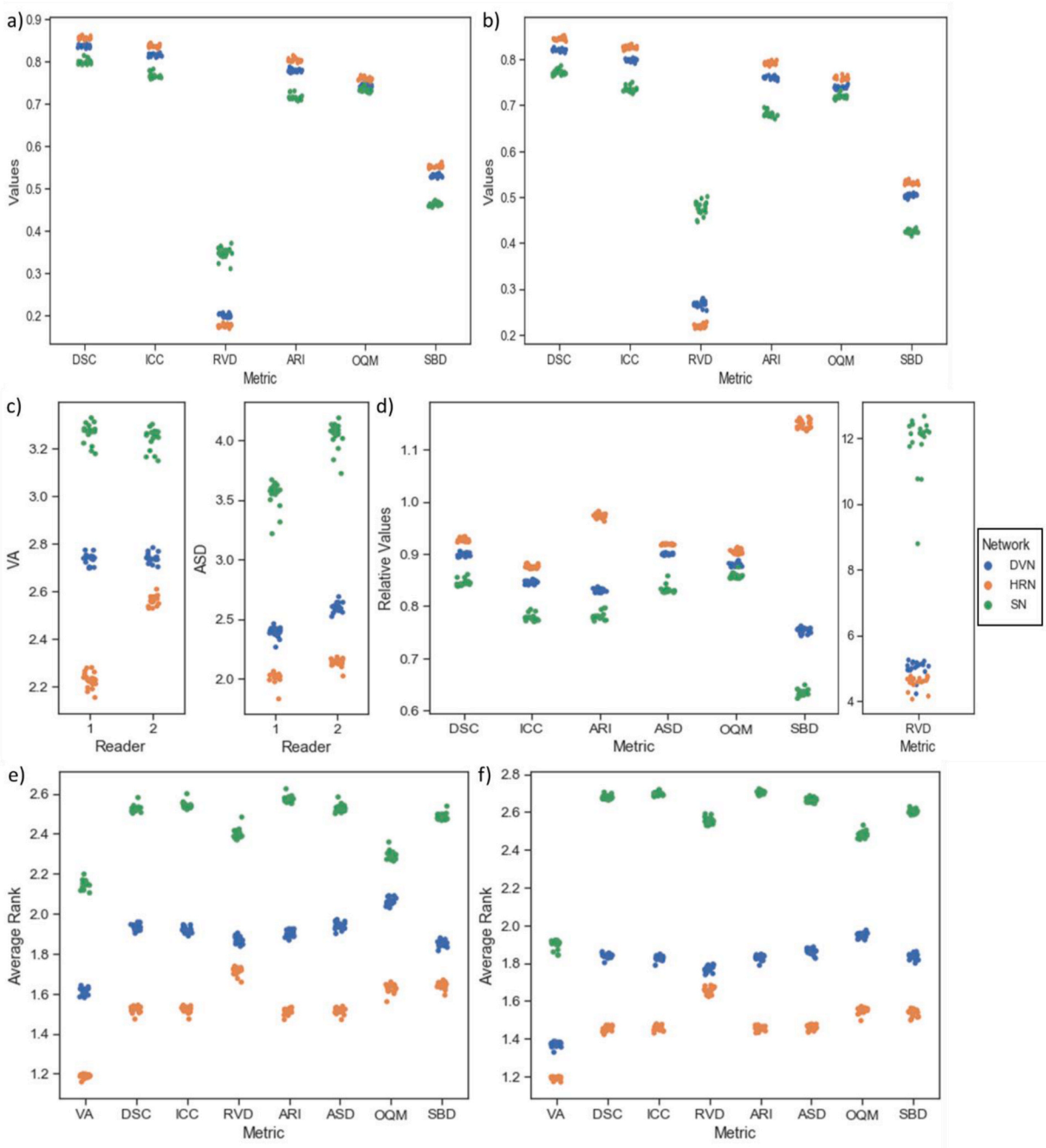

Fig. 5. Variations in values, relative values, and rank of selected metrics for 3 different ranking methods, using leave-one-out evaluation across the three networks of DVN (blue), HRN (orange), and SN (green). The scatterplots show the average values of 7 selected metrics for (a) reader 1, and (b) reader 2, and (c) VA and ASD only for both readers using ranking method 1, (d) variation in relative values of 7 selected metrics using ranking method 2, and the variation in average rank values of 7 selected metrics with VA using ranking method 3 for (e) reader 1 and (f) reader 2.

