Development of a segmentation tool to measure subcutaneous implant volumes

Julien Adda
Department of Computer Science, EPFL, Switzerland

Abstract

Segmenting by hand an MRI composed of 850 000 voxels takes 90 minutes for a trained expert. This task is a bottleneck in many phases of clinical research, where many acquisitions need to be made to validate the certainty of their results. The startup Volumina, which is developing a 3D scaffold that is intended to act as a prosthesis in the human body, is required to segment many MRIs. This time-consuming task justifies the need to develop an automatic and robust model capable of segmenting MRIs. This paper aims to present and compare the different segmentation models that were trained to solve Volumina's task. The goal was to segment MRI of mice to reconstruct and calculate the volume of the injected implant. The constraint was to develop a reliable model that would not be affected by the various shapes of the implant nor the contrast attenuation due to time between the implant and its surrounding. Several models were tested, starting from a simple thresholding algorithm and converging to deep learning networks containing millions of parameters. A focus was made to compare look-a-like U-Net models, such as the models proposed by the framework nnU-Net ("no-new-net"). U-Net, is a fully automated, end-to-end neural architecture that has shown great promise in most online segmentation challenges. Many different parameters can be tuned in the U-Net, significantly affecting its accuracy. Hence, it is essential to have an interpretability approach when dealing with these complex architectures. The best model, a 3D nnU-Net, can reconstruct the implant in less than 320 seconds. Its average Dice accuracy is 0.915 on the test set, which translates to an average error of 5.7% for the predicted volume. This acceptable error is more or less constant during the four available timestamps of the data, indicating its robustness.

I. Introduction

60% of breast cancer patients do not reconstruct their breasts after tumor excision because available procedures are either too invasive or do not bring stable results. The startup Volumina is developing a safe and stable scaffold, Adipearl, that would enable the natural repair of 3D soft tissues in one injection. To create this product, it is crucial to determine the evolution of the injected volume with respect to time. Fortunately, Adipearl is easily recognizable in MRI acquisition, as it is brighter than the other elements of the body. This non-invasive technique allows biologists to retrieve the relation between time and the shape of the implant by acquiring several MRIs in time. However, labeling by hand the MRIs is time-consuming, which justifies the need to develop an automatic and robust model capable of segmenting the MRIs.

Several techniques were tested, including simple algorithms such as Multi-Otsu Thresholding and Canny edge detection. A deeper analysis was made by training Conventional Neural Network models based on U-Net [1]. A simple 2D architecture with a Binary cross entropy loss function was compared to a 2D and 3D nnU-Net proposed by the Applied Computer Vision Lab (ACVL) of Helmholtz Imaging [2]. The models were trained to do binary pixel-wise segmentation on MRI .nii files. The labels to predict are binary masks, where the pixels representing the implant are ones and zero otherwise. The output of the models is a binary .nii file, which allows us to compute the volume of the implant.

II. Data Analysis

A. Data: MRIs and Ground Truth

The dataset provided by Volumina SA is composed of 57 MRI T2 weighted images. They were acquired on a 14.1T MRI system at CIBM (located on EPFL's campus in Lausanne, Switzerland). The data are MRIs of the lower back of 15 mice, where Volumina's gel, Adipearl, was injected. An MRI acquisition was made one day, three weeks, three months, and six months after the injections for each mouse. Each MRI, is a .nii file of dimensions (46, 192, 96), with a volume of 0.0142 mm^3 for each voxel. The ground truth is also .nii files of the same dimensions. They contain 1 or 0, indicating the presence of a voxel of Adipearl or the background, respectively.

The ground truth was annotated by hand, slice per slice, by two expert biologists of the Volumina's company. There was little to no difference between their two annotations. Hence, only one was kept.

B. Evolution of the implant with respect to time

Because of the metabolism of the mouse and interns constraints caused by its movement, the shape of the implant tends to evolve a lot with respect to time. In Figure 9, the shape of the implant of the same mouse is shown at different timestamps. We can observe that the canal where the gel was injected is vanishing, and the global shape is getting rounder.

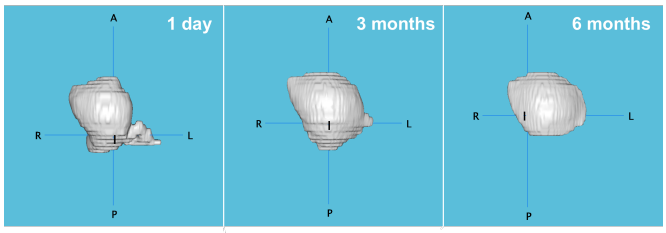

Fig. 1: Evolution of the implant's shape injected in mouse 97-779G. The volume after 1 day, 3 months and 6 months is respectively 431.1, 375.7 and 260.4 mm^3.

