Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

Comparing methods of detecting and segmenting unruptured intracranial aneurysms on TOF-MRAS: The ADAM challenge

Kimberley M. Timmins[a,*], Irene C. van der Schaaf[b], Edwin Bennink[a], Ynte M. Ruigrok[c], Xingle An[d], Michael Baumgartner[e], Pascal Bourdon[f,g], Riccardo De Feo[h,i], Tommaso Di Noto[j,k], Florian Dubost[l], Augusto Fava-Sanches[m], Xue Feng[n], Corentin Giroud[l], Inteneural Group[o,†], Minghui Hu[p], Paul F. Jaeger[e], Juhana Kaiponen[i], Michał Klimont[o,q], Yuexiang Li[r], Hongwei Li[s,t], Yi Lin[r], Timo Loehr[s], Jun Ma[u], Klaus H. Maier-Hein[e,v], Guillaume Marie[j,k], Bjoern Menze[t,s], Jonas Richiardi[j,k], Saifeddine Rjiba[w,f], Dhaval Shah[x], Suprosanna Shit[x], Jussi Tohka[i], Thierry Urruty[f,g], Urszula Walińska[o], Xiaoping Yang[u], Yunqiao Yang[y], Yin Yin[p], Birgitta K. Velthuis[b], Hugo J. Kuijf[a]

[a] *Image Sciences Institute, University Medical Center Utrecht, Utrecht, the Netherlands*
[b] *Department of Radiology, University Medical Center Utrecht, Utrecht, the Netherlands*
[c] *Department of Neurology and Neurosurgery, UMC Utrecht Brain Center, University Medical Center Utrecht, Utrecht, the Netherlands*
[d] *China Electronics Cloud Brain (Tianjin) Technology CO., LTD, 300309 PR China*
[e] *Division of Medical Image Computing, German Cancer Research Center, Heidelberg, Germany*
[f] *Xlim Laboratory, University of Poitiers, UMR CNRS 7252, Poitiers*
[g] *Poitiers University Hospital, CHU, Poitiers*
[h] *Sapienza Università di Roma, 00184 Rome Italy*
[i] *A.I. Virtanen Institute for Molecular Sciences, University of Eastern Finland, 70210 Kuopio, Finland*
[j] *Department of Radiology, Lausanne University Hospital*
[k] *University of Lausanne, Rue du Bugnon 46, Lausanne, CH, 1011*
[l] *Zelos Mediacorp, Rotterdam, the Netherlands*
[m] *Institute of Neuroradiology, University Hospital LMU, Munich, Germany*
[n] *University of Virginia, Biomedical Engineering, Thornton Hall, P.O. Box 400259, Charlottesville, VA, USA, 22904-4259*
[o] *Inteneural Networks, Warsaw, Poland*
[p] *Union Strong (Beijing) Technology Co. Ltd., DaZu Plaza T3-901, No. 2 Ronghua South Road, Beijing Economic Technological Development Area, Beijing, China, 100176*
[q] *Department of Radiology, Poznań University of Medical Sciences, Poznań, Poland*
[r] *Tencent Jarvis Lab, Shenzhen, China*
[s] *Department of Computer Science, Technical University of Munich*
[t] *Department of Quantitative Biomedicine, University of Zurich.*
[u] *Department of Mathematics, Nanjing University of Science and Technology, Nanjing, 210094 PR China*
[v] *Pattern Analysis and Learning Group, Department of Radiation Oncology, Heidelberg University Hospital, Heidelberg, Germany*
[w] *Canon Medical Systems, France*
[x] *Department of Informatics, Technische Universität München, Munich, Germany*
[y] *Huazhong University of Science and Technology, Wuhan, China*

ARTICLE INFO

Keywords:
Challenge
Segmentation
Detection
Aneurysms
Angiography

ABSTRACT

Accurate detection and quantification of unruptured intracranial aneurysms (UIAs) is important for rupture risk assessment and to allow an informed treatment decision to be made. Currently, 2D manual measures used to assess UIAs on Time-of-Flight magnetic resonance angiographies (TOF-MRAs) lack 3D information and there is substantial inter-observer variability for both aneurysm detection and assessment of aneurysm size and growth. 3D measures could be helpful to improve aneurysm detection and quantification but are time-consuming and

[*] Corresponding author.
E-mail address: k.m.timmins@umcutrecht.nl (K.M. Timmins).

[†] Inteneural Group consists of collaborators: Mikolaj A. Pawlak and Marek Kraft of Poznan University of Medical Sciences.

https://doi.org/10.1016/j.neuroimage.2021.118216.
Received 29 January 2021; Received in revised form 11 May 2021; Accepted 26 May 2021
Available online 27 May 2021.

would therefore benefit from a reliable automatic UIA detection and segmentation method. The Aneurysm Detection and segMentation (ADAM) challenge was organised in which methods for automatic UIA detection and segmentation were developed and submitted to be evaluated on a diverse clinical TOF-MRA dataset.

A training set (113 cases with a total of 129 UIAs) was released, each case including a TOF-MRA, a structural MR image (*T1, T2* or FLAIR), annotation of any present UIA(s) and the centre voxel of the UIA(s). A test set of 141 cases (with 153 UIAs) was used for evaluation. Two tasks were proposed: (1) detection and (2) segmentation of UIAs on TOF-MRAs. Teams developed and submitted containerised methods to be evaluated on the test set. Task 1 was evaluated using metrics of sensitivity and false positive count. Task 2 was evaluated using dice similarity coefficient, modified hausdorff distance (95th percentile) and volumetric similarity. For each task, a ranking was made based on the average of the metrics.

In total, eleven teams participated in task 1 and nine of those teams participated in task 2. Task 1 was won by a method specifically designed for the detection task (i.e. not participating in task 2). Based on segmentation metrics, the top two methods for task 2 performed statistically significantly better than all other methods. The detection performance of the top-ranking methods was comparable to visual inspection for larger aneurysms. Segmentation performance of the top ranking method, after selection of true UIAs, was similar to interobserver performance. The ADAM challenge remains open for future submissions and improved submissions, with a live leaderboard to provide benchmarking for method developments at https://adam.isi.uu.nl/.

1. Introduction

Approximately 3% of the world general population have an unruptured intracranial aneurysm (UIA) (Vlak et al., 2011). For some risk groups they are even more common, with a prevalence of approximately 10% in individuals with a positive family history for aneurysmal subarachnoid haemorrhage (aSAH) (Bor et al., 2014). Rupture of an intracranial aneurysm causes an aSAH which is a severe type of stroke. Approximately one-third of patients die, and another third have long-term, life-changing disabilities (Keedy, 2006; Nieuwkamp et al., 2009). During screening, it is important that UIAs are detected early, to allow for a treatment decision to be made. From diagnosis, the risk of growth and rupture of the UIA can be determined based on accurate measurement and assessment (Backes et al., 2017; Greving et al., 2014). If an aneurysm has high risk of rupture it will be treated preventively. Aneurysms with a lower rupture risk will be followed-up with imaging and carefully monitored to assess aneurysm growth which is an important determinant for aneurysm rupture (Backes et al., 2015). This allows informed treatment decisions to be made (Wardlaw and White, 2000). Due to the increasing availability and quality of brain imaging, the number of incidentally discovered UIAs is increasing, and follow up imaging is usually performed (Brown and Broderick, 2014; Nakagawa et al., 2019). Also, screening for UIAs with MRA is increasing with knowledge of risk factors for UIA presence. Screening for UIAs with MRA has been shown to be cost-effective in persons with a positive family history for aSAH and in persons with autosomal dominant polycystic kidney disease (Bor et al., 2010; Flahault et al., 2018; Hopmans et al., 2016). The most common imaging techniques for monitoring UIAs are contrast-enhanced computed tomography angiography (CTA) and non-contrast 3D time-of-flight magnetic resonance angiography (TOF-MRA). TOF-MRA is well suited for routine follow-up imaging as it does not need contrast agent or radiation (Lane et al., 2015).

The detection and measurement of UIAs can be difficult and it has been reported that approximately 10% of all UIAs are missed during screening (Forbes et al., 1996; Kim et al., 2017; Keedy, 2006; White et al., 2000). Detection is particularly difficult for small UIAs and detection by radiologists from MRAs of UIAs < 5 mm on MRAs can have a sensitivity as low as 35% (White et al., 2001). However, detection by radiologists is improving as MRA scan resolution is increasing, especially with higher field strengths (HaiFeng et al., 2017; Wrede et al., 2017). In clinical practice, aneurysm detection is performed by a radiologist carefully searching through the axial slices of the TOF-MRA, often combined with coronal and sagittal multi-planar reconstructions, a maximum intensity projection (MIP) or 3D volume reconstruction, before making 2D size measurements of the aneurysm.

As more individuals are followed-up or screened, the speed of clinical workflow could be increased with automatic methods of detection and quantification of UIAs from TOF-MRAs. However, it is important that these methods do not compromise the accuracy of human observers for the detection and measurement of UIAs. Automated volumetric segmentation of UIAs would enable 3D quantification of UIAs and may aid the prediction of UIA rupture risk. For example, it is known that the shape of an UIA, such as non-spherical and lobular shape, are related to an increase in growth and rupture risk (Backes et al., 2017; Lindgren et al., 2016; Raghavan et al., 2009). Furthermore, quantified shape measurements of the UIAs may aid in models assessing treatment complication risk (Ji et al., 2016).

There are numerous different methods for the (semi-) automatic detection and segmentation of UIAs. Semi-automatic methods include, defining the neck of the aneurysm where it attached to the parent vessel, before segmenting the aneurysm (Cardenes et al., 2011). The shape of the aneurysm has been used in some UIA detection techniques, including using blobness filters (Hentschke et al., 2012) and shape analysis of the surface of the vessel segmentations (Arimura et al., 2006; Bizjak et al., 2021; Lawonn et al., 2019). Furthermore, multiple deep learning techniques for UIA detection have been developed with high accuracy (Faron et al., 2019; Nishimori et al., 2018; Park et al., 2019). However, most methods are developed for CTA or Digital Subtraction Angiography (DSA) 2D images (Duan et al., 2019; Sulayman et al., 2016) and are for UIA detection only. The segmentation of UIAs is a difficult problem as UIAs can occur at many different locations and positions relative to the vessels.They are small and can vary greatly in shape and configuration. TOF-MRAs can also vary significantly during the time between baseline and follow-up scans, due to the use of different scanners, protocols, field strengths and field of view. This all leads to a basic requirement for accurate UIA detection and segmentation methods on TOF-MRA.

The Aneurysm Detection And segMentation (ADAM) Challenge described in this paper provides an overview of methods to fully automatically detect and segment UIAs from clinical TOF-MRA images (Timmins et al., 2020). The aim was to compare methods and assess the performance over clinical data from an in-house test set. Evaluation was performed by ranking the methods against each other, for both the detection and segmentation of UIAs, by determining detection and segmentation metrics. This paper provides an overview of the challenge including the organisation, the results, a detailed evaluation of methods submitted and their performance on the test data. This paper follows the structure outlined in the Biomedical Image Analysis challengeS (BIAS) guidelines for transparent reporting of biomedical image analysis challenges (Maier-Hein et al., 2020).

2. Material and methods

2.1. Challenge Organisation

The results of the ADAM Challenge 2020 were presented at the 23rd International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) on October 8th, 2020. From 3rd April 2020, participants could register on the website (http://adam.isi.uu.nl/) to participate in the challenge. They could download a training dataset (for full details on the data, see Section 2.3) to train and develop fully automatic methods for the challenge. Participants were also allowed to use their own training data, as long as they referenced this in their method descriptions. Once trained, methods were containerised by participants with Docker (Merkel, 2014) and submitted to the organiser. Examples and instructions are provided on the website (http://adam.isi.uu.nl/methods/). The containerisation allowed easy evaluation of the methods, guaranteeing it could be run on our platform. Submitted containers were run on an individual training case from the training dataset, containing UIAs, and the results were sent back to the participant for verification. If technical issues or bugs occurred, teams were allowed to resubmit a new version with the bugs fixed.

The final verified, submitted methods were evaluated on a test set of images (see Section 2.3) using evaluation code that was made publically available (https://github.com/hjkuijf/ADAMchallenge). If the method required, NVIDIA Titan Xp GPUs were used for evaluation. The deadline for submission for consideration for the challenge leaderboard at MICCAI was 17th August 2020 and the results and awards were announced at the MICCAI conference (8th October 2020). However, the challenge continues to remain open for submissions, with an up-to-date online leaderboard to allow for benchmarking of the methods. The ADAM challenge was advertised on the MICCAI website, various social media platforms, and via email to previous MRBrainS and WMH challenge participants (Kuijf et al., 2019; Mendrik et al., 2015).

2.2. Mission of the challenge

The ADAM Challenge consists of two tasks. Task 1 had the aim of automatic detection of UIAs on TOF-MRAs. Task 2 was for a method that could perform automatic segmentation of UIAs on TOF-MRAs. Participants could submit to one or both tasks, and methods submitted to task 2 were also assessed for task 1. The target cohort is the term used to describe the patient group of which data would be acquired for the final application of the submitted methods (Maier-Hein et al., 2020). For the ADAM Challenge the target cohort was any patient undergoing a clinical brain TOF-MRA to screen for the presence of an UIA. To reflect the clinical setting, some MRA scans were negative (i.e. a patient without any diagnosed UIAs) and some scans had more than one UIA. A patient in the target cohort may be scanned for the following reasons: (1) follow-up scans of patients with diagnosed UIA(s), with or without additional treated aneurysms; and (2) patients screened for positive family history of UIAs or aSAH. The challenge cohort is the term used to describe is the patient group of which the challenge data was acquired, for both the training and the test datasets (Maier-Hein et al., 2020). The challenge cohort consists of a subset of patients, who had an available TOF-MRA, from a cohort of patients at the University Medical Center (UMC), Utrecht with at least one diagnosed UIA and cohorts of persons screened for UIAs because of a positive family history for aSAH. The assessment aim of the challenge is to find a method that performs optimally for the automatic detection and segmentation of UIAs from the TOF-MRAs in the challenge cohort test dataset.

