Cover Page

1) Title of the paper:

Synthetic vascular image selection for deep learning based cerebral bifurcation classification

2) authors' affiliation and address:

LTeN, UMR-6607, Polytech' Nantes, France

&

INSERM, UMR-1087, l'institut du thorax, Nantes, France.

3) e_mail address:

Florent.Autrusseau@univ-nantes.fr

4) Conference & Publisher information:

https://embc.embs.org/2025/

5) bibtex entry:

```
@InProceedings{EMBC25_Quality,
  author    = {F. Autrusseau and R. Nader and M. {El Hassouni} and A. Nouri and N. Mansouri and V. {L'Allinec} and R. Bourcier},
  booktitle = {47th Intl. Conf. of the IEEE Engineering in Medicine and Biology Society (EMBC) },
  date      = {2025-07-14},
  title     = {Synthetic Vascular Image Selection For Deep Learning Based Cerebral Bifurcation Classification},
}
```

Synthetic vascular image selection for deep learning based cerebral bifurcation classification*

Florent Autrusseau[1,2], Rafic Nader[1,3], Mohammed El Hassouni[4], Anass Nouri[5,8], Nesrin Mansouri[1], Vincent L'Allinec[1,6] and Romain Bourcier[1,7]

***Abstract*— Deep learning algorithms rely heavily on large datasets to efficiently perform various pattern recognition tasks. However, collecting ground truth datasets, which include the necessary annotations for training neural networks, is often a challenging and labor-intensive process. Sometimes, in order to alleviate the labeling burden, while still providing high quality augmented data, synthetic models are being used. A properly designed synthetic model can prove very efficient for various pattern recognition tasks, subject to a thorough mimicking of the actual ground truth. However, when exploiting synthetic images, one might encounter significant drawbacks: how can we ensure that the synthetic data contributes positively to the neural network training process ? Is the synthetic image of sufficient quality to be useful ? Or should it be discarded from the training dataset (as it may lessen the CNN learning ability) ?**

In this work, we propose to run a subjective experiment to assess the similarities between vascular bifurcations, we can hence sort various bifurcations, may they be ground truth portions of the vascular tree as acquired on Magnetic Resonance Angiography (MRA) - Time of Flight (ToF) acquisitions, or synthetically modeled bifurcations. Once the subjective experiment was set up and conducted, we tested various objective quality measures to assess the fidelity of the synthetic images. More precisely, these automatic quality estimation metrics were used to remove any malfunctioning synthetic model from the training dataset. A CNN is finally trained on a bifurcation classification task with either the full training set or its reduced version (after quality filtering).

I. INTRODUCTION

A. Motivation and context

This work is a follow-up of our previous study [1], focusing on cerebral vascular bifurcation classification. The Circle of Willis (CoW) is an arterial ring-like structure located at the base of the brain. It is composed of a set of cerebral arteries, splitting into numerous bifurcations. These cerebral bifurcations are of particular interest to neuro-radiologists, as this is where most intracranial aneurysms (ICA) occur. An aneurysm is an outpouching of the arterial wall, a swelling, or bulge of the artery. Specifically, about 85% of all intracranial aneurysms occur on the circle of Willis, and more particularly onto 15 bifurcations (shown with yellow discs in Figure 1). The percentages within the gray discs represent the chance of an aneurysm to occur on the given bifurcation. Thus, knowing that most of the ICAs occur on these particular locations, it is crucial to be able to automatically detect (and classify) these bifurcations. Several works have been devoted to the classification of CoW arteries [2], [3], such a classification is not an easy task, as there is a well known significant variability in the shape of the CoW [4], [5]. Indeed, some arteries may be completely missing for some patients. Although the classification of

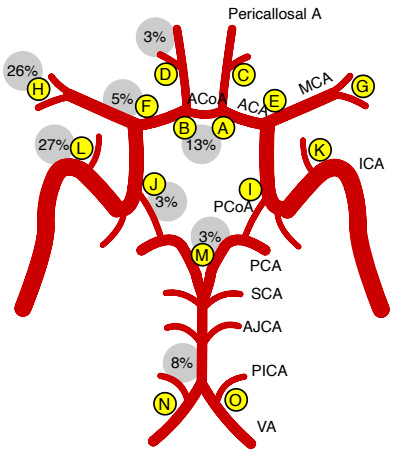

Fig. 1. Schematic representation of the CoW.

CoW arteries has widely been studied, very few works have focused on the CoW bifurcation classification [6]. We have presented in [1] a method aiming to detect and classify 13 bifurcations composing the Circle of Willis. Basically, a U-Net architecture [7], [8] allows to segment the cerebral vascular tree, then 3D patches were extracted around each bifurcation (detected via skeletonization, and 3D graph representation), ultimately, the bifurcation classification was performed through a 3D CNN.

In a later study [9], in the aim to cope with reduced training datasets, we have proposed a fully synthetic vasculature model (VaMos) [10], which goal is to mimic as best as possible the various geometric features of the arterial bifurcations, as well as the surrounding background. Initially, VaMos started as a humanization of mice vascular trees for cerebral arteries segmentation [11], then, it later evolved

*This work was supported by the RHU-ANR project "eCAN" #ANR-23-RHUS-0013, project VaMos4CoDE Prematuration CNRS, and PHC Toubkal 23/162.

