Comparison between threshold-based and deep learning-based bone segmentation on whole-body CT images.

Noémie Moreau[a,b], Caroline Rousseau[c,d], Constance Fourcade[a,b], Gianmarco Santini[b], Ludovic Ferrer[c,d], Marie Lacombe[d], Camille Guillerminet[d], Pascal Jezequel[c], Mario Campone[c,d], Nicolas Normand[a], and Mathieu Rubeaux[b]

[a]Université de Nantes, CNRS, LS2N, F-44000 Nantes, France
[b]Keosys Medical Imaging, Nantes, France
[c]University of Nantes, CRCINA, INSERM UMR1232, CNRS-ERL6001, Nantes, France
[d]ICO Cancer Center, Nantes - Angers, France

ABSTRACT

Objectives: Bone segmentation can help bone disease diagnosis or post treatment assessment but manual segmentation is a time consuming and tedious task in clinical practice. In this work, three automatic methods to segment bone structures on whole body CT images were compared. Methods: A threshold-based approach with morphological operations and two deep learning methods using a 3D U-Net with different losses, one with a cross entropy/Dice loss and the second with a Hausdorff Distance/Dice loss, were developed. Ground truth bone segmentations were generated by manually correcting the results obtained with the threshold based method. The automatic bone segmentations were evaluated using a Dice score and Hausdorff distance. Visual evaluation was also performed by a medical expert. Results: Dice scores of 0.953, 0.986 and 0.978 were achieved for the Threshold-based method and the two deep learning methods, respectively. Visual evaluation showed that the deep learning method with a Hausdorff Distance/Dice loss performed the best.

1. DESCRIPTION OF PURPOSE

Whole body CT imaging is widely used for diagnosis, staging and follow-up of patients with malignancies. Additionally, bone segmentation, which provides a lot of information to assess bone disease, is used to evaluate metastatic tumor burden in cancer like in prostate or breast cancer.

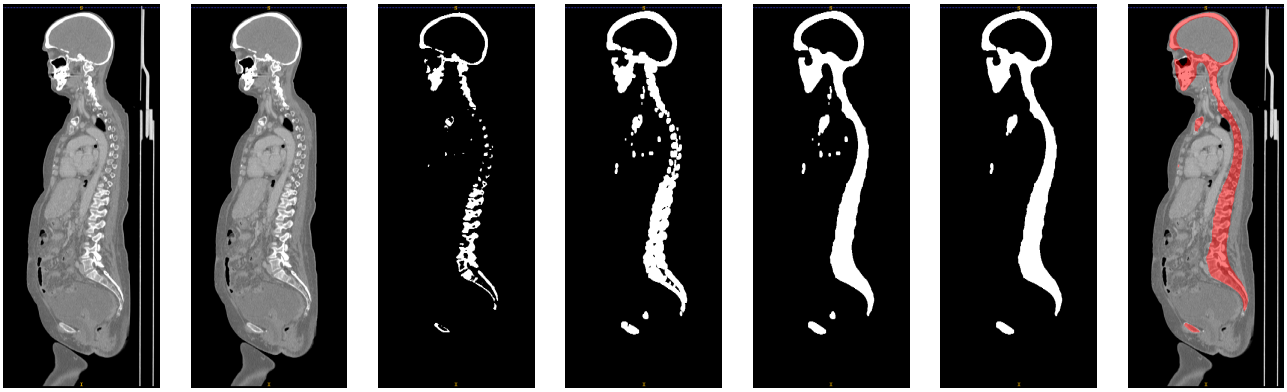
In prostate cancer, a method assessing this burden on bone scans is already used in clinical practice and is quantified with the Bone scan index (BSI) [1]. A new method on ^{68}Ga-PSMA PET/CT images developed by Bieth et al. showed promising results [2].

