Universal Lesion Segmentation: Identification of difficult cases

Max de Boer-Blazdell
max.deboer-blazdell@ru.nl
s1055222

Fenna Rasing
fenna.rasing@ru.nl
s1033690

Mats Robben
mats.robben@ru.nl
s1054883

Lucia Rust
lucia.rust@ru.nl
S1053676

***Abstract*—With the global cancer burden growing over the years, the workload of radiologists also increases. The rise of AI in the medical field includes automatic segmentation models, which have been adopted to help save time for radiologists. However, there is still a need for models with better accuracy, especially for rare or difficult to segment lesion types. The Diagnostic Image Analysis Group at Radboud University has launched the ULS23 challenge for this purpose. They developed a baseline model using nnU-Net and composed a dataset of a combination of fully- and partially annotated lesions for a wide range of lesion types. This project attempts to analyse the more difficult cases, by literature research, participating in a reader study and analysing the results of the baseline model. It is found that the more difficult cases mainly comprise of smaller lesions that blend in more with the background.**

***Index Terms*—nnU-Net, Universal Lesion Segmentation**

I. Introduction

It is predicted that the global cancer burden will grow significantly [11], leading to an increased workload for radiologists. This workload comprises of CT examinations, which are used for diagnosis, and monitoring during and after cancer treatment [5]. To navigate this workload successfully, there is an interest in the application of computer assistance, particularly through automatic segmentation models. These models aim to improve the inter-observer variability, which is associated with manual lesion annotations [9]. Traditionally, automatic segmentation of lesions was done with, among other methods, thresholding, region growing, and heuristic edge detection algorithms. Recently, the use of deep learning has been dominating this field, where U-Net [8] in particular has made some significant improvements to the traditional methods [1]. As input for the deep learning segmentation algorithms, the volume-of-interest (VOI) of the image where the lesion exists is used. Lesion selection happens by single-click selection, detection models, or bounding-box annotation. Given the selected lesions, the AI-driven segmentation models will provide 3D segmentation masks, which will save time and help with further analyses for the radiologists [9].

While AI-driven automatic tumor segmentation models have made significant improvements, particularly for lesion types that are more clearly visible like those found in the liver, kidney, or lungs, there remains a need for comprehensive models for other lesion types, which are frequently missed in clinical practice [7]. Part of the reason is that some lesion types are under-represented in the dataset, which could be because (1) they are very rare, or (2) they are hard to detect. Additionally, these lesions must first be identified and annotated by radiologists, which means that they should be able to detect them. However, as Souza et al. citesouza2005volume have shown, small lesions are often hard to detect, in part because a partial volume effect (where structures may not be visible because the resolution is too low) occurs, and hence it is difficult to say where a lesion starts and ends [10]. Identifying such cases could be a step towards improving overall model performance.

The evolution of Universal Lesion Segmentation (ULS) models requires a diverse training dataset. However, many current models rely on datasets with limited annotations and accessibility of ground-truth segmentation masks, which hinders reproducibility [6]. Furthermore, most ULS models are not publicly released, limiting integration into clinical workflows or validation by other researchers.

The ULS23 challenge has been launched by the Diagnostic Image Analysis Group at the Radboud University in order to encourage development of innovative models that can handle more diverse lesion types and encourage more publicly available models and data in this area. The training dataset provided by the ULS23 challenge contains a combination of 3D-segmented and partially 2D-annotated lesions to enhance the generalizability. The partially annotated dataset contains lesions of a wider variety, some of which are more difficult to segment. The test and validation datasets were chosen to contain lesions of a diverse range to ensure the applicability of the models. The challenge is evaluated based on the segmentation performance, inference speed and segmentation consistency. The baseline provided by the ULS23 challenge uses the nnU-Net Framework [4]. This method is a convolutional network that encodes and decodes information. In order to use both the fully and partially annotated data for training, the model was trained on predicted 3D pseudo-masks for the partially annotated data, making it a mixed-supervision model [3].