4.3. Network performance ranking

All three ranking methods yielded consistent ranking, with HRN being first followed by DVN and SN. RVD, SBD and ARI only yielded larger differences (>0.1) between readers and across networks using ranking method 1, with ASD and RVD yielding the largest differences. For ranking method 2, selecting the appropriate minimum value for infinite-ranged metrics, such as ASD, may be difficult, but the differences in values were greater than that of ranking method 1. Ranking method 3 yielded relatively similar differences across networks and between readers. Ranking method 1 yielded the smallest separation of the networks, with only ARI, SBD, RVD, and ASD showing clearer separation. For ranking method 2, only SBD showed large differences between networks, while RVD and OQM showed overlap in the relative values, thus affecting the ranking outcome. Although ranking method 3 showed a clear separation of the 3 networks, the variation within the network was larger and discrimination of network ranking was poorer for RVD and OQM. Maier-Hein's results [3] showed that single metric rankings (ranking method 1) are significantly more robust compared to when ranking is performed after the aggregation (ranking method 3). Our results indicated that ranking method 1 resulted in overlap in values and the ranking of the networks, and was greatly affected by the reader generating the GTs, while ranking method 3 produced greater separation, but also increased variation within the network. Ranking method 2

yielded greater separation than ranking method 1 but was not suitable for some metrics, such as RVD and OQM. Only SBD yielded consistently good discrimination of the networks for all three ranking methods.

4.4. Identification of metric features

Taha and Hanbury [10] recommended using metrics based on medical intent, such as using volumetric-based metrics when measuring tumor volume changes. The delineation requirements differ based on medical relevance, target location, disease types, and/or treatment methods and thus making it difficult to compare methods across journals [1]. Traditional metrics may be appropriate for healthy structure segmentation or single large anatomical segmentation, which differs mostly at the boundaries. Yet, they may not evaluate the important features of the output segmentation or account for medical consequences, but merely measure the geometric differences [1,14]. The delineation of the boundary is critical and thus, distance-based metrics are often used to measure the distance between the MSs and GTs. Kim et al. [11] proposed a spatial-distance-based metric, BLD, which is extended to include penalties for over- and under-estimation, leading to MSI [1]. However, both BLD and MSI did not show a high correlation with visual assessment compared to traditional metrics. Popovic et al. [9] proposed CF, a metric that could differentiate over- and under-estimation to account for different clinical applications. However, CF showed poor correlation with visual assessment and could not reflect plus-minus-sign agreement better than the readers, when compared to RVD. Shi et al. [12] proposed OQM, by combining spatial-overlap-based evaluation with human vision system interpretation, and penalizing inside holes, added regions, etc. OQM did yield perfect interpretation for fixed shapes, but performed slightly poorer than traditional metrics with clinical imaging data (prostate segmentation), and showed overlap in ranking across the networks during leave-one-out evaluation. Yeghiazaryan et al. [2] combined the 2 important aspects, spatial-distance, and spatial-overlap, and propose symmetric-boundary metrics. SBD and SBJ showed good correlation with visual assessment, though slightly poorer than traditional metrics, and good rank discrimination. However, neither SBP, SBTP nor SBTN outperformed the related traditional metrics.