There is also a significant variation in the global volume with respect to time. As shown in Figure 2, after three weeks, there is, on average, an increase of 23.8% due to inflammation. Afterward, the volume decreases. Across the 15 mice, the evolution of the volume with respect to time seems to behave in the same way.

C. Train/validation/test set splitting

For each acquisition, Volumina's biologist acquired two different MRIs by varying some parameters of the MRI machine. This leads to a different ground truth because of a difference in the brightness of the implant. This means that we have at our disposal 114 MRIs (2 times 57 acquisitions). The dataset is split into a train, test, and validation set in the respective proportions 53%, 14%, and 33%.

The data is carefully split to avoid bias by respecting an equal distribution of timestamps for the three groups. The duplicates of the

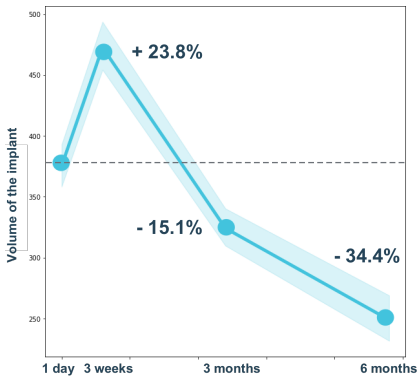

Fig. 2: Average volume of 15 mice with the standard deviation for the timestamp: 1 day, 3 weeks, 3 months and 6 months.

acquisition made on the same mouse and identical timestamps are in the same group. Additionally, even though the shape tends to change with respect to time, there is still a high correlation between an MRI of the same mouse at different timestamps. Hence, each mouse is present in only one of the three groups.

* * *

Highlighting the implant's variation in shape and volume reinforces the importance of studying its evolution, as the injected scaffold is intended to act as a prosthesis. Several techniques were tested and are presented in the following sections to develop an automatic and robust model capable of segmentation the MRIs.

To compare the results, a common metric, called the Dice similarity coefficient, was used to quantify the accuracy of prediction for each model. The metric measures the similarity between two sets of data. The value of a Dice coefficient ranges from 0, indicating no spatial overlap between two sets of binary segmentation results, to 1, indicating complete overlap. The Dice formula, given two sets, A and B (pixels predicted as the implant in an image and the actual implant pixels), is defined as:

$$Dice(A, B) = \frac{2 * |A \cap B|}{|A| + |B|}$$

Additionally, every model presented in this paper was trained and tested on the same data split.

III. Trivial and Thresholding models

A. Trivial model: baseline

As seen previously, the evolution of the volume seems to be equivalent for every mouse. The idea is to take advantage of the fact that the quantity injected into the patient is known at the injection time. The trivial model would then apply the percentage of variation to this volume for specific timestamps determined on the train set. For three weeks, three months, and six months, we have a variation of volume with respect to the initial volume of +23.8%, -15.1%, and -34.4%, respectively.

This model is not robust and not scalable. It is only used as a baseline to be compared with future models.

B. Thresholding model: MTOMO

Pixels labeled as the injected gel are 82.5% brighter than the average pixel, and for most images, the implant is a compact convex shape. Several techniques were used to take advantage of these properties.

A thresholding algorithm, called Multi-Otsu threshold, was used on the images. It separates the pixels of an input image into several different classes according to their intensity of the gray levels. For the MRI, the brightest group was labeled as the implant and the other groups as the background.

A gamma correction technique was tested on the images beforehand to emphasize the difference. The dark regions were made lighter by using a gamma smaller than 1. Several gamma values were tested, and 0.4 gave the best results.

To take into account the desired morphological shape, an edge detection algorithm such as the Canny edge detector was tested. The Canny edge detector is a multi-stage algorithm that uses the first derivative of a Gaussian to detect an edge in an image. This technique allowed to isolate the implant from its surrounding.

And finally, morphological functions were used to achieve better results, such as filling holes and reducing the shape of the detected implant.

Combinations of threshold and edge detection algorithms and morphological functions were tested on the test set. Figure 3 illustrates some of the methods previously described.

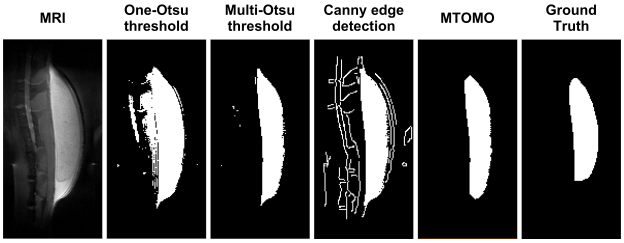

Fig. 3: Different methods applied on slice number 25 of mouse 94-795G acquired at Day 1

The overall results on the test set are poor, and only the best combination, called MTOMO (multithresh Otsu method with additional morphological operations), will be compared with the other models in section results V. The volume is usually overestimated, and the models cannot detect the implant at its extremities. MTOMO implementation is detailed in the Jupyter notebook called *"semi-automatic-methods.ipynb"*, and so are the other methods.

