

# An efficient data strategy for the detection of brain aneurysms from MRA with deep learning

Youssef Assis, Liang Liao, Fabien Pierre, René Anxionnat, Erwan Kerrien

## ► To cite this version:

Youssef Assis, Liang Liao, Fabien Pierre, René Anxionnat, Erwan Kerrien. An efficient data strategy for the detection of brain aneurysms from MRA with deep learning. The MICCAI workshop on Data Augmentation, Labeling, and Imperfections, Oct 2021, Strasbourg, France. pp.226-234, 10.1007/978-3-030-88210-5\_22 . hal-03391884

HAL Id: hal-03391884

<https://hal.univ-lorraine.fr/hal-03391884>

Submitted on 21 Oct 2021

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Copyright

# An efficient data strategy for the detection of brain aneurysms from MRA with deep learning

Youssef Assis<sup>1</sup>, Liang Liao<sup>2</sup>, Fabien Pierre<sup>1</sup>, René Anxionnat<sup>2,3</sup>, and Erwan Kerrien<sup>1</sup>

<sup>1</sup> University of Lorraine, CNRS, LORIA, F-54000 Nancy, France

<sup>2</sup> Department of Diagnostic and Therapeutic Neuroradiology, University of Lorraine, Nancy, France

<sup>3</sup> IADI, INSERM U1254, University of Lorraine, Nancy, France

**Abstract.** The detection of intracranial aneurysms from Magnetic Resonance Angiography images is a problem of rapidly growing clinical importance. In the last 3 years, the raise of deep convolutional neural networks has instigated a streak of methods that have shown promising performance. The major issue to address is the very severe class imbalance. Previous authors have focused their efforts on the network architecture and loss function. This paper tackles the data. A rough but fast annotation is considered: each aneurysm is approximated by a sphere defined by two points. Second, a small patch approach is taken so as to increase the number of samples. Third, samples are generated by a combination of data selection (negative patches are centered half on blood vessels and half on parenchyma) and data synthesis (patches containing an aneurysm are duplicated and deformed by a 3D spline transform). This strategy is applied to train a 3D U-net model, with a binary cross entropy loss, on a data set of 111 patients (155 aneurysms, mean size  $3.86\text{mm} \pm 2.39\text{mm}$ , min  $1.23\text{mm}$ , max  $19.63\text{mm}$ ). A 5-fold cross-validation evaluation provides state of the art results (sensitivity 0.72, false positive count 0.14, as per ADAM challenge criteria). The study also reports a comparison with the focal loss, and Cohen’s Kappa coefficient is shown to be a better metric than Dice for this highly unbalanced detection problem.

**Keywords:** Brain aneurysm detection · data sampling · CNN.

## 1 Introduction

Intracranial aneurysms are local dilations of the cerebral blood vessels. Their rupturing accounts for 85% of subarachnoid hemorrhages (SAH), and is related to high mortality and morbidity rates [13]. The generalization of radiologic examinations in the diagnostic process has exposed the detection of unruptured aneurysms as a problem of increasing clinical importance. The need to browse through 3D Computed Tomography Angiography (CTA) or Magnetic Resonance Angiography (MRA) data in an ever increasingly time-constrained clinical setting however leads to inevitable errors. The innocuity of 3D Time-of-Flight (TOF) MRA makes it particularly suited for screening, even though the detection of small aneurysms ( $< 5\text{mm}$ ) may be challenging [8]. A reliable automated method would be a valuable asset to assist radiologists in their clinical routine.

