Estimating Uncertainty in nnU-Net for Cardiac MRI Segmentation

Daniel Katz, Natalie Mendelson

Introduction

The goal of this project is to address the lack of uncertainty estimation in nnU-Net, a powerful image segmentation model (Fabian Isensee, 2020). Uncertainty estimation is crucial to identify potential failures or limitations of nnU-Net. In this report, we present a method to estimate nnU-Net uncertainty for cardiac MRI scans. The motivation behind this work stems from the need for affective identifying and eliminating poorly segmented images or regions within an image thus, having reliable and confident segmentation results in cardiac imaging, where accuracy and uncertainty play significant roles.

Previous work in uncertainty estimation for deep learning models has predominantly focused on probabilistic methods such as Bayesian inference and Monte Carlo dropout., This project explores an alternative approach using cyclic learning rate, ensemble predictions, entropy and statistical calculation to estimate uncertainty in nnU-Net.

Methods

Data set

Images: Contains cardiac T1-weighted images for 201 patients, 5 slices per patient and 11 T1-weighted images per slice (Hossam El-Rewaidy, 2018). Dataset is publicly available here.
Labels: Manual contours for Epi and Endocardial contours are provided for each T1- weighted image. Total of ~11.5K images and labels.

nnU-Net

nnU-Net is a deep learning-based segmentation method that automatically configures itself, including preprocessing, network architecture, training, and post-processing for any new task in the biomedical domain. This automatic configuration faced us with new challenges with approaching the core of the model and understanding its complicated work.
According to Isensee, Fabian, et al. "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation." *Nature methods* 18.2 (2021): 203-211. , givena new segmentation task, dataset properties are extracted in the form of a 'dataset fingerprint'. Then, a set of heuristic rules models parameter interdependencies and operates on this fingerprint to infer the data-dependent 'rule-based parameters' of the pipeline. These are complemented by 'fixed parameters', which are predefined and do not require adaptation. Up to three configurations are trained in a five-fold cross-validation.
Finally, nnU-Net automatically performs empirical selection of the optimal ensemble of these models and determines whether post-processing is required.