Of the 7 selected metrics, DSC, ICC, ARI, RVD, and ASD are traditional metrics, and all yielded good correlation with visual assessment, with ICC achieving the highest correlation with VA (Fig. 4). These metrics are also commonly employed in challenges [3], but may yield poor discrimination of network ranking, and are affected by the test data sampled. Moreover, spatial-overlap-based metrics (e.g. DSC), show very little difference in values when the GT's volume is large and conversely when GT's volume is small, making it difficult to compare networks efficiently. Distance-based metrics (e.g. AHD), ranges from 0 to infinity, is affected by border definition, and image spatial resolution and quality, thus making the comparison of different images difficult [14]. Although SBD yielded slightly poorer correlation with VA than traditional metrics, it improved discrimination of network rank, while yielding relatively small variation within the network. Assuming a good agreement between readers, metrics that are more sensitive to MS errors should be preferred, since they can evaluate the MSs to a greater degree [14]. This was clearly reflected in ranking method 1, where average SBD values showed greater differences between the 2 readers (Fig. 5). Moreover, SBD showed clear discrimination regardless of the ranking method employed. This may be an added feature required in the evaluation of networks during future challenges.

4.5. Limitations of the study

The previously proposed metrics were adapted where possible but may not truly reflect what was proposed by the authors, especially for cases with no segmentations. In our study, cases with no segmentation yielded 0 for metrics with finite range and NaN for infinite cases (e.g. ASD). Such cases were few and only accounted for about 1% of 1197 slices used for evaluation. A small number of 20 test subjects were used for evaluation. However, the number is similar to the median number of test data used in challenges [3]. Moreover, the prostate volumes range is large, which allowed for sufficient variations during the leave-one-out evaluation. Only two readers with different amount of experience and background carried out the visual assessment and annotations of the GT masks, which is fewer than the median number of annotators used in challenges [3]. However, visual assessment was not carried out in these challenges and the task is time-consuming and tedious, thus limiting the availability of readers for this study. The evaluation was carried out in 2D (slice-by-slice) and the results may differ from that evaluated in 3D for the clinical data. However, evaluation in 2D slices generated a large dataset, which resulted in more consistent averaging and ranking of the networks. Evaluation was only performed on clinical prostate MR images and more evaluation is required using other clinical data (e.g. ultrasound) or data with other clinical application (e.g. tumor segmentation) for our results to be applicable for all medical segmentation.

Here, we selected only 9 previously proposed metrics with 4 different measures that we thought would improve segmentation evaluation. Many other metrics have been proposed but were not included in this study. Only three networks were compared in this study, which yielded consistent ranking across the selected metrics. Our results are different from other authors [3,5], and this is may be due to evaluation using 2D slices compared to 3D images, as in the work performed by Maier-Hein [3] and Taha and Hanbury [5], and the MSs generated by the three networks, whereby the MSs generated by SN were obviously poorer than DVN and HRN. HRN also yielded better segmentation close to the boundary of the prostate while DVN yielded slightly more crude segmentation, but the segmented prostate volume is as good as that generated by HRN [18]. Comparing networks yielding closer MSs would be better for evaluation. However, inferences can still be drawn from the variations in network ranking within a network and across the three networks using the leave-one-out evaluation.

5. Conclusions

Nine previously proposed segmentation evaluation metrics were compared with 24 traditional metrics to identify features or metrics required for clinical segmentation evaluation. Dual negative-positive visual assessment grading was employed for visual assessment to evaluate over- and under-estimation of the segmentations. It yielded good agreement between 2 readers of $R = 0.783$ and 0.764 for absolute (0–5) and all grading (−5 to 5). Information theoretic-based (MI and VOI) and pair-counting-based (PRI and ARI) metrics, probabilistic-based metrics, (ICC, KAP, AUC, and MARK), and error of spatial overlap-based metrics (GCE) were affected by matrix size. All selected spatial-overlap, spatial-distance-based and hybrid metrics were not affected by matrix size. DSC, ICC, ARI, ASD, and OQM yielded evaluation results of synthetic shapes consistent with visual expectations. RVD showed better plus-minus-sign agreement with visual assessment than reader agreement and had the best correlation with VA. DSC, ICC, ARI, ASD, OQM, SBD, and RVD showed good correlation ($|\rho| \geq 0.7$) with VA using Spearman rank-order correlation with ICC consistently ranked first. These 7 metrics were also yielded consistent ranking using all three ranking methods with VA. SBD showed slightly poor correlation with visual assessment and poorer synthetic shape evaluation than corresponding traditional metrics, but exhibited clear rank discrimination using leave-one-out evaluation for all ranking methods. OQM showed slightly poorer correlation than traditional metrics and poor rank discrimination but the best synthetic shape evaluation. The selected traditional metrics yielded a good correlation with visual assessment and good fixed shapes evaluation results but relatively poor network rank discrimination, with ICC performing the best. Our findings show that the good rank discrimination could be investigated as an additional metric feature required for better network performance evaluation.