2.3. Challenge data sets

A total of 254 brain TOF-MRA scans were included with 282 untreated UIAs. The training dataset provided to participants consisted of 113 training cases, while the test dataset consisted of 141 cases, where each case contained a TOF-MRA and a structural image (either *T1*-, *T2*-weighted or FLAIR). All MRIs were performed at the UMC Utrecht, the Netherlands, on a variety of Philips scanners with field strength of either 1, 1.5 or 3T. The MRAs had an in-plane voxel spacing range of (0.195–1.04) mm and slice thickness range of (0.4–0.7) mm, without a set acquisition protocol. This was due to the clinical nature of the data and that it was taken from several studies across a long period of time (between 2001 and 2019). The subjects with UIAs ($N = 53$) had a median age of 55 years (range 24–75 years), with 75% of subjects being female. A subset ($N = 156$) of the dataset includes two scans from the same subject, both a baseline and a >6 month follow-up scan, to reflect the real clinical data. The UIAs ranged in size, with a median maximum diameter of 3.6 mm and a range from 1.0–15.9 mm. 25% ($N = 52$) of the scans contain multiple UIAs and 28% of the scans contained treated (either coiled or clipped) UIAs ($N = 59$). The median age of the population without UIAs was 41 years (range 19–61 years) and 65% were female. This reflects the clinical setting, as UIAs are more common in females and the older generation (Vlak et al., 2011). The dataset was realistic and diverse, reflecting different standard clinical protocols used between MR-scanners and over time.

2.3.1. Training and test data

Subjects were randomly split into training and test sets and it was ensured that both sets contained an adequate number of scans without any UIAs. Every case in the dataset contained one TOF-MRA and one structural (*T1*/*T2*/FLAIR) MR image of the same subject. The training dataset consisted of 113 cases: 93 cases containing at least one untreated, UIA (35 baseline and 35 follow-up cases of the same subject and 23 cases of unique subjects) and 20 cases of subjects without UIAs. The test dataset consists of 141 cases: 115 cases containing at least one untreated UIAs (43 baseline and 43 follow-up cases of the same subject and 29 cases of unique subjects) and 26 cases of subjects without UIAs. The training data is available on the challenge website and requires a registration and acceptance of our terms of distribution. An example of a provided training case can be seen in Fig. 1. A specific validation set was not provided and it is up to the participants to decide their own train/validation set split. Statistical tests were performed to ensure both training and test sets had a fair distribution of scans. An unpaired t-test was used to assess this difference in age, maximum diameter, and number of UIAs, number of treated UIAs, pixel spacing and slice spacing. Gender was assessed using Fisher's exact test, and the Chi-square test was used to assess location and magnetic field strength. The location categories used were: anterior cerebral or communicating artery (ACA/ACoA), the internal carotid artery (ICA), posterior communicating artery (PCoA), middle cerebral artery (MCA) and posterior circulation.

2.3.2. Pre-processing

All images were pre-processed with N4 bias-field correction (Tustison et al., 2011). The structural image was aligned to the corresponding TOF-MRA using the elastix toolbox for image registration (Klein et al., 2010). The transformation parameters used were provided with the training data. Both original and pre-processed data was provided to the registered participants.

2.3.3. Annotation procedure

All UIAs were diagnosed on the scans as part of clinical routine. The UIAs were manually segmented from the original TOF-MRAs using in-house developed software implemented in MeVisLab (MeVis Medical Solutions AG, Bremen, Germany). A contour was drawn around the outline of the UIA, on all axial slices of the MRA. The parent vessel and any branching vessels were excluded from the annotation and annotations were always drawn starting from the UIA neck to the UIA dome. An experienced interventional neuro-radiologist (> 10 years of experience) trained a second rater with considerable experience in medical image analysis and annotation software, but not specifically UIAs. The trained second rater annotated all images in the dataset. Finally, the first and

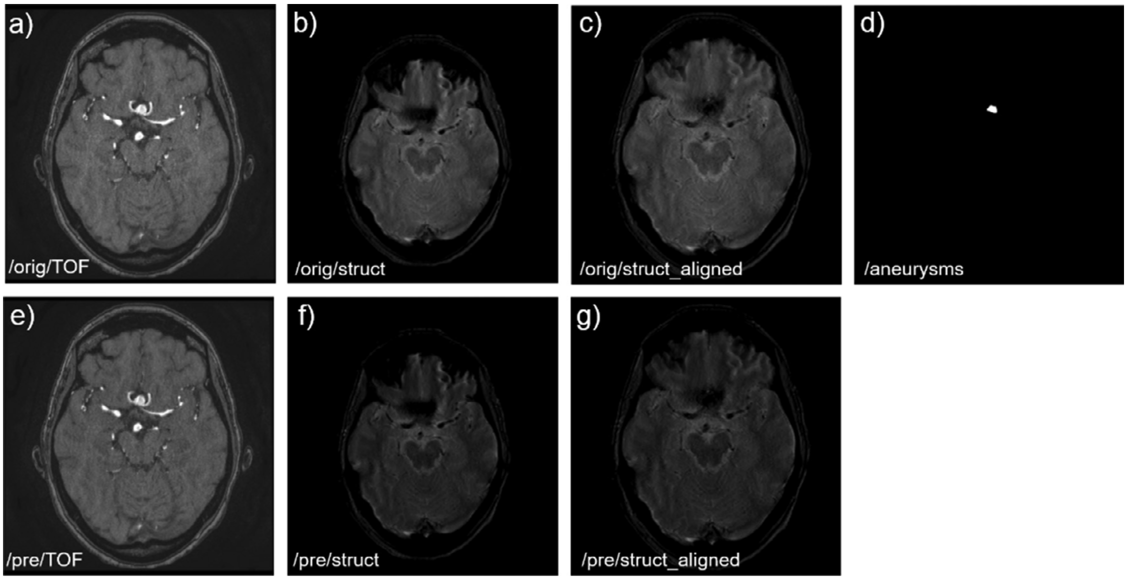


Fig. 1. An example training case.
Top Row: a) Original TOF-MRA, b) Original structural MR image c) original structural MR image aligned to TOF-MRA in a) using registration parameters derived from elastix (Klein et al., 2010) d) binary aneurysm image derived from annotations as described in Section 2.3.3
Bottom Row: e) pre-processed TOF-MRA using n4 bias field correction (Tustison et al., 2011), f) structural MR image preprocessed using n4, g) pre-processed structural MR image aligned to TOF-MRA using pre-determined registration parameters.

second rater assessed the full dataset together and made required modifications to the annotations in consensus to form the official ground truth data set. During annotation, the raters had access to the structural image and a radiologist report made at the time of the scan, indicating the location and size of the UIA. The same annotation procedure was performed for all treated UIAs and dilated to create a slightly larger mask for exclusion of treated aneurysm.

The resulting annotations were converted to binary masks and voxels were considered part of the UIA if they were >50% inside the contour. Untreated UIAs were given the label 1, treated UIAs label 2 and background was labelled 0. From the binary image, the centre of mass and maximum diameter of each of the untreated UIAs were determined in voxel coordinates in the corresponding TOF-MRA image space. This was provided in a text file for each training case.

2.4. *Assessment method*

2.4.1. *Metrics and ranking*

Task 1 and task 2 were evaluated separately using different metrics. All submitted methods for task 2 were also evaluated for task 1, where the centre of mass of 3D connected components in the image was used to determine the detection metrics.

For task 1, methods were evaluated by determining two detection metrics: (1) Sensitivity and (2) False Positive Count (the total number of false positives per scan). The sensitivity gives a measure of how many detected UIAs correspond to true UIAs, ensuring we optimise to detect as many of the UIAs as possible. False positive count balances the sensitivity ensuring not too many falsely identified UIAs are detected, which would not aid the radiologist.

For task 2, methods were evaluated by determining three segmentation metrics: (1) Dice Similarity Coefficient (DSC), (2) Modified Hausdorff Distance (MHD) (95th percentile) and (3) Volumetric Similarity (VS) (Taha and Hanbury (2015).) DSC describes how much the prediction and ground truth segmentations overlap. If there was no detection of UIAs, then the DSC was zero. MHD is a distance metric which is sensitive to the shape of the segmentation. This is important when segmenting UIAs as the shape may be used to assess rupture risk. MHD was only calculated where there was any detection of UIAs by the method, if there was no detection then it was ignored. VS assesses the similarity in volume of the predicted and ground truth segmentation. Accurate volume segmentation is important for UIAs for growth assessment.

Individual UIAs were defined as 3D connected components. A detection was considered positive when the predicted coordinate was within the maximum diameter of the location of the centre of mass of the ground truth UIA.

A similar ranking was performed for both tasks. Teams were ranked per metric. The rankings were averaged to achieve the overall ranking per task. For each team, each metric was averaged over all test scans containing UIAs, other than false positive count, which was evaluated over all test scans, independent of UIA presence. Next, for each average metric, the participating teams were ordered from best to worst. The metrics were scaled linearly to a number between 0 (corresponding to the best team) and 1 (worst team) and then averaged to obtain a single 'rank'. For task 1 the two detection ranks were averaged, and for task 2, the three segmentation ranks were averaged. For task 2, average interobserver segmentation metrics were also found based on measurements made by two separate observers, on a subset of the scans.

2.4.2. *Further analyses*

To evaluate the performance and approach of each method, more analyses were performed beyond the ranking procedure. In this way, we could determine if there were particular factors that affected the results including both the method approaches and the data characteristics. This included investigating the different method approaches, UIA size dependence, intra-subject variance and assessing train vs test performance.

2.4.2.1. Method analyses. Based on the ranking of the method, a detailed look at each method could be performed to see and characterise similarities and differences between the performances. This was performed to investigate if some methods performed significantly better than others and if method design had an influence on performance. Bootstrapping was performed to compute 95% confidence intervals for each metric and ranking for each team. 2,000 random samples were taken from the test set with replacement. If confidence intervals did not overlap, methods were considered to have significantly different performance. Furthermore, the STAPLE algorithm (Warfield et al., 2004) was used to ensemble first, all of the segmentations from each method and second, the segmentations from the top 3 teams in task 2. Segmentation metrics and rankings were determined for these STAPLE ensemble method results and compared to the individual team performances.

2.4.2.2. Segmentation performance of true UIAs. To assess the segmentation performance of the methods, the segmentation metrics were determined for only the true detected UIAs, excluding any false positives. This was done in order to imitate how the tool could be used in clinical practice; as a radiologist will only select a correctly detected UIA for segmentation. To make a similar scenario, it was assessed first if the predicted segmentation overlapped with the ground truth segmentation. Connected component analysis was performed on the predicted segmentation. If a connected component overlapped with the ground truth segmentation, it remained and all other connected components (false positives) were removed. Segmentation metrics were determined for the remaining connected component relating to the true UIA. This was performed for each predicted segmentation by each team and a mean of the metrics and a ranking was made for each team.

2.4.2.3. Detection performance on negative scans. When screening for UIAs, some scans will be negative if a patient does not have UIAs. A well performing method should have a low false positive rate on the negative scans, as no true UIAs exist in these scans. Twenty-six scans of the test set did not have any UIAs, and the performance of each method on these scans was evaluated by determining the average false positive count. The average false positive count in negative scans was compared to the average false positive count in all scans in the test set containing true positives.

2.4.2.4. Size of UIAs. It was thought that the size of aneurysm would affect the performance of methods, as it is known that detection rates from visual inspection are lower for smaller aneurysms (Wardlaw and White, 2000). The relationship between the size of the UIAs and the detection and segmentation performance was investigated. Both sensitivity and DSC were assessed for each team in four different size quartiles based on the maximum UIA diameter.

2.4.2.5. Intra-subject analyses. Both the training and test data contained a subset of baseline and follow-up scans of the same subject. As this is common in clinical practice, it is vital that a measurement method should perform to a similar standard for both baseline and follow-up imaging, even though the two scans may differ in scanner type, acquisition protocol and quality. An accurate measure of the volume difference between follow-up and baseline scans is important to be able to detect growth of the UIA. To assess if the method could detect growth, the difference in volume between baseline and follow-up ground truth segmentations was determined (ground truth volume change). This was compared to the difference in volume of follow-up and baseline predicted segmentations by each method (predicted volume change). These measurements were only assessed for detected true UIAs, where the UIA was detected on both baseline and follow-up scan by the method. Similarities between the two volume change measurements indicate how reliable the measurement of the method is and this was assessed using Kendall's rank correlation measure (Kendall, 1938). Kendall's tau indicates how well two values correspond, where 1 indicates a strong agreement, 0 indicates no association and -1 indicates a strong disagreement.

Furthermore, a method that performs well, and to the same standard, in both baseline and follow-up scans is required. The intrasubject performance of each team was investigated by comparing the evaluation metric for the baseline scan to the metric at the follow-up scan. A Wilcoxon-signed rank test was used to compare the two values for each team. This was performed for sensitivity, to assess detection performance, and DSC and volumetric similarity for segmentation performance.

2.4.2.6. Train vs test performance. To assess performance differences between the training and test data, all methods were re-run on the training set and detection and segmentation metrics were determined. Performance should be similar to that of the test set and a large increase in performance indicates that the method may not be very generalisable to unseen data. A similar ranking of methods was made and this performance was compared to the performance of the methods on the test set.

All data analyses were conducted using pandas (McKinney, 2010), scipy (Virtanen et al., 2020), seaborn (Michael Waskom and the seaborn development team, 2020) and pingouin (Vallat, 2018) toolboxes with Python 3.7.

3. Results

3.1. Training and test data

There were no statistically significant differences between the cases of the training and test datasets in age ($p = 0.20$), sex ($p = 1$), maximum diameter of the UIA ($p = 0.58$), number of UIAs ($p = 0.32$), number of treated UIAs ($p = 0.45$), magnetic field strength of the scanner ($p = 0.11$), in-plane voxel spacing of the scan ($p = 0.43$), slice thickness of the scan ($p = 0.78$).