[1,2]F. Autrusseau is with the Institut du Thorax (ITX) and LTeN, Univ. of Nantes, France. <FirstName>.<LastName>@univ-nantes.fr [1,3]R. Nader is now with Astek-Direction de la Recherche et de l'Innovation (and was with ITX when this work was completed) [1]N. Mansouri is with the ITX, Univ. of Nantes, France. [4]A. Nouri is with the Setime Lab., Univ. Ibn Tofail, Morocco and Univ. Caen Normandie, ENSICAEN, GREYC, France. [5,8]M. El Hassouni is with the Mohammed V Univ. in Rabat, Morocco. [1,6]V. L'Allinec is with the ITX, Univ. of Nantes, and Univ. hospital, Angers, France. [1,7]R. Bourcier is with the Univ. Hospital in Nantes and ITX lab., Nantes, France.

into a fully synthetic vascular tree model aiming for pattern recognition tasks (bifurcation classification or intracranial aneurysm detection and segmentation). Indeed, this synthetic model proved very useful for both ICA detection and bifurcation classification [9]. However, despite the quite efficient modeling offered by VaMos, it may happen that some generated synthetic images fail to faithfully represent their ground truth. For various reasons (poor segmentation, strong noise amplitude, unsuitable arterial thickness modification, etc.) some arteries could be missing, thus leading to reduced learning performances from the bifurcation classification neural network.

This paper is devoted to the identification of any malfunctioning synthetic models. Let us now present in details the scope of this work, and the various steps that are needed to efficiently filter out any failed synthetic 3D patch.

B. Scope of the current work

As previously explained, our main objective in this work, is to be able to identify any synthetic models that actually failed, and hence, may not be suitable to efficiently train a given Convolutional Neural Network on a pattern recognition task. Ideally, the best way to determine which synthetic patch generation went wrong would be to ask a group of human observers (preferably experts in the field) to assess the quality or similarity between image pairs. Such a process is commonly exploited in the image quality assessment research topics. Indeed, such approaches are used to estimate the perceived quality of compressed images for instance [12].

However, such subjective experiments are rarely employed (if ever) on medical images, as their inherent features (noisy data, low contrast, multi-dimensionality) make it quite difficult to assess the quality. Moreover, due to the complexity of medical images, naive observers can hardly be enrolled during subjective tests, expert observers (doctors of medicine or even radiologists) should be recruited for such a task.

We have nevertheless designed a subjective similarity measure protocol, which aim is to sort various 3D bifurcations with decreasing similarities to a given ground truth. Obviously, such a test cannot be conducted on a whole training dataset, typically made up of thousands of images. Traditionally, subjective tests, involving human observers, are commonly used to design Objective Quality Metrics (OQMs), *i.e.* computer programs, aiming to predict the subjective score provided by humans assessors. In the context of classical (natural) image quality assessment, numerous OQMs have been designed [13], alternatively, one can also find various methods that are particularly applicable to 3D point clouds or 3D meshes [14], [15]; however, very few are devoted to medical images [16], [17], [18].

In this work, we are not actually looking for an efficient objective quality metric (*i.e.*, having the ability to accurately predict a quality score), but we are more precisely interested in finding a metric that would allow us to discard the worst matches. Indeed, when removing the worst synthetic patches from the training dataset, we might be able to improve the CNN performance. Here, we consider bifurcation classification as the final application (as already studied in [1], [9]), but a similar reasoning could be applied to aneurysm detection, CoW classification, or even thrombosis diagnosis, etc.

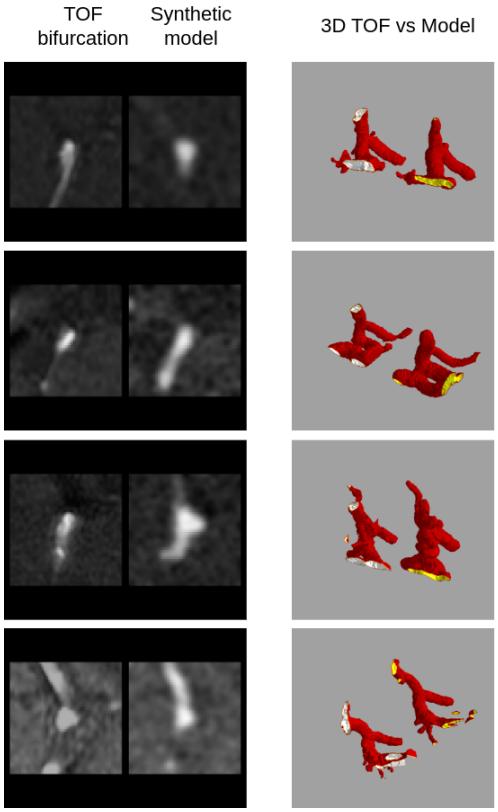

Fig. 2. An example of the diversity we can find in bifurcation geometry.

Figure 2 gives a few examples of cerebral bifurcations (label #E in Fig. 1) from various patients, along with their synthetic models. We can observe a significant variability in their shapes (number of branches, diameters, angles, tortuosity, *etc.*); and hence strong geometric distortions which might be disadvantageous when testing quality metrics, as we will see later on in this work (sec. III).