However, in breast cancer, no method based on bone scan has shown clinical relevance [3] but literature shows that tumor response can be assessed by evaluating the size and intensity of bone metastases on PET scans [4]. Our final goal is to develop a new method to automatically segment bone and bone lesions on PET/CT images, calculate a new index and prove its clinical relevance to evaluate the tumor response in the context of metastatic breast cancer. The present paper describes the first steps carried out to achieve this goal: the bone segmentation.

As shown in a previous work, the Dice score may not be the best index to evaluate bone segmentation, due to its low sensibility to small scattered false segmentations, which can be physiologically significative [5]. However, this type of errors can be enhanced by the Hausdorff Distance, as it takes into account the distance to the ground truth in its formulation. This led to think that a Hausdorff Distance-based loss integrated in a deep learning scheme for bone segmentation could give better results.

This work compares three approaches to segment the bone structure: one threshold-based and two deep learning based methods with different loss functions.

Further author information: (Send correspondence to Noémie Moreau)
N. Moreau: E-mail: noemie.moreau@keosys.com



(a) (b) (c) (d) (e) (f) (g)

Figure 1. Bone segmentation steps using the threshold-based method. (a) original CT, (b) removal of the table, (c) first threshold between 200 and 3000 HU, (d) dilation operation, (e) iterative hole filling, (f) connected components, (g) erosion and final result on the CT.

2. METHODS

2.1 Data

Fifty patients were recruited in the prospective $EPICURE_{seinmeta}$ metastatic breast cancer study (NCT03958136). Imaging data were acquired in two sites using different imaging systems. For this work, we only used the collected CT images. Bone segmentations were generated using the threshold-based method and then manually corrected by 4 non-specialist image processing researchers to be further used as ground truth. For the deep learning approaches, 50 patients, split into three folds according to the recruiting site and the extent of skeletal metastatic involvement, were used for training.

2.2 Threshold-based method

The bone structures have specific Hounsfield values allowing the use of threshold and morphological operations to segment them. The method presented here is inspired by the work of Banik, Rangayyan, and Boag [6]. Figure 1 shows the different steps of bone segmentation. First, the exam table is removed from the original CT to avoid its segmentation as bone. Then, the bone structures are segmented by applying a threshold between 200 and 3000 Hounsfield Unit (HU), and by carrying out dilation and iterative hole filling tasks. As the bone structure is completely connected by joints, a connected component filter was used to remove small errors found in the heart region as shown in the Figure 1-f. The last step is an erosion operation to prevent over-segmentation.

2.3 Deep learning methods

2.3.1 Network architecture

The 3D U-Net from the framework "no new net" (nnU-Net) [7] was used in the two deep learning methods. This framework achieved state-of-the-art performances on recent challenges like KiTS2019 [7]. Moreover, it allows to automatically set a number of hyper-parameters given informations such as input data volume and memory consumption (see figure 2). It also enables easy integration of new architectures and methods, such as new loss functions. Except for the loss functions, all network parameters used in this paper can be found in the original nnU-Net publication [7].

In this work, we compare the performance of the standard 3D nnU-Net with a combined Cross Entropy/Dice loss, to a tuned 3D nnU-Net using a Hausdorff Distance/Dice loss.