In this paper, we present and advise on modifications to the baseline model. As part of our project, we participated in a reader study where we were asked to fully segment lesions in 3D in the Thorax-Abdomen area. Using experience from this reader study, literature research, and results from the trained baseline model, we found the cases where the baseline model performs significantly worse. Based on these difficult cases we will advised on modifications to the baseline model, where we hypothesize that emphasizing on the difficult cases in training of the baseline model might improve performance.

II. Methods and Materials

In this section, we will explain the data used in this project, the baseline model of the ULS23 challenge, the reader study we performed and the experiments conducted in this project. The code used for the analysis and additional material can be found at https://github.com/MatsRobben/ULS_AIMI_Group12.

A. Data

The training data used for this project is the 3D novel fully annotated dataset provided by the ULS23 challenge [3]. This thus excludes the partially annotated training data, and part of the fully annotated dataset. We decided to exclude this part of the training dataset because we wanted the model to train relatively quickly and focues on more difficult lesion types. Specifically, our training dataset contains the lesion VOI's of 750 lesions of various types from DeepLesion [12], 744 bone lesions and 124 pancreas lesions provided by the RadboudUMC (see Fig.1. for an overview of the different types of lesions). The DeepLesion data was gathered by 3D segmentation done by medical students in the axial plane using measurement information. The final label was the majority mask. The lesions that are included in this dataset were selected by hard-negative mining with a standard 3D nnUnet. The data from the RadboudUMC were selected from radiological reports, where the final 3D segmentation was provided by an expert radiologist. The bone lesion dataset is included as this type of lesion has different characteristics than lesions from soft-tissue, making it more difficult to distinguish by models trained only on data of soft-tissue lesions. The pancreatic lesion dataset is included because this type of lesion is also often difficult to segment. All data is pre-processed to ensure the VOI's have the right dimensions (256x 256y 128z voxels). If the VOI's did not have the right dimensions, the volumes were padded.

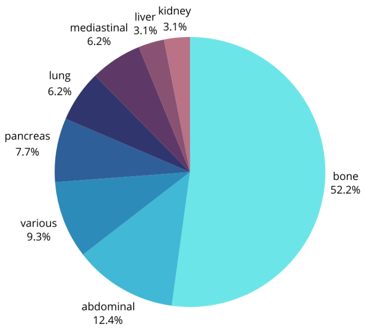


Fig. 1: The ULS23 challenge provided training data with a distribution of annotated lesion types. Medical students annotated the data from the DeepLesion dataset (750 lesions of various types) in the axial plane using measurement information. Expert radiologists annotated the data provided by RadboudUMC (744 bone lesions and 124 pancreas lesions).

The test dataset used for this challenge is comprised of data from the RadboudUMC and the Jeroen Bosch Ziekenhuis. The dataset contains a diverse range of lesion types. These types are mammaries, peritoneum, liver, lung, bones, pancreas, colon, kidney, and more. The segmentation masks were acquired by specifically trained annotators, and corrected by a radiologist. Before the annotators provided the 3D segmentations, an experienced radiologist checked the measurements of the lesion.

B. Baseline

The baseline used in the ULS23 challenge is based on nnU-Net [4]. This network is a self-configurating version of the original U-Net [4]. Both U-Net and nnU-Net were originally designed for biomedical image segmentation, so they are very suitable for the task presented in ULS23. However, U-Net has a disadvantage as it requires expert knowledge to manually tune the model for the task and dataset. In the case of the ULS23 challenge, this means that the model would require separate tuning for every dataset used, which would require a lot of effort to do properly and would be very difficult to optimize. Therefore, no new U-Net (nnU-Net) was chosen for this challenge as it is self-configurating. This means that the nnU-Net tunes the parameters by itself, based on some rules. Another advantage of the self-configuration is its speed. The authors of nnU-Net minimized the amount of empirical choices to be made, so it requires almost no compute resources outside of standard model training, making it a good choice for the ULS23 challenge as limited resources are available. Furthermore, because of the published recipe for self-configuration, the model is more white-box than the average neural network.