II. Subjective and Objective Quality Assessment

In the following, we first introduce a new subjective protocol (in section II-A), for which we have enrolled several human observers, then, we experiment various objective quality metrics, of different kinds (sec. II-B). Next, (within sec. III) we will investigate whether the accuracy of the bifurcation classification may be increased by filtering out unreliable images thanks to any of the tested objective quality metrics.

A. Subjective Experiment

Classical subjective quality assessment protocol cannot be straightforwardly applied to medical images. As previously mentioned, such images being noisy, low contrast, and in three dimensions, it is quite difficult for an observer to efficiently judge any resemblance between a pair of images.

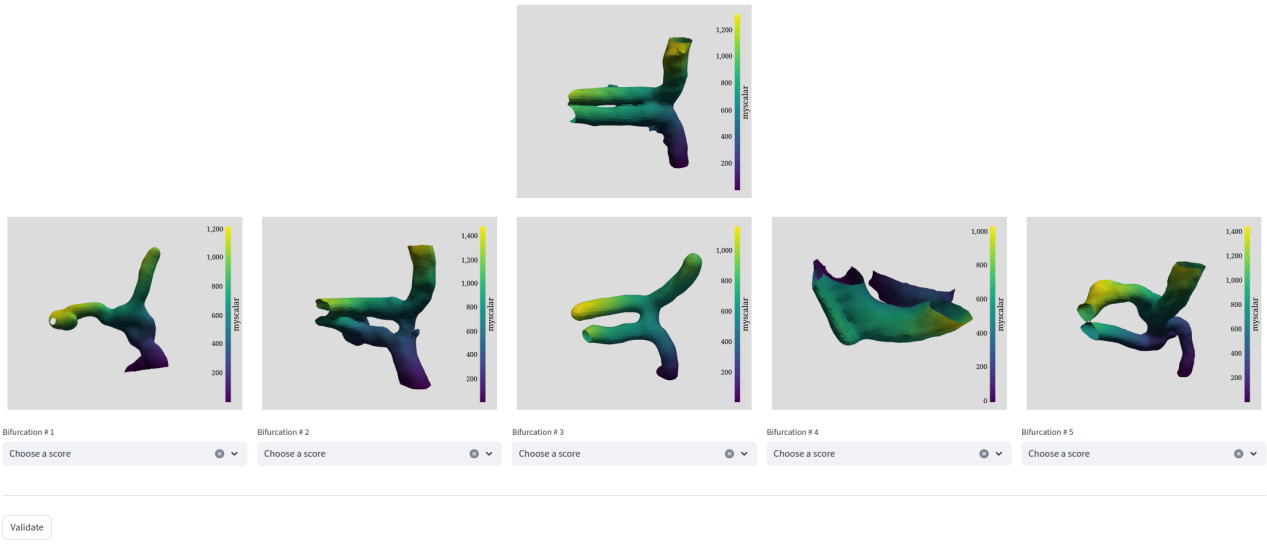

Fig. 3. Layout of the subjective protocol.

We have thus opted for a simplified protocol in which the observers would be asked to rank images by decreasing similarity. The protocol operates as follows: a reference 3D bifurcation is displayed on the top of the screen, and below, aligned onto a second row, a series of 5 candidate bifurcations are shown (Figure 3 shows a typical example of such a display). The subjective experiment runs in a web browser[1], the angiographic images (MRA-ToF) were converted to 3D meshes for a simplified display and manipulation. Underneath each candidate bifurcation, a select box allows the observer to rank the images, from the most similar to the most different image. In order to avoid any repeating patterns possibly induced by the observers' fatigue, all the displays are randomized, *i.e.* within each display screen the positions of the five candidate images are scrambled, and in the course of the experiment, the consecutive displays are also shown in a random order. Hence, the observers do not see the same images at the same time or at the same location. Overall, 480 candidate bifurcations were tested (and compared to 96 reference bifurcations of interest). Specifically, the bifurcations were extracted from the MRA-ToF acquisition within 2 cm wide cubes (50 $voxels$). In order to avoid ending up with a test being too tedious and too long, we had to split the experiment into 4 different sessions. This way, for each test session, the observers had to evaluate 24 consecutive displays (showing a reference bifurcation along with its 5 candidates, as seen on Fig. 3), which took about 20 to 30 minutes per observer and per test session. Among the 480 candidate bifurcations, 240 were gathered from different patient's MRA acquisitions (different from the reference), and 240 were generated by our synthetic vascular model. Overall, 31 human observers have been enrolled (students and staff from the Universities of Nantes, France and Kenitra, Morocco), before running the test, they were briefed about the purpose of our study, and the protocol was explained. Among the 31 human observers, 10 are considered experts in the field as they are involved in project dealing with cerebral arteries (aneurysms, vascular segmentation, biology of the arterial cells, etc.) and hence are all well aware of the CoW geometrical configuration. No significant differences were found between the subjective scores issued by naive and experts subjects. The observers were able to zoom (in/out), shift, and rotate the bifurcation within each 3D panel, they were thus able to align (when possible) the 6 displayed bifurcations, before sorting them by decreasing order of similarity. They were instructed to judge the overall shape of the bifurcations (*i.e.* number of branches, as well as their lengths and orientations).

