Effect of Different Uncertainties in Medical Image Segment Error Estimation

Interactive Segmentation of 3D Medical Images

Sungjin Kim

Supervisors: Klaus Hildebrandt, Nicolas Chaves-de-Plaza

EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfillment of the Requirements
For the Bachelor of Computer Science and Engineering
June 25, 2023

Name of the student: Sungjin Kim
Final project course: CSE3000 Research Project
Thesis committee: Klaus Hildebrandt, Nicolas Chaves-de-Plaza, Thomas Abeel

Abstract. Although automated segmentation of 3D medical images produce near-ideal results, they encounter limitations and occasional errors, necessitating manual intervention for error correction. Recent studies introduce an active learning pipeline as an efficient solution for this, requiring user corrections only on some of the most uncertain parts of the automatically segmented image. It does so by combining different uncertainty fields, which are various ways to quantify possible errors. We investigate into its individual uncertainty fields and their combination scheme in attempt to validate its methods. Additionally, we replace its methods for estimating uncertainty with another common way to do so, called the ensemble method, to test possible improvements at uncertainty estimation. Results of this research validates the combination method of the active learning pipeline, and shows weak advantages but strong disadvantages of the ensemble method when compared to the combined method of the active learning pipeline.

1 Introduction

State-of-the-art in automatic medical image segmentation produce results that resemble the quality of the ones drawn by human experts [1]. These advancements in segmenting specific entities in 3D images have made it convenient for clinicians to extract crucial anatomical structures efficiently, i.e. labeling a specific organ from a CT scan image of a person. However, such automatic segmentations still exhibit certain limitations and occasionally introduce errors in the segmentation process. Thus, manually checking and correcting these errors remains necessary, a laborious and time-consuming task, impeding the widespread adoption of fully automated segmentation methods in clinical practice.

To address this challenge, an interactive iterative pipeline was proposed in [2]. This algorithm uses the idea of *Active Learning* (AL) [3] framework, querying certain parts in the automatically segmented image with the most uncertainty for users to manually correct only on such parts, maximizing the effect of a single input instance. The pipeline in detail, works by first taking user-given labels for a single plane slice of the 3D input image. The automatic segmentation algorithm then labels voxels in the image according to the predicted classification, where it is also generated an uncertainty field that represents the lack of confidence in the classification of each voxel. Based on the uncertainty field, the pipeline selects a slice with the most inaccuracy, where the user then provides correction inputs using 2D segmentation tools. This process is repeated by feeding additional user-generated inputs into the automatic segmentation algorithm again, until the user is satisfied with the segmentation results. This would ensure that the quality of the segmentation would eventually reach the clinical standards of the user while requiring minimal amount of manual effort. We refer to this pipeline as the baseline from now on, as this forms the basis of the research.

In this paper, we focus on the generation of the uncertainty field — the estimation of incorrectness in the classification produced by automatic segmentation algorithms. Specifically, since the uncertainty field generation in the baseline works by combining different uncertainty terms, we attempt to investigate the roles of different terms in the baseline and their combination scheme, as [2] does not show extensive evaluation for these. Additionally, we apply the concept of one of the current methodologies for quantifying uncertainties to compare their performance compared to the method in the baseline. Current methodologies can be categorized into four high-level modalities [4]. Firstly, *Deterministic Methods* uses a single processing network such as evidential deep learning [5] to estimate uncertainty, a simple and efficient way to do so. Secondly, *Bayesian Neural Networks* on the other hand, varies the parameters of a deep learning model [6] to estimate uncertainty by creating multiple outputs and quantifying its distribution. Thirdly, *Ensemble Methods* uses multiple models to make predictions, where the variance of individual predictions serve as a measure of uncertainty [7]. Finally, *Test-Time Data Augmentation* augments the test data to assess uncertainty from the outputs created from feeding the augmented test data [8]. From the four modalities we choose the ensemble method, as it is not too simple, do not require a neural network, and can be evaluated using data with minimum pre-processing.

This leads to the following research questions: *What role does different uncertainties and the combination scheme play in the results of the baseline? What other ways of uncertainty affect the performance of estimating errors?* To answer the questions, we first implement the baseline pipeline to explore its uncertainty terms. Then we implement an application of the ensemble method to experiment if such method can improve the estimation of errors.

We use the following structure for this paper to discuss the findings throughout this research. Section 2 describes the methodology we use to generate uncertainty fields, focusing on the qualities of an uncertainty field, how the baseline creates them, and how we introduce a new way to create them. Section 3 then describes the procedures and metrics we use to evaluate different uncertainties, along with the results and its analysis. Finally, we conclude the paper by reflecting on issues that may render as problematic, and suggesting what could be done to extend this research.