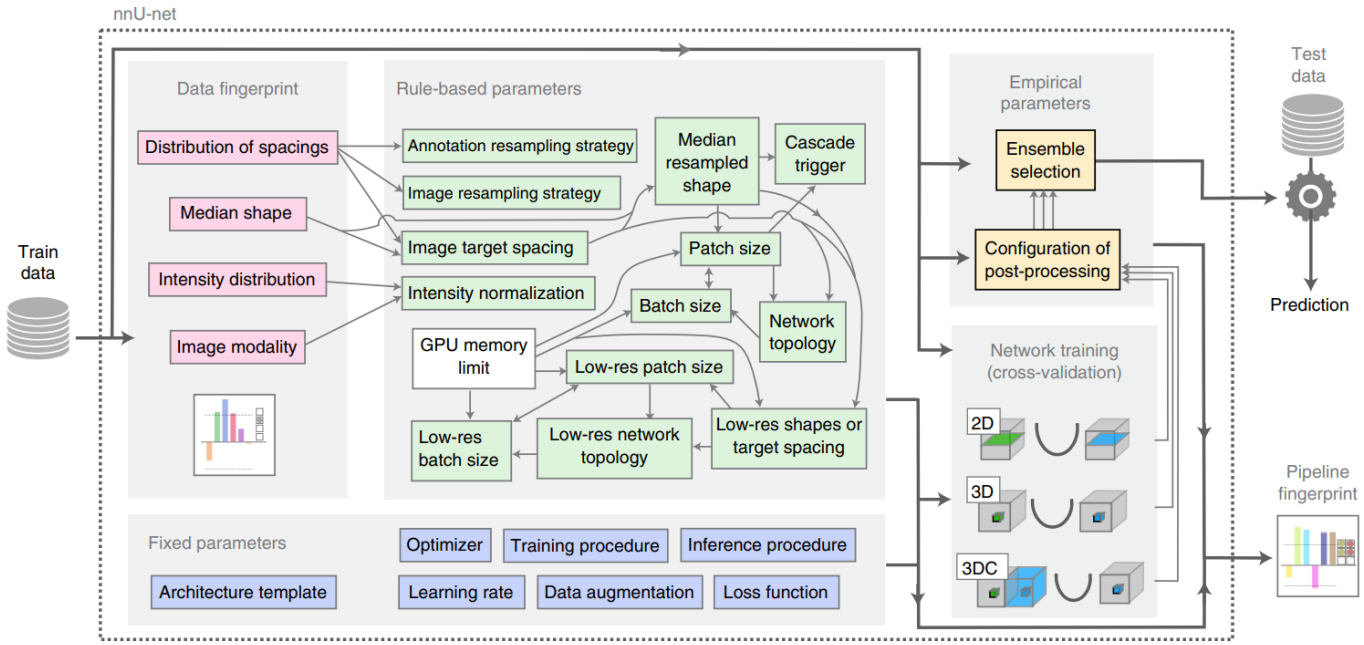


figure 1: Automated scheme configuration of nnUnet.

Approach

Our research was based on the publication titled "Efficient Bayesian Uncertainty Estimation for nnU-Net" by Zhao, Yidong, et al., presented at the 25th International Conference on Medical Image Computing and Computer Assisted Intervention–MICCAI 2022 in Singapore. The proceedings were published by Springer Nature Switzerland in 2022.

In our approach, we employed an algorithmic method that involved modifying the learning rate in a cyclic manner during training. This modification allowed the model to converge to multiple minima instead of a single minimum. In the original nnUnet, the training was performed for 1000 epochs, but we extended this duration and defined 4 cycles.

$$\boldsymbol{T}_{num\ of\ total\ epoches} = 1200$$
$$\boldsymbol{M}_{num\ of\ cycles} = 4$$
$$\boldsymbol{T_c} = \frac{T}{M}\,, \qquad \boldsymbol{t_c} = t \bmod T_c$$
$$\boldsymbol{lr(t)} = \begin{cases} \alpha_r & ,t_c = 0 \\ \alpha_0 \left[1 - \dfrac{\min(t_c, \gamma T_c)}{T}\right] & ,t_c > 0 \end{cases}$$
$$we\ defined:$$
$$\boldsymbol{\alpha_r} = 0.1\,, \alpha_0 = 0.01\,, \gamma = 0.8$$

The main idea behind our approach was to extract different sets of checkpoints from each minimum and examine whether the model provided consistent predictions or showed uncertainty. By comparing the predictions across these checkpoints, we aimed to determine if the model agreed on a particular outcome or if it expressed uncertainty.

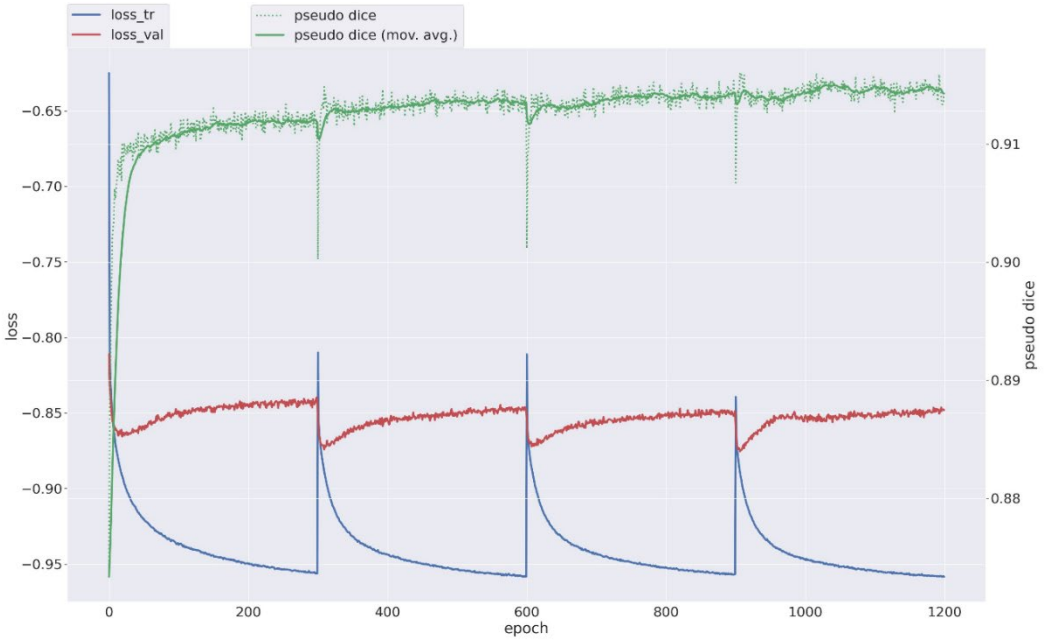


figure 2: showing the cyclic lr application – alternating dice, training loss and validation loss.