3.2. Challenge submission

Over 250 users registered for the challenge on the website, and 11 teams submitted methods. Two teams submitted only under task 1, for the detection of UIAs, and nine teams submitted under task 2, for the segmentation of UIAs. Results, presentations, posters and a brief description of all submitted methods can be found on the challenge website (http://adam.isi.uu.nl/results/results-miccai-2020/). The inference code submitted in Docker containers for the challenge is also available for most methods on DockerHub (https://hub.docker.com/orgs/adamchallenge).

3.2.1. Task 1 Submissions

MiBaumgartner submitted a 3D neural network based on the Retina U-Net architecture (Jaeger et al., 2018). The decoder was extended to incorporate semantic segmentation information and followed by a Path Aggregation Network (Liu et al., 2018) to generate the features used for the detection prediction. (Baumgartner et al., 2020)

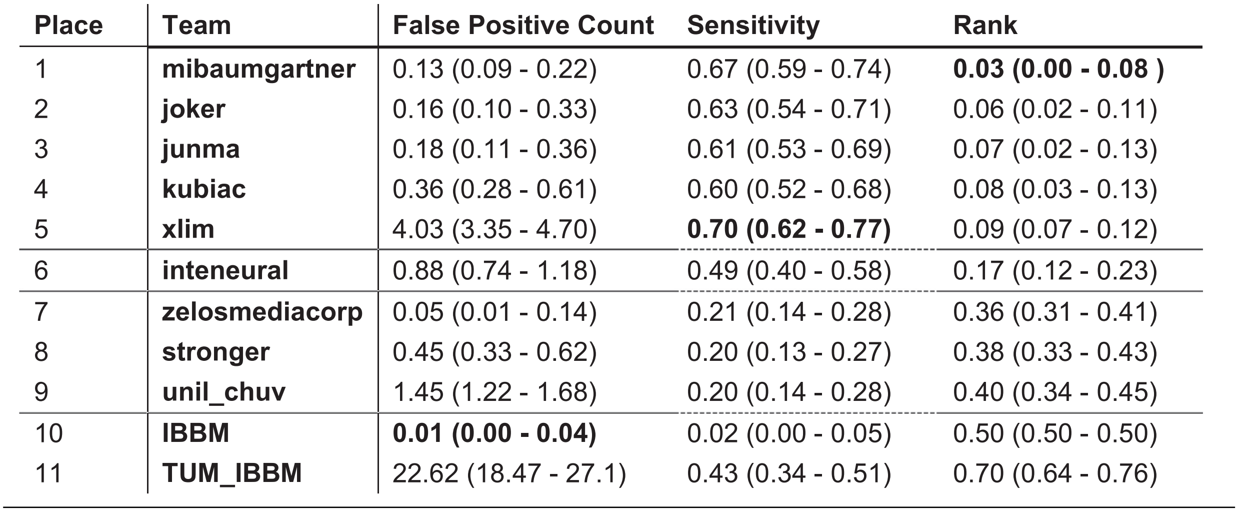
Unil_chuv submitted a 3D U-Net (Ronneberger et al., 2015) which was patch trained using patches selected based on landmark points from a registered vessel atlas (Mouches and Forkert, 2019). Both the ADAM dataset and an in-house dataset for training. On inference, patches were evaluated only if they were within a set distance from the registered landmark points and had a minimum intensity. A maximum number of four false positives were allowed based on the average brightness of the connected components. (Di Noto et al., 2021, 2020)

3.2.2. Task 2 submissions

IBBM submitted a 2D convolutional neural network with TriWinged-Net architecture based on the BtrflyNet (Sekuboyina et al., 2018). MIPs of the MRAs were made in all three orientations (axial, coronal and sagittal) with each view as a different input branch. These are encoded separately before being concatenated in the centre of the network. From this, there were three corresponding decoding branches, to provide segmentation masks for each view which were, finally, recombined to form the full segmentation volume. (Shit et al., 2020)

Inteneural submitted a method including three 2D neural networks with U-Net architecture based on EfficientNet (Tan and Le, 2019) that were pre-trained using ImageNet (Fei-Fei et al., 2010). Each network was fine-tuned for one axis: axial, coronal and sagittal with 2 input channels: raw TOF signal and blood vessel segmentation, which was

Table 1a
Task 1: Average metrics and ranking for each team, with the lowest (best) rank placing highest in the table. Each value is provided as a mean of all scans (95% confidence interval, determined using bootstrapping). The dotted lines indicates groups of methods that can be considered to have statistically different ranking from the other groups as their 95% ranking confidence intervals do not overlap.



Place	Team	False Positive Count	Sensitivity	Rank
1	**mibaumgartner**	0.13 (0.09 - 0.22)	0.67 (0.59 - 0.74)	**0.03 (0.00 - 0.08)**
2	**joker**	0.16 (0.10 - 0.33)	0.63 (0.54 - 0.71)	0.06 (0.02 - 0.11)
3	**junma**	0.18 (0.11 - 0.36)	0.61 (0.53 - 0.69)	0.07 (0.02 - 0.13)
4	**kubiac**	0.36 (0.28 - 0.61)	0.60 (0.52 - 0.68)	0.08 (0.03 - 0.13)
5	**xlim**	4.03 (3.35 - 4.70)	**0.70 (0.62 - 0.77)**	0.09 (0.07 - 0.12)
6	**inteneural**	0.88 (0.74 - 1.18)	0.49 (0.40 - 0.58)	0.17 (0.12 - 0.23)
7	**zelosmediacorp**	0.05 (0.01 - 0.14)	0.21 (0.14 - 0.28)	0.36 (0.31 - 0.41)
8	**stronger**	0.45 (0.33 - 0.62)	0.20 (0.13 - 0.27)	0.38 (0.33 - 0.43)
9	**unil_chuv**	1.45 (1.22 - 1.68)	0.20 (0.14 - 0.28)	0.40 (0.34 - 0.45)
10	**IBBM**	**0.01 (0.00 - 0.04)**	0.02 (0.00 - 0.05)	0.50 (0.50 - 0.50)
11	**TUM_IBBM**	22.62 (18.47 - 27.1)	0.43 (0.34 - 0.51)	0.70 (0.64 - 0.76)

performed using Jerman filter (Jerman et al., 2015). A loss function including both a generalised dice loss (Sudre et al., 2017) and boundary loss (Kervadec et al., 2021) was used. The final prediction was determined as an average of the evaluated models' outputs. (Walińska et al., 2020)

Joker submitted a 3D fully-convolutional neural network based on no new U-Net (nnUNet) (Isensee et al., 2021). Group Normalisation (Wu and He, 2018) was used instead of Batch Normalisation and leaky ReLU was used. A Dice ranking loss was used for training. Predictions were made by four separately trained models and ensembled using majority voting. (Yang et al., 2020)

JunMa submitted a 3D fully-convolutional neural network based on no new U-Net (nnUNet) (Isensee et al., 2021). Networks were trained using five-fold cross validation and two different loss functions: Dice loss and cross entropy, and Dice loss with topK loss (Berrada et al., 2018) because the two losses have been proven to be robust on highly imbalanced segmentation tasks (Ma et al., 2021). At prediction, the five models with optimum performance were ensembled. (Ma, 2020; Ma and An, 2020)

Kubiac submitted an ensemble of 18 neural networks with three network variants: A two path dual resolution fully convolutional neural network and two U-Net (Ronneberger et al., 2015) style architectures with two paths including contextual information in both the encoding and decoding path (Hilbert et al., 2020) trained on different loss functions. The loss functions were the sum of cross entropy, (generalised) Dice loss (Sudre et al., 2017) and boundary loss (Kervadec et al., 2021). (De Feo et al., 2020)

Stronger submitted an ensemble method of three models, where each model included a segmentation and a classification stage. The segmentation stage was based on a patch-trained 3D U-Net (Zhou et al., 2019). The classification consisted of a 3D convolutional neural network to distinguish between true and false positives. (Hu et al., 2020)

TUM-IBBM submitted a U-Net based architecture with MRA and aligned structural image as different input channels (Li et al., 2018). Two networks were trained on sagittal and coronal slices and during testing, voxelwise predictions of both models were averaged. (Loehr et al., 2020)

Xlim submitted a hybrid two input neural network: one for 3D patches and the second for the corresponding maximum intensity projection of the patches (Nakao et al., 2018). The two paths are brought together with a final concatenation layer. The patches consist of vessels only, segmented from the MRAs using an intensity and morphological transform based method. (Rjiba et al., 2020)

Zelosmediacorp submitted a 3D fully convolutional neural network with a U-Net like architecture (Ronneberger et al., 2015) trained on patches centred on the average UIA position. Twelve networks were trained on four different training and validation splits, and the best of four networks were selected to form an ensemble that averaged the outputs of each network on the test set. Monte-Carlo dropout (Wang and Manning, 2013) was used for both training and inference. (Giroud and Dubost, 2020)

Further, more in-depth descriptions of each method can be found on the website (http://adam.isi.uu.nl/results/results-miccai-2020/).

3.3. Metrics and rankings

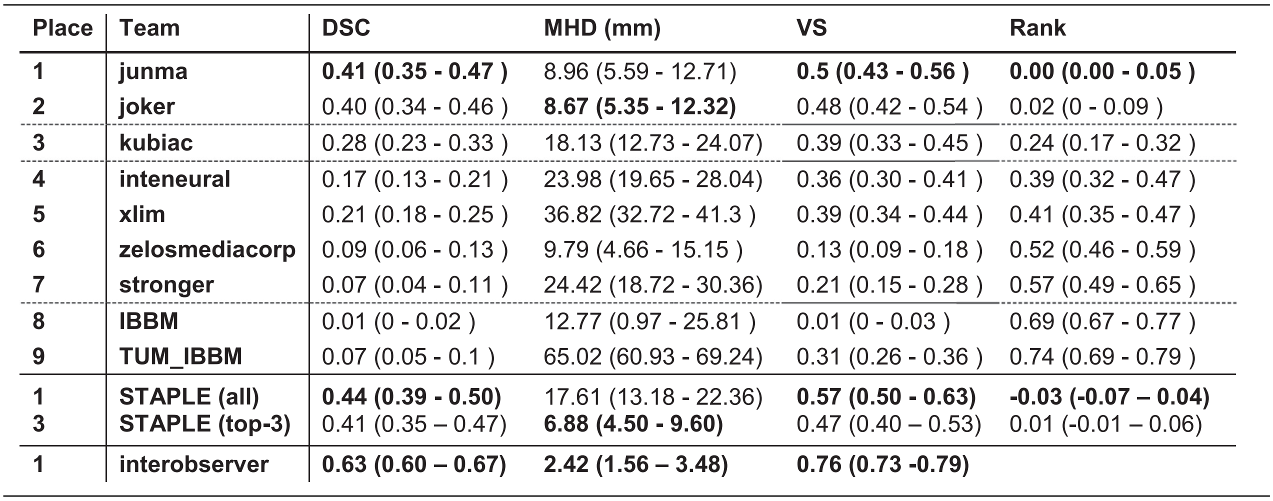
The mean performance of each participating team for task 1 is shown in Table 1a) and for task 2 is shown in Table 1b). The dotted lines indicate groups of methods that can be considered to have statistically different ranking from the other groups as their 95% ranking confidence intervals do not overlap. Figs. 2 and 3 are bar charts and boxplots to show the distribution of metrics for each team. For task 1 the method of **xlim** performed best for sensitivity and the method of **IBBM** performed best for false positive count. Based on the overall ranking (equal weighting of both metrics) **mibaumgartner** performed the best for task 1. For task 1, **mibaumgartner, joker, junma** and **kubiac** had overlapping bootstrapped confidence intervals for rank and thus were considered to have not substantially different performance from each other. For task 2, **junma** had the best DSC and VS and **joker** had the best MHD. Based on the overall ranking (equal weighting of all three segmentation metrics) **junma** performed the best for task 2. For task 2, **junma** and **joker** performed statistically significantly better than any other methods based on the bootstrapped confidence intervals being non-overlapping with any other methods. The bottom row of Table 1b) indicates the interobserver agreement of two observers. This was assessed as a mean over 144 scans (72 paired baseline-follow-up scans). The average metrics are much higher than any submitted method. An example segmentation of team **junma** can be seen in Appendix A, Fig. 1.

3.4. Further analyses

3.4.1. Method analyses

All 11 submissions for both tasks used deep learning techniques for the detection and/or segmentation of the UIA and information about the methods is provided in Table 2. The U-Net (Ronneberger et al., 2015)

Table 1b
Task 2: Average metrics and ranking for each team, and the brackets contain the 95% confidence interval determined using bootstrapping. The dotted lines indicates groups of methods that can be considered to have statistically different ranking from the other groups as their 95% ranking confidence intervals do not overlap. STAPLE (all) and STAPLE (top-3) are the average metrics and ranking of the segmentation from the STAPLE algorithm of all and the top-3 methods, respectively. Interobserver are the metrics comparing manual segmentations of two different observers on a subset of the scans. DSC: Dice Similarity Coefficient; Modified Hausdorff Distance: MHD; VS: Volumetric Similarity.



Place	Team	DSC	MHD (mm)	VS	Rank
1	junma	**0.41 (0.35 - 0.47)**	8.96 (5.59 - 12.71)	**0.5 (0.43 - 0.56)**	**0.00 (0.00 - 0.05)**
2	joker	0.40 (0.34 - 0.46)	**8.67 (5.35 - 12.32)**	0.48 (0.42 - 0.54)	0.02 (0 - 0.09)
3	kubiac	0.28 (0.23 - 0.33)	18.13 (12.73 - 24.07)	0.39 (0.33 - 0.45)	0.24 (0.17 - 0.32)
4	inteneural	0.17 (0.13 - 0.21)	23.98 (19.65 - 28.04)	0.36 (0.30 - 0.41)	0.39 (0.32 - 0.47)
5	xlim	0.21 (0.18 - 0.25)	36.82 (32.72 - 41.3)	0.39 (0.34 - 0.44)	0.41 (0.35 - 0.47)
6	zelosmediacorp	0.09 (0.06 - 0.13)	9.79 (4.66 - 15.15)	0.13 (0.09 - 0.18)	0.52 (0.46 - 0.59)
7	stronger	0.07 (0.04 - 0.11)	24.42 (18.72 - 30.36)	0.21 (0.15 - 0.28)	0.57 (0.49 - 0.65)
8	IBBM	0.01 (0 - 0.02)	12.77 (0.97 - 25.81)	0.01 (0 - 0.03)	0.69 (0.67 - 0.77)
9	TUM_IBBM	0.07 (0.05 - 0.1)	65.02 (60.93 - 69.24)	0.31 (0.26 - 0.36)	0.74 (0.69 - 0.79)
1	STAPLE (all)	**0.44 (0.39 - 0.50)**	17.61 (13.18 - 22.36)	**0.57 (0.50 - 0.63)**	**-0.03 (-0.07 – 0.04)**
3	STAPLE (top-3)	0.41 (0.35 – 0.47)	**6.88 (4.50 - 9.60)**	0.47 (0.40 – 0.53)	0.01 (-0.01 – 0.06)
1	interobserver	**0.63 (0.60 – 0.67)**	**2.42 (1.56 – 3.48)**	**0.76 (0.73 -0.79)**	

Table 2
Submitted methods sorted on their final ranking per task, with highest placed ranking first, and information about method design.