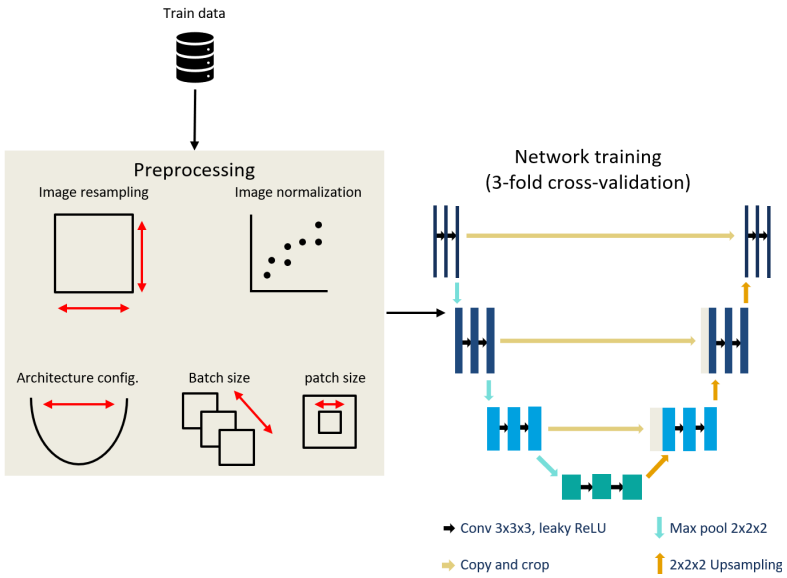


Figure 2. Preprocessing of the data with image resampling and normalization included in the nnU-Net framework. It automatically sets the batch size (min. size of 2), patch size and number of pooling operations, while maximizing the amount of spatial context. (Figure inspired from [7].)

Cross Entropy and Dice loss The sum of the cross entropy loss (L_{CE}) and the dice loss (L_{DCS}) is the loss used in the standard nnU-Net [7]. It is defined by Drozdal et al. [8] as:

$$L_{CE} = \sum_i p_i log(q_i) + (1 - p_i) log(1 - q_i))$$

$$L_{DCS} = -\frac{2\sum_i p_i q_i}{\sum_i p_i + \sum_i q_i}$$

with p the ground truth and q the predicted segmentation.

Hausdorff Distance and Dice loss The Hausdorff Distance is rarely used as a loss function because it is not differentiable. However, Karimi and Salcudean developed three alternative methods to approximate the Hausdorff distance, make it differentiable and use it as a loss [9]. Only the one using the Distance Transform (see Figure 3) will be used in this work as it is the closest approximation of the Hausdorff distance:

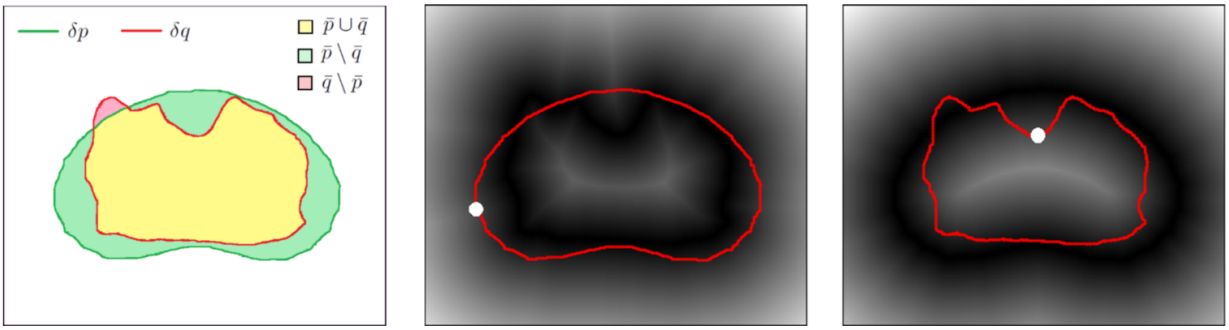
$$L_{HD} = \frac{1}{|\Omega|} \sum_{\Omega} ((p - q)^2 \circ (d_p^\alpha + d_q^\alpha))$$

with with p the ground truth and q the predicted segmentation, d_p and d_q their respective distance transforms, $\circ$ the Hadamard product and Ω the grid on which the image is defined. α determines how strongly the loss penalize the larger errors, the same value as in the original paper was used : $\alpha = 2.0$ [9].

The Hausdorff distance loss allows to weight the segmentation errors with their distance from the ground truth contrary of the Dice score that give importance to the volume error.

As the Hausdorff distance loss focuses only on the largest error which could lead to unstable learning, the loss is balanced with a Dice loss term: $L_{HD+DCS} = L_{DCS} + \lambda L_{HD}$.