The first computer-assisted detection (CAD) system reported in the literature [1] was based on traditional image processing. Recently, convolutional neural networks (CNNs) have proven their superior performance in visual tasks, including medical image analysis. Detecting brain aneurysms is very challenging because aneurysms are scarce (a few tens to hundreds of positive voxels among millions in the MRA data), and their number is indefinite a priori. Therefore, it has been but very recently that deep learning approaches have been investigated in that context. First, several 2D approaches have been proposed in the literature [12,17], but all most recent approaches are fully 3D. The performance of the dual-path multiscale DeepMedic model was deemed promising [14] but as a complement to an expert reader [6]. Last year, the ADAM challenge enabled an objective comparison of a variety of other 3D approaches. The leading 3 methods for the detection task were based on 3D U-net [3] to evade the problem of the indefinite number of aneurysms through the generation of a heat map. The scarcity, and thereby high class imbalance, was tackled either through the loss function and/or the model. The Dice similarity coefficient, and the Binary Cross Entropy (BCE) and TopK losses were combined to form an ensemble loss in [11]. A different ensemble loss approach was taken in [18]. Four models based on the No New-Net were trained and the final segmentation was decided in a majority voting amongst these models. The leader method [2] focused on the model architecture with a Retina U-net model that aggregates an encoder network and a feature pyramid network to guide the high-resolution detection with strong semantic features at low resolution. The actual impact of all these variants was questioned by the recent emergence of a new leader with a vanilla 3D U-net model trained with a combined Dice and BCE loss, and a prediction based on an ensemble of 5 models [19]. We believe a major difference dwells in the samples generation. If the first 3 methods used large patches ( $\{192,224,256\} \times 256 \times 56$  voxels), this last one used  $128^3$  patches with a rich data augmentation process. Another limitation that we see is the small size of the dataset and the drudgery of the voxel-wise annotation.

The current study focuses on the data strategy to generate input samples that are better designed to tackle the class imbalance and reduced data set issues. We used a vanilla 3D U-net in all our experiments but small patches ( $48^3$ ) were used. Aneurysm annotation is approximate but fast. This approximation is claimed to be precise enough for the detection task. A combination of guided sample selection and sample synthesis is also proposed, and BCE and focal losses are compared. Finally, we advocate for the use of Cohen’s Kappa as a better metric than Dice in this class imbalance situation.

## 2 Materials and methods

### 2.1 Dataset and data annotation

A total of 111 TOF-MRA examinations (56 females, 55 males) were collected at our medical institution between April 2015 and January 2020. The criterion for inclusion was the presence of an aneurysm. All aneurysms are saccular. Criteria for exclusion were any pre-treated aneurysm and large aneurysms ( $> 20$ mm). The images were acquired on a 3T scanner (GE Healthcare) with the following parameters: TR = 28 ms, TE = 3.4 ms, slice thickness = 0.8 mm, FOV = 24, flip angle =  $17^\circ$ , 4 slabs (54

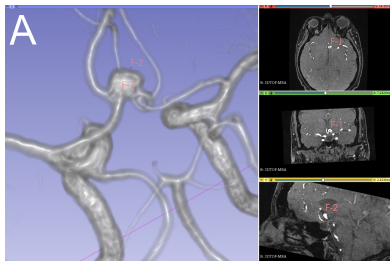


Fig. 1: Aneurysm annotation as an approximate sphere with 2 points in Slicer

slices/slab), acquisition time = 6 min 28 s, resulting in  $512 \times 512 \times 254$  volumes with a  $0.47 \times 0.47 \times 0.4\text{mm}^3$  voxel size. Each DICOM data was anonymized and converted to NIfTI format on the clinical site before processing. Each examination contained from one (81/111) to five aneurysms (1 case) for a total of 155 aneurysms with a mean diameter of  $3.86 \text{ mm} \pm 2.39 \text{ mm}$  (min: 1.23 mm, max: 19.63 mm). These were mostly small aneurysms since 60 were below 3 mm and 66 between 3-5 mm, which makes it a challenging dataset.

Previous works rely on databases where aneurysms have been segmented voxel-wise. This annotation is both tedious and tainted with intra- and inter-rater variability. Since we only aim at detecting aneurysms, we deployed a less accurate but much faster annotation: each aneurysm was annotated, by a radiologist with 10 years of experience, by placing two points, one at the center of the neck and the other at the dome, so as to define a sphere that approximated the aneurysm sack. 3D Slicer software was used [7] to place points in volume rendering view (see Fig. 1)

## 2.2 Model implementation

Our software code was written in Python (3.8.5) using Keras (2.4.3). It was based on D.G. Ellis’s open-source 3D U-net implementation [5] with 4 layers. Our first investigations with the available regular patch sampling did not show convergence of the model. As a consequence, the patch generator described below was plugged as input to the model. The following hyperparameters were used: 100 epochs, constant learning rate =  $10^{-4}$ , BCE loss, Adam optimizer, batch size=10, with batch normalization. Each input volume was normalized between 0 and 1.