Team	Task	Place	Architecture	2D/3D	Segmentation Loss function	Ensemble[1]	Use of structural image[2]	Data Augmentation[3]	Post Processing[4]
mibaumgartner	1	1	Retina U-Net + Path Aggregation Network	3D	Dice and Cross Entropy	√	√	C, M, R, S	FPS
unil_chuv	1	9	U-Net	3D	Dice			C, M, R	FPC, L, M
junma	2	1	U-Net (nn-Unet)	3D	Dice and Cross entropy or Top-K	√		C, M, R, S	
joker	2	2	U-Net (nn-Unet)	3D	Dice ranking	√	√	C, E, M, R, S	
kubiac	2	3	Multi resolution U-Net style network and CNN classifier	3D	Cross entropy, (generalised) Dice and Boundary	√	√	T	L
inteneural	2	4	Efficientnet-b1	2D	Generalised Dice and Boundary	√			FPC, FPS
xlim	2	5	AneurysmNet	2D (MIP) and 3D	Dice				FPS
zelosmediacorp	2	6	U-Net	3D	Dice	√		M, R, T	FPS
stronger	2	7	U-Net and CNN classifier	3D	Dice and Cross Entropy	√		M, R, T	
IBBM	2	8	TriWingedNet	2.5D	Dice				
TUM_IBBM	2	9	U-Net	2D	Dice	√	√	M, R, SH	FPS

[1] Ensemble was used at any point of the method, either for training and/or inference. Different ensembles were used including combing models: with different validation splits, different loss functions and different architectures

[2] Use of the structural image as input for the models

[3] Augmentation of training data: C = contrast augmentation, E = elastic deformation, M = mirroring, R = rotation, S = scaling, SH = shearing, T = translation

[4] Post-Processing: FPC = false positive reduction based on count, FPS = false positive reduction based on size/volume, L = location dependent inference, M = merge neighbouring detections/segmentation

was the most common architecture with 72% (8/11) submissions using a U-Net style architecture for at least part of their method. The top two ranking segmentation methods used nnU-Net (Isensee et al., 2021) as the base for their approach. Seven methods used 3D approaches, including the top 5 ranking methods. All methods incorporated the Dice loss in their loss function for training, however **junma** and **joker,** the top-ranking segmentation methods, also incorporated topK loss (Berrada et al., 2018). Ensembles were commonly used, and appeared to boost performance with the top 5 methods for task 1 and 2 using an ensemble. Ensembles were used by different teams in various ways for example: with different validation splits, different loss functions and different architectures before combining the trained models. **Unil_chuv** was the only team to use an external, in-house dataset for training. 8/11 teams use augmentation of the training data and 7/11 teams used post-processing techniques to reduce the number of false positives.

3.4.2. Segmentation Performance of true UIAs

To evaluate segmentation performance, average segmentation metrics were determined for all teams for only the true UIAs that were detected, as displayed in Table 3. A similar ranking was made as for task 2 based on these metrics. It was observed that this ranking changed the placing of the teams, as is shown by the red brackets and arrows. However, the top 3 teams remained unchanged in position. The box plots of the segmentation metrics for each team over detected UIAs only is shown in Appendix B, Fig. 1.

Table 3
The mean segmentation metrics of each team evaluated only on the detected true UIAs. The arrows and brackets in red signify the difference between the original task 2 ranking (Table 1b), and the ranking based only on the detected UIAs. All values are quoted as means with 95% confidence intervals determined by bootstrapping in brackets. Table is ordered with the highest placed ranking first. DSC: Dice Similarity Coefficient; Modified Hausdorff Distance: MHD; VS: Volumetric Similarity.

Place	team	DSC	MHD (mm)	VS	Rank
1	**junma**	0.64 (0.59 - 0.68)	2.62 (2.12 - 3.31)	0.71 (0.65 - 0.76)	0.00 (0.00 - 0.14)
2	**joker**	0.60 (0.55 - 0.66)	2.95 (2.42 - 3.66)	0.66 (0.60 - 0.72)	0.11 (0.02 - 0.25)
3	**kubiac**	0.45 (0.39 - 0.51)	4.95 (3.82 - 6.28)	0.53 (0.45 - 0.6)	0.53 (0.25 - 0.70)
4	**xlim (↑ 1)**	0.40 (0.35 - 0.44)	6.55 (5.42 - 7.83)	0.58 (0.52 - 0.64)	0.61 (0.32 - 0.80)
5	**stronger (↑ 2)**	0.39 (0.27 - 0.5)	5.87 (3.94 - 8.06)	0.54 (0.36 - 0.71)	0.63 (0.25 - 0.94)
6	**zelosmediacorp**	0.40 (0.30 - 0.50)	5.63 (4.29 – 7.00)	0.49 (0.37 - 0.62)	0.66 (0.30 - 0.87)
7	**IBBM (↑ 1)**	0.30 (0.11 - 0.47)	5.47 (2.00 - 12.17)	0.49 (0.11 - 0.82)	0.74 (0.23 – 1.00)
8	**inteneural(↓ 4)**	0.34 (0.28 - 0.41)	5.76 (4.34 - 7.64)	0.42 (0.34 - 0.50)	0.80 (0.40 - 0.96)
9	**TUM_IBBM**	0.31 (0.24 - 0.38)	8.44 (6.85 - 10.25)	0.56 (0.48 - 0.65)	0.83 (0.45 - 0.93)

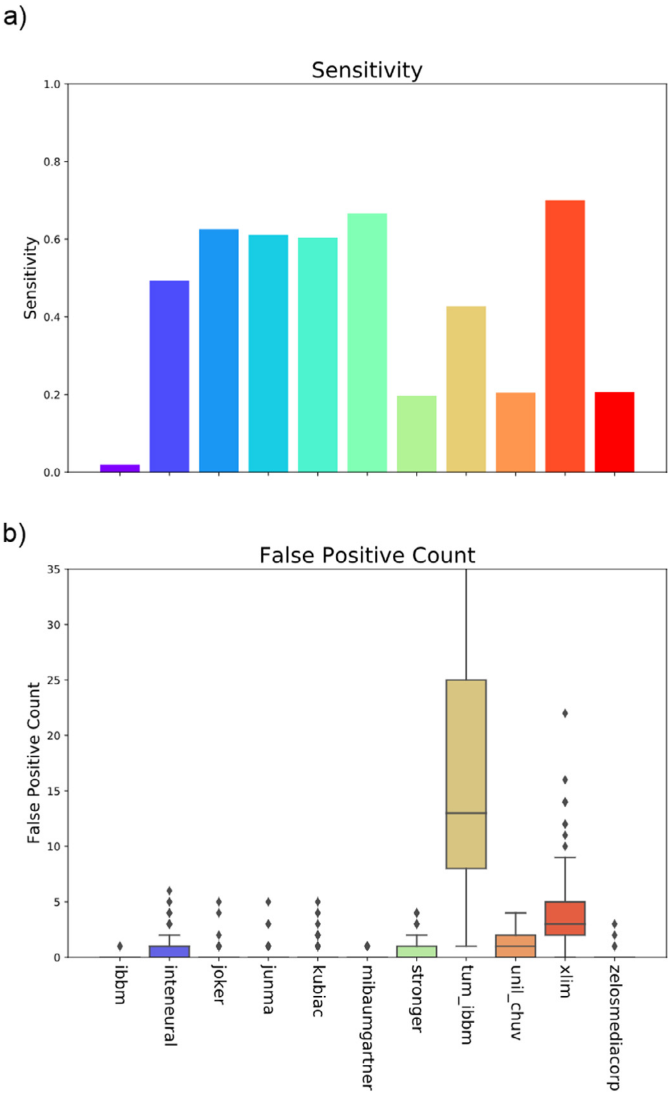


Fig. 2. Sensitivity and False Positive Count for all teams for all scans in the test set.
a) Bar chart of sensitivity of all teams for task 1, taken as an average across all scans in the test set b) Box plot of total false positive count per scan of all teams for all scans in the test set

3.4.3. *Detection Performance on Negative Scans*

The average false positive count over all scans containing no true UIAs was determined (Appendix C, Table 1). This can be compared to the average false positive count for all scans with true UIAs. Teams **IBBM, zelosmediacorp, junma** and **joker** all have a zero false positive count for the scans containing no UIAs. All teams have a smaller false positive count per scan for the negative scans, compared to the positive scans containing true UIAs. **IBBM** and **zelosmediacorp** have a low false positive count for positive scans (0.02 and 0.06 respectively), but they also had a very low true positive count. **Junma** and **joker** have a substantially higher false positive count for positive scans (0.22 and 0.20 respectively).

3.4.4. *Size of UIAs*

The detection and segmentation performance improved with the size of the UIA. Fig. 4 shows the increase in sensitivity with increasing UIA diameter, when assessing the UIA diameter in four quartiles. This was represented as the mean sensitivity over all teams for each UIA. The error bar shows the 95% confidence interval of the mean. In Appendix D, Fig. 1, it can be seen how the sensitivity of each individual team varies with size of UIA. Fig. 5 a) and b) demonstrate that the segmentation performance also increased with UIA size. In 5a) the median DSC over all teams for each UIA was plotted against the individual UIA diameter. In 5b), the UIA diameter is again split into four quartiles and the mean DSC over all teams for each UIA was included. DSCs for individual teams were plotted in Appendix D, Fig. 2.

3.4.5. *Intra-subject analyses*

Table 4 shows the volume change measurements, the ground truth measurements and the predicted measurements for each team, and how well they agree using the Kendall's tau correlation measure. All measurements are taken only for true UIAs with a positive detection in both baseline and follow-up scans. This means that the ground truth volume is also different as it is taken as a mean over a different set of scans. The median ground truth difference over all baseline and follow-up scans was 2.9 µl. Team **IBBM** was not included, as less than 5 true UIAs were detected for both baseline and follow-up scans. **Junma** were found to have the highest statistically significantly agreement between ground truth and predicted volume change (Kendall's tau > 0.5, p < 0.05). **Inteneural** had a Kendall tau < 0, which indicates there was some disagreement between ground truth and predicted volume change. **Stronger** and **TUM_IBBM** had values for Kendall's tau which were close to zero, suggesting that there is no association between ground truth and predicted volume change for these methods.

The performance of each method was evaluated between baseline and follow-up scans using the Wilcoxon rank test, the results of which can be seen in Appendix E, Fig. 1. For sensitivity, DSC and volumetric similarity, all methods had $p >= 0.05$ suggesting that performance was not different between baseline and follow-up subjects.

Table 4
Comparison of volume change measurements (median (IQR)) for ground truth and predicted segmentations with correlation measure, Kendall's tau (*p* value). Volume change measurement was determined as the volume of the follow-up volume minus the baseline volume in µl. Note that the ground truth volume is different for each team, as it is evaluated only over true UIAs that were detected in both baseline and follow-up scans by the method.

Team	Ground Truth Volume Change(µl)	Predicted Volume Change(µl)	Kendall's tau (p value)
inteneural	3.8 (-0.2, 11.6)	0.4 (-2.1, 10.4)	-0.10 (0.57)
joker	5.0 (-0.3, 13.3)	1.8 (-3.7, 7.3)	0.42 (<0.05)
junma	4.2 (0.7, 12.5)	0.2 (-3.0, 10.7)	0.54 (<0.05)
kubiac	2.9 (-0.6, 11.8)	1.5 (-1.3, 6.9)	0.17 (0.19)
stronger	12.1 (-13.9, 14.8)	-0.7 (-4.3, 12.6)	0.06 (0.92)
TUM_IBBM	4.0 (-0.2, 13.7)	-6.4 (-27.4, 15.0)	0.09 (0.53)
xlim	2.9 (-1.4, 10.2)	-9.5 (-22.1, 18.1)	0.44 (<0.05)
zelosmediacorp	8.5 (-12.8, 13.3)	5.5 (-1.9, 13.1)	0.42 (0.11)

3.4.6. Train vs test performance

All the submitted methods were also evaluated on the training data. The results can be seen in Appendix F, Tables 1a and 1b; which correspond to Tables 1a and 1b in the main text. As expected, the results on the training data are generally better than on the test data. For task 1, the overall ranking remains roughly similar, with some teams going up or down a few places. This could suggest that some methods generalise less well to the unseen test data, resulting in a lower performance on the test data as compared with the training data. For task 2, the top-4 ranking methods remained the same order of ranking as when assessed on the training data. All methods show a considerable drop in performance when assessed on the test set, relative to the training set. This suggests that the methods submitted for task 2 do not generalise well to the test data set.

4. Discussion

This paper presents the results and analysis of the Aneurysm Detection and segMentation Challenge held at the international conference of Medical Image Computing and Computer Assisted Intervention (MICCAI) in October 2020.

Two methods perform significantly better than all other methods for both tasks: (1) detection and (2) segmentation of UIAs on TOF-MRAs. Although the results are encouraging for automated UIA detection and segmentation methods, there is still room for substantial improvement. Compared to visual UIA detection from MRAs, the sensitivity of the submitted methods is, on average, lower than quoted in literature (HaiFeng et al., 2017; White et al., 2000). The submitted segmentation methods also show a lower performance than the two observers in this study. Future developments will hopefully bring new and updated methods that are closer in performance to manual methods.