IV. U-Nets: Conventional Neural Network model

Manual segmentation is a heavy workload, but besides being fast, a fully automated solution needs to be robust and precise. The previously seen models do not match these criteria. A deep learning approach was undertaken, as they are known for their excellent performance for segmentation tasks, particularly the U-Net architectures. U-Net, proposed in 2015, is an image segmentation technique developed primarily for medical image analysis that can precisely segment images using a small amount of training data [1].

The U-Net architecture consists of a contracting path (encoder) and an expansive path (decoder). The encoder extracts features of different levels through a sequence of convolutions, ReLU activations, and max poolings, allowing to capture the context of each pixel. The number of feature channels is doubled at each down sampling step, and the dimensions are reduced. The encoder is learning to transform the input image into a feature representation. The decoder is a symmetric expanding path that upsamples the result to increase the resolution of the detected features. Therefore, the expanding path consists of a sequence of up-convolutions and concatenations with

the corresponding feature map from the contracting path, followed by ReLU activations. The resulting network is almost symmetrical, giving it a u-like shape. In the U-Net architecture, skip-connections are added to skip features from the contracting path to the expanding path to recover spatial information lost during downsampling. Additionally, it is experimentally validated that these additional paths benefit model convergence [3]. To obtain the prediction on an input image, the model's output is passed in a sigmoid activation function, which gives each pixel a probability that it is part of the implant. A threshold of 0.5 is applied to label the pixel either 1 or 0. The original architecture of the U-Net, which contains 23 convolutional layers, is presented in Figure 5.

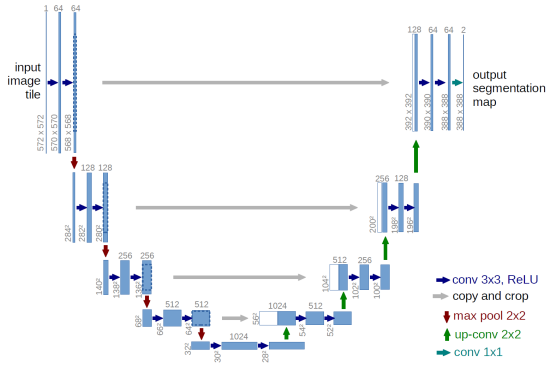

Fig. 4: U-Net architecture. The arrows denote the different operations. Blue box corresponds to a multi-channel feature map and white boxes represent copied feature maps. The number of channels is denoted on top of each box. The image pixel size is provided at the lower left edge of the box.

The architecture can be modified to become a 3D U-Net capable of segmenting 3-dimensional images. The core structure is the same, having both a contracting and expansive path. However, all 2D operations are replaced with corresponding 3D operations (3D convolutions, 3D max pooling, and 3D up-convolutions).

For our task, several U-Net look-alike models were tested. At first, a simple 2D U-Net model was implemented from scratch (IV-A). Secondly, a self-configuring method, called nnU-Net, was trained (IV-B). It is a framework developed and maintained by the Applied Computer Vision Lab (ACVL) of Helmholtz Imaging. Both the 2D and 3D versions were tested.

A. Simple 2D U-Net

The U-Net had to be slightly modified in order to use it for our particular task. As it is a 2D model, each MRI was sliced on its first dimensions, transforming each input into 46 images of dimensions $192 \times 96 \times 3$. In consequence, the dimensions of the input images in the model were changed to $192 \times 96 \times 3$, as the original U-Net was designed for images of size $572 \times 572 \times 3$, and the output to only one channel since we need only the probability of one of the classes (pixel-wise binary classification problem). The padding was modified to *"same"* to avoid shrinking when doing convolutions, and batch normalization was added after each ReLU activation to speed-up training. In total, there are 15 million trainable parameters.

The models' weights are initialized randomly, and the batch size is set to 5 images. Each image was normalized, and both the mean and the standard deviation were fixed. Regarding the loss, the PyTorch *"BCEWITHLOGITSLOSS"* [4] was chosen to deal with our unbalanced data set (97.1% of pixels are labeled as background):

$$L(\mathbf{x},\mathbf{y}) = -\frac{1}{N \cdot |I|} \sum_{n=1}^{N} \sum_{i \in I} [p_c y_{n,i} \cdot \log \sigma(x_{n,i}) + (1 - y_{n,i}) \cdot \log(1 - \sigma(x_{n,i}))] \quad (1)$$

With N being the number of images in a batch, I being the number of pixels, $y_{n,i}$ being the true value (0 or 1) of the pixel i of the image n, and $x_{n,i}$ being its prediction. The weight positive answer p_c was set to 10 (several values were tested). As rare classes could end up being ignored because they are underrepresented during training, the weight of the positive answer is used to limit false negatives (predicting background instead of the implant). The loss combines a Sigmoid layer and the BCELoss in one single class. It is more numerically stable; we take advantage of the log-sum-exp trick for numerical stability.