In the original paper [9], λ is updated every epoch to give equal weights to each term of the loss : it is the ratio between the mean of L_{HD} and the mean of L_{DCS} over the batch size from the previous epoch. However, experiments showed that the training was still unstable with the automatic parameters set by the nnU-Net. For this reason, λ was fixed to 0 between the 1^{st} and the 250^{th} epoch. Between 250^{th} and the 750^{th} epoch, $\lambda = \text{mean}(L_{DCS})/\text{mean}(L_{HD})/500 * (epoch - 250)$. After the 750^{th} epoch, $\lambda = \text{mean}(L_{DCS})/\text{mean}(L_{HD})$.



(a) (b) (c)

Figure 3. Figures from [9] which explain the computation of the Hausdorff distance with distance transforms. (a) Overlap between the ground truth $\bar{p}$ and the predicted segmentation $\bar{q}$. (b) Distance transform d_q with δp overlaid in red. The white point is the location of $hd(\delta p, \delta q)$. (c) Similar to (b) with distance transform d_p and δq overlaid in red, for $hd(\delta q, \delta p)$.

2.4 Evaluation

2.4.1 Quantitative evaluation

Quantitative evaluation was obtained by computing the mean recall, mean precision, mean Dice score, mean Hausdorff distance and the maximum Hausdorff distance across all test cases.

Dice score The Dice score evaluates the degree of overlap between the ground truth and the result segmentations. An average Dice score as computed over all testing patients. A global Dice score was also performed by combining all the testing cases in a single one. The Dice score is the most common criteria to evaluate the performance of automatic segmentation methods. However for large volume like the bone, small errors can not weigh enough in the computation. This leads to really different visual results with the same Dice score depending on how close to the ground truth the errors are located.

Hausdorff distance The Hausdorff distance is an indicator of the largest segmentation error. It is computed between the boundaries of the ground truth and the result segmentations. The Hausdorff distance is the longest distance between one point in one of the two boundaries to its closest point in the other boundary. It is also a common criteria to evaluates the performance of automatic segmentation methods.

For the Dice score and the Hausdorff distance, a paired t-test (significance level at $p = 0.05$) was performed to evaluate the significance of the difference between methods.

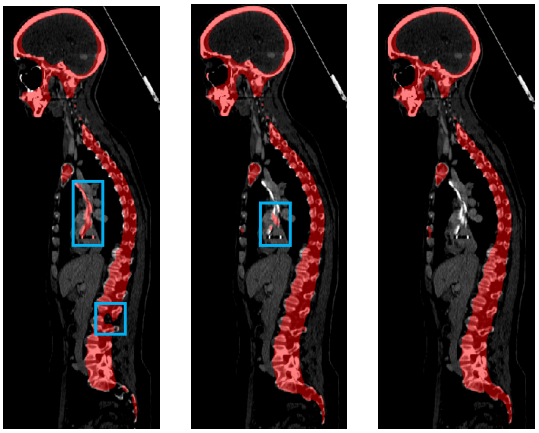
2.4.2 Qualitative evaluation

Qualitative evaluation was performed visually by a medical expert. For all cases, the expert was asked to choose the best segmentation result between the three proposed (one for each method). For each input CT volume, the three segmentations along with the original CT were shown simultaneously using the Keosys Viewer [10]. The expert could visualize the volumes in the three axes and roam through the slices. He could choose to show any combination of the three segmentations at a time. Each segmentation was randomly named with a different number in each case. If the difference between two segmentations was too small, the expert could choose both of them. If a segmentation presented to many errors, the expert could judge it non acceptable.

3. RESULTS

This work compares the results of three bone segmentation methods:

1) A threshold-based method with thresholding and morphological operations. 2) A standard 3D nnU-Net with a Cross entropy/Dice loss ($U\text{-}Net_{CE+DCS}$). 3) A tuned 3D nnU-Net with a Hausdorff distance/Dice loss($U\text{-}Net_{HD+DCS}$).