The entire subjective dataset was thus split into four smaller sets, the scores from 13 observers were collected for set #1, subjective scores from 14 observers composed set #2, 15 for set #3 and finally, set #4 was gathered from 14 different observers. Some of the observers ran several test sessions. Inevitably, when running such a subjective experiment, and especially here, on 3D volumes, some observers may exhibit some inconsistencies with the others. We have thus computed dendrograms separating the observers' scores (see Figure 4) in order to discard incoherent observers from the analysis. Similarly, the subjective dataset (480 candidate bifurcations collected from 96 references) had to be manually composed without any *a priori* knowledge on the actual subjective match with the reference. We thus had to rule out all displays (composed of the 6 test bifurcations) presenting incoherent subjective scorings. Whenever the average standard deviation across all observers and all five images of a given display was above 1.0, the said display (reference bifurcation along with its 5 test candidates) was removed from the subjective dataset. Ultimately, our subjective dataset was composed as

[1]The subjective test was developed in Python, using Streamlit along with the PyVista / stpyvista libraries (`https://streamlit.io/`, `https://docs.pyvista.org/`, `https://github.com/edsaac/stpyvista`)

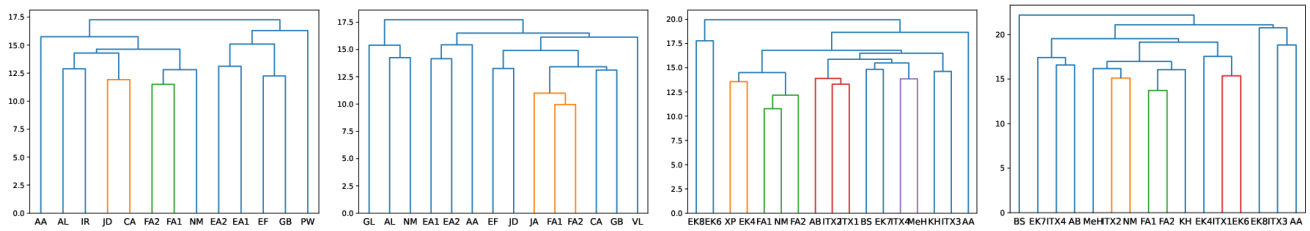

Fig. 4. Observer selection process, using clustering via dendrograms, with euclidean distance computation.

follows: 8, 10, 14 and 12 observers for each test set (when significant distances were witnessed among observers, the ones being located at the extremities of the dendrogram tree were discarded), and 320 test images (composed of 157 patients, and 163 models). Generally speaking, the aim of a subjective experiment is to determine which objective quality metric would best predict observers' assessment of quality or, in our case, bifurcation rankings.

B. Objective quality metrics

Obviously, we cannot rely on a subjective experiment in a normal use-case scenario. It is thus crucial to come up with an Objective Quality Metric (OQM), *i.e.* a software, being able to estimate the subjective score as provided by the human observers. Such metrics can be very diverse in their design as well as in their purpose.

OQMs mostly fall into one of these five categories:

- Statistical metrics,
- Perceptual quality metrics,
- Correlation based metrics,
- 3D-meshes metrics,
- Neural network-based quality evaluation.

As their name implies, statistical metrics would basically compare some statistical features between two given images. Correlation-based metrics (such as the ones being used in 3D registration methods), might be useful in our working scenario due to the inherent geometric distortions. 3D-mesh based quality metrics might also be of interest, and finally, quality metrics based on neural networks can also be of a great help to estimate the subjective scores (especially when geometric distortions are to be considered).

For still 2D natural images and videos, perceptual quality metrics are legion, advanced human visual system features are modeled in order to come up with an accurate estimation of the perceived quality. Things are much more complicated for 3D, noisy, medical images. Indeed, very few OQMs are efficient against geometric distortions, *i.e.* if there is any rotation, shift, or any kind of geometric transform between the reference and test image (especially when it comes to assess 3D volumes/meshes). Therefore, unfortunately, very few quality metrics have been specifically designed for (and are particularly efficient on) medical images. Even from a subjective quality evaluation perspective, our working scenario is very complex, as it basically piles up many problems that may hamper the performances of an objective quality evaluation. Indeed, our dataset is composed of noisy 3D medical images, including actual ToF acquisitions, and synthetic models, and, above all, the bifurcations are collected from different patients. Knowing the significant variability in the shape of the Circle of Willis, very strong geometric distortions are to be expected between the test bifurcations.

In this work, we have tried 12 different quality metrics, three were statistical metrics (PSNR, SSIM [19] and Mattes mutual information [20]), two were correlation metrics (ANTS Correlation [21], and Normalized Cross Correlation - NCC [22]), 5 were mesh-based quality metrics (PCQM [23], RR_NSS-L1 [24], RR_NSS-L2 [24], FR_Kdtree [25] and FMPD [26]), we have also computed the Hausdorff distance (mesh-based) between both inputs [27] and finally, we have also tested the no-reference Multi-Modal Point Cloud Quality Assessment (MM_PCQA) metric [28]. This metric, pre-trained on the Waterloo Point Cloud dataset [29], leverages point-based network encodings of 3D patches and image-based neural network encodings of 2D projections.