The Stochastic Gradient Descent was replaced with the Adam Optimizer [5], known to converge faster during training. The learning rate set initially to 0.1 is multiplied by 0.8 every 50 epochs. These values were tuned through Grid Search, maximizing the Dice coefficient on the validation test.

The model was trained for 250 epochs and lasted 5h, but the best model was around 100 epochs (overfitting was seen afterward).

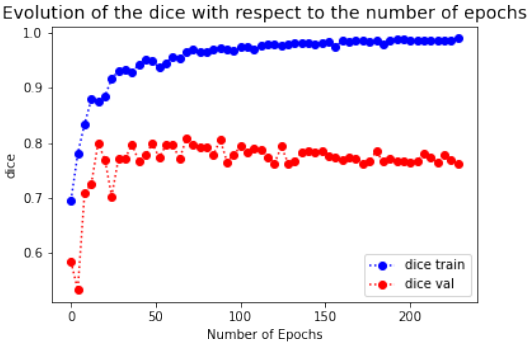

Fig. 5: Training of model 2D U-Net: Dice coefficient with respect to the epochs for the train and the validation set.

Training was done on 96× Intel(R) Xeon(R) Gold 6342 CPU @ 2.80GHz with 500 Gb of RAM. The CPUs are located at the Medical Image Processing Lab in Geneva.

B. nnU-Net: 2D nnU-Net and 3D nnU-Net

nnU-Net refers to a robust and self-adapting framework based on 2D and 3D vanilla U-Nets. It is a deep learning-based segmentation method that automatically configures itself.

There are three types of parameters that need to be determined: fixed, Rule-based and empirical. The fixed parameters are predetermined. They are, for example, the architecture design decisions or the training scheme. The Rule-based parameters are configured according to the data fingerprint (low dimensional representation of dataset properties). It includes dynamic network adaptation (input dimension), target spacing, and resampling or intensity normalization. And finally, the empirical parameters are determined empirically by monitoring validation performance after training.

The architecture is similar to a U-Net. It contains plain convolutions, instance normalization and leaky ReLU for the activation functions. Two computational blocks per resolution stage are used in both encoder and decoder. The down-sampling is done with

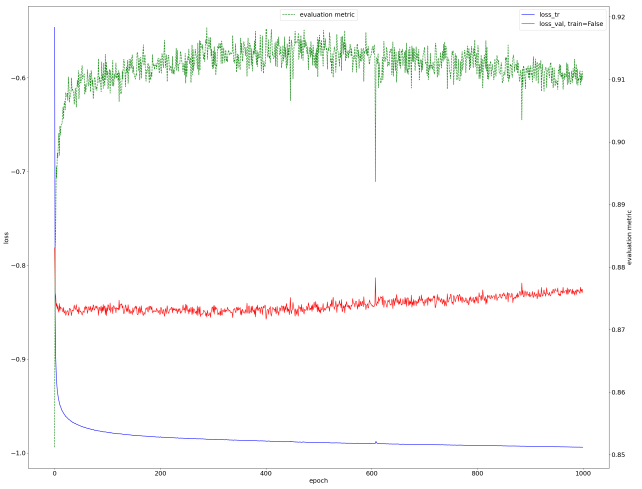

Fig. 6: Training of model nnU-Net 2D for fold number 4

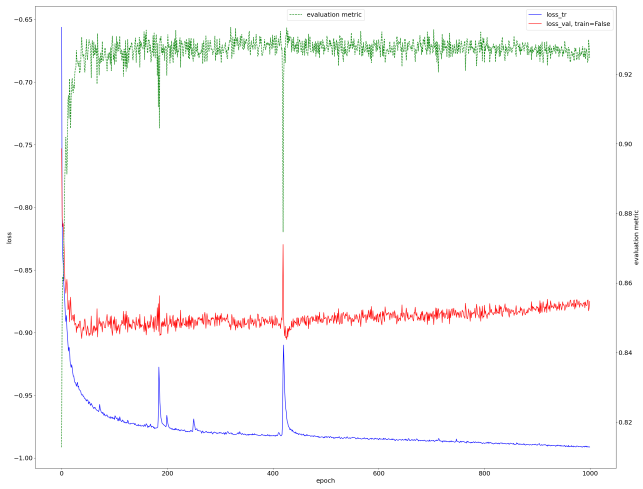

Fig. 7: Training of model nnU-Net 3D for fold number 2

strided convolutions (stride $>$ 1), and the up-sampling is done with transposed convolutions.

The framework offers three separate configurations: 2D U-Net, 3D full resolution U-Net, and 3D U-Net cascade. Only the first two models were used because the 3D U-Net cascade is beneficial when the input data is bigger than the 3D full resolution U-Net's patch size, which was not the case.