(a) (b) (c)

Figure 4. Bone segmentation results for the three methods on a patient with heart calcification. (a) $Threshold-based$: false segmentation of the heart and under-segmentation of the vertebrates, (b) $U\text{-}Net_{CE+DCS}$: over-segmentation of the heart , (c) $U\text{-}Net_{HD+DCS}$: best segmentation result.

Results of the quantitative evaluation are shown in table 1. All three methods have a Dice score superior to 0.90 but the threshold-based method is significantly worse than the two other methods in terms of Dice score (p-value $\leq$ 0.001 for both methods) and worse than the $U\text{-}Net_{HD+DCS}$ method in terms of Hausdorff distance (p-value $=$ 0.011). The two deep learning methods are significantly different in term of Dice score (p-value $\leq$ 0.001) but not in terms of Hausdorff distance (p-value = 0.29) even if the mean Hausdorff distance and the maximum Hausdorff distance are better with the $U\text{-}Net_{HD+DCS}$ method.

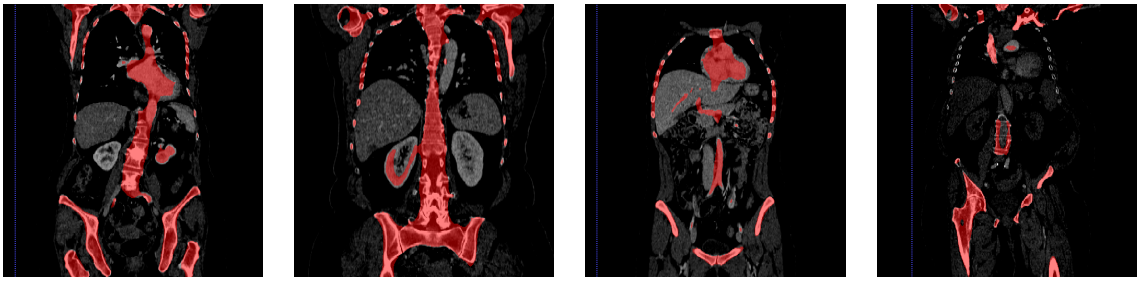
When conducting the qualitative assessment, the expert never chose the threshold-based segmentation, as it presents some segmentation errors in the heart or kidneys and under-segmentation in the vertebrates. Moreover, the threshold-based segmentation was assessed as not acceptable for 18 cases out of 50 as too many segmentation errors occur. For this 18 cases, a mean Dice of 0.947 was calculated. Figure 5 shows non acceptable segmentation results on 4 patients with their respective Dice score.
The expert did not observe any significant difference between the two deep learning approaches except for some cases (7 out of 50) where small volumes distinct from the bone were segmented by the $U\text{-}Net_{CE+DCS}$ like the small part of the heart in Figure 4. Overall, the $U\text{-}Net_{HD+DCS}$ method was always considered as providing the best results, in a tie for 43 out of 50 with $U\text{-}Net_{CE+DCS}$ method and 0 out of 50 with the threshold-based method.

Table 1. Quantitative evaluation

Methods	**Mean Recall**	**Mean Precision**	**Mean Dice**	**Global Dice**	**Mean HD**	**Max HD**
$Threshold-based$	0.949	0.963	0.955	0.953	103.5	187.0
$U\text{-}Net_{CE+DCS}$	0.988	0.984	0.986	0.986	49.99	131.0
$U\text{-}Net_{HD+DCS}$	0.980	0.975	0.978	0.978	39.59	112.73

4. DISCUSSION AND CONCLUSION

This work compared three approaches to segment the bone in whole-body CT images. All three methods gave good quantitative results but the visual analysis showed that deep learning methods performed best. This work also showed that a network with Hausdorff/Dice-based loss gave better results than a CE/Dice loss. Even if the quantitative results were not significantly different, the visual evaluation showed for a few patients that the $U\text{-}Net_{CE+DCS}$ method sometimes over-segment small volumes distinct from the bone, which is not the case with the $U\text{-}Net_{HD+DCS}$ method. This is consistent with a previous work that showed that a network trained with a CE/Dice loss tends to over-segment part of the heart in the presence of calcifications which shows the same Hounsfield value as the bone [5].