2 Methodology

2.1 Preliminaries

Notation For clarity, we define the notations used in the baseline, where $\Omega \subset \mathbb{R}^3$ represent the three-dimensional domain of the image, $y : \Omega \to \{0, 1\}$ the classification label (0 for background and 1 for foreground), $I : \Omega \to \mathbb{R}$ the image intensity function, $p_1 : \Omega \to \mathbb{R}$ the foreground classification probability, and $U : \Omega \to \mathbb{R}$ the uncertainty field.

Inputs and Outputs of Uncertainty Field Algorithm We also define the inputs and outputs for an algorithm that creates an uncertainty field. Assuming a voxel $x \in \Omega$, the algorithm takes as input $I(x)$ the input image intensities, $y(x)$ the partially erroneous segmentation labeling, and $p_1(x)$ the probability of

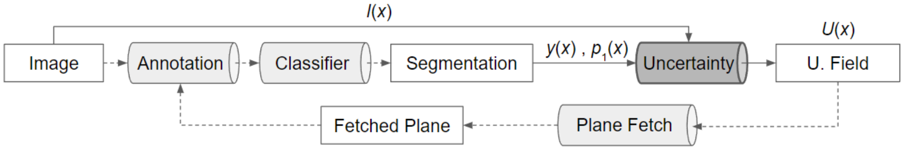


Fig. 1: Visualization of the baseline pipeline. The uncertainty field generation algorithm takes as input $I(x)$ from the image, and $y(x)$ and $p_1(x)$ from the segmentation predicted by the classifier to produce the uncertainty field $U(x)$.

x being classified as 1 in the generated segmentation, to generate the uncertainty field U where the values between 0–1 indicate the lack of confidence the segmentation has for each x. Refer to Fig. 1 to see how this process takes place in the entire pipeline of the baseline.

Effective Uncertainty Field An effective uncertainty field corresponds well to the actual difference between the generated segmentation and the ground truth. Uncertainty should ideally be minimal for voxels where the segmentation is correct, and maximal for those incorrect. Failing to do so may cause fault in the next step of the pipeline, making it query a slice that is perceived to be the most inaccurate, but in fact is one of the accurate slices when compared to the ground truth. Since the goal is to minimize the number of user inputs, it is necessary to maximize the correctness of the uncertainty field to also maximize the effectiveness of a single user input.

2.2 Baseline Uncertainty Field

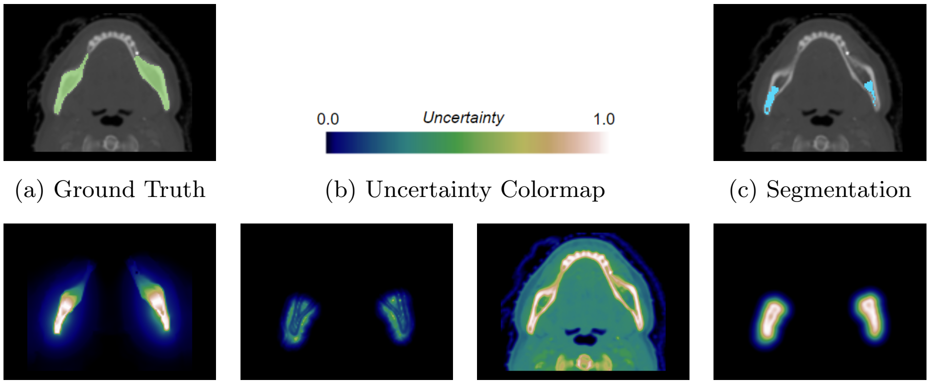
The baseline uncertainty field utilizes the expression below to generate uncertainty for each voxel x in the image domain, which is a weighted sum of four energy terms.

$$U(x, y) = \lambda_E U_E(x, p_1(x)) + \lambda_B U_B(x, y) + \lambda_R U_R(x, y) + \lambda_S U_S(x, y), \tag{1}$$

where U_E represent the entropy energy based on classification probabilities, U_B the boundary energy based on image gradient around segmentation boundaries, U_R the regional energy based on intensity distribution for different labels, and U_S the smoothness energy based on the neighboring voxels around x. Fig. 2 shows examples of each energy term for different organ segmentations.

Entropy Energy quantifies the degree of unpredictability of the classification labeling by calculating the entropy of p_1(x), which directly relates to the segmentation.

$$U_E(x, p_1(x)) = -p_1(x) \log_2 p_1(x) - (1 - p_1(x)) \log_2(1 - p_1(x)) \tag{2}$$



(a) Ground Truth

(b) Uncertainty Colormap

(c) Segmentation

(d) Entropy E.

(e) Boundary E.