A full volume could be predicted by patch reconstruction: the initial volume was resampled to an isotropic 0.4 mm voxel size; predictions were computed for patches that cover the entire volume; and the resulting volume was resampled to the original resolution. To avoid border effects due to convolutions on small patches, an overlap of 8 voxels was considered between neighboring patches and only the central  $32 \times 32 \times 32$  part of the patches were juxtaposed to cover the final volume.

## 2.3 Patch generation and data augmentation

The discriminative power of a classifier depends on its capacity to statistically model both the background (negative samples) and the foreground (positive samples). In large patches approaches, an aneurysm is present in most patches, which requires

Table 1: Comparison of model variations: *Model0* is our proposed model. Dice and  $\kappa$  (Kappa) coefficients were evaluated on the validation set at the end of training. Other ADAM and patch-wise metrics were measured on the test set (11 patients).

Model	Validation set		ADAM metrics		Patch-wise metrics			
	Dice	$\kappa$	Sensitivity	FPs/case	TP	FP	FN	TN
<i>Model0</i>	0.339	0.665	0.970	0.454	14	5	2	2194
<i>Model1</i>	0.089	0.527	0.803	0.190	12	2	4	2197
<i>Model2</i>	0.038	-1.21e-8	0	0	0	0	16	2199
<i>Model3</i>	0.434	0.772	0.879	1.545	14	11	2	2188
<i>Model4</i>	0.245	0.589	0.833	1.0	13	8	3	2191

healthy patients in the database. In our small patch approach, negative (aneurysm-free) patches are very common outside the aneurysm surroundings, but multiple instances need to be extracted from each patient data to build reliable statistics on the background. On the opposite side, only one single positive patch exists for each aneurysm. Adapted data sampling is an efficient strategy to handle this problem [9]. Our first sampling strategy then consists in duplicating 50 times each positive patch, centered on each aneurysm. A variety of shapes are synthesized by applying a random distortion to each duplicate: each control point on a  $3 \times 3 \times 3$  lattice enclosing the patch, except the central point, is moved randomly by 4 mm in all 3 space directions and patch voxel locations in the original volume are computed using cubic spline interpolation. But class imbalance also emanates from the vascular information that only represents 3 to 5% of the background signal. In order to guide the model to discriminate between healthy and pathological vessels, our second sampling strategy consists in taking half of the negative samples centered on a blood vessel. The 100 brightest voxels were selected as patch centers. The other 100 centers were randomly selected within voxel values between the 20th and the 80th percentiles. Patch overlap was avoided by enforcing a minimum 20 mm distance between any two patch centers.

As a result we used 200 negative patches, and 50 positive patch duplicates. We used  $48 \times 48 \times 48$  patches with an isotropic voxel size of 0.4 mm, closest to the nominal resolution, so that patches were cubes with a side length of 19 mm. Data augmentation was applied in the process with random rotations by 0 to  $180^\circ$  and shifts by 10 mm in all 3 space directions.

## 2.4 Metrics and performance evaluation

The training was monitored with the Dice coefficient. However, since the aneurysms are scarce and small, this metric lacks sensitivity to detection errors. We also computed Cohen’s Kappa coefficient ( $\kappa$ ) [4], that is more robust to class imbalance. These metrics are computed voxel-wise on the collection of input patches.

The performance of a model was evaluated on a test set, using mean sensitivity and FP count/case scores as defined for Task 1 in the ADAM challenge [16]. The connected components (CC) in both ground truth and predicted full volumes were labeled. A True Positive (TP) is a CC in the ground truth that contains the center of gravity of a predicted CC. A False Negative (FN) is a CC in the ground truth with

no such predicted CC. A False Positive (FP) is a predicted CC whose center is not contained in any ground truth CC.

However, the above metrics do not enable True Negative count, and thereby prevent from computing specificity. Thereafter, we also computed patch-wise statistics (no positive duplicate): a patch is considered positive if it contains a positive voxel, else it is negative. This enables to compute a full confusion matrix.