Dice score	0.88	0.96	0.90	0.94
	(a)	(b)	(c)	(d)

Figure 5. Bone segmentation results for the threshold-based method on 4 patients with their respective Dice score.

In terms of Dice score, all three methods achieved mean and global Dice score superior to 0.90 which is generally considered as a good segmentation result that should be confirmed by visual evaluation. However, the results shows that it was not the case for the threshold-based method with 18 cases assessed as unacceptable as too many segmentation errors appeared. Figure 5 shows 4 patients with some of the most frequent encountered errors: false segmentation of the heart, kidney and under-segmentation of the vertebrates. These errors are not well taken into account by the Dice score as only one patient has a Dice score less than 0.90. This confirms that the Dice score is not the best index to evaluate bone segmentation: the bone volume is large compared to other organs volume, which reduces the weight of their false segmentations.

In conclusion, in the case of bone segmentation, the Hausdorff/Dice-based loss improves the results as the dice score is not the best evaluator index and therefore the Hausdorff Distance part of the loss allow the network to consider all type of errors.

ACKNOWLEDGMENTS

This work is partially financed through "Programme opérationnel regional FEDER-FSE Pays de la Loire 2014-2020" noPL0015129 (EPICURE) and by the French National Association for Research and Technology (ANRT, CIFRE grant number 2019/0432).

The $EPICURE_{seinmeta}$ study was approved by the French Agence Nationale de Sécurité du Médicament et des produits de santé (ANSM, 2018-A00959-46) and the Comité de Protection des Personnes (CPP) IDF I, Paris, France (CPPIDF1-2018-ND40-cat.1), and a written informed consent was signed by each participant.

References

[1] M. Imbriaco et al. "A new parameter for measuring metastatic bone involvement by prostate cancer: The bone scan index". In: *Clinical Cancer Research* 4.7 (1998), pp. 1765–1772.

[2] M. Bieth et al. "Exploring New Multimodal Quantitative Imaging Indices for the Assessment of Osseous Tumor Burden in Prostate Cancer Using 68Ga-PSMA PET/CT." In: *Journal of Nuclear Medicine* (2017).

[3] M. Colombié et al. "Évaluation d'une méthode de quantification de la masse métastatique osseuse par mesure automatisée du Bone Scan Index, dans le suivi thérapeutique des cancers du sein". In: *Médecine Nucléaire* 4101.6 (2013), pp. 233–273.

[4] Ukihide Tateishi et al. "Bone Metastases in Patients with Metastatic Breast Cancer: Morphologic and Metabolic Monitoring of Response to Systemic Therapy with Integrated PET/CT". In: *Radiology* 247.1 (2008), pp. 189–196.

[5] N Moreau et al. "Deep learning approaches for bone and bone lesion segmentation on 18FDG PET/CT imaging in the context of metastatic breast cancer". In: *Engineering in Medicine and Biology Conference* (2020).

[6] S. Banik, R. Rangayyan, and G. Boag. *Landmarking and Segmentation of 3D CT Images*. Morgan & Claypool, 2009.

[7] Fabian Isensee et al. “Automated Design of Deep Learning Methods for Biomedical Image Segmentation”. In: *arXiv preprint* arXiv:1904.08128 (2020).

[8] M. Drozdal et al. “The importance of skip connections in biomedical image segmentation”. In: *Deep Learning and Data Labeling for Medical Applications* (2016), pp. 264–271.

[9] D. Karimi and S. E. Salcudean. “Reducing the Hausdorff Distance in Medical Image Segmentation With Convolutional Neural Networks”. In: *IEEE Transactions on Medical Imaging* 39.2 (2020), pp. 499–513.

[10] *Keosys Medical Imaging Viewer*. https://www.keosys.com/read-system/. Accessed: 2020-08-18.