(f) Regional E.

(g) Smoothness E.

Fig. 2: Visualization of the four energy terms. (a) shows in green an ideal segmentation based on ground truth, where (c) shows in blue an erroneous segmentation. (b) shows the colors mapped to different uncertainty values for the four baseline energy terms (d–g) which predict the error of (c).

The expression above produces low entropy when $p_1(x)$ is either near 0 or 1, meaning that the classification is highly predictable - the blue region in Fig. 2 (d). This changes when the probability approaches 0.5 as the classification becomes less predictable, producing high entropy and therefore uncertainty.

Boundary Energy checks if there are sharp edges in the input image around the classification boundaries of the segmentation, as classification based on *Random Walker* (used in baseline) tends to become inaccurate for images without sharp edges [9].

$$U_B(x,y) = \delta(D_s(x,y))\frac{1}{1+|\nabla I(x)|^{\alpha}} \tag{3}$$

Here, the delta function $\delta(d)$ taking as input $D_s(x,y)$ the distance between x and the nearest classification boundary, approaches zero as the distance increases; which is why in Fig. 2 (e) the values are in proximity of the classification boundary. By multiplying this with the inverse of the image intensity gradient magnitude, the boundary energy produces high uncertainty when the gradient is weak (smooth surfaces) for voxels in proximity.

Regional Energy predicts voxel classification using its image intensity by comparing to the normal distributions of the intensities of other voxels with the same labels, using the expression below:

$$U_R(x,y) = p(Y=y(x)|I(x)) = \frac{p(I(x)|Y=y(x))}{p(I(x)|Y=0)+p(I(x)|Y=1)} \tag{4}$$

Here, $p(I(x)|Y=0)$ and $p(I(x)|Y=1)$ are estimated by the normal distribution probability densities of the intensities of voxels labeled as 0 and 1, respectively.

Voxels with similar intensities with voxels already labeled with the same label would produce high regional energy. This highlights voxels that are likely to be labeled, contributing to finding unlabeled voxels that should be labeled. Note that regional energy does not depend on distances, thus it may render uncertainties for most voxels, like shown in Fig. 2 (f).

Smoothness Energy calculates its result based on the surface area of the segmentation for voxels around x. Below is the expression used for the calculation, where N_x represents the neighboring voxels around x.

$$U_S(x, y) = \iiint_{N_x} \delta(D_s(x, y))\, dV \tag{5}$$

This idea is similar to boundary energy, however the calculation measures how close the neighbouring voxels are to the classification boundaries. The more neighbours are close to the boundaries, the higher the energy would be, like depicted in Fig. 2 (g).

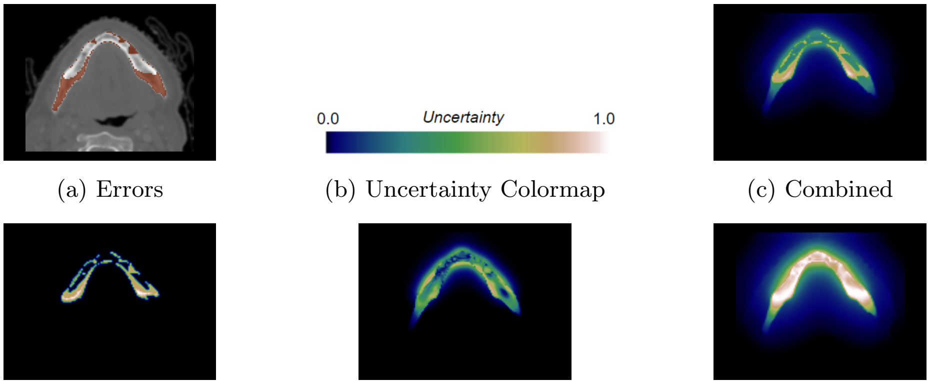
2.3 Ensemble Method Uncertainty Field

As stated in the introduction section, an ensemble method uncertainty estimation uses multiple models to make segmentations, where the variance of individual predictions (or the *Ensemble*) serve as a measure of uncertainty. This accounts for the possibility that the initial prediction is an outlier, measuring its uncertainty using other predictions. In our case, after generating the initial segmentation we additionally create an ensemble of classification probabilities $p_1(x)$ and classification labels $y(x)$ according to the probabilities, by generating predictions while changing the parameters of the classifier algorithm (further discussed in Section 3.1 - Implementation). Such parameters control the behavior of the algorithms, especially for areas with the hardest decision, thus producing different classification boundaries in the most uncertain areas. When the ensemble is collected, we first check if most predictions in the ensemble make the same classification, then measure the variance of the predicted probabilities, and finally incorporate entropy energy from equation (2) calculated from the mean of the predicted probabilities.