The five first cited OQMs operate on grayscale images and are full-reference, *i.e.* need the whole reference and distorted image as inputs. The remaining seven are based on 3D meshes evaluation. RR_NSS is a reduced-reference method that assesses the quality of 3D meshes by comparing natural statistics extracted from features of both reference and distorted meshes [24]. RR_NSS-L1 and RR_NSS-L2 are variants of this method, using L1 and L2 norms, respectively, to measure the similarity between these statistical representations. The FR-Kdtree method accurately measures how much a test mesh has been distorted, comparing to its reference. This is performed by searching for the closest points between the two meshes, using a technique called the k-d tree [25]. FMPD (Fast Mesh Perceptual Distance), is a full reference metric that considers the mesh local roughness measure derived from Gaussian curvature while assessing the similarity between a reference and a distorted 3D mesh [26].

Commonly, when running a joint Objective-Subjective experiment, a good way to express the prediction accuracy, is simply to plot the subjective Mean Opinion Score (MOS) versus the predicted MOS (MOSp) provided by the metrics (see Fig. 6). Ideally, one would expect to have a somewhat linear distribution on the MOS *vs.* MOSp plot. Pearson Correlation Coefficient (PCC), Spearman Rank Correlation Coefficient (SRCC), or Root Mean Square Error (RMSE) are the most widely used performances metrics to assess the OQMs. The subjective experiment we propose here is not quite related to quality, but rather, we aim to rank the

test bifurcations by similarities, hence the above performance metrics might not present a high aptitude in this task. The ultimate performance test would be to assess the pattern recognition task (bifurcation recognition) when potentially failed bifurcations models or unusual patient geometry are encountered, and discarded (see sec. III-B).

III. EXPERIMENTAL RESULTS

A. Objective vs Subjective scores

When running a subjective experiment, as a first analysis step, it is crucial to ensure that the subjective scores are correctly distributed and span the whole quality range. Indeed, a poorly distributed dataset might induce important issues when looking for an efficient objective quality metric that would faithfully estimate the perceived quality. Moreover, as already explained, the composition of our dataset is a bit peculiar, it includes both some actual bifurcations acquired on MRA-ToF images, and also some synthetically modeled images. Hence, it is important to make sure that the two types of images are also well distributed within the subjective quality range.

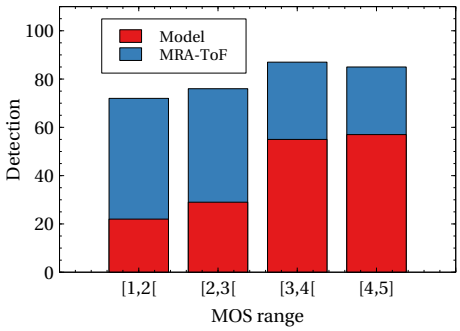

Fig. 5. Repartition of the true bifurcations *vs.* synthetically modeled ones.

We show on Figure 5 for each quality step, how the actual ToF and modeled bifurcations are distributed. It appears that the synthetic images are globally quite well assessed by the human observers, although their ranking seems to be slightly lower than the actual ToF bifurcations. Now that we have ensured that the subjective scores are well distributed, let us evaluate the ability of some common quality metrics and correlation-based measures to predict the subjective scores. As previously explained, we have thus computed the SRCC, the PCC, and RMSE between the subjective MOS, and the predicted scores. Table I shows the performances of each tested OQM on the subjective dataset. Correlation are given in absolute values, as no matter if the correlation is positive (similarity or quality measure) or negative (dissimilarity evaluation). As can be observed on this Table, it seems that not a single OQM is able to very accurately predict the subjective scores. In fact, Figure 6 appears to confirm this statement, as we can observe that the MOS *vs.* MOSp plots are quite erratic, the distribution is very wide, and quite far away from being linear. We only show SSIM and Mattes MI here, but all other OQMs exhibit a similar behavior. However, as we will witness in the next subsection (section III-B),

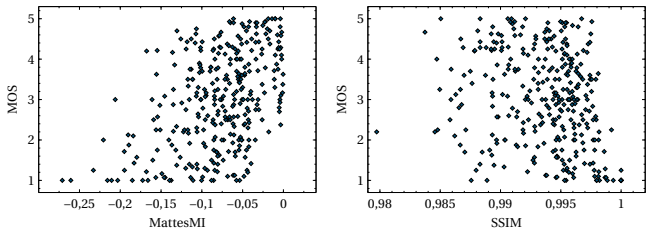

Fig. 6. Two examples of MOS *vs.* MOSp plots (Mattes Mutual Information and SSIM).

such a weak ability to predict the subjective scores might not hinder the quality-based filtering. Again, what is of utmost importance in this work, is the metrics ability to discard the worst quality models.