4.1. Top ranked methods

Mibaumgartner placed first in task 1 for detection of UIAs and did not participate in the second segmentation task. The method focuses on the detection task, by outputting bounding boxes from which a centre of mass was derived, as opposed to performing semantic segmentation. This is different to all other submitted methods. **Mibaumgartner** opts to still include semantic segmentation information by using Retina U-Net (Jaeger et al., 2018), before classifying and regressing anchor boxes using a Path Aggregation Network (Liu et al., 2018). **Mibaumgartner** did not discriminate between treated and untreated UIAs, using both as foreground voxels for training, which was different from other methods. This may have aided detection by giving more examples as some aneurysms treated with coils may look similar to untreated UIAs. As treated UIAs were masked on evaluation, this did not negatively affect the performance. Furthermore, **mibaumgartner** used both the structural MR images and the MRAs, which may have aided in the performance of the model by incorporating more information. Although **mibaumgartner** has the highest overall ranking, it does not achieve the highest sensitivity or lowest false positive count.

For task 1 and task 2, the methods of **junma** and **joker** showed comparable performance, both ranking above the other methods. Both use a 3D U-Net architecture based on the no new net (nnUnet). The nnUnet is an "out-of-the box tool for state-of-the-art segmentation" which is an open-source deep learning segmentation framework that automatically adapts to new datasets. In December 2019, the nnUNet performed optimally or on par with the best methods in 19 different biomedical image analysis challenges, including the KiTS challenge (https://kits19.grand-challenge.org/), the largest challenge at MICCAI 2019 (Isensee et al., 2019). **Joker** made some small changes to the model, including using group normalisation instead of batch normalisation, although this did not appear to make much difference to its overall performance. **Joker** also used the structural images as input for training.

4.2. Method analyses

All top 3 methods for each task used an ensemble of trained models for prediction and in total 7/11 submitted methods used an ensemble. It is known that ensembles of deep learning models can aid in both image classification (Krizhevsky et al., 2017) and segmentation tasks (Kamnitsas et al., 2018; Kuijf et al., 2019). In general, ensemble methods were made up of models trained on different train/validation data splits or cross-validation. Winning team **junma** trained using five fold cross-validation and two different loss functions, before selecting the optimal five trained networks (based on DSC) to ensemble. **Joker** used an ensemble of four networks, which included networks trained for different classes in the scan (both treated and untreated UIAs) as well as including the structural MRI scans in two of the networks. The STAPLE analysis confirms that ensembles perform well, with an ensemble of all segmentations from all methods achieving the best ranking. STAPLE using an ensemble segmentation of the top three teams for task 2, **junma, joker** and **kubiac,** performs better than **joker** and **kubiac** individually but **junma** still remains the highest ranking.

In addition to **joker,** the methods of **mibaumgartner, kubiac** and **TUM_IBBM** also use the structural images in their method suggesting that the networks may benefit from having the information contained in the structural images when detecting and segmenting UIAs. Other teams use the structural images to aid in patch selection for training.

The volume of an UIA is a very small percentage of the volume of a whole TOF-MRA, and in most MRAs only one UIA is present. As a result of this unbalanced problem, most methods chose to use ground truth knowledge for the patch selection, choosing a particular proportion of training patches to contain an UIA. Only two methods, **inteneural** and **xlim,** perform vessel segmentation on the TOF scans before performing UIA detection/segmentation. However, both methods are middle ranking (0.39, 0.41 respectively), suggesting that vessel segmentation may not help much in UIA detection or segmentation.

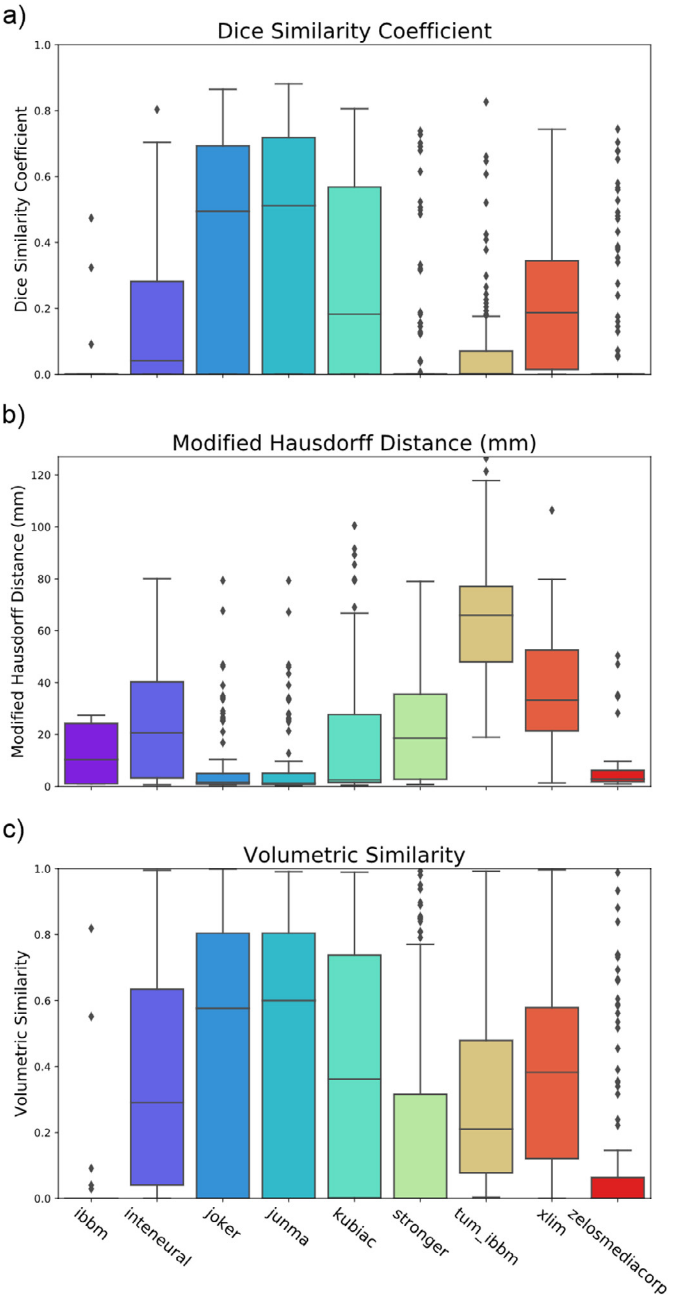


Fig. 3. Box plots of metrics for all teams for task 2.
a) Dice Similarity Coefficient (DSC) b) Modified Hausdorff Distance (MHD) c) Volumetric Similarity for all scans containing a UIA. Each point in the box plots is the metric evaluated on one scan in the test set for each method. The centre line shows the median metric of all scans.

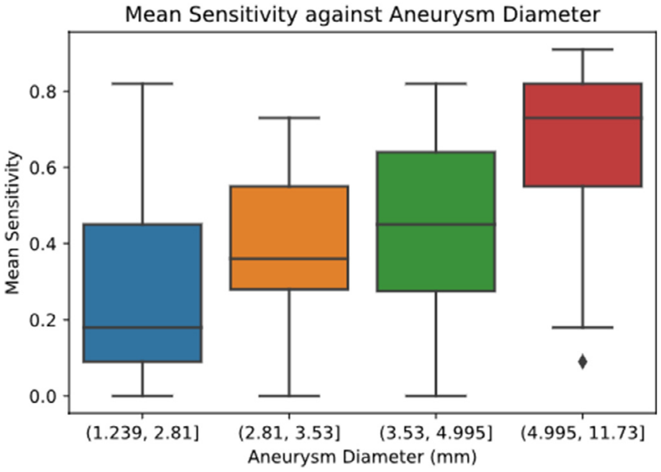


Fig. 4. Sensitivity of methods for UIA of different sizes.
Sensitivity of all teams for each UIA as a function of maximum UIA diameter in mm, when separating UIA diameter into four quartiles. Each point included in the box plot is the mean sensitivity of all teams across each UIA.

Almost all task 2 segmentation methods used dice loss in some form for training their networks. This is a calculated choice, as dice is one of the metrics on which we evaluate the submitted solutions. Some methods use the generalised dice loss (Sudre et al., 2017), which has proven to be reliable for unbalanced problems, and others in combination with other loss functions such as cross-entropy, topK and boundary loss (Kervadec et al., 2021). The winning method **junma** used an ensemble of methods trained using dice + cross entropy and dice + topK loss. **Kubiac** and **inteneural** both included the boundary loss in their loss functions for training their models. By including boundary loss, the models are trained to minimise the distance between the predicted and ground truth segmentations. This reduces the problems associated with regional based metrics, such as Dice, for highly imbalanced data. **Kubiac** and **inteneural** have similar performance for task 2 (rankings 0.24 and 0.39 respectively) and this may be due to the similar architecture and loss function used.

Many teams performed post-processing to only accept positive detections of a certain number of voxels, within a range that was common in the training dataset. Further, some teams even limited the maximum number of true positives found based on probability, size or intensity of the predictions. This aided in the challenge ranking, as we explicitly evaluated on false positive count. This can be seen for example by **xlim,** with a mean false positive count of 4.03 but a sensitivity of 70%, meant their ranking was lower than if they had perhaps used a further false positive reduction method.

4.3. Segmentation Performance of true UIAs

The top three teams in task 2, **junma, joker** and **kubiac**, also ranked top for segmentation performance of true UIAs only. **Junma** with a DSC of 0.64 is slightly higher than the interobserver DSC of 0.63. The MHD and VS are comparable to the MHD and VS of the interobserver, with all 95% confidence intervals overlapping. This suggests that the automatic segmentation method performance is on par with the manual segmentation, once the true UIA has been identified. This method could be used in the clinical research or routine, whereby a radiologist would only need to select an UIA, from a small population of candidate UIAs, and segmentation of the correct UIA could be performed.

4.4. Detection performance on healthy scans

Top performing teams **junma** and **joker** also perform well on scans without true UIAs, and have an average false positive count of 0 for such scans. This would be ideal for in the clinic by not wrongly identifying UIAs, and providing radiologists with more work to censor these falsely identified UIAs. Team **IBBM** and **zelosmediacorp** also had a false positive count of 0, however, their overall detection performance (sensitivity) across all scans, including those with positive UIAs, was poor.

4.5. Size of UIAs

Overall, it was clear that both detection and segmentation performance was better for all methods for larger UIAs, as both sensitivity and DSC increased with UIA diameter (Spearman's coefficient = 0.47 and 0.42 respectively). Not surprisingly, smaller UIA are more difficult

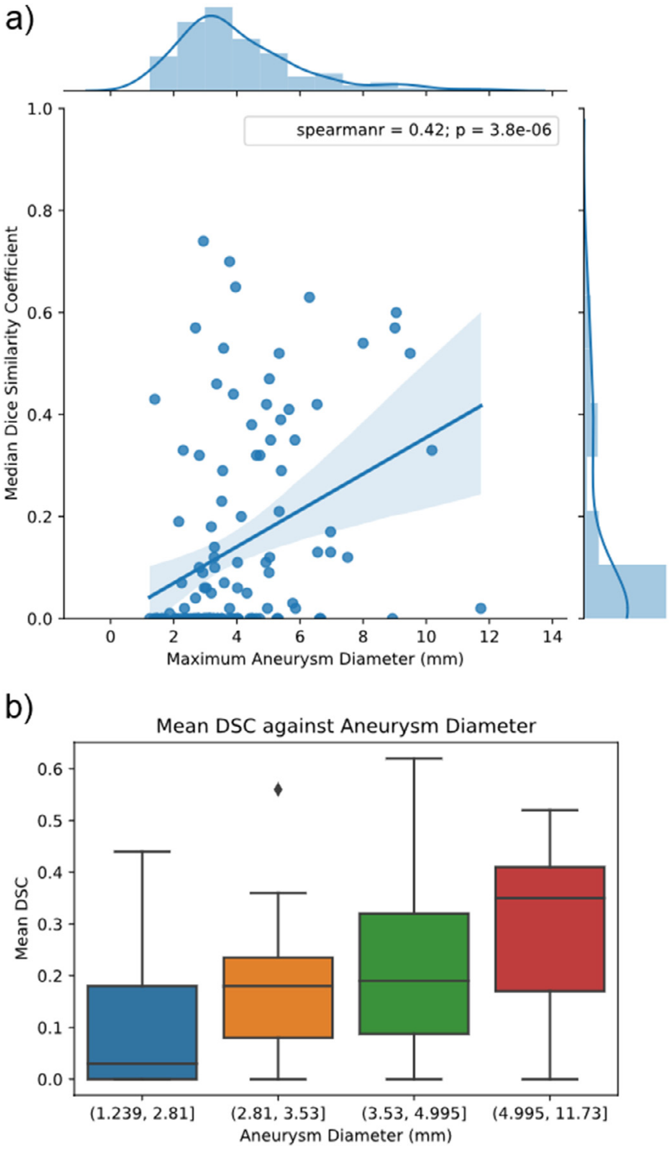


Fig. 5. Dice Similarity Coefficient (DSC) as a function of aneurysm diameter a) Median Dice Similarity Coefficient (DSC) of all teams for each UIA as a function of maximum UIA diameter in mm, b) Mean DSC of all teams for each UIA as a function of maximum UIA diameter in mm, when separating UIA diameter into four quartiles.

to detect, which is also consistent with studies investigating visual detection of aneurysms. White et. al. (White et al., 2000) cite an average of 87% sensitivity for detecting UIAs on MRAs by radiologists, of which sensitivity is 38% for UIAs < 3 mm and 94% for UIA > 3 mm. From the results, it can be seen that the lower quartiles of diameter have a comparable sensitivity. **Xlim** has the highest sensitivity with 71% for UIAs with diameter > 3.54 mm and < 4.98 mm and 95% for UIAs > 4.98 mm. As such, this method may be suitable for detection of larger UIAs with performance that is on par with human visual inspection. We assessed segmentation using DSC, which is a difficult measure for small objects and is limited by voxel sizes of the images. For small UIAs, with few voxels, the overlap will be less likely and this results in a smaller DSC.