The optimizer is a stochastic gradient descent with a high initial learning rate (0.01) and a considerable Nesterov momentum (0.99). The learning rate is reduced during the training using the 'polyLR' schedule, which is an almost linear decrease to 0. The loss combines the Dice loss and a cross-entropy loss (the loss terms are averaged). It is well suited to address the class imbalance and provides stability during training.

A variety of data augmentation techniques is applied during training, such as rotation and scaling, gaussian noise, gaussian blur, change in brightness and contrast, gamma augmentation and mirroring.

The models are trained in a 5-fold cross-validation with a fixed length of 1000 epochs for each configuration. A custom five training/validation set was created that respected the rules seen in section II-C. During the training, the framework keeps in memory the weights that gave the best Dice accuracy on the validation set. At inference time, the framework aggregates the result of the five-fold models. The following figures (6 and 7) show the evolution of the validation and training loss with respect to the number of epochs as well as the average Dice coefficient on the validation set.

The training of the 2D and 3D model took respectively 29 and 100 hours per fold that last each 1000 epochs. Training was done on a GPU NVIDIA A30 with 24GB of RAM. It is located at the Medical Image Processing Lab in Geneva. The framework and its instructions can be found on the GitHub page of the Division of Medical Image Computing, German Cancer Research Center (DKFZ) [6].

V. Results

A. Dice coefficient

The results of the four models on the test set are presented in Table I. 38 MRIs of 5 mice are in the test set. The distribution of these MRIs per timestamp is indicated at the bottom of the Table.

The implant is located in the center of the MRI, in the slice range of 10 to 35. The 2D and 3D nnU-Net were trained on the entire range of slices, but the 2D U-Net and the MTOMO model were trained on a subset of the slice range to avoid a significant proportion of 2D images with no implant. To compare the models, the results are the average of the Dice coefficient per slice from range 9 to 36. The results are presented per timestamp, and the overall columns show the average on the whole test set. 95% confidence intervals are present to show the robustness of each model.

Several key points can be deducted from the results:

- **MTOMO and 2D U-Net are not robust:** we can see a significant drop in the Dice coefficient for a particular timestamp. For MTOMO, its accuracy is 16% less than its overall results for time stamp three months, and for 2D U-Net, it is a drop of 12% for time stamp one day. This can also be seen in the 95% confidence interval in the overall results. We can see that they are twice as big for MTOMO and 2D U-Net compared to the two nnU-Net models.
- **The two nnU-Net are significantly better than MTOMO and 2D U-Net:** the overall Dice accuracy is, on average, 18% better for the two nnU-Net than the other models. The best score for each timestamp belongs to the two nnU-Net (in bold in the Table).
- **3D nnU-Net is the best model, but not by a big margin:** 3D nnU-Net reports the best Dice coefficient in the overall score and in all the timestamps. However, the difference with the second-best model, 2D nnU-Net, seems to be very little. There is only a 1.1% increase in the overall score, and the 95% confidence intervals are globally in the same range.

B. Volume difference

The predicted volume is calculated by counting the voxels labeled 1 in the prediction .nii file. It is then multiplied by the volume of a voxel. For each MRI, the predicted volume is computed, and the following formula is used to quantify the accuracy of the prediction:

$$\text{Relative difference} = \left(\frac{\text{predicted volume} - \text{real volume}}{\text{real volume}} \right) * 100$$

Table II summarizes the results: the average and the absolute average of the relative difference are shown, so as the maximum error in percent that each model made on the test set. Additionally,

Model	1 Day	3 Weeks	3 Months	6 Months	Overall
MTOMO	0.722 ± 0.038	0.810 ± 0.026	0.592 ± 0.042	0.700 ± 0.043	0.706 ± 0.019
2D U-Net	0.692 ± 0.047	0.813 ± 0.036	0.824 ± 0.034	0.817 ± 0.041	0.787 ± 0.020
2D nnU-Net	0.910 ± 0.026	0.921 ± 0.021	0.889 ± 0.026	0.899 ± 0.025	0.905 ± 0.012
3D nnU-Net	$\mathbf{0.920 \pm 0.023}$	$\mathbf{0.929 \pm 0.019}$	$\mathbf{0.902 \pm 0.023}$	$\mathbf{0.904 \pm 0.026}$	$\mathbf{0.915 \pm 0.011}$
Number of 3D MRI	10	10	10	8	38
Number of 2D images	280	280	280	224	1064

TABLE I: Average of the Dice coefficient with respect to the timestamps and the overall score on the test set. 95% confidence intervals are also shown.

the standard deviation is shown and was calculated by taking the sum of the squares of the relative differences and divided by the number of samples.