B. CNN Performances increase

In this work, we shall not present in detail the bifurcation classification method. This latter has been discussed at length in our previous work [1]. Here, we have been using the exact same neural network, but this work being more of a feasibility study, we have restricted the set of bifurcations to be classified. Here, we aim to recognize only bifurcations #A, #B, #E, #F, #K, #L and #M (see yellow labels in Fig. 1). As previously explained in [1], bifurcations #C and #D are rarely present in the acquired MRA-ToF, and lead to an improper learning (this statement also stands for bifurcations #N and #O); this is commonly due to the MRI technician cropping the MRA-ToF to the very center of the Circle of Willis when an aneurysm is know to be located there, and thus leaving out the peripheral portions. Moreover, contrary to what we might expect from Figure 1, bifurcations #I and #J are actually very close to #E and #F, and hence, when detecting the latter, we actually have very good chances to also detect the former. We have also decided to leave bifurcations #G and #H out of this analysis, as these two were the ones leading to most classification errors in our previous analysis. These two bifurcations are actually the source of some disagreement even among neuro-radiologists.

The training set was gathered from 14 ToF acquisitions, hence if all 7 bifurcations were actually present within each ToF volume, we would have ended up having $14 \times 7 = 98$ original bifurcations, but due to missing arteries, our training dataset is composed of 84 original bifurcations only. For each reference bifurcation, the VaMos model was launched with varying parameters, spline coefficient (tortuosity adjustments) were set as 3, 6, 9 and 12. For each parameter, the model was executed three times (in order to generate a wide variety of models, *i.e.* varying diameters, angles, background noise, ...). Ultimately, the training dataset was thus composed of 84 original bifurcations and 980 synthetic models.

In this study, we have tested 12 different OQMs on our subjective dataset (falling in all 5 categories mentioned in sec. II-B). In order to ensure a fair comparison between all quality metrics, we have removed the 100 bifurcations presenting the lowest predicted quality. The training and test

	MM_PCQA	PCQM	FMPD	PSNR	RR_NSS-L1	SSIM	Hausdorff	ANTSCor	NCC	RR_NSS-L2	MattesMI	Kdtree
SRCC	0.0453	0.0736	0.1219	0.1544	0.2260	0.2730	0.3254	0.3864	0.3878	0.4102	0.4281	0.4895
PCC	0.0148	0.0054	0.1342	0.2055	0.2553	0.2513	0.3406	0.4400	0.4454	0.4094	0.4609	0.5022
RMSE	2.8753	3.2211	3.1130	56.0894	4.4587	2.3289	15.6470	3.5308	2.9897	2.8795	3.2901	2.8792

TABLE I

PERFORMANCES METRICS (SRCC, PCC, AND RMSE) FOR EACH TESTED OQM. CORRELATION ARE GIVEN IN ABSOLUTE VALUES, OQMS ARE SORTED BY INCREASING SRCC

	Training set		**Test set**
	(ToF)	(Synthetic)	(ToF)
#A	13	156	15
#B	11	132	14
#E	14	154	15
#F	14	154	15
#K	10	120	13
#L	9	108	15
#M	13	156	14
Sum	**84**	**980**	**101**

TABLE II

COMPOSITION OF BOTH THE TRAINING AND TEST DATASETS (NUMBER OF CROPPED PORTIONS OF THE MRA-TOF REPRESENTING EACH LABEL).

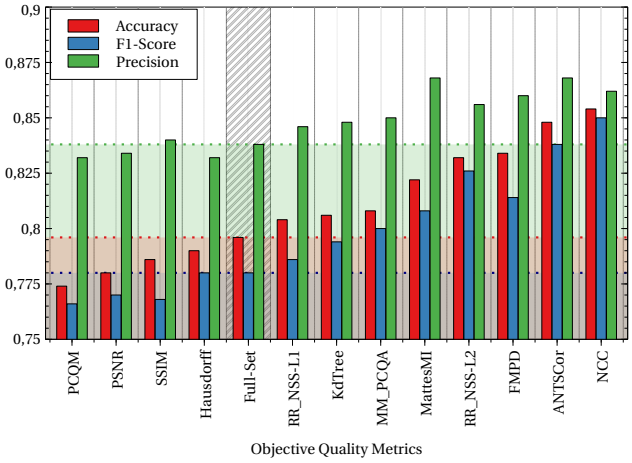

Fig. 7. Performances of the bifurcation classification CNN (Accuracy, F1-Score and Precision).

datasets are composed as detailed in Table II. We have managed to have each and every considered bifurcation label uniformly distributed. The training step was performed through 5 folds. Figure 7 presents the classification performances (Accuracy, Precision, F1-Score) reached thanks to each and every tested quality metric. In this figure, we have sorted the quality metrics by increasing CNN accuracy. Besides all tested OQMs, the plot also features the performances achieved when using the entire (unfiltered) dataset ("Full-Set") highlighted within the hatched area. Hence, all OQMs positioned to the left of this hatched area actually result in a loss of performances (in terms of accuracy only here), whereas all metrics positioned to the right, lead to improved performances. If the ranking was applied on the other metrics (increasing F1-Scores or Precision), the order of the OQMs would remain fairly similar. For an easy comparison of any improvement or deterioration, the horizontal dotted lines (along with the filled area) project this "Full Dataset" baseline throughout the x-axis. Wherever a bar is above its corresponding dotted line (*i.e.* of a corresponding color), this means we were able to improve this particular performance. Obviously, when a given bar is below the dotted line (of similar color), the performance was decreased. As can be observed on this plot, the correlation-based metrics (ANTSCor and NCC) offer the best performance increase (with respect to all three performances metrics), whereas SSIM, PSNR and PCQM actually lessen the overall performances. This does not quite come as a surprise, as the correlation metrics are relatively resilient to geometric distortions (to a certain extent), while also comparing the geometric shape (which the statistical metrics cannot achieve).