### 3 Experiments and Results

#### 3.1 Ablation study

A first set of experiments aimed at evaluating the relevance of various parts of the model and the patch sampling strategy described in sections 2.2 and 2.3, which will be denoted as *Model0*. 4 variants were tested.

BCE is very sensitive to class imbalance [15]. In order to see the effectiveness of our data strategy to counter class imbalance, *Model1* was trained the same as *model0* but using the focal loss [10], that was designed to focus the training on the minority class. *Model2* only generates 5 duplicates (instead of 50) for each positive patch. *Model3* uses 50 duplicates but without random distortion applied. *Model4* only considers 100 background patches (instead of 200, 50 on vessels, 50 outside).

The data set was split into 3 sets used for: training (78 cases, 70%), validation (22 cases, 20%) and test (11 cases, 10%). All models were trained with the training set, and were monitored with the validation set. Table 1 reports the results of this study. Dice and  $\kappa$  are observed on the validation set at the end of training. Other performance metrics (see Sec. 2.4) are computed on the test data set.

No real improvement could be observed with the focal loss (*Model1*), and BCE is even more sensitive, which demonstrates the efficacy of our sampling strategy. The Dice score remained very low because the predicted CC were very small. However,  $\kappa$  could better capture the relatively good performance of this model. *Model2* did not converge: class imbalance is indeed an issue. *Model3* provided good results but with too many FPs, due the lack of diversity in the aneurysm shapes shown to the model during training. The excess in positive voxels leads to larger TPs, which explains the better Dice and  $\kappa$  scores. Finally, the class balance is improved in *Model4*, but it underperforms *Model0* because the sample size for the negative patches is too small to reliably model background statistics.

#### 3.2 5-fold validation

The global performance of our proposed model (*Model0*) was assessed using 5-fold validation. 5 models were trained, each time with 4 subsets for training and leaving one subset for test. Predictions were generated for each patient in each test set, providing a prediction for each patient. The mean aneurysm diameters in the 5 splits were: 3.82 mm, 3.74 mm, 3.96 mm, 3.84 mm and 3.93 mm.

Fig. 2 displays the Free-response Receiver Operating Characteristics (FROC) curve: It reports the sensitivity and FP count scores, as per ADAM, computed on all 111 patients in our dataset for various detection thresholds. By adjusting the detection threshold, and by comparison with abc method, our method achieves a sensitivity

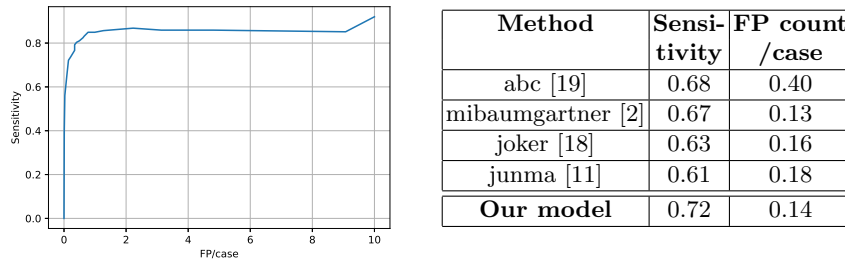


Fig. 2: (left) FROC curve for our model: AUC=85.24%. (right) Comparison with 4 leading methods in ADAM challenge (in decreasing order).

of 0.80 @ 0.40 FP/case, and, with mibaumgartner method, a sensitivity of 0.70 @ 0.13 FP/case. The Area Under Curve (AUC) is 85.24%. The optimal detection threshold was determined as the closest point to the upper left corner. Our model reaches a sensitivity of 0.72@0.14 FP/case.

## 4 Discussion

In this study, in order to determine the impact of our data strategy we voluntarily used a vanilla 3D U-net model with a BCE score and simple optimization process (e.g. fixed learning rate and number of epochs). The focus was put on the data to assess the impact of various aspects of their preparation on learning.