Let us define U_{Ens} the ensemble uncertainty field, where $x \in \Omega$ represents a voxel in the 3D image domain, $U_E : \Omega, \mathbb{R} \to \mathbb{R}$ the baseline entropy energy, $N_F : \Omega \to \mathbb{Z}$ the voxel-wise number of foreground predictions, $N \in \mathbb{Z}$ the number of all predictions (size of the ensemble), $p_v : \Omega \to \mathbb{R}$, the variance of the ensemble classification probabilities, and $p_m : \Omega \to \mathbb{R}$ the mean probability of the ensemble.

$$U_{\text{Ens}}(x) = \lambda_1 U_E(x, \frac{N_f(x)}{N}) + \lambda_2 p_v(x) + \lambda_3 U_E(x, p_m(x)) \tag{6}$$

Here the first term calculates the entropy of the percentage of foreground labels, quantifying the non-unanimity of the predictions; we call this the labeling entropy. The second term measures the variance of the probabilities in the ensemble, which also does the same but in a non-thresholded way. Finally the third term computing the baseline entropy energy of the mean probability of the ensemble to measure lack of confidence in the predictions. Fig. 3 illustrates examples of each terms and the final result.



(a) Errors

(b) Uncertainty Colormap

(c) Combined

(d) $H(N_f(x)/N)$

(e) $p_v(x)$

(f) $H(p_m(x))$

Fig. 3: Visualization of the ensemble uncertainty (c). (a) shows in red the difference between the ground truth and the initial segmentation, where (b) shows colors mapped to the values of the (intermediate) results of the ensemble uncertainty (c–f). (d) shows the labeling entropy, (e) the probability variance, and (f) the baseline entropy energy.

3 Experimental Setup and Results

3.1 Setup

We investigate the four baseline uncertainty energy terms from equations (2–5), and add up the terms to produce a combined baseline uncertainty field. We then compare the uncertainty field generated from the ensemble method to the combined baseline uncertainty field.

Implementation We first define the implementation details of the segmentation: we use the *Random Walker* (RW) method [9] to provide probabilistic segmentation, where each voxel is labeled based on their classification probabilities (i.e. classified as foreground when $p_1(x) \geq 0.5$). We set the only parameter of the RW algorithm as $\beta = 0.1$, since it produces reasonable results compared to other values.

Baseline Uncertainty Implementation uses distance transform for calculating $D_s(x, y)$, the expression $e^{-\frac{x^2}{2\alpha}}$ where $\alpha = 8$ for approximating δ-functions in equation (3) and (5), and $\alpha = 0.5$ in equation (3). The purpose of α parameters is to calibrate for the distance between voxels in the 3D image domain, and such values worked good in our test cases. We use different combinations of equation (1) listed in Table 1 to test different energy terms.

Table 1: Variations of equation (1) for combining energy terms

Description	Expression
Entropy Energy	U_E
Boundary Energy	U_B
Regional Energy	U_R
Smoothness Energy	U_S
All Energies (same as [2])	$0.80U_E + 0.05U_B + 0.15U_R + 0.00U_S$

Ensemble Uncertainty Implementation first generates a prediction using RW from above, which we consider as the final segmentation. We then create 9 additional predictions while taking a random nonzero value for β from a normal distribution with $\mu = 0.1$ and $\alpha = 0.05$, forming an ensemble of size $N = 10$ combined with the initial prediction. We use $\lambda_1 = 0.5$ and $\lambda_2 = \lambda_3 = 0.25$, emphasizing labeling entropy as classification labels have direct relation to errors.

Evaluation We evaluate the uncertainty fields by measuring their correspondence with the actual error. We do so qualitatively by performing manual evaluation where we compare segmentation errors and uncertainty fields using 3D visualization tools. Additionally, the following paragraphs describe some quantitative evaluation metrics we use for this paper.

Receiver Operating Characteristic (ROC) Curves illustrate the ability of a thresholded classifier to make correct decisions, which in this case is the capability of the uncertainty field to mark areas with actual errors as uncertain. To do such analysis for the uncertainty fields we create, we apply thresholding using the mean value of the uncertainty field, as it appear to be a logical split of the continuous uncertainty into binary uncertainty when manually comparing the two for fields with different uncertainty distributions. We then compare the binary uncertainty field with the error image (difference between the ground truth and the segmentation) in order to compare correct and incorrect predictions.

Mutual Information $I(X; Y)$ measures the correlation between two events X and Y, in a set of data involving the two. Below describes the expression to calculate $I(X; Y)$, where $P(x)$ represent the probability of $x \in X$ occurring, and

$P(x, y)$ the joint probability of x and y occurring simultaneously.