When training on the full dataset across all objective quality metrics (OQMs), the model maintains consistently high precision, but the F1-score exhibits significant variability. This indicates low sensitivity (recall), as the model correctly avoids false positives but fails to identify all true positives. The model, only predicts a class when it is highly confident. This can be attributed to ambiguity in the bifurcations or poor representation caused by certain alterations in the synthetic model. These issues likely make some instances harder for the model to classify correctly.

By using correlation-based metrics such as ANTSCor or NCC, the images selected for training likely offer a clearer representation of their respective classes, minimizing ambiguity and noise. This enables the model to better learn the discriminating features of the bifurcation, resulting in a noticeable improvement in the F1-score. The consistently high precision proves the model continues to effectively avoid false positives.

IV. DISCUSSION AND CONCLUSION

This work is dedicated to a joint subjective and objective quality evaluation of 3D cerebral vascular arteries and bifurcations. In fact, behind this quality assessment framework, we aim to optimize the learning step of a bifurcation classification method. A Convolutional Neural Network is being used to classify the cerebral bifurcations along the Circle of Willis, we feed this CNN for its training step with both actual (patient) acquisitions from magnetic resonance angiography images and with synthetically modeled patches. Indeed, we have designed a fully synthetic model (VaMos) able to replicate Time of Flight images and subsequently generate massively annotated datasets. However, although such

modeled patches are globally functioning rather well, as they can significantly improve the classification performances, we want to automatically remove any bad representations of the synthetic model at the CNN input. In other words, we aim to clean up the training dataset for an optimized learning. We have thus designed a subjective protocol where human observers are asked to rank a set of candidate 3D bifurcations with respect to a given reference. Once such a subjective dataset gathered, we have identified a set of Objective Quality Metrics (computer programs, aiming to predict a subjective score) that might best estimate the Mean Opinion Scores provided by the observers. Identifying a good performance quality metric might help us to filter out any malfunctioning synthetic model. Unfortunately, our dataset is composed of very difficult images to deal with; Indeed the input images are 3D, noisy, of low contrast, bad quality and most importantly with strong geometric distortions. Hence, very few OQMs have ever been designed considering all these constraints. However, although the subjective quality appeared to be very challenging, we could nevertheless identify a few metrics that were able to efficiently discard some of the most unreliable synthetic models. We could witness such an efficient filtering by scrutinizing the performances improvement reached by the classification CNN on OQM-based cleaned up input datasets. Although the observed improvements were quite modest (Accuracy going from 0.796 for the Full-set baseline to 0.854 when filtering images with NCC, Precision going from 0.838 to 0.868 or F1-Score leaping from 0.780 to 0.850), still, it is encouraging to witness possible improvements via the use of OQMs. A properly designed quality metric (specifically intended to deal with the various constraints at hand) might lead to more significant performance improvement. Such a work will be the subject of a forthcoming study.

REFERENCES

[1] R. Nader, R. Bourcier, and F. Autrusseau, "Using deep learning for an automatic detection and classification of the vascular bifurcations along the circle of willis," *Medical image Analysis*, 2023.

[2] F. Dumais, M. P. Caceres, F. Janelle, K. Seifeldine, N. Arès-Bruneau, J. Gutierrez, C. Bocti, and K. Whittingstall, "eICAB: A novel deep learning pipeline for circle of willis multiclass segmentation and analysis," *NeuroImage*, vol. 260, p. 119425, 2022.

[3] K. Yang, F. Musio, Y. Ma, *et al.*, "Benchmarking the CoW with the TopCoW challenge: Topology-aware anatomical segmentation of the circle of willis for CTA and MRA," *arXiv:2312.17670v3*, 2024.

[4] L. B. Hindenes, T. Ingebrigtsen, J. G. Isaksen, A. K. Håberg, L.-H. Johnsen, M. Herder, E. B. Mathiesen, and T. R. Vangberg, "Anatomical variations in the circle of willis are associated with increased odds of intracranial aneurysms: The tromso study," *Journal of the Neurological Sciences*, vol. 452, p. 120740, 2023.

[5] N. Stojanović, A. Kostić, R. Mitić, L. Berilažić, and M. Radisavljević, "Association between circle of willis configuration and rupture of cerebral aneurysms," *Medicina (Kaunas)*, vol. 55(7), 2019.

[6] I. Essadik, A. Nouri, R. Touahni, R. Bourcier, and F. Autrusseau, "Automatic classification of the cerebral vascular bifurcations using dimensionality reduction and machine learning," *Neuroscience Informatics*, 2022.

[7] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: learning dense volumetric segmentation from sparse annotation," in *Intl. Conf. on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2016, pp. 424–432.