C. Reader study

As previously mentioned, there are still a number of lesion types that are difficult to identify [7]. In the reader study [2], we were given a partial CT volume, in which one image would contain a long-axis and short-axis that indicate where the lesion is. Using this box, we were tasked to create a 3D mask of the lesion by annotating the lesion in other slices of the CT volume. Performing this task helped us better understand the problem at hand, given that we are not medical experts and thus had no prior experience with annotating CT scans.

Within this task, we each annotated the same 20 partial CT volumes. For each partial CT volume, the Intersection over Union (IoU) metric was calculated between the annotated images and the provided true segmentations. Next to this, a combined annotation was created, by only labelling pixels as containing a tumor if we all individually also labelled that pixel. This is a similar approach to how the DeepLesion data was annotated. Finally, the IoU between the combined annotation and the true segmentations was calculated.

D. Experiments

To investigate difficult cases, we first re-trained the baseline model [3] on the novel data described in the Data section.

After training, the DICE score was calculated for each image in this set. Then, the images with a lowest DICE score were selected for further inspection. Our goal with this approach is to determine what types of lesions our model struggles with. Then, we would be able to compare this against other models trained for a similar task. For each of these images, the size of the lesion and amount of connected components was recorded. These results were compared against the same objective measurements of the whole validation set to ensure that conclusions would be supported by the whole dataset.

Next, the images were subjectively analysed by visual inspection using ITK-SNAP [13]. The provided true segmentations were compared to the segmentations performed by the trained model. We looked at the location of the true segmentation, whether it blends into the background (similar colour to surrounding structures) and the number of lesions predicted by the model. With both objective and subjective measures, changes to the training pipeline can be made. For example, these difficult types to predict cases could be shown to the model more then once, which could help the model in better predicting these difficult cases. Because these cases represent a very small percentage of the total dataset, we should not run the risk of overfitting on these cases, hence overall performance should increase.

III. Results

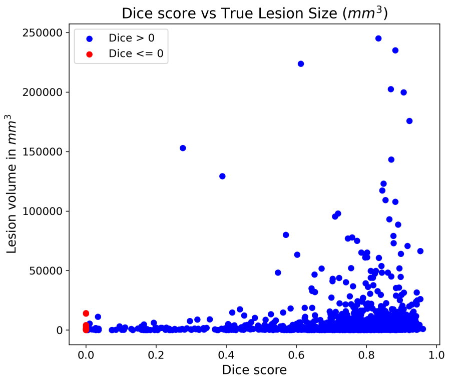


Fig. 2: The DICE score plotted against the true size of the lesion. The lesion size is measured in mm^3.

A. Objective analysis

The objective analysis shows that lesions with a DICE score of 0 have a smaller size and that bigger lesions have a relatively higher DICE score, see Fig.2. Furthermore, the histogram of DICE scores in Fig.3 shows that there are two peaks. The first peak is at zero, these are the lesions where the prediction is completely off. The other peak is around 0.7-0.9. In between these peaks the frequency of lesions (with a DICE score between 0.1 and 0.6) is much lower. Due to the large discrepancy between the groups we decided to further investigate the 100 cases in group one.

Finally, Fig.4 with the connected component analysis shows that in the worst 100 cases, there are multiple cases where the model predicts multiple lesions, while the true number of components is always one. We can also see that of the 100 cases there are about 40 that do not have any components, so no segmentation was given.

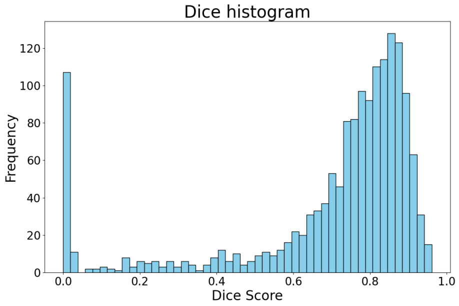


Fig. 3: Histrogram visualizing the frequency of different DICE scores

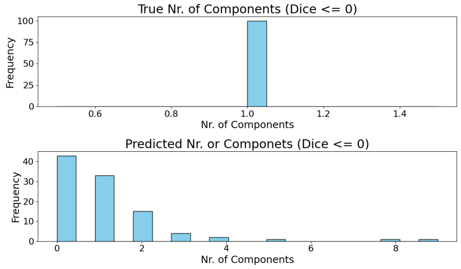


Fig. 4: Histogram visualizing the number of lesions predicted in the 100 worst cases (with a DICE score of 0) vs the number of true lesions.