Ensembling –

We extracted multiple checkpoints from the nnU-Net model for each minimum. Specifically, we selected the last 10 checkpoints from each cycle. With these checkpoints loaded, we explored two different approaches for inference.

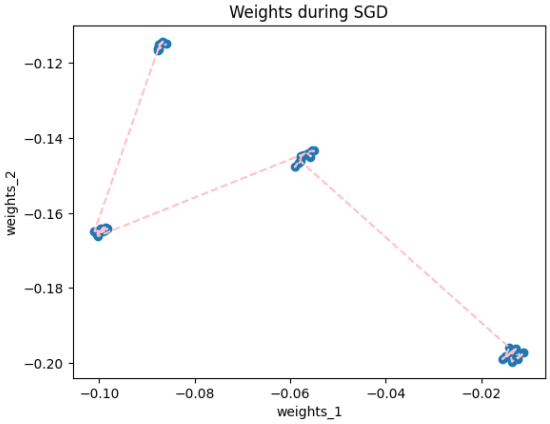


Figure 3 : Example of weights location during training with cyclic learning rate.

1) **SGD Bayesian Inference**

To estimate the probability of the test output $\boldsymbol{y_{test}}$ given the test input $\boldsymbol{x_{test}}$ and the dataset $\boldsymbol{D}$, we approximated it as the average of individual checkpoint probabilities:

$$(\mathbf{1})\ \ \boldsymbol{p(y_{test}|x_{test}, D)} \approx \frac{\mathbf{1}}{\boldsymbol{n}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{p(y_{test}|x_{test}, w_{ti})}\ \ ; \boldsymbol{n - number\ of\ checkpoints}$$

$$\boldsymbol{w_{ti}} \in \boldsymbol{W}\ , \boldsymbol{such\ that\ W} = \{\boldsymbol{w_t}|\boldsymbol{\gamma T_c} < \boldsymbol{t}\ \leq \boldsymbol{T_c}\}$$

By applying this approach, we obtained two probability maps: one representing the mean probability of each pixel belonging to the foreground (class 1), and the other indicating the probabilities of each pixel being part of the background (class 2). (Note: This output reflects the nn-Unet classic method, as it was initially designed as a multiclass model.)

Entropy Calculation –

After obtaining the mean probability maps through ensembling (by considering all saved checkpoints), we calculated the entropy value for each pixel using the following formula:

$$(2)\ \boldsymbol{H\left(y_{test}^{ij}\right) = -\sum_{k=1}^{C} p(y_{test}^{ij} = k|\, x_{test}, D) \log_2 p(y_{test}^{ij} = k|\, x_{test}, D)}$$

$$\boldsymbol{such\ that\ C - number\ of\ classes}$$

This calculation yielded an entropy value for each pixel. Pixels with high confidence were represented by darker shades (approaching zero), while those with uncertainty were colored and more noticeable.

Normalization:

To obtain a single uncertainty score per image, we required normalization of the total entropy value. The regular mean approach was not suitable for us due to varying segmentation sizes across different images. As a solution, we generated a dilated image and subtracted it from the original image, isolating only the segmentation contours. We then normalized the sum of entropy values by considering the area of these contours.

Let –

$(3)\ \boldsymbol{p_{class\,1}(y_{test}|x_{test}, D) \approx \frac{1}{n}\sum_{i=1}^{n} p_{class\,1}(y_{test}|x_{test}, w_{ti})}$ - be the mean image of foreground probabilities.

We defined a $threshold = 0.5$ and converted it to a binary image I.

$(4)\ \boldsymbol{H_{total} = \sum_i \sum_j H(y_{test})_{ij}}$ represent the total entropy value.

$(5)\ \boldsymbol{I^* = I \oplus A}$; $\boldsymbol{A\ is\ a\ structing\ element}$.

$(6)\ \boldsymbol{I_{normaliztion} = I - I^*}$

$$(7)\ \boldsymbol{Uncertainty\ score = \frac{H_{total}}{\sum_{i,j} I_{normaliztion}}}$$

This normalization process ensured that the uncertainty score provided a consistent measure, accounting for the varying sizes of segmentations across different images.