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} P(x,y) \log_2 (\frac{P(x,y)}{P(x)P(y)}) \tag{7}$$

As a thresholded analysis loses substantial amount of information that comes from the differences in uncertainty values, we compute the correlation between a raw uncertainty field of continuous values and a voxel-wise error image [10]. We use 20 bins to make the calculation possible while maintaining much of the continuous property of the raw uncertainty field.

Data We take 12 CT images of patients' head and neck from [11], from which we focus on two organs: the mandible and the parotid gland. For each image, we create two trimmed images that bounds closely to the organs by introducing some margin to the ground truth segmentation of the organs. For each trimmed images, we randomly select a 2D slice to label the slice according to the ground truth, and create 10 predictions as described above in the *Implementation* section. Here the first predictions serve as final segmentations such that the result of the segmentations are the same as without using the ensemble, as we focus on improving the estimation of uncertainty than the segmentation itself.

We now divide the data into four categories, based on the organ and the Dice Similarity Coefficient (DSC) [12] of the final segmentation: Mandible - High DSC, Mandible - Low DSC, Parotid - High DSC, and Parotid - Low DSC. This would provide context when evaluating uncertainty fields, enabling us to reflect on the circumstances of interesting results. The randomly created segmentations appeared to cluster around DSC of 0.6 and 0.2 for both organs, which are approximately the mean DSC of the high and low DSC categories above.

3.2 Baseline Uncertainty Results

We compare the four baseline energy terms and the combined uncertainty using different combinations of the terms, according to Table 1. We do this for the four categories of the data from the above section; Fig. 4 shows some snapshots of the resulting uncertainty fields.

Qualitative Analysis From Fig. 4 we see a general weakness in entropy, boundary and smoothness energies in the second row, where they fail to detect errors caused by the area of the foreground prediction being too small - having low DSC due to underprediction. Still, we notice two of the energy terms with unique capabilities: entropy energy seems to be the most accurate match to the errors when compared to the other standalone energies, while regional energy shows strength in highlighting areas that are physically far away from the segmentation, contrary to the general weakness described above. However, we see that it highlights too much of the background as the intensity difference between the foreground and the background decreases. Overall, such findings explains the

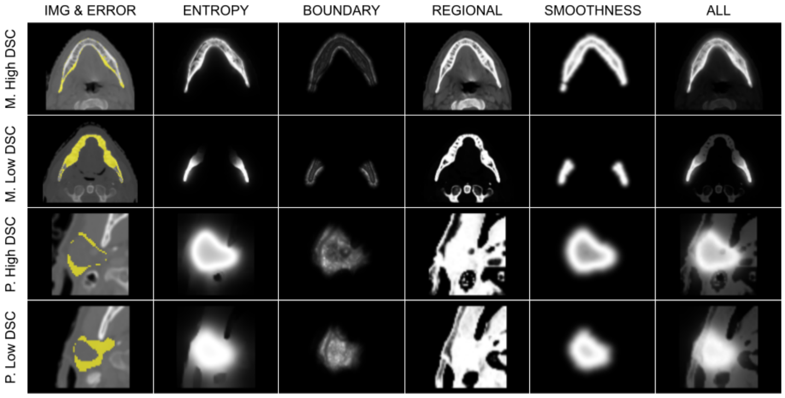


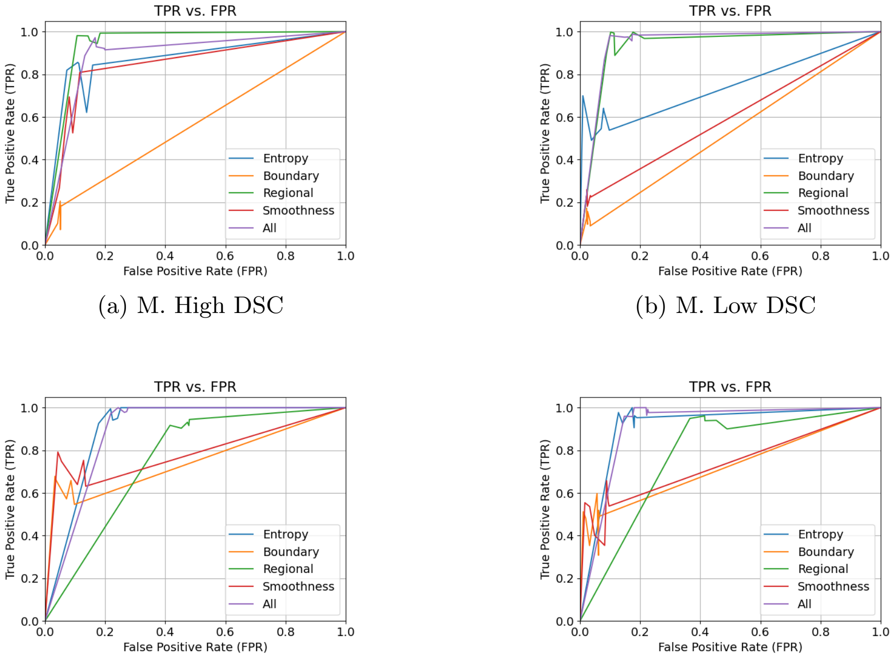
Fig. 4: Example snapshots of uncertainty fields with different baseline energy combination, where the leftmost column shows the original image along with the segmentation errors marked in orange. The intensity of the grayscale images on the other columns represent the uncertainty magnitude.