4.6. Intra-subject Analyses

Comparing volume change between ground truth and predicted segmentations, found that different methods performed differently. **Junma** had the best agreement between ground truth and predicted volume changes (Kendall's tau > 0.5), suggesting that can accurately measure volumetric change and growth. **Junma** had the best segmentation performance overall which could explain the volumetric change agreement. For some methods there was disagreement or almost no association between the predicted and ground truth volume changes, suggesting that these methods are not appropriate for measuring volumetric growth. It was also noted that the actual volumetric change was very small, and none of the aneurysms showed considerable growth between baseline and follow-up. The small volumetric change may explain the low volumetric change agreement of all methods. Based on the segmentation metrics and Wilcoxon rank test, the methods performed similarly for both baseline and follow-up scans. One variable that may have affected the intrasubject performance, was the train, test and validation splits between the methods, as many methods did not take baseline-follow-up pairs into account.

4.7. Train vs test performance

Most methods, for both tasks, had a considerably lower performance on the test data than on the training data. This suggests that these methods did not generalise well to the unseen data. Reasons for this could be in the method design, the training/validation data splits, aneurysm sizes, or not taking into account the baseline-follow-up pairs. The distribution of aneurysm and scan characteristics is similar between the training and test sets, ensuring that the training data is representative of the test data. Nevertheless, some features such as aneurysm shape or the configuration with respect to the parent vessel were difficult to take into account, as they can vary considerably between patients. This reflects the true clinical nature of the data set, but ideally methods should be able to detect and segment UIAs, even on unseen examples.

4.8. Future work

Overall, further improvement is necessary to be comparable to manual clinical standards for UIA detection and segmentation. All methods performed worse for smaller UIAs and as small UIAs are often overlooked by radiologists, this would be a main aspect for improvement of the methods. Furthermore, with increased screening studies, detection of small UIAs would be beneficial to speed up workflow and to learn more about the prevalence of UIAs in the general population. The best detection method used a network specifically designed for detection as opposed to semantic segmentation. The other submitted methods appear limited for detection with most using a generic semantic segmentation method. This suggests that a "brute force" technique, by just applying a standard U-Net architecture, may not be optimal for this problem. Instead, future developments should think out of the box. It was also noted that few methods use information from the structural images to aid in their methods. Perhaps some prior knowledge of, for example, the location, shape and size of the UIA would aid in the method performance. The dataset was a true clinical dataset, with a mixture of scan parameters, and although this makes it technically challenging, a method that performs well over the whole test set would be very convenient to have for clinical use. For larger aneurysms, the top-ranked detection methods had a performance that was on par with human visual detection suggesting that these methods could be used for the detection of larger UIAs.

The method of **junma** showed promising segmentation performance on the true UIAs. This suggests that a semi-automatic workflow allowing a radiologist to identify the location of the UIA and then using the model of **junma** as an accurate method of UIA segmentation may already be of use in current clinical practice. In future work, incorporating this segmentation method, with an improved detection method, may lead to an optimal automatic detection and segmentation method for UIAs.

5. Conclusions

The provided results were presented at the 23rd International Conference of MICCAI 2020. Methods for UIA detection and segmentation are encouraging but require further development before being able to be accurately used to detect, segment and quantify UIAs automatically, to the same level as a radiologist. However, detection methods may be suitable for use for larger aneurysms. Furthermore, segmentation performance of the top ranking method suggests it may be suitable for UIA segmentation after manual selection of the true UIA. The ADAM challenge remains open for submission of both new and improved methods .

6. Data availability

Training data and results are available at http://adam.isi.uu.nl/ . Scripts for evaluation of methods can be found at: https://github.com/hjkuijf/ADAMchallenge .

The test set is not publicly available, as it is kept secret for evaluation purposes of the submitted methods. The inference code submitted in Docker containers for the challenge is also available for most methods, whose teams gave permission, on DockerHub (https://hub.docker.com/orgs/adamchallenge).

7. Credit author statement

Kimberley M. Timmins; Conceptualization, Methodology, Software, Formal analysis, Data Curation, Writing - Original Draft, Visualization, Project administration, Irene C. van der Schaaf; Conceptualization, Resources, Data Curation, Writing - Review & Editing, Supervision, Funding acquisition, Edwin Bennink; Conceptualization, Methodology, Software, Writing - Review & Editing, Ynte M. Ruigrok; Resources, Writing - Review & Editing, Supervision, Funding acquisition, Xingle An, Michael Baumgartner, Pascal Bourdon, Riccardo De Feo, Tommaso Di Noto, Florian Dubost, Augusto Fava-Sanches, Xue Feng, Corentin Giroud, Inteneural Group, Minghui Hu, Paul F. Jaeger, Juhana Kaiponen, Michał Klimont, Yuexiang Li, Hongwei Li, Yi Lin, Timo Loehr, Jun Ma, Klaus H. Maier-Hein, Guillaume Marie, Bjoern Menze, Jonas Richiardi, Saifeddine Rjiba, Dhaval Shah, Suprosanna Shit, Jussi Tohka, Thierry Urruty, Urszula Walińska, Xiaoping Yang, Yunqiao Yang, Yin Yin; Methodology, Software, Writing - Review & Editing Birgitta K. Velthuis; Resources, Writing - Review & Editing, Supervision, Hugo J. Kuijf; Conceptualization, Methodology, Software, Writing - Review & Editing, Supervision, Funding acquisition.

Acknowledgements

We acknowledge the support from the Netherlands Cardiovascular Research Initiative: An initiative with support of the Dutch Heart Foundation, CVON2015-08 ERASE and CVON2018-02 ANEURYSM@RISK. We further acknowledge support from the focus area Applied Data Science, Utrecht University. **YMR** has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (PRYSM, grant agreement No. 852173). Team **inteneural** wish to acknowledge members of the Inteneural Group: Mikolaj A. Pawlak MD PhD, Marek Kraft PhD, Michał Mikołajczak MSc and Dominik Pieczyński MSc for their contributions. Team **joker** acknowledges support from the Key-Area Research and Development Program of Guangdong Province, China (No. 2018B010111001), National Key R&D Program of China (2018YFC2000702) and the Scientific and Technical Innovation 2030-"New Generation Artificial Intelligence" Project (No. 2020AAA0104100). Team **junma** acknowledges support from the National Natural Science Foundation of China (No. 11971229). Team **kubiac** acknowledges support from European Social Fund (grant S21770), Academy of Finland (grant 316258) and CSC - IT Center for Science, Finland, for computational resources. Team **mibamgartner** acknowledges support from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) (grant 410981386).

Appendices

Appendix A. Segmentation example

For the top row, one slice of the MRA is shown, where the segmentation of the ground truth and predicted segmentation is similar, shown by the large overlap. In the bottom row, the predicted segmentation is much smaller than the ground truth and there is little overlap. The **junma** method segmented better in the centre of the aneurysms than at the edge of the aneurysm.

Fig. A1

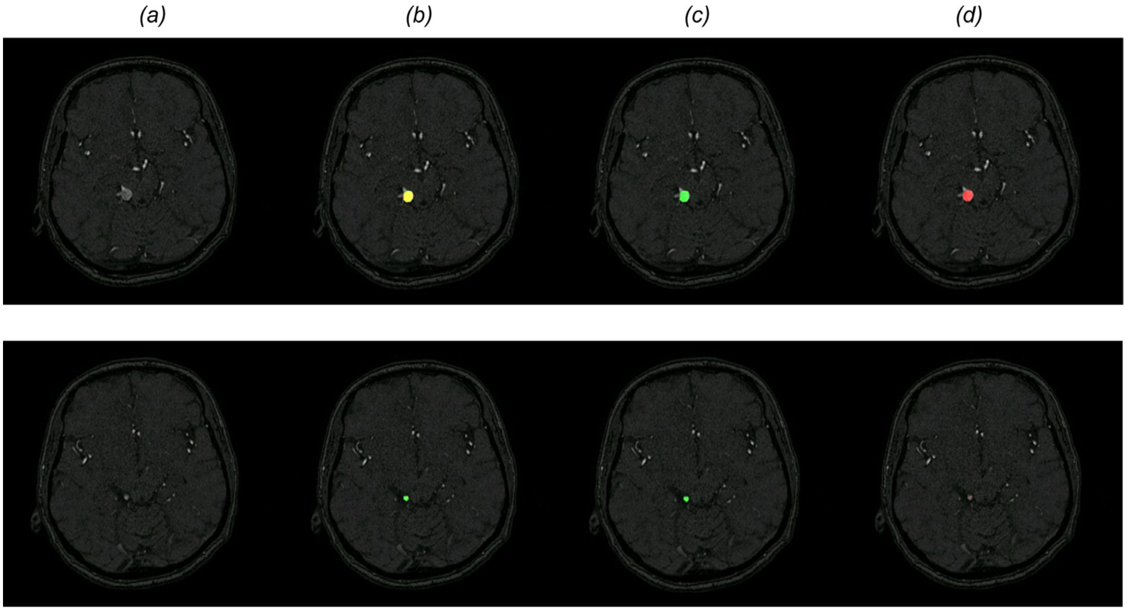


Fig. A1. Segmentation of team junma on an example test case.
Figure 1: slices of the TOF-MRA of the same test case, to show how the segmentation from the Junma method varied from the ground truth segmentation. Columns a) no segmentation overlaid, b) both segmentations overlaid in yellow, c) ground truth segmentation overlaid in green, d) predicted segmentation overlaid in red.

.

Appendix B. Segmentation Performance of true UIAs

Fig.B1

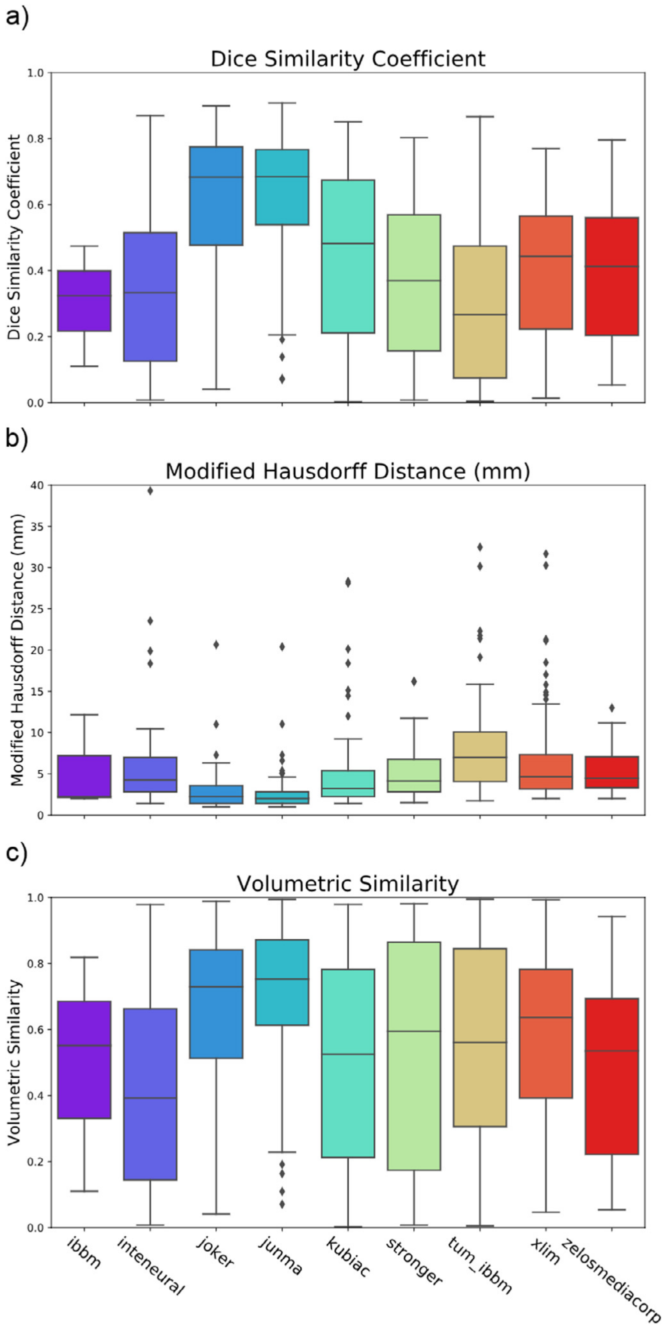


Fig. B1. Segmentation Performance of all detected UIAs
Boxplots showing the distribution of segmentation metrics for all teams for task 2, taken only true UIAs detected by the method.

.

Appendix C. Detection Performance on Healthy Scans

Table C1

Table C1
Average number of false positive count per scan: Column 1: all scans in the test set; column 2: all negative scans without true UIAs; and column 3: all positive scans containing true UIAs. Table is sorted in ascending order of the first column (number of false positives in all scans).

