EPiC Series in Health Sciences
Volume 8, 2026, Pages 139–142
Proceedings of The 25th Annual Meeting of the International Society for Computer Assisted Orthopaedic Surgery

# Quantifying Segmentation Uncertainty to Evaluate the Quality of ML Generated Deltoid Masks in Shoulder Arthroplasty Patients

Rakesh Raushan [1], Ashish Singh[1], Likitha Shetty[1], Josie Elwell[1], Christopher P Roche[1], Vikas Kumar[1], Hamidreza Rajabzadeh-Oghaz[1*]
[1] Exactech, Inc. Gainesville, FL, USA
rakesh.raushan@exac.com, ashish.singh@exac.com, likitha.shetty@exac.com, josie.elwell@exac.com, chris.roche@exac.com vikas.kumar@exac.com, hamid.oghaz@exac.com

**Abstract**

Despite the growing development of image-based machine-learning models, their integration into clinical practice remains limited. A significant barrier to adoption is the reliability of these models' predictions. This study demonstrates the use of uncertainty analysis to evaluate output of a CT-based model trained to segment deltoid muscles in shoulder arthroplasty patients. By quantifying uncertainty through metrics such as entropy, mutual information, and variance, we created 46 distinct image-level uncertainty scores for 108 good-quality and 100 low-quality segmentation outputs. In addition, these uncertainty scores were used to train a Gaussian Naïve Bayes model to identify low-quality cases, and the results were compared with those from single-metric thresholding. The results show that boundary 75 percentile entropy is the most predictive single uncertainly parameters (accuracy: 68%, recall: 68%, precision: 67%) while the trained model outperformed all single predictive metrics (accuracy: 78%, %, recall: 76%, precision: 78%). Our study indicates a uses case of utilizing uncertainty analysis to identify segmentation outputs that may require further manual correction, which will increase the trust, and potentially help for clinical adoption of ML segmentation models.

## 1 Introduction

Shoulder arthroplasty relies on the assessment of radiological images (X-Ray, CT, and MRI scans) to evaluate the quality and integrity of bones and muscles. However, the current clinical practice for such evaluations is predominantly based on visual inspection and subjective qualitative evaluation. In recent years, advances in artificial intelligence (AI) have enabled the use of machine learning models

to evaluate musculoskeletal structures more objectively. A previous study shows an automated pipeline for segmentation and quantification of deltoid characteristics in shoulder CT scans with high accuracy [1]. However, as suggested in the article, an inaccurate segmentation necessitates manual corrections. Currently, the process of identifying such inaccurate or low-quality segmentation relies on visual inspection, limiting the scalability due to manual intervention leading to high cost and time requirements.

To address these challenges, uncertainty estimation has emerged as a promising approach [2, 3]. Uncertainty maps, derived from the probability outputs of segmentation models (e.g., unthresholded SoftMax outputs), provide additional insights into the confidence of segmented masks. These maps can highlight regions requiring correction, serving as guidance for technicians and reducing the reliance on purely manual inspection. Uncertainty estimates can be quantified into meaningful metrics and be summarized as an image-level single quality indicator (QI) score, providing a standardized way to assess segmentation reliability. In this study, we aim to test the performance of utilizing different 46 uncertainty metrics (such as Entropy and top 2 probability difference) for identifying low quality masks. Furthermore, using a database of expert evaluated and labeled deltoid mask, we train a model to increase performance of identifying low quality segmented masks.

# 2 Methods

A previously trained and clinically validated model for segmentation of deltoid muscles using 3D CT scan images was utilized for this study [1]. The model was previously used to segment pre-operative CT images of 4,009 primary shoulder arthroplasty patients [4]. All segmented images were visually evaluated by two observers familiar with anatomy of deltoid and classified to a. acceptable, b. not acceptable, where not acceptable are those that, the model could not accurately capture the boundary of deltoid muscle. A sample of accepted and not accepted cases are shown in Figure 1. For this study, we randomly selected 108 segmented masks (gender: 74 F, 31 M, 3 NA; age:71±8 yrs) with acceptable quality and 100 segmented masks (gender: 37 F, 54 M, 9 NA; age:69±9 yrs) with poor quality or rejected cases. Utilizing the unthresholded SoftMax outputs, we calculated pixel-level uncertainty measures entropy, top2 diff probability, and measured an image-level and then mask boundary level score by quantifying these uncertainties with metrics *Percentiles* (P25, P50, P75): percentile value of uncertainty in boundary region; *Mean* : Mean value of uncertainty in boundary region; *Standard Dev*: Standard dev of uncertainty values in the boundary region; En-o-Ec: Entropy of the expected class, Ex-e-Ec: Expected entropy of expected class. The dataset was split into training and test sets in an 80:20 ratio. The training set was further divided into training and validation subsets using a 90:10 split. A check for collinearity was done for the features and those with high correlation were eliminated by keeping those that has higher performance of testing cohorts. The training data were used to train the Gaussian Naïve Bayes model, while the threshold cut-off was determined using the validation set. For the remaining test dataset, performance of model was compared against single feature thresholding.