high weight of entropy energy and moderate weight in regional energy for equation (1) from the implementation details of [2], as it uses regional energy to add extra information to the seemingly best energy: the entropy energy. The rightmost column where it sums up all energies similarity to such strategy seems to show the best performance, which is to be verified through quantitative analysis below.

Quantitative Analysis Table 2 shows the average results for each data category and uncertainty combination. True positive rates (TPR) represent the ratio between voxels estimated correctly as erroneous from the uncertainty field and all voxels proven to be erroneous through comparison with the ground truth, while false positive rates (FPR) represent the ratio between voxels incorrectly estimated as erroneous and all voxels proven to be correct. Mutual information (MI) represent the correlation between uncertainty field strengths and actual error. Along with the ROC curves in Fig. 5, we verify all observations made from the qualitative analysis process, especially that the combination of all energies shows good results throughout all data categories, on both the thresholded and non-thresholded evaluation.

Table 2: Evaluation of different baseline uncertainty combination

Data	M. High DSC			M. Low DSC			P. High Dsc			P. Low DSC		
Metric	TPR↑	FPR↓	MI↑	TPR↑	FPR↓	MI↑	TPR↑	FPR↓	MI↑	TPR↑	FPR↓	MI↑
						Mean						
Entropy	0.80	0.12	0.04	0.58	0.06	0.05	0.96	0.22	0.11	0.94	0.16	0.12
Boundary	0.15	**0.05**	0.01	0.12	**0.03**	0.01	0.62	**0.07**	0.09	0.47	**0.04**	0.07
Regional	**0.97**	0.15	**0.06**	**0.96**	0.14	**0.09**	0.92	0.46	0.05	0.93	0.44	0.05
Smooth	0.57	0.08	0.03	0.22	**0.03**	0.01	0.71	0.09	0.09	0.51	0.07	0.07
All	0.92	0.17	**0.06**	**0.96**	0.15	**0.09**	**0.99**	0.25	**0.13**	**0.98**	0.20	**0.14**
					Std. Deviation							
Entropy	0.09	0.03	0.01	0.07	0.03	0.02	0.03	0.03	0.04	0.05	0.03	0.03
Boundary	0.05	0.00	0.01	0.03	0.00	0.00	0.05	0.03	0.01	0.09	0.02	0.01
Regional	0.02	0.03	0.01	0.05	0.04	0.01	0.01	0.02	0.02	0.02	0.04	0.03
Smooth	0.18	0.02	0.02	0.02	0.01	0.00	0.06	0.04	0.01	0.10	0.03	0.01
All	0.03	0.02	0.01	0.04	0.04	0.02	0.01	0.02	0.04	0.02	0.03	0.04



(a) M. High DSC

(b) M. Low DSC

(c) P. High DSC

(d) P. Low DSC

Fig. 5: ROC curves comparing true positive rates and false positive rates of different baseline uncertainty combinations, for each data categories. Each point represents a data instance.

3.3 Ensemble Uncertainty Results

Now we compare the uncertainty field created from the ensemble method with the combined baseline uncertainty field from above (last row from Table 1), using the same data categories. Fig. 6 shows snapshots of the resulting uncertainty fields.