Funding sources

This study was supported by the National University Health System (NUHS) Center Grant Seed Funding, Singapore [NUHSCGSF/2019/07].

Author contributions

Ying-Hwey Nai: Conceptualization, Methodology, Software, Validation, Writing - Original Draft, Supervision, Project administration, Funding acquisition. **Bernice W. Teo:** Formal analysis, Writing - Original Draft, Visualization. **Nadya L. Tan:** Visualization. **Sophie O'Doherty:** Methodology, Validation. **Mary C. Stephenson:** Writing - Review & Editing, Funding acquisition. **Yee Liang Thian:** Validation, Writing - Review & Editing, Funding acquisition. **Edmund Chiong:** Writing - Review & Editing, Funding acquisition. **Anthonin Reilhac:** Writing - Review & Editing, Supervision, Funding acquisition.

Declaration of competing interest

None.

Acknowledgments

We would like to acknowledge Mr. Koby Yi Wei Chua from Anglo-Chinese Independent, Singapore for his help in creating the prostate masks. We would also like to acknowledge Dr. Wynne Yuru Chua and Dr. Bertrand Wei Leng Ang from the National University Hospital (NUH), Singapore, for teaching prostate segmentation.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compbiomed.2021.104497.

References

[1] H. Kim, J.I. Monroe, S. Lo, M. Yao, P.M. Harari, M. Machtay, J.W. Sohn, Quantitative evaluation of image segmentation incorporating medical consideration functions, Med. Phys. 42 (2015) 3013–3023, https://doi.org/10.1118/1.4921067.

[2] V. Yeghiazaryan, I. Voiculescu, Family of boundary overlap metrics for the evaluation of medical image segmentation, J. Med. Imag. 5 (2018) 1, https://doi.org/10.1117/1.jmi.5.1.015006.

[3] L. Maier-Hein, M. Eisenmann, A. Reinke, S. Onogur, M. Stankovic, P. Scholz, T. Arbel, H. Bogunovic, A.P. Bradley, A. Carass, C. Feldmann, A.F. Frangi, P. M. Full, B. van Ginneken, A. Hanbury, K. Honauer, M. Kozubek, B.A. Landman, K. März, O. Maier, K. Maier-Hein, B.H. Menze, H. Müller, P.F. Neher, W. Niessen, N. Rajpoot, G.C. Sharp, K. Sirinukunwattana, S. Speidel, C. Stock, D. Stoyanov, A. A. Taha, F. van der Sommen, C.W. Wang, M.A. Weber, G. Zheng, P. Jannin, A. Kopp-Schneider, Why rankings of biomedical image analysis competitions should be interpreted with care, Nat. Commun. 9 (2018), https://doi.org/10.1038/s41467-018-07619-7.