# 3 Results

Frothy six single metrics were measured as single score that quantify segmentation quality, summary of top performed metrics is reported in Table 1. After removing redundant features by identifying the collinearity, 13 parameters remained. Table 1 describes the performance of different metrics as well as trained model on identifying cases with low quality. Among single metrics P75 of Entropy had the highest performance (accuracy of 68.27%, recall 68.0%, precision of 66.67%), while the model could outperform all single metrics (accuracy of 78%, recall 76%, precision of 78).

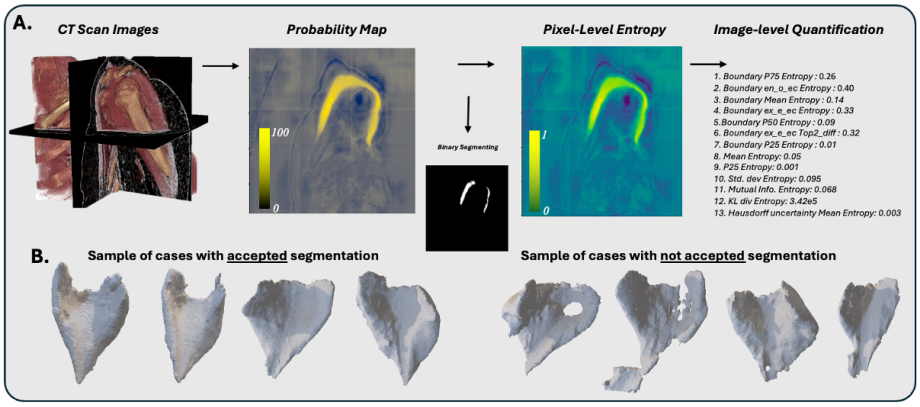


**Figure 1:** A. Workflow of quantifying uncertainty for segmentation masks**.** B. Sample of selected acceptable and not acceptable cases.

**Table 1**: Report of uncertainty analysis measurements.

| | *Metrics* | *50 Percentile of low-quality Case* | *50 Percentile of high-quality Case* | *P-value* | *Threshold (50 Percentile of 300 samples)* | *Accuracy* | *Recall* | *Precision* | *F-score* |
|---|---|---|---|---|---|---|---|---|---|
| *1* | *B P75 Entropy* | *0.269* | *0.248* | *<0.05* | *0.255* | *68.27* | *68.00* | *66.67* | *67.33* |
| *2* | *B En-o-Ec Entropy* | *0.407* | *0.384* | *<0.05* | *0.393* | *68.27* | *69.00* | *66.35* | *67.65* |
| *3* | *B Mean Entropy* | *0.141* | *0.129* | *<0.05* | *0.134* | *68.75* | *68.00* | *67.33* | *67.66* |
| *4* | *B Ex-e-Ec Entropy* | *0.327* | *0.302* | *<0.05* | *0.314* | *67.79* | *69.00* | *65.71* | *67.32* |
| *5* | *B P50 Entropy* | *0.096* | *0.071* | *<0.05* | *0.082* | *68.27* | *68.00* | *66.67* | *67.33* |
| *6* | *B Ex-e-Ec Top2_diff* | *0.322* | *0.294* | *<0.05* | *0.308* | *66.83* | *67.00* | *65.05* | *66.01* |
| *7* | *B P25 Entropy* | *0.016* | *0.010* | *<0.05* | *0.013* | *63.46* | *65.00* | *61.32* | *63.11* |
| *8* | *Mean Entropy* | *0.048* | *0.044* | *<0.05* | *0.046* | *63.94* | *64.00* | *62.14* | *63.05* |
| *9* | *P25 Entropy* | *0.001* | *0.0007* | *<0.05* | *0.0009* | *63.94* | *54.00* | *65.06* | *59.02* |
| *10* | *Std. dev Entropy* | *0.091* | *0.089* | *<0.05* | *0.090* | *62.02* | *61.00* | *60.40* | *60.70* |
| *11* | *Mutual Info. Entropy* | *0.064* | *0.063* | *<0.05* | *0.0635* | *53.37* | *60.00* | *51.28* | *55.30* |
| *12* | *KL div Entropy* | *2.79e5* | *2.45e5* | *<0.05* | *2.64e5* | *60.58* | *61.00* | *58.65* | *59.80* |
| *13* | *HU Mean Entropy* | *0.003* | *0.002* | *<0.05* | *0.0025* | *61.06* | *74.00* | *57.36* | *64.63* |
| *14* | *Q-indicator Model* | *-* | *-* | *-* | *-* | ***78.0*** | ***76.0*** | ***78.0*** | ***77.0*** |