First, a rough, but fast annotation was employed, which enables to rapidly label a large number of MRA volumes. Besides, a small patch approach was chosen. Small, non-intersecting patches are assumed independent, which allows for an efficient exploitation of even a small set of original MRA images (111). Second, we proposed an adapted data sampling process in two steps. On one side, guided sampling: The negative (aneurysm-free) patches are extracted by half centered on blood vessels and the other half elsewhere. We have shown that 200 patches were better able than 100 to capture the background statistics (*Model0* vs *Model4*). On the other side, data synthesis: The positive patches are duplicated 50 times, which enables to counter the high class imbalance (*Model0* vs *Model2*), and various shapes are synthesized by applying random non-rigid distortions, which describes the foreground statistics more accurately and enables a reduction of FPs (*Model0* vs *Model3*).

The proposed model has a sensitivity of 0.72, with a FP count/case of 0.14. FROC analysis showed that it is competitive with the best current leading methods in the ADAM challenge. Our method will have to be adapted to the ADAM challenge conditions for a definite comparison to be made. Future efforts will aim at further reducing FPs. Tests with the Focal Loss (*Model1*) generated smaller CC, which reduced the FP score but at the expense of sensitivity. Besides cases that are easy for a radiologist to discard, the most challenging FPs are located where a small artery, close to the resolution limit, branches onto a large artery. These are mistaken for small aneurysms (see Fig. 3, left). Indeed the performances of our model are lower on small aneurysms. Of all 155 aneurysms, 34 were not detected (FN). But 18 of these FNs had a diameter below 2 mm and 10 more were below 3 mm. The sensitivity of our model is 0.53

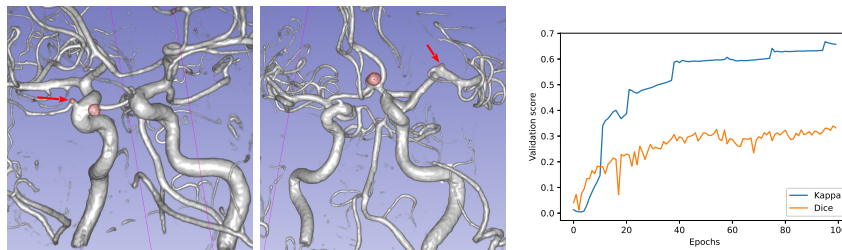


Fig. 3: (left) Branching of small arteries (here, ophtalmic artery, see arrow) may be mistaken for an aneurysm (predicted CC are in red, annotation points are present). (middle) Example of an overlooked aneurysm (arrow). (right) Typical surge of  $\kappa$  is a good predictor of the final convergence (around epoch 10).

for aneurysms smaller than 2 mm, and reaches 0.89 for larger aneurysms. But this difficulty in detecting small aneurysms is inherent to MRA [8]. Note that during the visual review of the results by a radiologist with 30 years of experience, 8 FPs proved to be actual aneurysms that had been overlooked during the initial annotation (see Fig. 3, middle).

In our experiments, we observed a typical sudden surge in  $\kappa$  score that was correlated to a satisfactory convergence of the training phase (see Fig. 3, right). We interpreted it as a better sensitivity of  $\kappa$  over Dice to even small intersections between prediction and ground truth volumes.

## 5 Conclusion

In this paper, we presented an efficient data sampling strategy to detect intracranial aneurysms from MRA images, that is able to reach a state of the art sensitivity of 0.72 at 0.14 FP/case. Joining forces will hopefully decrease the number of FPs to design a more specific classifier. A future extension of this work is combining this data strategy with more sophisticated architectures and loss functions whose efficacy has been demonstrated, in particular by the ADAM challenge. Furthermore, a current work in progress investigates the  $\kappa$  score as a loss function, to leverage its capacity to assess the quality of a classifier despite the highly class imbalance problem.

## Acknowledgments

We want to thank the Grand Est region and the regional and university hospital center (CHRU) of Nancy in France for funding this work. Experiments presented in this paper were carried out using the Grid’5000 experimental testbed, being developed under the INRIA ALADDIN development action with support from CNRS, RENATER and several Universities as well as other funding bodies (see <https://www.grid5000.fr>).

