EPiC Series in Health Sciences

Volume 7, 2024, Pages 211–215

Proceedings of The 24th Annual Meeting of the International Society for Computer Assisted Orthopaedic Surgery

Automatic Assessment of AI-produced 3D Medical Image Segmentations of the Scapula using Deep Learning

Garance Thoviste[1], Lhoussein Axel Mabrouk[1], Fabrice Bertrand[1] and Clément Daviller[1]

[1] Blue Ortho, an Exactech company, Meylan, France

garance.thoviste@blue-ortho.com, clement.daviller@blue-ortho.com

Abstract

Artificial intelligence (AI) and machine learning (ML) take an ever-growing place in medical care. Anatomical segmentation and reconstruction is one of the fields where ML reveals to be very efficient. Yet, verification of ML results still requires human verification and correction especially on pathologic morphologies. We propose an automatic assessment of AI-generated scapular reconstructions. Based on deep learning (DL), it separates predictions requiring little to no revision from predictions where corrected voxels represent more than 1% of the scapula, with an accuracy of 80%.

1 Introduction

Preoperative planning and surgery guidance of total shoulder arthroplasty (TSA) using 3D computed tomography (CT) reconstruction have proven to increase accuracy of glenoid placement even for experienced surgeons [1] [2]. 3D reconstructions of the scapula have transformed how surgeons visualize glenoid deformity and plan corrective surgery [3].

Generating the 3D scapular model is a tedious, time-consuming process that is generally performed manually by trained technicians. Though, recent advances in deep learning have made it possible to automatically segment CT scans with great accuracy using convolutional neural networks (CNN) [4]. However, human reviewing is still necessary to guarantee reconstructions accuracy required by a medical device. This is especially true for cases presenting metal artefacts, severe glenoid deformity or significant osteophyte formation.

In the following, we introduce and evaluate the accuracy of a DL-based model, designed to assess automated ML-reconstructions. It provides an index indicating the reliability of the produced segmentation, enabling the full automation of cases on which the AI is equivalent to trained technicians.

2 Material and Methods

2.1 Data

This study includes 144 shoulder CT scans from past TSA cases, that were split into 3 subsets:

- Training: 2240 images from 85 different scans
- Validation: 201 images from 9 different scans
- Test: all 12,800 images from 50 scans

Each scan was resized to comprise 256 slices. Each slice was associated to a scapula segmentation prediction as well as a correction mask. The predictions, in the form of probability maps, were generated by an ensemble of 2D models. Correction masks highlights the predictions pixels that were modified (either added or removed) by internal experts to create the final segmentation (considered to be the ground truth).

Slices were meticulously selected by hand to maintain consistent coherence in the correction masks, ensuring the optimal representation of all types of prediction errors. These errors, whether attributed to image quality, patient anatomy, or the presence of metal artifacts, were adequately reflected in each subset.

2.2 Model

The evaluation model is based on a CNN architecture [5] and takes as input an image from a CT scan and the AI-generated scapula segmentation of that image. It outputs the probability for each pixel to be modified by the operator. The correction masks predicted by the model are used to calculate for each case a Correction Rate (CR), which corresponds to the ratio of the modified volume (added and removed) to the actual bone volume.

2.3 Evaluation

The model accuracy was assessed over the test dataset, formed by 50 cases distributed as follows:

- 25 cases with a CR <= 1%, classified as acceptable
- 25 cases with a CR > 1%, to be reviewed.

The assessment was done at different levels by comparing:

- the difference between the CR obtained with the reference correction masks CR_{ref} and the CR obtained with the model's predictions CR_{pred}.
- the agreement between the classification prediction ("to review" or "acceptable") and the ground truth.

3 Results

3.1 CR error

The model's median CR error on the test dataset was ±0.87%. For the 17 cases with a predicted CR < 1%, the error drops to 0.27%. Figure 1 shows the difference between CR_{pred} and CR_{ref} for cases belonging to this dataset.