Model	Results in % on the test set: Average	Standard deviation	Absolute average	Max error	Hours of training	Time to predict[1]
Baseline	9	0.023	11	41.7	0h	1s
MTOMO	-30.14	0.252	40.5	74.04	0h	30s
2D U-Net	**3.14**	0.012	14.15	44.48	5h	16s
2D nnU-Net	-7.65	0.014	8.90	27.69	29h	146s
3D nnU-Net	-3.85	**0.005**	**5.74**	**17.79**	100h	320s

TABLE II: Relative difference in the volume for the 5 models.
1: how much does it take for a model to predict one $46 \cdot 192 \cdot 96$ MRI

Ideally, the average would be centered around 0. If it is negative, our model is under-sampling its prediction. In other words, it predicts a smaller volume than the actual volume. It is the case for the two nnU-Net (-7.65% and -3.85%). Regarding the absolute average, the closest to 0 the better.

Overall, we find in Table II the same trend as for the Dice coefficient results. This is not surprising as they are, by definition, related.

MTOMO and 2D U-Net have a large absolute average, indicating their incapacity to predict the volume accurately.

What is surprising is the significant improvement between the 3D nnU-Net and the 2D nnU-Net. Even though the difference in the Dice coefficient was little, we can see a decrease of 44% in the absolute average and a drop of 9.9% in the worst prediction. The average and standard deviation are also better.

The results of the baseline model, defined in section III-A, are also shown. The 3D nnU-Net has better results in every category, which indicates that the model has successfully learned its task.

The penultimate column shows the hour of training of each model. Even though some models were trained on GPUs rather than CPUs, it indicates the model's complexities. The accuracy seems to increase as the model gets more complex, which is expected.

The last column, *"Time to predict"* is the time it takes to predict one MRI and generate the binary .nii file. The prediction was made on a MacBook Pro (2019) macOS Monterey with a CPU speed of 2.4 GHz and a Ram capacity of 16 Go. Labeling by hand 46 images of dimension 192×96 takes 90 minutes for an expert. In comparison, 3D nnU-Net achieves this task in 320 seconds.

VI. Model comparisons

To explore the strength and differences of each model, a comparison analysis was made.

A. Dice per slice

The Dice coefficient per each slice of each model is shown in Figure 8.

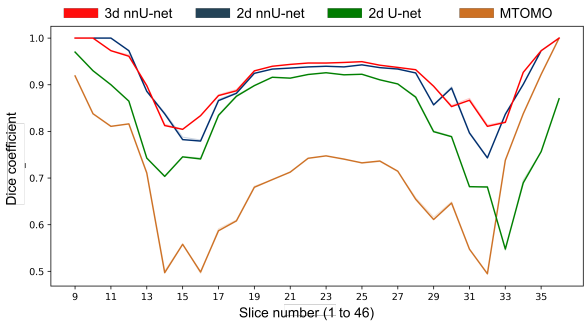

Fig. 8: Evolution of the Dice with respect to the slice for the 4 models.

The perfect model would be a straight line at level 1, which would indicate a Dice coefficient of 1 for all the slices.

This inverted U-shape is present among the four models but at different positions with respect to the y-axis. The better the model, the higher the position of the inverted U-shape. This indicates that the models are struggling to predict accurately at the extremities (slice 13 to 19 and slice 30 to 35) but achieve a high score in the middle (slice 19 to 30). As the implant is centered in the middle, the drop in accuracy at the extremities is because the implant is ending. The surface to detect is getting smaller, which implies a diminishing contrast between the rare implant pixels and the numerous background pixels and more atypical shapes that the model is not used to detect.

The 3D nnU-Net seems to be better for detecting the implant's extremities than the 2D nnU-Net. By working in 3D, the model can rely on the previous slices to help locate the implant at the extremities. On the other hand, as the 2D model is treating each slice independently, it cannot anticipate a decrease in volume. Figure 9 illustrates this particular strength of the 3D model. We can see its capacity to predict the end of the implant accurately compared to the 2D model.

B. Volume error per slice

Besides having a very similar Dice score, 3D nnU-Net can better predict the volume compared to 2D nnU-Net. To illustrate this phenomenon, the Figure 10 shows the average error and the absolute average error in number of pixels per slice of the two models. The Dice scores of the two models are also plotted on the same Figure.

On average, both models predict a smaller number of voxels at every slice than the ground truth (the values of the two average dash

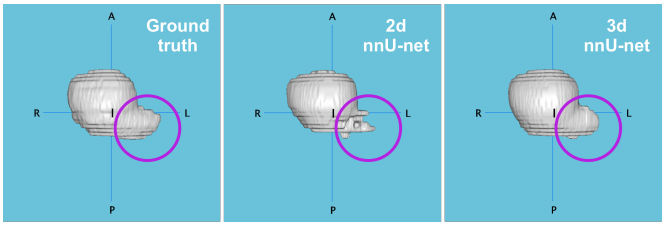

Fig. 9: Prediction comparison between the 2D nnU-Net and 3D nnU-Net on sample number 97-815G at week 3. The Dice coefficients are, for the 2D and 3D nnU-Net, 0.916 and 0.944, and the predicted volume errors are 11.1% and 0.7%.