Team	All Scans	Negative Scans	Positive Scans
IBBM	0.01	0.00	0.02
zelosmediacorp	0.05	0.00	0.06
mibaumgartner	0.13	0.08	0.15
joker	0.16	0.00	0.20
junma	0.18	0.00	0.22
kubiac	0.36	0.04	0.43
stronger	0.45	0.38	0.47
inteneural	0.88	0.58	0.95
unil_chuv	1.45	1.54	1.43
xlim	4.03	4.08	4.02
TUM_IBBM	22.62	22.5	22.65

Appendix D. Size Analyses

Fig. D1, Fig. D2

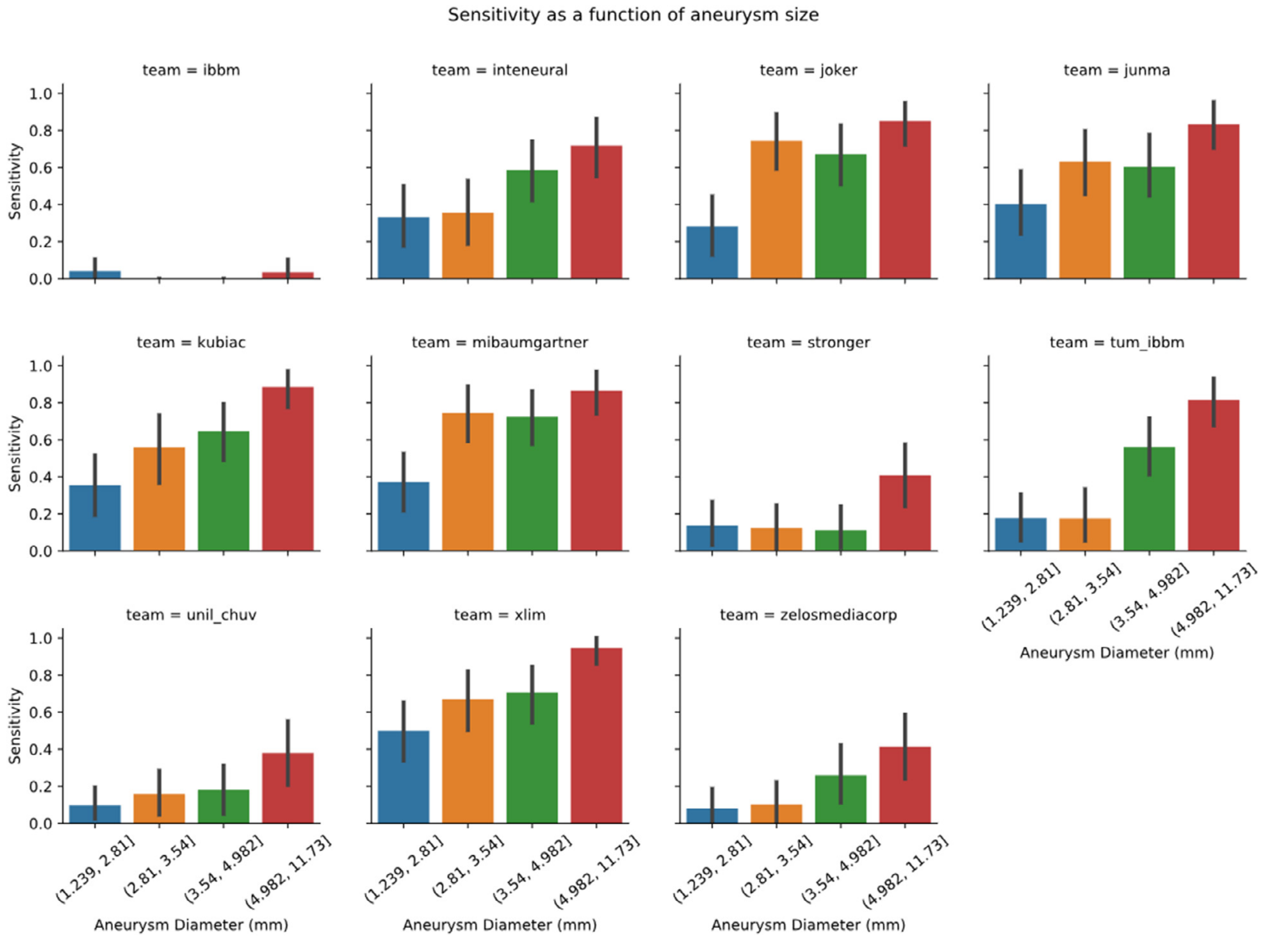


Fig. D1. Sensitivity of each method assessed in different size UIAs
Sensitivity of each method based on all scans split into four quartiles of maximum aneurysm diameter.

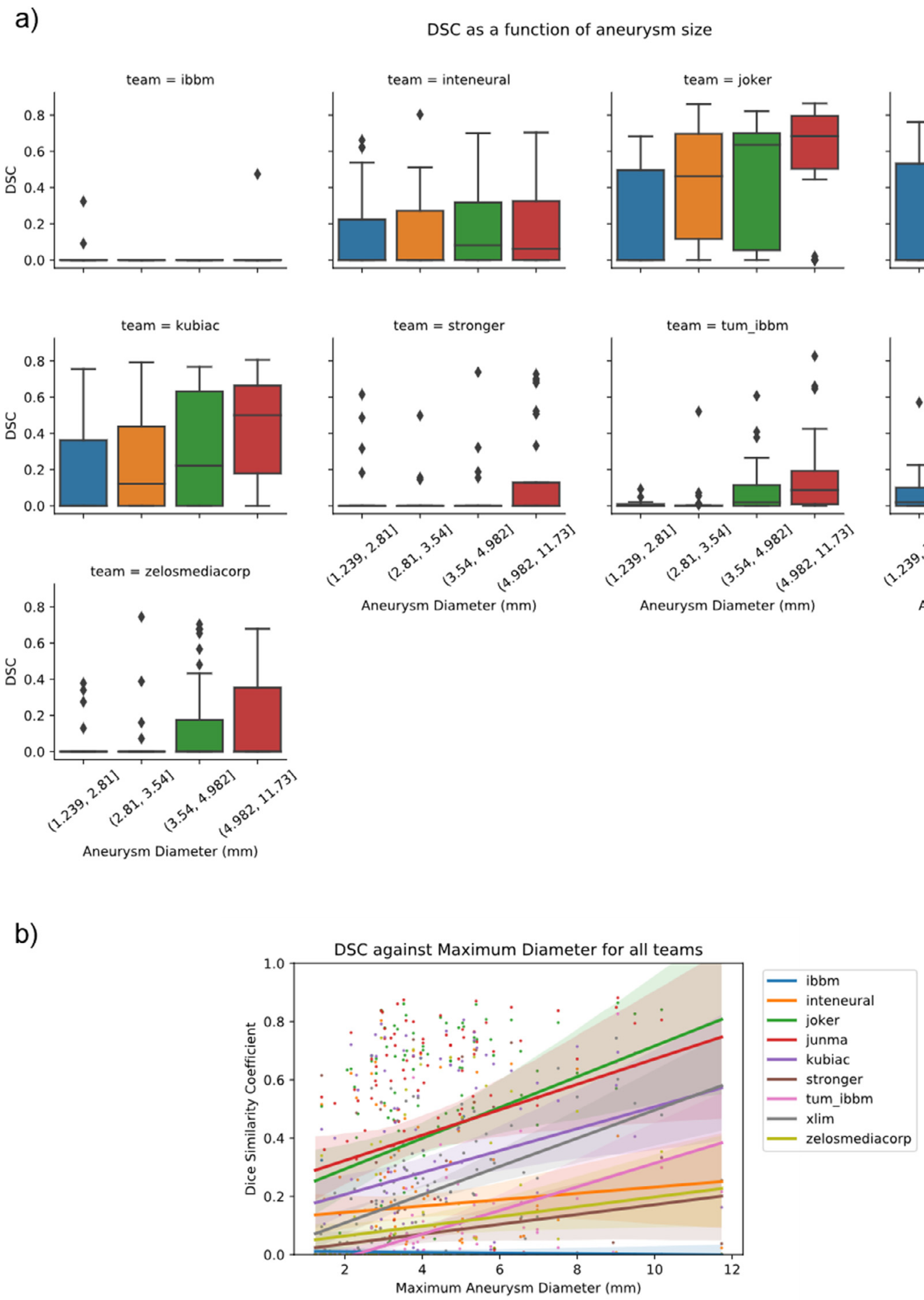


Fig. D2. a) Dice Similarity Coefficient (DSC) of each method assessed in different size UIAs.
Dice Similarity Coefficient (DSC) of each method as a function of maximum aneurysm diameter when split into quartiles
Figure D2) b) Dice Similarity Coefficient (DSC) against UIA Maximum Diameter
DSC of each method a function of maximum aneurysm diameter for each UIA in all scans. Each point represents the DSC of one scan for one method.

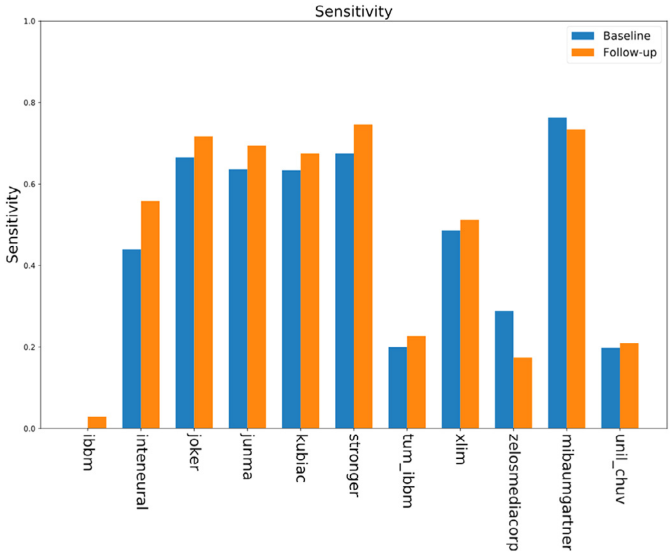


Fig. D2. Continued

Appendix E. Intrasubject Performance

Fig. E1 and Fig.E2

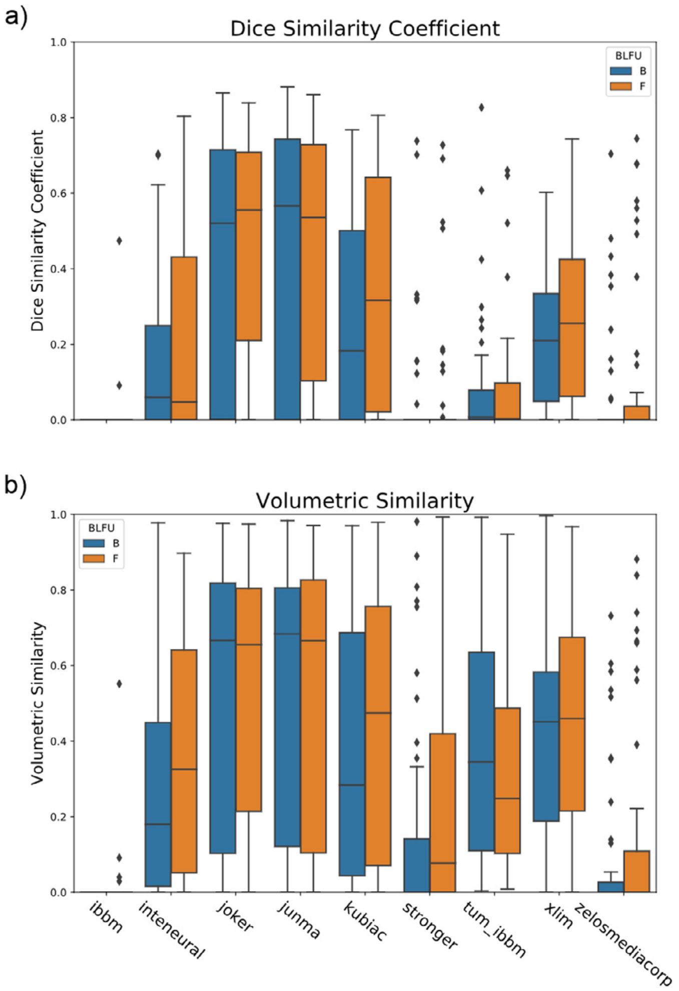


Fig. E1. Sensitivity for Baseline and Follow-up scans
Mean sensitivity of each subject for each team for all baseline and follow-up scans in the test set.

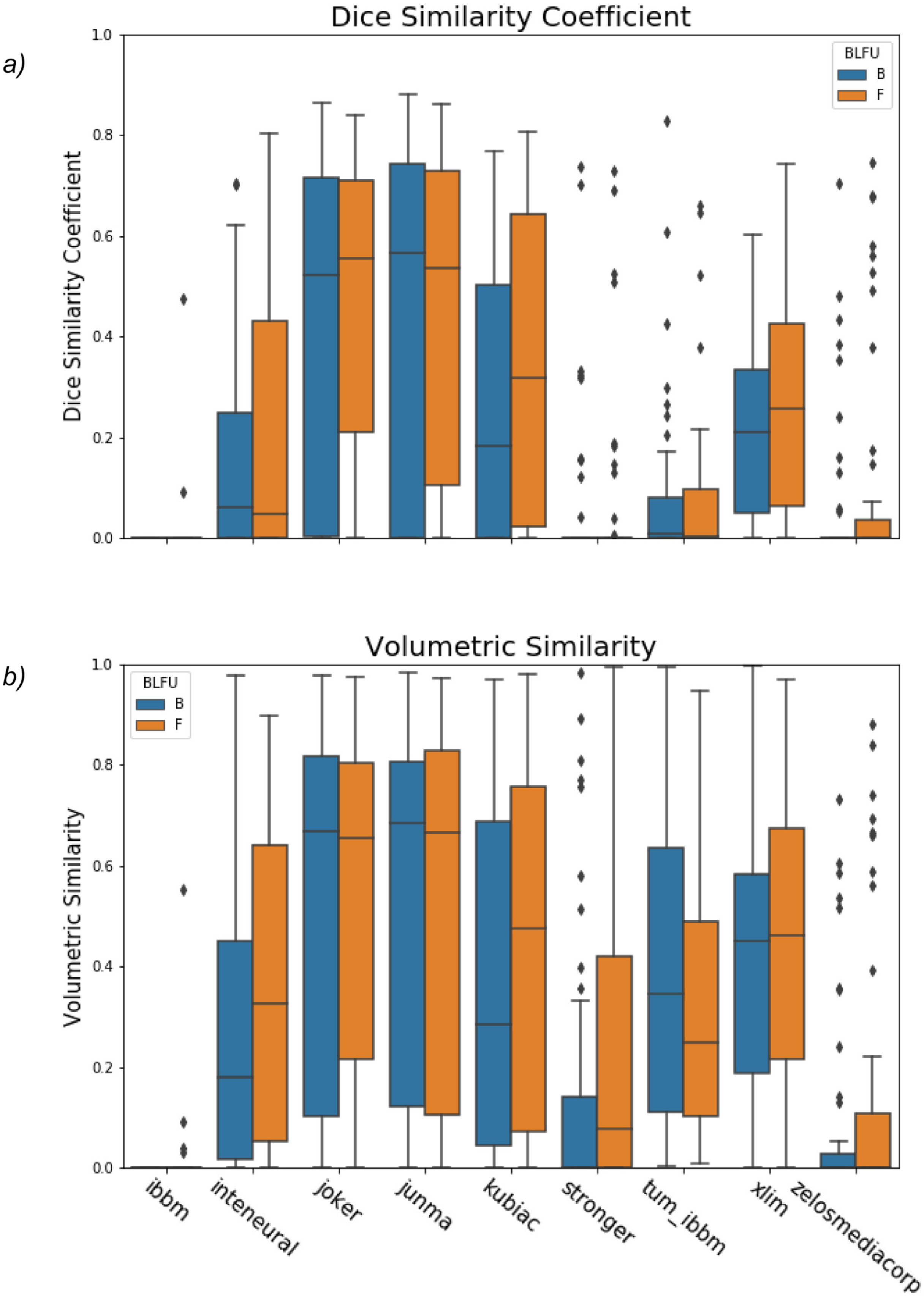


Fig. E2. Segmentation Metrics for Baseline and Follow-up Scans
DSC and VS of each subject for each team for baseline and follow-up scans of all scans in the test dataset.