2) Checkpoints total Entropy

As an alternative approach, we explored the calculation of entropy using 40 different checkpoints, focusing solely on the segmentation probability map obtained from each checkpoint. The entropy calculation is defined as follows:

$$(8)\ \boldsymbol{H\left(y_{test}^{ij}\right) = -\sum_{k=1}^{n} p_{class1}(y_{test}^{ij} = k|\, x_{test}, D) \log_2 p_{class1}(y_{test}^{ij} = k|\, x_{test}, D)}$$

$$n - total\ num.of\ checkpoints$$

This method allows us to assess the level of agreement among the checkpoints regarding the segmentation alone. Bright pixels indicate areas where the checkpoints exhibit disagreement, while dark pixels indicate areas of agreement.

To obtain the uncertainty score, we followed the same normalization procedure as previously explained (Normalization). This ensured a consistent scaling of the output, enabling a meaningful quantification of uncertainty.

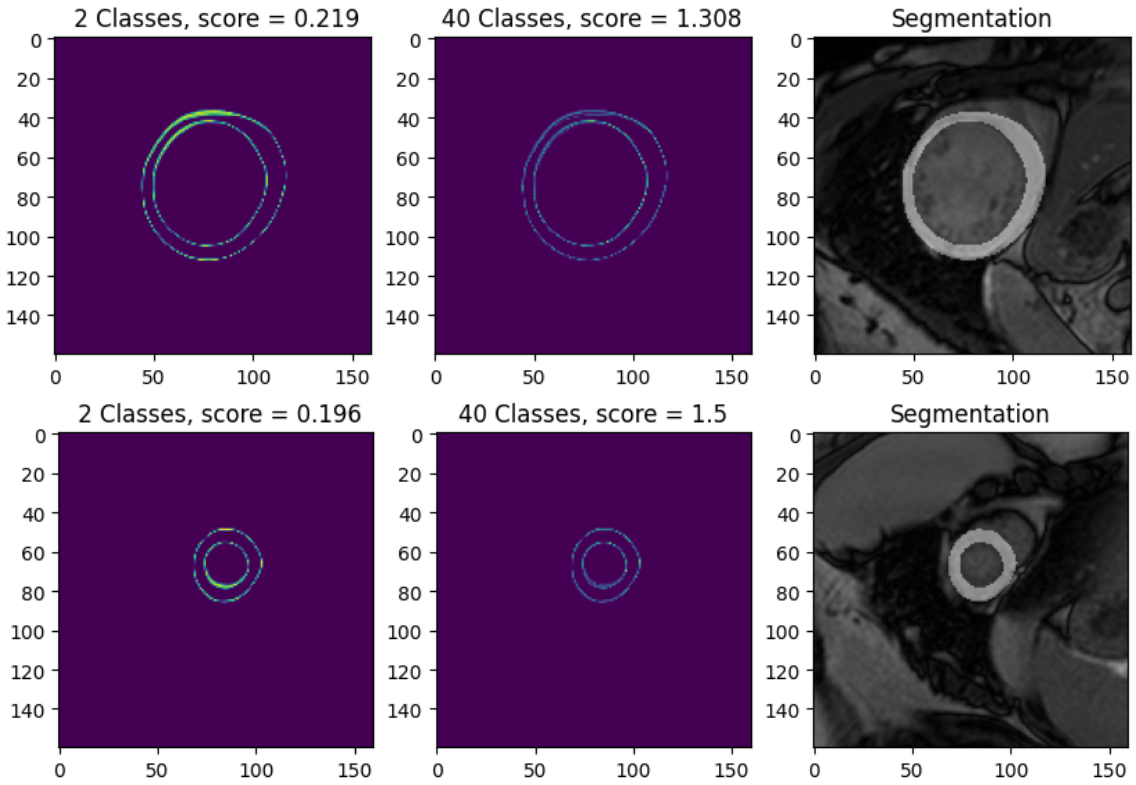
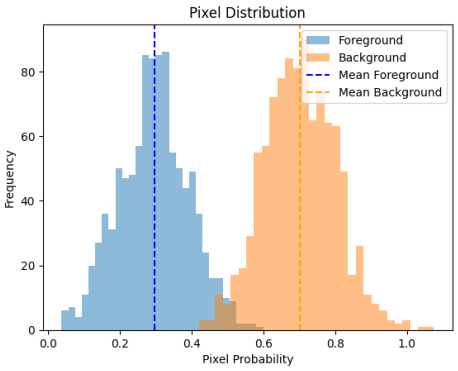


Figure 4 :

Examples of Uncertainty Maps using the 1st and 2nd methods. The highlighted points represent pixels with uncertainty, while the purple pixels indicate areas where segmentation is certain.