B. Subjective analysis

In table I, a summary of the subjective analysis of the 100 worst cases is shown. Here, the amount of almost invisible, in-blending, partially in-blending and standing out lesions are shown. We can see that most of these 100 cases are difficult to accurately predict, made even more difficult by their small size. Additionally, in 13 cases, multiple lesions were predicted.

TABLE I: Counts of degree of visibility of the true lesion in the images, subjectively categorized. These counts are performed on the images with a DICE score of 0 (100 total images).

Almost invisible	Blends in	Partially blends in	Stands out
5	54	22	19

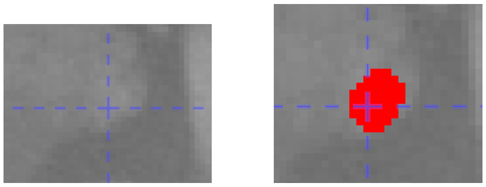
C. Reader study

In table II, the results of the reader study are shown. Here, the individual DICE score, and the DICE score of the combined annotations are shown. For a more detailed description of how these scores are calculated, see section II-C. It is important to note that these scores are calculated over 20 trials (a trial is one partial CT volume containing a lesion), and that all annotators had no previous experience with annotating lesions.

TABLE II: DICE score of individuals compared against provided true segmentations and combined annotations against provided true segmentations. Scores were calculated over 20 trials.

	Fenna	Lucia	Mats	Max	Combined
Average	0.624	0.570	0.617	0.566	0.486
Standard Deviation	0.158	0.152	0.153	0.127	0.153
Maximum	0.818	0.811	0.843	0.754	0.728
Minimum	0.265	0.279	0.294	0.188	0.165

While performing this task, we noted that there were a few cases in which lesions were particularly troublesome to identify. In particular, small lesions were hard to detect and it was often unclear whether they did not belong to some larger structure. Another troublesome case is when a lesion would look nearly identical to surrounding structures, for example as can be seen in figure 5.



(a) Lesion without segmentation mask.

(b) Lesion with segmentation mask (red).

Fig. 5: Example of lesion that blends into the background. The lesion is located in the centre of the blue cross.

IV. Discussion

ULS models usually struggle with rare lesion types, smaller lesions and lesions where it is unclear where they begin or end [10]. In our objective analysis of baseline model results, we discovered that lesions with a DICE score of zero are smaller than most lesions with a higher DICE score. Furthermore, this analysis shows that in some cases the model also predicts multiple lesions, where there should only be one. This may be because there are multiple lesions present in that VOI, but the task was to only segment one. This phenomena was also found in our subjective analysis, where we saw that in some cases the model segmented multiple different lesions. We think future work could focus on this lacking aspect of the current model, by only selecting one lesion during post-processing or by incorporating a penalty in the loss function when selecting multiple lesions. Additionally, our finding of the subjective analysis is that in many predicted lesions with a DICE score of zero, the visibility of the lesion is bad, corresponding to what we found in the literature [10]. Low visibility can also be the reason that the DICE score is either in the zero group or the high (DICE score close to 1) preforming one (Fig.3). The model either finds the lesion and thus does an okay job in segmenting it or it does not find the correct location of the lesion. Our experience with the reader study confirms that smaller lesions are in general more difficult to segment. Results of the reader study also show us that lesion segmentation is a difficult task where there is a lot of variability between annotators. However, this conclusion is made based on 20 annotations made by non-medical students, so it is questionable whether we can generalize this. Unfortunately, we were unable to test our hypothesis of improving performance by focusing on more difficult cases in training. This can be investigated in future work.