[4] B.H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, L. Lanczi, E. Gerstner, M.A. Weber, T. Arbel, B.B. Avants, N. Ayache, P. Buendia, D.L. Collins, N. Cordier, J.J. Corso, A. Criminisi, T. Das, H. Delingette, Ç. Demiralp, C.R. Durst, M. Dojat, S. Doyle, J. Festa, F. Forbes, E. Geremia, B. Glocker, P. Golland, X. Guo, A. Hamamci, K. M. Iftekharuddin, R. Jena, N.M. John, E. Konukoglu, D. Lashkari, J.A. Mariz, R. Meier, S. Pereira, D. Precup, S.J. Price, T.R. Raviv, S.M.S. Reza, M. Ryan, D. Sarikaya, L. Schwartz, H.C. Shin, J. Shotton, C.A. Silva, N. Sousa, N. K. Subbanna, G. Szekely, T.J. Taylor, O.M. Thomas, N.J. Tustison, G. Unal, F. Vasseur, M. Wintermark, D.H. Ye, L. Zhao, B. Zhao, D. Zikic, M. Prastawa, M. Reyes, K. Van Leemput, The multimodal brain tumor image segmentation benchmark (BRATS), IEEE Trans. Med. Imag. 34 (2015), https://doi.org/10.1109/TMI.2014.2377694, 1993–2024.

[5] A.A. Taha, A. Hanbury, Cloud-based benchmarking of medical image analysis, Cloud-Based Benchmarking Med. Image Anal. (2017) 87–105, https://doi.org/10.1007/978-3-319-49644-3.

[6] T. Heimann, B. Van Ginneken, M.A. Styner, Y. Arzhaeva, V. Aurich, C. Bauer, A. Beck, C. Becker, R. Beichel, G. Bekes, F. Bello, G. Binnig, H. Bischof, A. Bornik, P. M.M. Cashman, Y. Chi, A. Córdova, B.M. Dawant, M. Fidrich, J.D. Furst, D. Furukawa, L. Grenacher, J. Hornegger, D. Kainmüller, R.I. Kitney, H. Kobatake, H. Lamecker, T. Lange, J. Lee, B. Lennon, R. Li, S. Li, H.P. Meinzer, G. Németh, D. S. Raicu, A.M. Rau, E.M. Van Rikxoort, M. Rousson, L. Ruskó, K.A. Saddi, G. Schmidt, D. Seghers, A. Shimizu, P. Slagmolen, E. Sorantin, G. Soza, R. Susomboon, J.M. Waite, A. Wimmer, I. Wolf, Comparison and evaluation of methods for liver segmentation from CT datasets, IEEE Trans. Med. Imag. 28 (2009) 1251–1265, https://doi.org/10.1109/TMI.2009.2013851.

[7] M. Styner, J. Lee, B. Chin, M.S. Chin, O. Commowick, T. Hoai-Huong, V. Jewells, S. Warfield, 3D segmentation in the clinic: a grand challenge II at MICCAI 2008 - MS lesion segmentation, Midas J. (2008) 1–6.

[8] O. Maier, B.H. Menze, J. von der Gablentz, L. Häni, M.P. Heinrich, M. Liebrand, S. Winzeck, A. Basit, P. Bentley, L. Chen, D. Christiaens, F. Dutil, K. Egger, C. Feng, B. Glocker, M. Götz, T. Haeck, H.L. Halme, M. Havaei, K.M. Iftekharuddin, P. M. Jodoin, K. Kamnitsas, E. Kellner, A. Korvenoja, H. Larochelle, C. Ledig, J.H. Lee, F. Maes, Q. Mahmood, K.H. Maier-Hein, R. McKinley, J. Muschelli, C. Pal, L. Pei, J. R. Rangarajan, S.M.S. Reza, D. Robben, D. Rueckert, E. Salli, P. Suetens, C. W. Wang, M. Wilms, J.S. Kirschke, U.M. Krämer, T.F. Münte, P. Schramm, R. Wiest, H. Handels, M. Reyes, ISLES 2015 – A public evaluation benchmark for ischemic stroke lesion segmentation from multispectral MRI, Med. Image Anal. 35 (2017) 250–269, https://doi.org/10.1016/j.media.2016.07.009.