3) T-test statistic measure

In our final approach, we employed the T-test on the distribution of each pixel derived from all checkpoints. The underlying null hypothesis assumed that the foreground and background originated from the same distribution, indicating uncertainty in the pixel's classification. Conversely, if a difference was observed in the distribution of foreground and background, it suggested that the pixel came from a distinct distribution, implying certainty in its classification.



$$f_1 - the\ distribution\ of\ each\ pixel\ to\ be\ in\ class\ 1\ (foreground)$$

$$f_2 - the\ distribution\ of\ each\ pixel\ to\ be\ in\ class\ 2\ (background)$$

$$H_0 : \mu_1 = \mu_2\ , H_1 : \mu_1 \neq \mu_2$$

We calculated the P-values for each pixel using the T-test and identified any P-values exceeding the significance level of 5%. These significant P-values were highlighted in the resulting image, which served as the uncertainty map in this case.

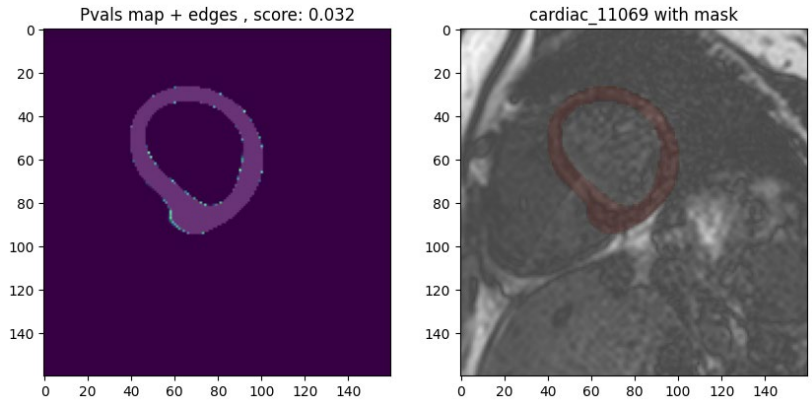


Figure 5 : Demonstrating an example of uncertain pixels using T-test

To obtain the uncertainty score, we followed the same normalization procedure as previously explained.

Results

Training: Besides the uncertainty measure we have achieved better results on dice and validation loss using the cyclic training and enlarged number of epochs.