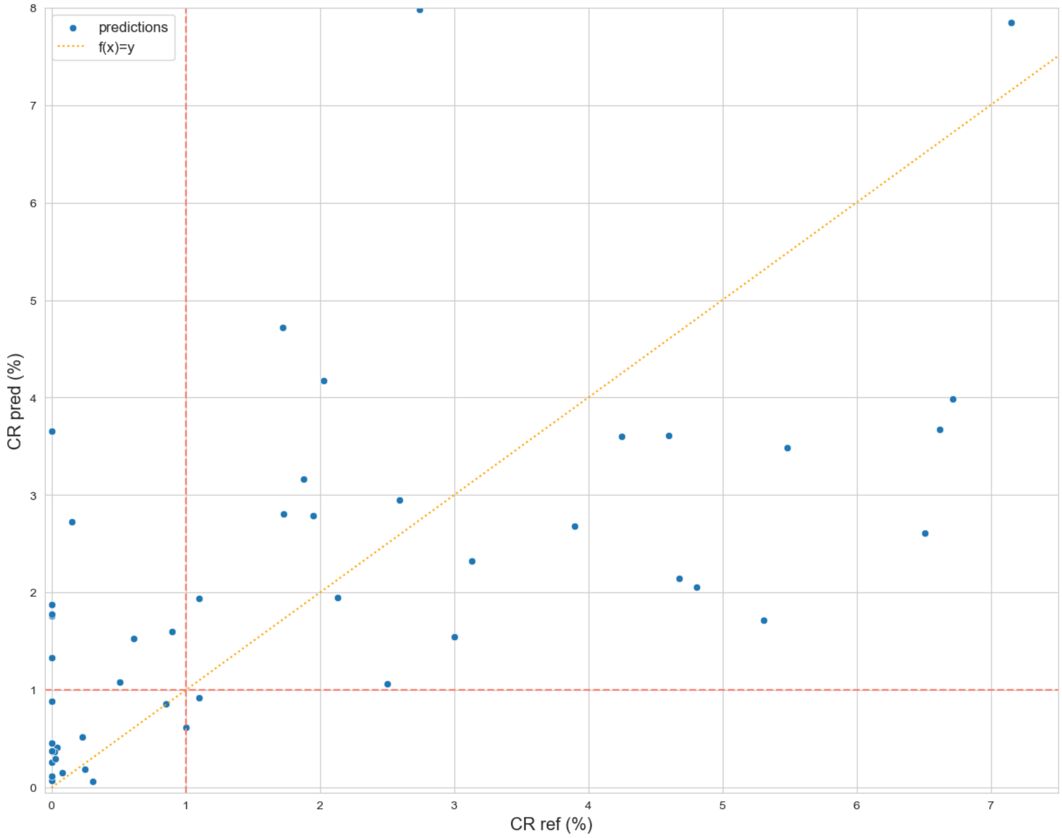

Figure 1: CR_{pred} expressed as a function of CR_{ref}. One outlier (CR_{ref} = 56.75% and CR_{pred} =40.07%) wasn't featured for visualization purposes. Dotted yellow line represents CR_{pred} = CR_{ref} and the red dashed lines show the CR threshold (1%) used for classification.

3.2 Classification

Table 1 presents the confusion matrix of the cases classification obtained with a $CR_{threshold}$=1%. The model's sensitivity was 96%, as only one case "To Review" was misclassified as "Acceptable" (CR_{ref} = 1.1 ; CR_{pred} = 0.92), and its specificity was 64%.

		Predicted classification			
	Total 50	To Review 33	Acceptable 17	**Accuracy = 0.8**	**F1 score = 0.83**
Actual classification	To Review 25	TP = 24	FN = 1	TPR = 0.96	FNR = 0.04
	Acceptable 25	FP = 9	TN = 16	FPR = 0.36	TNR = 0.64

Table 1: Confusion Matrix of the segmentation predictions classification

4 Discussion and Conclusion

Results indicate that the model tends to over-estimate the CR on acceptable cases, causing some to be misclassified. This is not necessarily problematic, as in a fully automated scenario our concern would be to avoid unreliable predictions being sent without being corrected. Thus, the parameters were tuned to favor sensitivity over specificity, so that only predictions with high confidence are accepted.

The test set contained features that the model had little to no exposure during training, such as very noisy scans or cases with contrast agent. But results and visual assessment of the predictions demonstrated that the model was still able to flag correction areas for human intervention.

Ensuring the accuracy of AI algorithms remains the primary obstacle in implementing fully automated models in healthcare. In this study, we showed that AI can be used to identify unreliable predictions for human review and correction.

Existing literature on unsupervised evaluation focuses primarily on anomaly detection [6] and uncertainty estimation [7] [8] [9], whereas our study introduces a method to quantify and categorize correction work.

Combined with the formerly developed automated scapula segmentation model [4], this model is anticipated to accelerate the scapula reconstruction processus. It should reduce the need for exhaustive manual oversight, moving closer to a fully-automated system. Predictions with low CR_{pred} could be sent to surgeons for pre-operative planning as is. With an ever-growing demand for TSA planning and navigation, this would optimize operator's work, focusing on challenging cases such as CT with implants.

For this last point, the CR threshold used to distinguish acceptable predictions would have to be tuned in a production setting to completely avoid false positives. Therefore, further work is planned to strengthen this study with more data.

References

[1] E. V. Cheung, "Computer navigation in shoulder arthroplasty," *Seminars in Arthroplasty: JSES,* vol. 33, no. 4, pp. 870-875, December 2023.

[2] A. Greene, S. Polakovic, C. Roche and Y. Dai, "Clinical Use of a Computer Assisted Anatomic Total Shoulder Arthroplasty System: An Analysis of 574 Cases," in *CAOS*, 2019.

[3] X. Fan, Z. Qiyang, T. Puxun, L. Joskowicz and X. Chen, "A review of advances in image-guided orthopedic surgery," *Physics in medicine and biology,* vol. 68, no. 2, 5 January 2023.

[4] G. Schmitt, A. Greene, S. Polakovic, N. Davis and F. Bertrand, "Results of a Machine Learning Algorithm for Automatic Three-Dimensional Segmentation of Computed Tomography Scans of the Shoulder," in *ORS Conference*, 2021.

[5] R. Venkatesan and B. Li, Convolutional Neural Networks in Visual Computing, T. &. F. Group, Ed., CRC Press, 2018.

[6] J. Yang, R. Xu, Z. Qi and Y. Shi, "Visual Anomaly Detection for Images: A Systematic Survey," *Procedia Computer Science,* vol. 199, pp. 471-478, 2022.

[7] D. Nie, L. Wang, L. Xiang, S. Zhou, E. Adeli and D. Shen, "Difficulty-Aware Attention Network with Confidence Learning for Medical Image Segmentation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019.

[8] A. N. Angelopoulos and S. Bates, "A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification," arXiv preprint arXiv:2107.07511., 2022.

[9] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U. R. Acharya, V. Makarenkov and S. Nahavandi, "A review of uncertainty quantification in deep learning: Techniques, applications and challenges," *Information Fusion,* vol. 76, pp. 243-297, 2021.