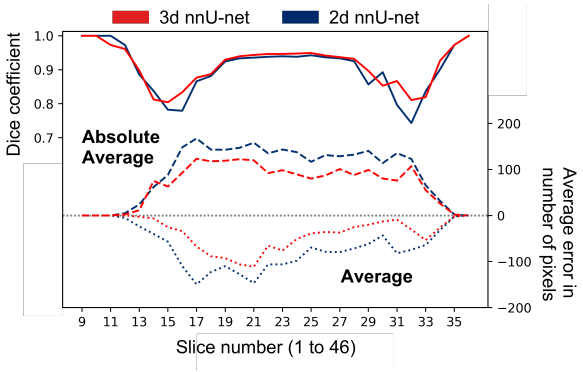

Fig. 10: Dice coefficient and average and absolute average error in number of pixels for the models 3D and 2D nnU-Net. The results are an average per slice from 9 to 35.

lines are negative). This is not surprising as we have seen that the two models were under-sampling the volume in Table II, but here, it is shown that it is the case at every slice.

Additionally, the model 3D nnU-Net is clearly performing better than the 2D nnU-Net as the red dash curve of the absolute average is constantly under the blue dash one. The interesting point is that the difference between the two absolute averages is constant with respect to the slices. This is not the case for the Dice coefficient, where there is a small difference in the middle and a big difference at the extremities.

Because the implant is centered in the middle, the implant's surface is bigger. Consequently, a slight error in the prediction will significantly affect the accuracy of the overall volume. This is one of the limitations of the Dice score; it does not consider the size of the ground truth. The 3d nnU-net has a slightly better dice score in a region with many implant pixels. Consequently, the error volume difference is significant between the two models.

VII. Future Research

One issue that should be addressed is the recurrent under-sampling of the 3D nnU-Net model. A possible solution would be to fine-tune the threshold of the last layer used on sigmoid function output. The model is making a lot of false negatives by predicting implant voxels as background. This is certainly due to the class imbalance and was addressed with the parameter positive weight for the 2D U-Net model. Putting its poor performance aside, it was the model with an average error closest to zero (Table II).

Secondly, increasing the test data is always a good solution to evaluate the robustness of the model. Our data is very homogeneous. The amount of injected Adipearl in each mouse is roughly the same, which means that the model has not been tested on smaller or bigger implants. A variation in the volume was highlighted in section II-B, but it was a variation limited to a couple of dozens of percentages. Additionally, because of the anatomy of the back of the mouse, the implant resulted in a simple oval shape. Even though data augmentation was done on the MRIs when training the nnU-Net models, the shape of the implant remained unchanged. The model is hence inexperienced with dealing with atypical and complex shapes.

Lastly, we highlighted the limitation of the Dice coefficient for our particular task. Further research could be done by implementing different losses and exploring the possibility of using another metric. An interesting comparison of losses is made in the following paper: *"A survey of loss functions for semantic segmentation"* [7]. It motivated the choice of using a binary cross-entropy with logits loss for the 2D U-Net model. Still, several losses could be implemented, and compare the results to our current one.

VIII. Conclusions

In this project, several models were tested to segment mice MRI. The models were trained to retrieve the shape of an implant by creating a binary MRI .nii file. The comparison of the models allowed us to show the drawbacks of some and the strengths of others. The more complex the model is, the better the results. Without surprise, the U-Net architecture gave the best results, in particular the 3D nnU-Net. Several ideas were presented for improvement, such as dealing with the under-sample behavior of the nnU-Net models. Additionally, the limitation of the Dice coefficient was shown for a task where the goal is to predict the volume as accurately as possible. Besides the low resolutions of the MRI and a small training set, the 3D nnU-Net achieved promising results. The segmentation is done in 320 seconds with an average Dice accuracy of 0.915 on the test set, which translates to an average error of 5.7% for the predicted volume. The model that can be run on a personal computer will allow the biologist of Volumina to gain valuable time.

If better results are desired, the biologist can still intervene on the predicted .nii file to correct potential mistakes. For a more realistic shape, the extremities of the MRI should be looked at, and to achieve a more accurate global volume, the intervention should be done in the middle of the implant. Nevertheless, this semi-automatic, which is not required for reasonable results, is a considerable upgrade time-wise compared to labeling by hand the entire MRI.