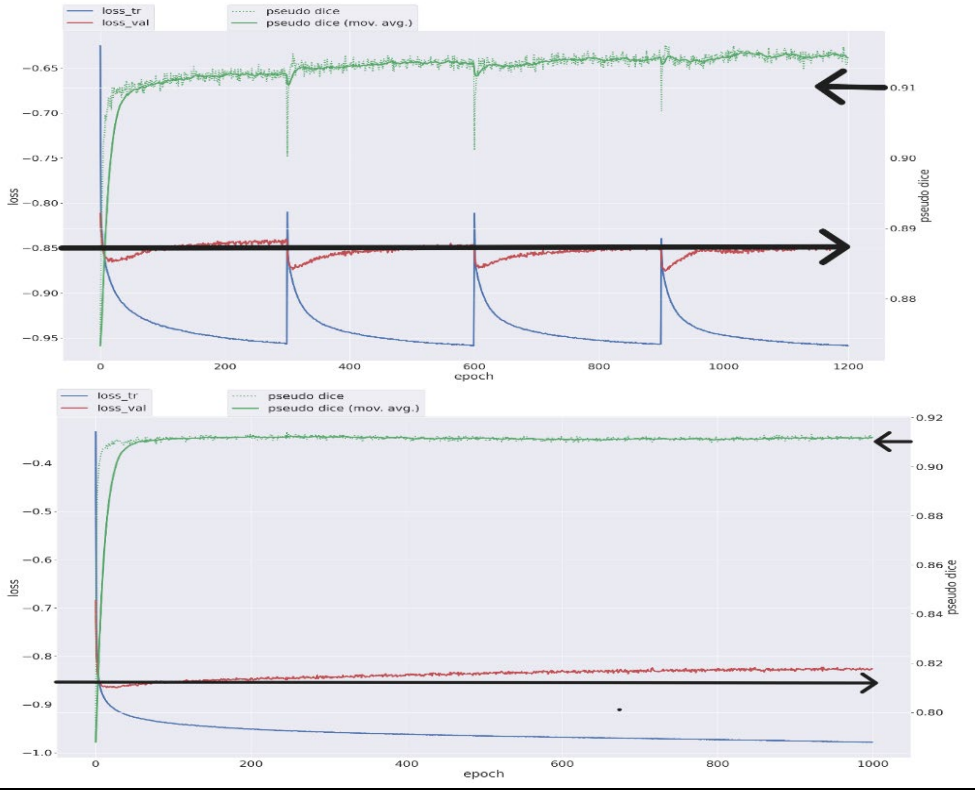


Figure 6: demonstrates the enhancements achieved, with the dice score exceeding 91% in our modified model (indicated by the upper arrows), while it remained below this threshold in the original model. Furthermore, the validation loss in the cyclic model has been reduced to a value smaller than -0.85, contrasting with the higher value observed in the original model (lower arrow).

Inference:

To validate the confidence of our methods and pick the best approach of those written above, the uncertainty scores obtained using the proposed method were compared to the segmentation accuracy measured by the Dice coefficient of the model prediction. A correlation analysis between the uncertainty scores and the Dice coefficient was performed.

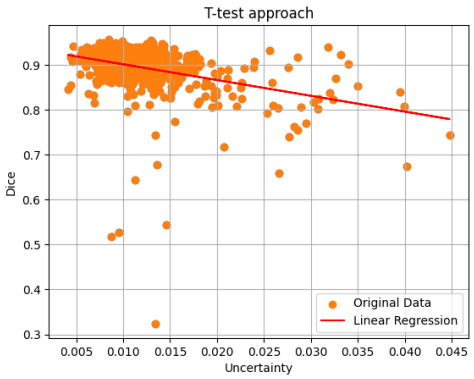


Figure 7:
Slope:-3.508
Intercept: 0.936
R-squared: 0.127
p-value: 7.5 e-18
Standard error: 0.394

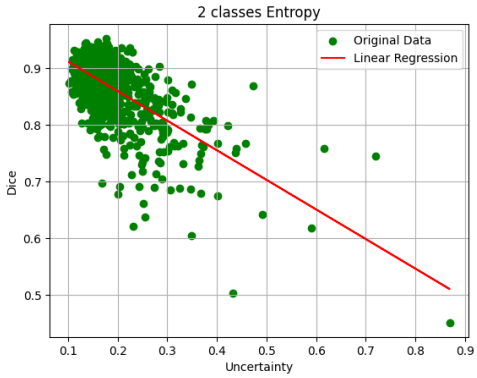


Figure 8:
Slope:-0.523
Intercept: 0.964
R-squared: 0.403
p-value: 9.58 e-114
Standard error: 0.02

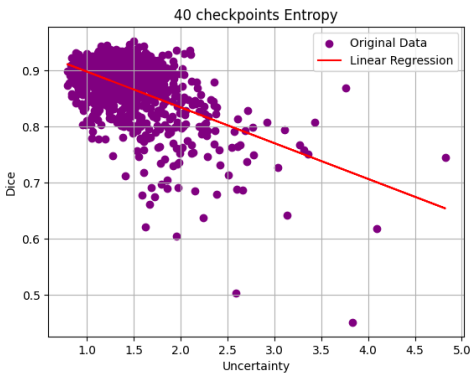


Figure 9:
Slope:-0.064
Intercept: 0.962
R-squared: 0.254
p-value: 2.01 e-65
Standard error: 0.003

As can be noticed above, although we got relatively low and insignificant R^2 for all methods. The best method to get uncertainty measure is according to (1) SGD Bayesian Inference – 2 classes Entropy. The results demonstrated a correlation between higher uncertainty scores and lower Dice coefficients, indicating that the estimated uncertainty aligns with segmentation accuracy. We used linear regression for estimating the correlation.