[8] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Intl. Conf. on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 234–241.

[9] R. Nader, V. L'Allinec, R. Bourcier, and F. Autrusseau, "Vascular models (VaMos): Application to bifurcation classification and aneurysm detection," in *Intl. Conf. on Pattern Recog. (ICPR)*, 2024.

[10] R. Nader, F. Autrusseau, V. L'Allinec, and R. Bourcier, "Building a synthetic vascular model: Evaluation in an intracranial aneurysms detection scenario," *IEEE Transactions on Medical Imaging*, 2024.

[11] S. Chater, N. Lauzeral, A. Nouri, Y. El Merabet, and F. Autrusseau, "Learning from mouse ct-scan brain images to detect mra-tof human vasculatures," in *43rd IEEE EMBC*, 2021.

[12] M. Testolina and T. Ebrahimi, "Review of subjective quality assessment methodologies and standards for compressed images evaluation," in *Applications of Digital Image Processing*, vol. 11842, Intl. Society for Optics and Photonics. SPIE, 2021, p. 118420Y.

[13] X. Min, G. Zhai, J. Zhou, M. C. Q. Farias, and A. C. Bovik, "Study of subjective and objective quality assessment of audio-visual signals," *IEEE Trans. on Image Processing*, vol. 29, pp. 6054–68, 2020.

[14] E. Alexiou, E. Upenik, and T. Ebrahimi, "Towards subjective quality assessment of point cloud imaging in augmented reality," in *IEEE 19th Intl Workshop on Multimedia Signal Processing*, 2017, pp. 1–6.

[15] G. Lavoué, "A multiscale metric for 3d mesh visual quality assessment," *Computer Graphics Forum*, vol. 30, 2011.

[16] R. Rodrigues, L. Lévêque, J. Gutiérrez, *et al.*, "Objective quality assessment of medical images and videos: review and challenges," *Multimed Tools Appl.*, 2024.

[17] A. Mason, J. Rioux, S. E. Clarke, A. Costa, M. Schmidt, V. Keough, T. Huynh, and S. Beyea, "Comparison of objective image quality metrics to expert radiologists' scoring of diagnostic quality of mr images," *IEEE Trans. Med Imaging*, vol. 39(4), pp. 1064–72, 2020.

[18] R. Thanki, S. Borra, N. Dey, and A. S. Ashour, *Medical Imaging and Its Objective Quality Assessment: An Introduction*. Cham: Springer International Publishing, 2018, p. 3–32.

[19] A. Horé and D. Ziou, "Image quality metrics: PSNR vs. SSIM," in *20th Intl Conference on Pattern Recognition*, 2010, pp. 2366–2369.

[20] D. Mattes, D. Haynor, H. Vesselle, T. Lewellen, and W. Eubank, "Non-rigid multimodality image registration," in *SPIE Medical Imaging*, July, 3rd 2001, pp. 1609–1620.

[21] A. Brian, T. Nicholas, S. Gang, C. Philip, K. Arno, and G. James, "A reproducible evaluation of ANTs similarity metric performance in brain image registration," *NeuroImage*, vol. 54(3), pp. 2033–44, 2011.

[22] J. P. Lewis, "Fast template matching," *Vision Interface*, pp. 120–123, 1995.

[23] G. Meynet, Y. Nehme, J. Digne, and G. Lavoué, "PCQM: A full-reference quality metric for colored 3D point clouds," in *Intl. Conf. on Quality of Multimedia Experience (QoMEX)*, Ireland, 2020.

[24] Z. Zhang, W. Sun, X. Min, T. Wang, W. Lu, W. Zhu, and G. Zhai, "A no-reference visual quality metric for 3d color meshes," in *IEEE Intl. Conf. on Multimedia & Expo Workshops*, 2021, pp. 1–6.

[25] N. Ray, D. Shevitz, Y. Li, R. Garimella, A. Herring, E. Kikinzon, K. Lipnikov, H. Rakotoarivelo, and J. Velechovsky, "Efficient kd-tree based mesh redistribution for data remapping algorithms," in *SIAM Intl. Meshing Roundtable 2023*. Springer Nature, 2024, pp. 25–41.

[26] K. Wang, F. Torkhani, and A. Montanvert, "A fast roughness-based approach to the assessment of 3d mesh visual quality," *Computers & Graphics*, vol. 36, no. 7, p. 808–818, 2012, augmented Reality Computer Graphics in China.

[27] N. Aspert, D. Santa-Cruz, and T. Ebrahimi, "Mesh: measuring errors between surfaces using the hausdorff distance," in *IEEE Intl. Conf. on Multimedia and Expo*, vol. 1, 2002, pp. 705–708 vol.1.

[28] Z. Zhang, W. Sun, X. Min, Q. Zhou, J. He, Q. Wang, and G. Zhai, "MM-PCQA: Multi-modal learning for no-reference point cloud quality assessment," *IJCAI*, 2023.

[29] Q. Liu, H. Su, Z. Duanmu, W. Liu, and Z. Wang, "Perceptual quality assessment of colored 3d point clouds," *IEEE Trans. on Visualization and Computer Graphics*, vol. 29, no. 8, pp. 3642–3655, 2023.