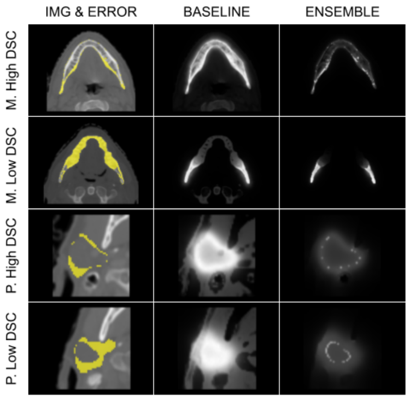


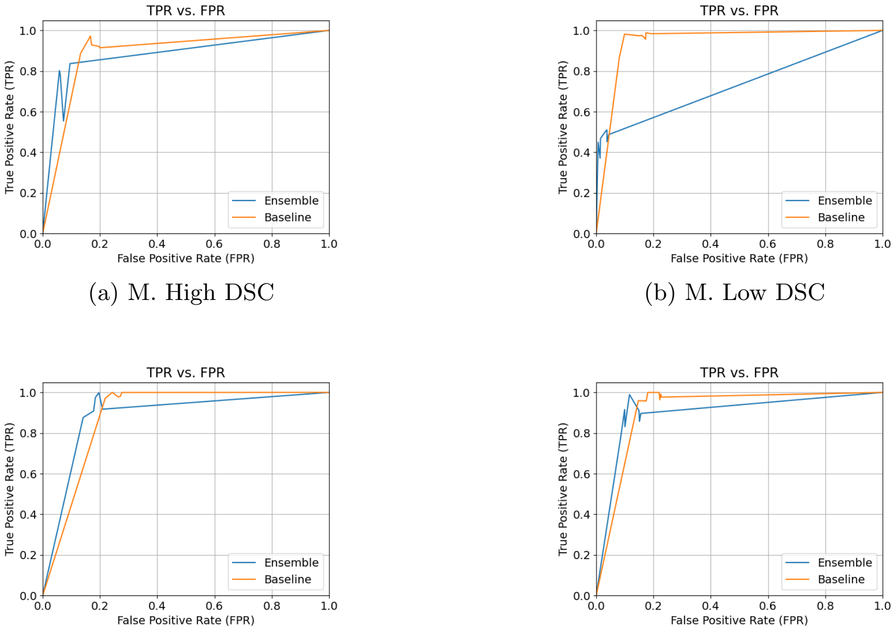
Fig. 6: Example snapshots of uncertainty fields created by the ensemble method and the combined baseline method, where the leftmost column shows the original image alone with the segmentation errors marked in orange. The intensity of the grayscale images on the other columns represent the uncertainty magnitude.

Qualitative Analysis From Fig. 6 we see that the ensemble uncertainty field emphasizes parts of the classification boundaries, as they are the regions where ensemble predictions differ the most. Consequently, we see that it matches the shape of the error images for segmentations with high DSC much better than the ones with low DSC, while the combined baseline method exhibit decent performance in all DSC levels.

Quantitative Analysis Similarly to Table 2, Table 3 shows the average results for each data category and uncertainty field types. Comparing TPR and FPR in Fig. 7 (a thresholded evaluation), we see that the ensemble uncertainty has slightly lower TPR than the baseline uncertainty concerning the segmentation of the mandible (which has bigger image size and more apparent edges than the parotid gland), where the difference stands out even more in low DSC segmentations. The non-thresholded mutual information evaluation also shows similar patterns. Overall, we verify our prediction from the qualitative analysis

Table 3: Evaluation of ensemble uncertainty vs. baseline

Data	M. High DSC			M. Low DSC			P. High Dsc			P. Low DSC		
Metric	TPR↑	FPR↓	MI↑	TPR↑	FPR↓	MI↑	TPR↑	FPR↓	MI↑	TPR↑	FPR↓	MI↑
					Mean							
Baseline	**0.92**	0.17	**0.06**	**0.96**	0.15	**0.09**	**0.99**	0.25	**0.13**	**0.98**	0.20	**0.14**
Ensemble	0.75	**0.07**	0.05	0.46	**0.03**	0.05	0.94	**0.18**	0.11	0.88	**0.12**	0.12
				Std. Deviation								
Baseline	0.03	0.02	0.01	0.04	0.04	0.02	0.01	0.02	0.04	0.02	0.03	0.04
Ensemble	0.10	0.01	0.02	0.04	0.01	0.01	0.05	0.02	0.04	0.06	0.03	0.04



(a) M. High DSC

(b) M. Low DSC

(c) P. High DSC

(d) P. Low DSC

Fig. 7: ROC curves comparing true positive rates and false positive rates of different methods to obtain uncertainty fields, for each data categories. Each point represents a data instance.

that the ensemble method performs better in higher DSC, but not better than the baseline uncertainty in our experiments. However, in all cases the FPR of the ensemble uncertainty is less than those of the baseline, meaning less of the correctly labeled voxels are being marked uncertain.

4 Discussion

Data The data we use for the evaluation of uncertainty fields only consists of CT scans of small organs in the jaw of the patients. Since the algorithm that generates the fields are set to produce decent results in such organs when implementing, uncertainty estimation for other organs in different images may not be effective. Considering the generation of segmentations, they tend to underpredict due to the characteristics of the RW algorithm, thus most errors occur from predictions being too small rather too big when compared to the ground truth. Additionally, when generating segmentations we provide (simulated) inputs only in planes along the same axis, which may reduce randomness in the creation of predictions and limit observations from different point of views. Finally we lack evaluations using near-perfect segmentations (> 0.9 DSC), missing information that may be significant.