**B,HU, En-o-Ec, and Ex-e-Ec refer to boundary, Hausdorff uncertainty, entropy of expected class, and expected entropy of expected class respectively.*

# 4 Discussion

While numerous studies have developed image-based machine learning models, very few have reached clinical practice. There are several studies proposing different methodology and metrics to evaluate the quality of image segmentation and mostly designed to identify areas of segmentation error or to guide human-in-the-loop annotation [5, 6]. The results of our study demonstrated the application of uncertainty analysis in identifying cases that were incorrectly segmented by our deltoid model. Utilizing this approach could improve machine learning models by prioritizing cases with higher uncertainty for correction and incorporating them into the training dataset. This study has some limitations. First, while the model in this study is designed to evaluate one of the soft-tissue segmentations, its applicability to other soft-tissue segmentations such as that of rotator cuff and bone segmentations such as of scapula and glenoid remains unvalidated. Second, although the study focuses on prediction of uncertainty, it does not address inherent bias due to underlying segmentation model. Third, our training and testing dataset was relatively small, and additional external validation is necessary to ensure consistent model performance across different populations and imaging centers. In future work, we plan to incorporate other uncertainty estimation methods, such as epistemic approaches, to assess uncertainties arising from the model's limited knowledge.

# 5 Conclusions

In this study, we demonstrated an application of uncertainty analysis to identify quality of deltoid segmentations. Integrating this approach into clinical workflows can streamline evaluations, improve model reliability, and prioritize cases requiring manual correction for iterative model development.

# Reference

[1] Rajabzadeh-Oghaz, H., Elwell, J., Kumar, V., Mabrouk, L., Daviller, C., Berry, D., Singh, A., Polakovic, S., Schoch, B., Roche, C.: Machine-Learning Model for Quantification of Deltoid Characteristics. Proc. 2024 Orthop. Res. Soc. (2024)

[2] Czolbe, S., Arnavaz, K., Krause, O., Feragen, A., "Is segmentation uncertainty useful?" CoRR, 2021.

[3] Whitbread, L., Jenkinson, M., " Uncertainty Categories in Medical Image Segmentation: A Study of Source-Related Diversity", unsure, 2022.

[4] Rajabzadeh-Oghaz, H., Kumar, V., Berry, D., Singh, A., Schoch, B., Aibinder, W., Gobbato, B., Polakovic, S., Elwell, J., Roche, C.P.: Impact of Deltoid CT Image Data on the Accuracy of Machine Learning Predictions of Clinical Outcomes After Anatomic and Reverse Total Shoulder Arthroplasty. J. Clin. Med. (2024)

[5] Jalal, N., Śliwińska, M., Wojciechowski, W., Kucybała, I., Rozynek, M., Krupa, K., Matusik, P., Jarczewski, J., Tabor, Z. "Evaluating Uncertainty Quantification in Medical Image Segmentation: A Multi-Dataset, Multi-Algorithm Study." Appl. Sci. 2024.

[6] Ye, A., Chen, Q. Z., & Zhang, A. (2023, November). Confidence contours: Uncertainty-aware annotation for medical semantic segmentation. In Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, 2023.