## References

1. Arimura, H., Li, Q., Korogi, Y., et al.: Computerized detection of intracranial aneurysms for three-dimensional MR angiography: Feature extraction of small protrusions based on a shape-based difference image technique. *Medical physics* **33**(2), 394–401 (2006)



2. Baumgartner, M., Jaeger, P., Isensee, F., et al.: Retina U-net for aneurysm detection in MR images. In: Automatic Detection And Segmentation Challenge (ADAM) (2020), <https://adam.isi.uu.nl/results/results-miccai-2020/participating-teams-miccai-2020/ibbm/>
3. Çiçek, Ö., Abdulkadir, A., Lienkamp, S., et al.: 3D U-net: learning dense volumetric segmentation from sparse annotation. In: International conference on medical image computing and computer-assisted intervention (MICCAI). pp. 424–432 (2016)
4. Cohen, J.: A coefficient of agreement for nominal scales. *Educational and psychological measurement* **20**(1), 37–46 (1960)
5. Ellis, D.: 3D U-net convolution neural network with Keras. <https://github.com/ellisdg/3DUNetCNN> (legacy branch, commit dc2d0604499298266e7aaf1db68603288bd34577) (2017)
6. Faron, A., Sichtermann, T., Teichert, N., et al.: Performance of a deep-learning neural network to detect intracranial aneurysms from 3D TOF-MRA compared to human readers. *Clinical neuroradiology* **30**(3), 591–598 (2020)
7. Fedorov, A., Beichel, R., Kalpathy-Cramer, J., et al.: 3D Slicer as an image computing platform for the quantitative imaging network. *Magnetic resonance imaging* **30**(9), 1323–1341 (2012), <https://slicer.org>, PMID: 22770690
8. Jang, M., Kim, J., Park, J., et al.: Features of “false positive” unruptured intracranial aneurysms on screening magnetic resonance angiography. *PloS one* **15**(9), e0238597 (2020)
9. Johnson, J., Khoshgoftaar, T.: Survey on deep learning with class imbalance. *Journal of Big Data* **6**(1), 1–54 (2019)
10. Lin, T.Y., Goyal, P., Girshick, R., et al.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision (ICCV). pp. 2980–2988 (2017)
11. Ma, J., An, X.: Loss ensembles for intracranial aneurysm segmentation: An embarrassingly simple method. In: Automatic Detection And Segmentation Challenge (ADAM) (2020), <https://adam.isi.uu.nl/results/results-miccai-2020/participating-teams-miccai-2020/junma-2/>
12. Nakao, T., Hanaoka, S., Nomura, Y., et al.: Deep neural network-based computer-assisted detection of cerebral aneurysms in MR angiography. *Journal of Magnetic Resonance Imaging* **47**(4), 948–953 (2018)
13. Shi, Z., Hu, B., Schoepf, U., et al.: Artificial intelligence in the management of intracranial aneurysms: current status and future perspectives. *American Journal of Neuroradiology* **41**(3), 373–379 (2020)
14. Sichtermann, T., Faron, A., Sijben, R., et al.: Deep learning-based detection of intracranial aneurysms in 3D TOF-MRA. *American Journal of Neuroradiology* **40**(1), 25–32 (2019)
15. Taghanaki, S.A., Abhishek, K., Cohen, J., et al.: Deep semantic segmentation of natural and medical images: a review. *Artificial Intelligence Review* pp. 1–42 (2020)
16. Taha, A., Hanbury, A.: Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. *BMC medical imaging* **15**(1), 1–28 (2015)
17. Ueda, D., Yamamoto, A., Nishimori, M., et al.: Deep learning for MR angiography: automated detection of cerebral aneurysms. *Radiology* **290**(1), 187–194 (2018)
18. Yang, Y., Lin, Y., Li, Y., et al.: Automatic aneurysm segmentation via 3D U-net ensemble. In: Automatic Detection And Segmentation Challenge (ADAM) (2020), <https://adam.isi.uu.nl/results/results-miccai-2020/participating-teams-miccai-2020/joker/>
19. Yu, H., Fan, Y., Shi, H.: Team abc. In: Automatic Detection And Segmentation Challenge (ADAM) (2020), <https://adam.isi.uu.nl/results/results-live-leaderboard/abc/>