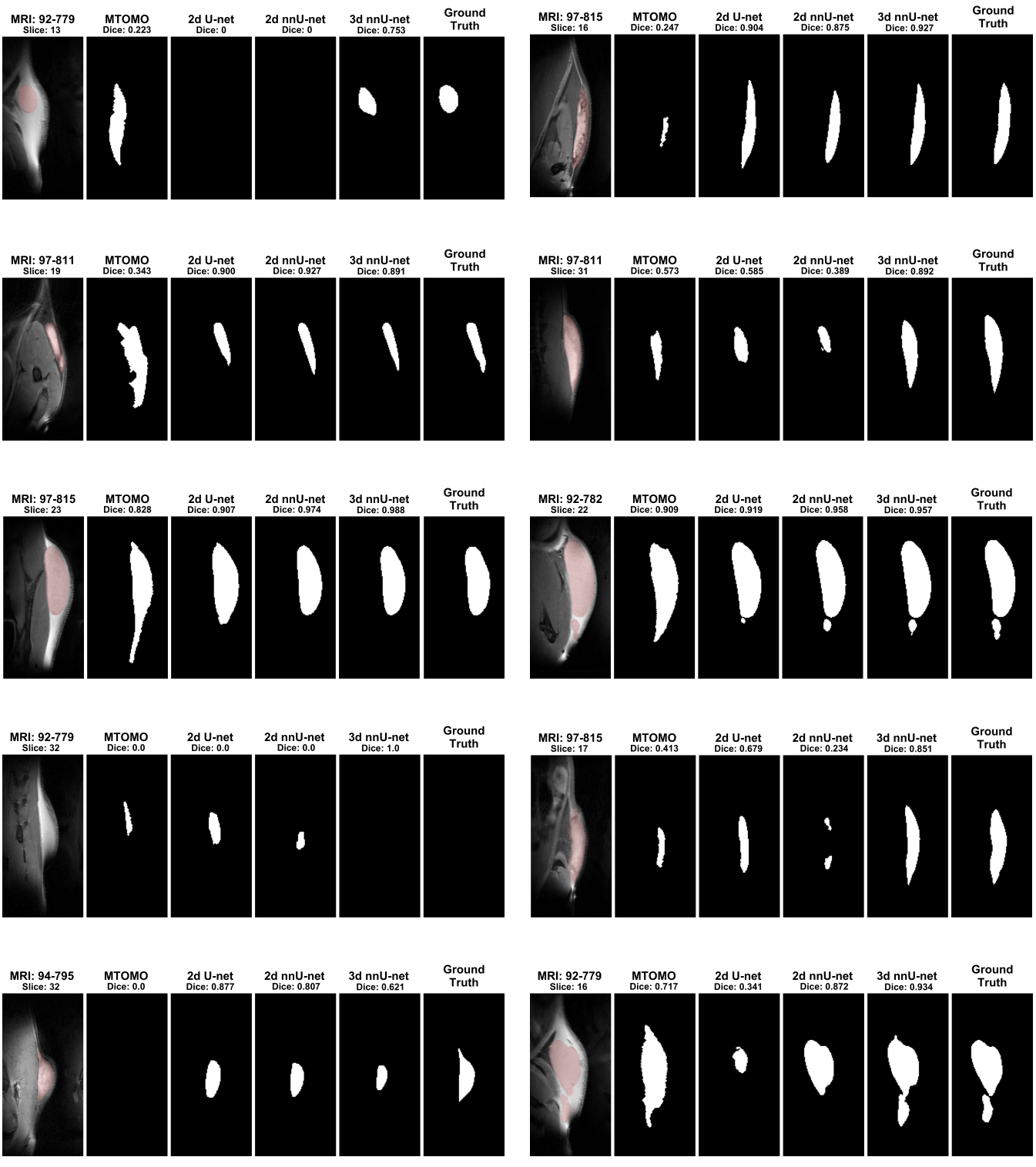

Fig. 11: Ten examples of prediction on the test set by the different models. The MRI is the first image, and the concerned slice and mouse id are written at the top. The red area is the implant to be segmented. The ground truth, annotated by hand, is the last image. The Dice accuracy is shown at the top of each image. We can see the robustness of the 3D nnU-Net compared to the other models.

REFERENCES

[1] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[2] F. Isensee, P. Jaeger, and S. e. a. Kohl. (2020) nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. [Online]. Available: https://doi.org/10.1038/s41592-020-01008-z

[3] M. Drozdzal, E. Vorontsov, G. Chartrand, S. Kadoury, and C. Pal. (2016) The importance of skip connections in biomedical image segmentation. [Online]. Available: https://arxiv.org/abs/1608.04117

[4] C. Torch, *BCEWITHLOGITSLOSS*, 2019. [Online]. Available: https://pytorch.org/docs/stable/_modules/torch/nn/modules/loss.html#BCEWithLogitsLoss

[5] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[6] (2020) Applied computer vision lab of helmholtz imaging: Github of nnu-net. [Online]. Available: https://github.com/MIC-DKFZ/nnUNet

[7] S. Jadon, "A survey of loss functions for semantic segmentation," 2020.