There are some limitations related to the data in this project. First of all, we only used part of the available training data due to computational constraints, this means that we are missing some lesion types that are present in the partially annotated and the remaining fully annotated data, but not in the novel data that we use. Future research may explore the effects of using all available training data. Additionally, there exists a class imbalance of lesion types as can be seen in Fig.1. This can lead to decreased performance on the test data set. This class imbalance could be tackled by sampling the training batch such that the ratio of lesion types is more even. Furthermore, the training data is used both for training the model and validating the model. This can lead to over-fitting on the training data and worse performance on the test data. This could have been avoided by training the model with a cross validation scheme, which was not implemented do to computational limitations. Another limitation is on the segmentation masks of the DeepLesion3D dataset. As mentioned, the segmentation was done by medical students, without feedback from a radiologist. There is thus no guarantee that this segmentation is correct, and may need checking by radiologists in future work. Another limitation in our analysis of the worst cases was that we did not know the lesion type of the given cases. This meant that we could not determine whether the worst cases were the more uncommon lesion types. This should be investigated in future studies.

For the reader study we also found a number of limitations and suggestions for future work. The largest oversight was the selected metric, which only considered the voxels for which every team member agreed. This is problematic as only one annotator can lower the overall score significantly. Instead, we propose that in the future we should select metrics that are less extreme, for example we could select the voxels by majority vote.

V. Conclusion

To conclude, we identified the difficult to predict cases of the baseline model used in the ULS23 challenge through a combination of a reader study, literature research and an objective and subjective analysis of the 100 images with the lowest DICE scores. We found, in confirmation with the literature research, that these difficult cases of lesions are relatively small and often blend in with the background, making them difficult to detect visually, as we also experienced during the reader study. Based on these findings, we made some suggestions for future research, such as testing whether emphasizing on the more difficult cases in the training batches would improve performance.

Acknowledgment

We would like to thank Dr. Alessa Hering for her help during the project. In particular, we would like to thank her for critical reflections on how our research was shaped and how the methods were implemented. In addition, we would like to thank group 6 of the AI and Medical Imaging course, and in particular Haoze Zhu, for helping us with the dataset preprocessing and helping us solve a difficult technical problem.

References

[1] Carlos E Cardenas et al. "Advances in auto-segmentation". In: *Seminars in radiation oncology*. Vol. 29. 3. Elsevier. 2019, pp. 185–197.

[2] Grand Challenge. *Universal Lesion Segmentation Training [AIMI]*. https://grand-challenge.org/reader-studies/universal-lesion-segmentation-training-aimi/. Online; accessed 27 May 2024.

[3] M.J.J de Grauw et al. "ULS23: Benchmark and Baseline for 3D Universal Lesion Segmentation in Computed Tomography". In: *Medical Image Analysis* (2024).

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

[5] Robert J McDonald et al. "The effects of changes in utilization and technological advancements of cross-sectional imaging on radiologist workload". In: *Academic radiology* 22.9 (2015), pp. 1191–1198.

[6] Jialin Peng and Ye Wang. "Medical image segmentation with limited supervision: a review of deep network models". In: *IEEE Access* 9 (2021), pp. 36827–36851.

[7] Yu Qiu and Jing Xu. "Delving into Universal Lesion Segmentation: Method, Dataset, and Benchmark". In: *European Conference on Computer Vision*. Springer. 2022, pp. 485–503.

[8] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation". In: *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer. 2015, pp. 234–241.

[9] Jiaqi Shao et al. "Application of U-Net and Optimized Clustering in Medical Image Segmentation: A Review." In: *CMES-Computer Modeling in Engineering & Sciences* 136.3 (2023).

[10] Andre Souza, Jayaram K Udupa, and Punam K Saha. "Volume rendering in the presence of partial volume effects". In: *IEEE transactions on medical imaging* 24.2 (2005), pp. 223–235.

[11] Hyuna Sung et al. "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries". In: *CA: a cancer journal for clinicians* 71.3 (2021), pp. 209–249.

[12] Ke Yan et al. "DeepLesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning". In: *Journal of medical imaging* 5.3 (2018), pp. 036501–036501.

[13] Paul A. Yushkevich et al. "User-Guided 3D Active Contour Segmentation of Anatomical Structures: Significantly Improved Efficiency and Reliability". In: *Neuroimage* 31.3 (2006), pp. 1116–1128.