[9] A. Popovic, M. de la Fuente, M. Engelhardt, K. Radermacher, Statistical validation metric for accuracy assessment in medical image segmentation, Int. J. Comput. Assist. Radiol. Surg. 2 (2007) 169–181, https://doi.org/10.1007/s11548-007-0125-1.

[10] A.A. Taha, A. Hanbury, Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool, BMC Med. Imag. 15 (2015), https://doi.org/10.1186/s12880-015-0068-x.

[11] H.S. Kim, S.B. Park, S.S. Lo, J.I. Monroe, J.W. Sohn, Bidirectional local distance measure for comparing segmentations, Med. Phys. 39 (2012) 6779–6790, https://doi.org/10.1118/1.4754802.

[12] R. Shi, K.N. Ngan, S. Li, The objective evaluation of image object segmentation quality. Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), Springer Verlag, 2013, pp. 470–479, https://doi.org/10.1007/978-3-319-02895-8_42.

[13] D.M.W. Powers, Evaluation: from precision, Recall and F-measure to Roc, informedness, markedness & correlation, J. Mach. Learn. Technol. 2 (2011) 37–63.

[14] V. Yeghiazaryan, I. Voiculescu, V. Yeghiazaryan, I. Voiculescu, Department of Computer Science an Overview of Current Evaluation Methods Used in Medical Image Segmentation CS-RR-15-08 an Overview of Current Evaluation Methods Used in Medical Image Segmentation, 2015. https://www.cs.ox.ac.uk/files/7732/CS-RR-15-08.pdf.

[15] S.M. Lee, J.H. Xin, S. Westland, Evaluation of image similarity by histogram intersection, Color Res. Appl. 30 (2005) 265–274, https://doi.org/10.1002/col.20122.

[16] E. Fernandez-Moral, R. Martins, D. Wolf, P. Rives, A new metric for evaluating semantic segmentation: leveraging global and contour accuracy, in: IEEE Intell. Veh. Symp. Proc. 2018-June, 2018, pp. 1051–1056, https://doi.org/10.1109/IVS.2018.8500497.

[17] Weblink for MedSD (Last date of reference: 29 Oct 2020): http://insightsoftwareconsortium.github.io/SimpleITK-Notebooks/Python_html/34_Segmentation_Evaluation.html.

[18] Y.-H. Nai, B.W. Teo, N.L. Tan, K.Y.W. Chua, C.K. Wong, S. O'Doherty, M. C. Stephenson, J. Schaefferkoetter, Y.L. Thian, E. Chiong, A. Reilhac, Evaluation of multimodal algorithms for the segmentation of multiparametric MRI prostate images, Comput. Math. Methods Med. 2020 (2020) 1–12, https://doi.org/10.1155/2020/8861035.

[19] E. Gibson, F. Giganti, Y. Hu, E. Bonmati, S. Bandula, K. Gurusamy, B. Davidson, S. P. Pereira, M.J. Clarkson, D.C. Barratt, Automatic multi-organ segmentation on abdominal CT with dense V-networks, IEEE Trans. Med. Imag. 37 (2018) 1822–1834, https://doi.org/10.1109/TMI.2018.2806309.

[20] W. Li, G. Wang, L. Fidon, S. Ourselin, M.J. Cardoso, T. Vercauteren, On the compactness, efficiency, and representation of 3D convolutional networks: brain parcellation as a pretext task, in: Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics). 10265 LNCS, 2017, pp. 348–360, https://doi.org/10.1007/978-3-319-59050-9_28.

[21] L. Fidon, W. Li, L.C. Garcia-Peraza-Herrera, J. Ekanayake, N. Kitchen, S. Ourselin, T. Vercauteren, Scalable multimodal convolutional networks for brain tumour segmentation, in: Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics). 10435 LNCS, 2017, pp. 285–293, https://doi.org/10.1007/978-3-319-66179-7_33.

[22] S. Vanbelle, A new interpretation of the weighted Kappa coefficients, Psychometrika 81 (2016) 399–410, https://doi.org/10.1007/s11336-014-9439-4.