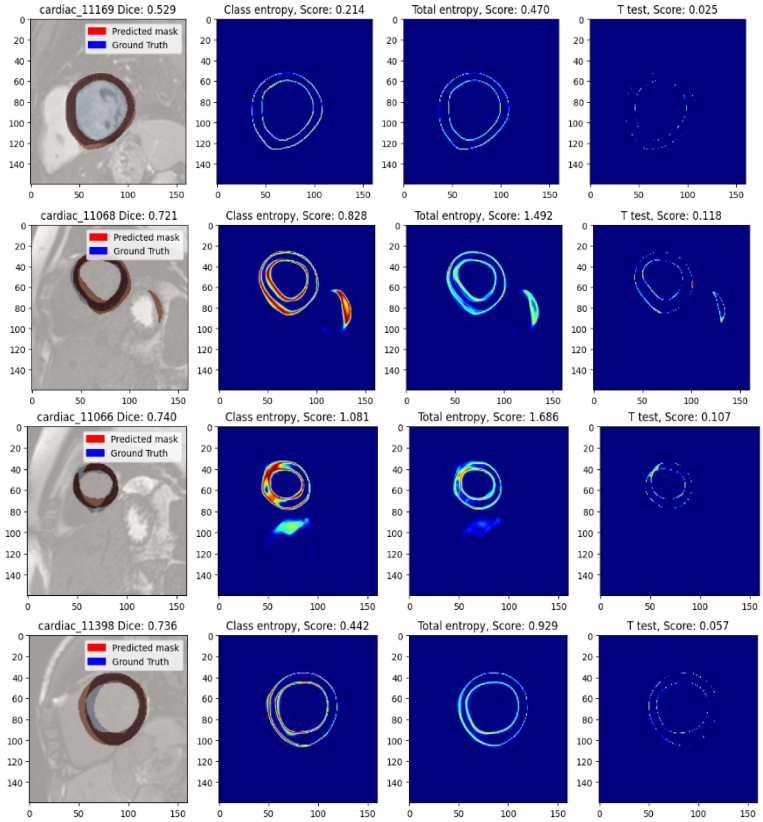
Where does the uncertainty metric fail to predict model performance?

We plotted some of the high uncertainty / low dice results (images are on next page) to see the cases where our method fails/ succeeds. Overall, we see that it works when the prediction is 'on the right track, but not there yet' or 'somehow right but not exactly', but it doesn't work when the prediction is completely wrong.

We can see in cardiac_11169 that's we had a total failure to predict, but the uncertainty is not high. We see this is a very unusual mask for model, the very apex of the heart. This part of our data is different (whole circle), and we think it doesn't have enough representation in the dataset. Augmentation might help. This is why all checkpoints failed in predicting it thus giving low uncertainty. On the other hand, in cardiac_11068 we can see a bad segmentation and very high uncertainty. Especially we see the extra bit that is predicted and gets high uncertainty on the maps.

In cardiac_11066 we see another example where the uncertainty map gives us valuable information. The prediction. We can see exactly where on the heart wall the prediction fails. Here is a good example of what we are looking to prevent in our work.

In cardiac 11398 we see another example of 'about the right prediction but not exactly'. Uncertainty is high and areas of bad prediction stand out.



Conclusion and Future Work

In conclusion, this project presents a novel method for estimating uncertainty in nnU-Net for cardiac MRI segmentation. By leveraging cyclic learning rate, ensemble predictions, and entropy calculation, the proposed method provides a measure of uncertainty that can indicate potential failures or limitations of nnU-Net in large-scale image segmentation tasks.

Future work can explore further enhancements to the uncertainty estimation method. This could involve investigating alternative ensemble techniques; for example - weighted ensemble with uncertainty, exploring different entropy calculations, different normalization or incorporating additional sources of uncertainty such as model heterogeneity or mask volume and mean distance. The applicability of the proposed method can also be extended to other medical imaging tasks or domains where uncertainty estimation is crucial for reliable decision-making.

Overall, this research contributes to improving the reliability and confidence in nnU-Net's segmentation results, enabling more informed decision-making in cardiac MRI analysis, and potentially benefiting a wide range of image segmentation applications.

References

[1] Isensee, Fabian, et al. "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation." *Nature methods* 18.2 (2021): 203-211.

[2] Zhao, Yidong, et al. "Efficient Bayesian Uncertainty Estimation for nnU-Net." *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part VIII*. Cham: Springer Nature Switzerland, 2022.

[3] Duy, Vo Nguyen Le, Shogo Iwazaki, and Ichiro Takeuchi. "Quantifying statistical significance of neural network-based image segmentation by selective inference." *Advances in Neural Information Processing Systems* 35 (2022): 31627-31639.

[4] Hossam El-Rewaidy, Maryam Nezafat, Jihye Jang, Shiro Nakamori, Ahmed S. Fahmy, and Reza Nezafat. "Nonrigid active shape model–based registration framework for motion correction of cardiac T1 mapping." Magnetic resonance in medicine (2018), doi: 10.1002/mrm.27068