Results We see in the qualitative visual analysis of the baseline uncertainty terms that boundary and smoothness terms seems to not add much information when estimating areas of errors. Entropy energy matches the shape of the erroneous areas the best as a standalone uncertainty, and we verify in the quantitative analysis that adding extra information from the regional energy to entropy energy does increase the performance, likely by enhancing the ability to highlight unlabeled areas in the segmentation that are likely to be classified as foreground. Therefore we also verify the scheme of the combination of different uncertainty terms from [2]. The ensemble method however, while it shows less of the false estimation for correct segmentations, it shows worse performance than the baseline in images that are bigger in size and has a lower DSC. From the pattern it shows, we expect a possibility for its superior performance in high DSC segmentations.

Reproducibility The methods introduced in this paper to generate uncertainty fields are well described and straightforward, thus making the replication of the algorithm itself an easy task. However, the algorithm assumes the RW algorithm as a prerequisite. When using another classifier for predicting segmentations, creation of the voxel-wise classification probability image is necessary. Specifically for the ensemble uncertainty estimation, the classifier require to be manipulated to create an ensemble of predictions.

5 Conclusions

In this paper we investigate into an already existing uncertainty estimation methods from an interactive pipeline to segment 3D medical images that uses an AL framework. We explore its different uncertainty models through isolated evaluations, and justify the combination scheme of the models they use in the AL pipeline. Additionally, we test a new method - the ensemble method - to assess its efficacy at estimating errors when replaced with the original method. Although in this paper the new method failed to prove superiority to the original method, future works may prove its advantages in different circumstances or simply find a better method.

5.1 Future Work

As discussed above, future works to be continued from this paper includes using a wider variety of images and ways to generate segmentations. Particularly, using a different classifier than the RW algorithm may benefit the utility of the ensemble method or other baseline uncertainties, as the new segmentations may exhibit less under-prediction and introduce new dimensions of errors that require detection.

References

1. Milletari, F., Navab, N., & Ahmadi, S. A. (2016, October). V-net: Fully convolutional neural networks for volumetric medical image segmentation. 2016 fourth international conference on 3D vision, *3DV* 565-571. IEEE.
2. Top, A., Hamarneh, G., & Abugharbieh, R. (2011, September). Active learning for interactive 3D image segmentation. In International Conference on Medical Image Computing and Computer-Assisted Intervention, 603-610. Springer, Berlin, Heidelberg.
3. Settles, B. (2010). Active learning literature survey (Tech. Rep. No. 1648). University of Wisconsin-Madison.
4. Zou, K., Chen, Z., Yuan, X., Shen, X., Wang, M., & Fu, H. (2023). A Review of Uncertainty Estimation and Its Application in Medical Imaging, 2–3.
5. Sensoy, M., Kaplan, L., & Kandemir, M. (2018). Evidential deep learning to quantify classification uncertainty. Proceedings of the 32nd International Conference on Neural Information Processing Systems, 3183–3193.
6. Maddox, W. J., Izmailov, P., Garipov, T., Vetrov, D. P., & Wilson, A. G. (2019). A simple baseline for bayesian uncertainty in deep learning. Advances in Neural Information Processing Systems, *32*.
7. Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. Advances in Neural Information Processing Systems, *30*.
8. Wang, G., Li, W., Aertsen, M., Deprest, J., Ourselin, S., & Vercauteren, T. (2019). Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. Neurocomputing, *338*, 34–45.
9. Grady, L. (2006). Random walks for image segmentation. IEEE transactions on pattern analysis and machine intelligence, *28(11)*, 1768-1783.
10. Judge, T., Bernard, O., Porumb, M., Chartsias, A., Beqiri, A., & Jodoin, P.-M. (2022). CRISP - Reliable Uncertainty Estimation for Medical Image Segmentation, 5–11.
11. Raudaschl, P. F., Zaffino, P., Sharp, G. C., Spadea, M. F., Chen, A., Dawant, B. M., . . . & Jung, F. (2017). Evaluation of segmentation methods on head and neck CT: Auto-segmentation challenge 2015. Medical Physics, *44(5)*, 2020-2036.
12. Zou, K. H., Warfield, S. K., Bharatha, A., Tempany, C. M., Kaus, M. R., Haker, S. J., Wells, W. M., 3rd, Jolesz, F. A., & Kikinis, R. (2004). Statistical validation of image segmentation quality based on a spatial overlap index. Academic radiology, *11(2)*, 178–189.