Appendix F. Train vs Test Performance

Tables F1a and F1b.

Table F1a

Task 1: Detection metrics and ranking assessed for all methods on training data. FP is average false positive count over all scans, sensitivity is average sensitivity over scans containing true UIAs. Difference is the average values of the test set subtracted from the average value of the training set.

	Train set			Test set			Difference (Train - Test)		
Team	FPcount	Sensitivity	rank	FPcount	Sensitivity	rank	FPcount	Sensitivity	rank
Kubiac	0.15	1.00	0.00	0.36	0.60	0.08	-0.21	0.40	-0.10
Mibaumgartner	0.04	0.96	0.03	0.13	0.67	0.03	-0.09	0.29	0.00
Junma	0.01	0.94	0.04	0.18	0.61	0.07	-0.17	0.33	-0.00
Joker	0.04	0.92	0.06	0.16	0.63	0.06	-0.12	0.29	0.00
Xlim	2.82	0.89	0.12	4.03	0.70	0.09	-1.21	0.19	0.03
inteneural	0.62	0.83	0.13	0.88	0.49	0.17	-0.26	0.34	-0.00
IBBM	0.03	0.56	0.31	0.01	0.02	0.50	0.02	0.54	-0.20
Zelosmediacorp	0.00	0.53	0.33	0.05	0.21	0.36	-0.05	0.32	-0.00
Stronger	0.45	0.47	0.38	0.45	0.20	0.38	0.00	0.27	0.00
unil_chuv	1.27	0.29	0.52	1.45	0.20	0.40	-0.18	0.09	0.12
TUM_IBBM	32.86	0.59	0.79	22.62	0.43	0.70	10.24	0.16	0.09

Table F1b

Task 2: Segmentation metrics and ranking assessed for all methods on training data. All values are averages over all scans containing true UIAs. DSC: Dice Similarity Coefficient, MHD: Modified Hausdorff Distance measured in mm, VS: Volumetric Similarity. Difference is the average values of the test set subtracted from the average value of the training set.

	Train set				**Test set**				**Difference (Train - Test)**			
Team	**DSC**	**MHD**	**VS**	**rank**	**DSC**	**MHD**	**VS**	**rank**	**DSC**	**MHD**	**VS**	**rank**
Junma	0.83	2.52	0.91	0.00	0.41	8.96	0.50	0.22	0.42	-6.44	0.41	-0.22
Joker	0.79	2.01	0.86	0.05	0.40	8.67	0.48	0.20	0.39	-6.66	0.38	-0.15
Kubiac	0.78	5.11	0.90	0.05	0.28	18.13	0.39	0.43	0.50	-13.00	0.51	-0.38
inteneural	0.73	14.58	0.84	0.15	0.17	23.98	0.36	0.95	0.56	-9.40	0.48	-0.80
Zelosmediacorp	0.40	7.71	0.48	0.46	0.09	9.79	0.13	0.06	0.31	-2.08	0.35	0.40
IBBM	0.29	8.08	0.41	0.55	0.01	12.77	0.01	0.02	0.28	-4.69	0.40	0.53
Stronger	0.21	22.06	0.34	0.70	0.07	24.42	0.21	0.47	0.14	-2.36	0.13	0.23
Xlim	0.27	35.29	0.32	0.76	0.21	36.82	0.39	4.02	0.06	-1.53	-0.10	-3.26
TUM_IBBM	0.11	63.50	0.30	1.00	0.07	65.02	0.31	22.65	0.04	-1.52	-0.00	-21.65

References

Arimura, H., Li, Q., Korogi, Y., Hirai, T., Katsuragawa, S., Yamashita, Y., Tsuchiya, K., Doi, K., 2006. Computerized detection of intracranial aneurysms for three-dimensional MR angiography: feature extraction of small protrusions based on a shape-based difference image technique. Med. Phys. 33, 394–401. doi:10.1118/1.2163389.

Backes, D., Rinkel, G.J.., Greving, J.P., Velthuis, B.K., Murayama, Y., Takao, H., Ishibashi, T., Igase, M., TerBrugge, K.G., Agid, R., Jaaskelainen, J.E., Lindgren, A.E., Koivisto, T., Von Und Zu Fraunberg, M., Matsubara, S., Moroi, J., Wong, G.K.C., Abrigo, J.M., Igase, K., Matsumoto, K., Wermer, M.J.H., Van Walderveen, M.A.A., Algra, A., 2017. ELAPSS score for prediction of risk of growth of unruptured intracranial aneurysms. Neurology 88, 1600–1606. doi:10.1212/WNL.0000000000003865.

Backes, D., Vergouwen, M.D.I., Tiel Groenestege, A.T., Bor, A.S.E., Velthuis, B.K., Greving, J.P., Algra, A., Wermer, M.J.H., Van Walderveen, M.A.A., Terbrugge, K.G., Agid, R., Rinkel, G.J.E., 2015. PHASES score for prediction of intracranial aneurysm growth. Stroke 46, 1221–1226. doi:10.1161/STROKEAHA.114.008198.

Baumgartner, M., Jaeger, P.F., Isensee, F., Maier-Hein, K.H., 2020. Retina U-Net for Aneurysm Detection in MR Images [WWW Document]. URL http://adam.isi.uu.nl/results/results-miccai-2020/participating-teams-miccai-2020/mibaumgarter/ .

Berrada, L., Zisserman, A., Kumar, M.P., 2018. Smooth Loss Functions for Deep Top-k Classification.

Bizjak, Ž., Likar, B., Pernuš, F., Špiclin, Ž., 2021. Modality agnostic intracranial aneurysm detection through supervised vascular surface classification. In: Drukker, K., Mazurowski, M.A. (Eds.), Medical Imaging 2021: Computer-Aided Diagnosis. SPIE, p. 21. doi:10.1117/12.2580868.

Bor, A.S.E., Koffijberg, H., Wermer, M.J.H., Rinkel, G.J.E., 2010. Optimal screening strategy for familial intracranial aneurysms: a cost-effectiveness analysis. Neurology 74, 1671–1679. doi:10.1212/WNL.0b013e3181e04297.

Bor, A.S.E., Rinkel, G.J.E., van Norden, J., Wermer, M.J.H., 2014. Long-term, serial screening for intracranial aneurysms in individuals with a family history of aneurysmal subarachnoid haemorrhage: a cohort study. Lancet Neurol. 13, 385–392. doi:10.1016/S1474-4422(14)70021-3.

Brown, R.D., Broderick, J.P., 2014. Unruptured intracranial aneurysms: Epidemiology, natural history, management options, and familial screening. Lancet Neurol. 13, 393–404. doi:10.1016/S1474-4422(14)70015-8.

Cardenes, R., Pozo, J.M., Bogunovic, H., Larrabide, I., Frangi, A.F., 2011. Automatic aneurysm neck detection using surface voronoi diagrams. IEEE Trans. Med. Imaging 30, 1863–1876. doi:10.1109/TMI.2011.2157698.

De Feo, R., Kaiponen, J., Tohka, J., 2020. Aneurysm Segmentation in the ADAM Challenge: KUBIAC team [WWW Document]. URL http://adam.isi.uu.nl/results/results-miccai-2020/participating-teams-miccai-2020/kubiac/ .

Di Noto, T., Marie, G., Tourbier, S., Alemán-Gómez, Y., Esteban, O., Saliou, G., Cuadra, M.B., Hagmann, P., Richiardi, J., 2021. Weak labels and anatomical knowledge: making deep learning practical for intracranial aneurysm detection in TOF-MRA.

Di Noto, T., Marie, G., Tourbier, S., Alemán-Gómez, Y., Saliou, G., Bach Cuadra, M., Hagmann, P., Richiardi, J., 2020. ADAM: Aneurysm Detection And segMentation Challenge Unil – CHUV Team [WWW Document]. URL http://adam.isi.uu.nl/results/results-miccai-2020/participating-teams-miccai-2020/unil_chuv/ .

Duan, H., Huang, Y., Liu, L., Dai, H., Chen, L., Zhou, L., 2019. Automatic detection on intracranial aneurysm from digital subtraction angiography with cascade convolutional neural networks. Biomed. Eng. Online 18, 110. doi:10.1186/s12938-019-0726-2.

Faron, A., Sijben, R., Teichert, N., Freiherr, J., Wiesmann, M., Sichtermann, T., 2019. Deep learning-based detection of intracranial aneurysms in 3D TOF-MRA. Am. J. Neuroradiol. 40, 25–32. doi:10.3174/ajnr.A5911.

Fei-Fei, L., Deng, J., Li, K., 2010. ImageNet: constructing a large-scale image database. J. Vis. 9, 1037. doi:10.1167/9.8.1037, –1037.

Flahault, A., Trystram, D., Nataf, F., Fouchard, M., Knebelmann, B., Grünfeld, J.-P., Joly, D., 2018. Screening for intracranial aneurysms in autosomal dominant polycystic kidney disease is cost-effective. Kidney Int 93, 716–726. doi:10.1016/j.kint.2017.08.016.

Forbes, G., Fox, A.J., Huston, J., Wiebers, D.O., Torner, J., 1996. Interobserver variability in angiographic measurement and morphologic characterization of intracranial aneurysms: A report from the International Study of Unruptured Intracranial Aneurysms. Am. J. Neuroradiol. 17, 1407–1415.

Giroud, C., Dubost, F., 2020. Probabilistic Segmentation and Detection of Aneurysm from brain MRA with an Ensemble of 3D Convolutional Neural Networks and Monte Carlo Dropout [WWW Document]. URL http://adam.isi.uu.nl/results/results-miccai-2020/participating-teams-miccai-2020/zelosmediacorp/ .

Greving, J.P., Wermer, M.J.H., Brown, R.D., Morita, A., Juvela, S., Yonekura, M., Ishibashi, T., Torner, J.C., Nakayama, T., Rinkel, G.J.E., Algra, A., 2014. Development of the PHASES score for prediction of risk of rupture of intracranial aneurysms: a pooled analysis of six prospective cohort studies. Lancet Neurol 13, 59–66. doi:10.1016/S1474-4422(13)70263-1.

HaiFeng, L., YongSheng, X., YangQin, X., Yu, D., ShuaiWen, W., XingRu, L., JunQiang, L., 2017. Diagnostic value of 3D time-of-flight magnetic resonance angiography for detecting intracranial aneurysm: a meta-analysis. Neuroradiology 59, 1083–1092. doi:10.1007/s00234-017-1905-0.

Hentschke, C.M., Beuing, O., Nickl, R., Tönnies, K.D., 2012. Automatic cerebral aneurysm detection in multimodal angiographic images. IEEE Nucl. Sci. Symp. Conf. Rec. 3116–3120. doi:10.1109/NSSMIC.2011.6152566.

Hilbert, A., Madai, V.I., Akay, E.M., Aydin, O.U., Behland, J., Sobesky, J., Galinovic, I., Khalil, A.A., Taha, A.A., Wuerfel, J., Dusek, P., Niendorf, T., Fiebach, J.B., Frey, D., Livne, M., 2020. BRAVE-NET: fully automated arterial brain vessel segmentation in patients with cerebrovascular disease.. Front. Artif. Intell. 3. doi:10.3389/frai.2020.552258.

Hopmans, E.M., Ruigrok, Y.M., Bor, A.S., Rinkel, G.J., Koffijberg, H., 2016. A cost-effectiveness analysis of screening for intracranial aneurysms in persons with one first-degree relative with subarachnoid haemorrhage. Eur. Stroke J. 1, 320–329. doi:10.1177/2396987316674862.

Hu, M., Feng, X., Yin, L., 2020. Unruptured Intracranial Aneurysm Segmentation from TOF-MRA Images Using Cascaded 3D Convolutional Neural Networks [WWW Document]. URL http://adam.isi.uu.nl/results/results-miccai-2020/participating-teams-miccai-2020/stronger/ .

Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H., 2021. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Methods 18, 203–211. doi:10.1038/s41592-020-01008-z.

Isensee, F., Jäger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H., 2019. Automated Design of Deep Learning Methods for Biomedical Image Segmentation. arxiv: http://arxiv.org/abs/1904.08128

Jaeger, P.F., Kohl, S.A.A., Bickelhaupt, S., Isensee, F., Kuder, T.A., Schlemmer, H.-P., Maier-Hein, K.H., 2018. Retina U-Net: Embarrassingly Simple Exploitation of Segmentation Supervision for Medical Object Detection.

Jerman, T., Pernuš, F., Likar, B., Špiclin, Ž., 2015. Beyond Frangi: an improved multiscale vesselness filter, in: Ourselin, S., Styner, M.A. (Eds.), . p. 94132A. https://doi.org/10.1117/12.2081147

Ji, W., Liu, A., Lv, X., Kang, H., Sun, L., Li, Y., Yang, X., Jiang, C., Wu, Z., 2016. Risk score for neurological complications after endovascular treatment of unruptured intracranial aneurysms. Stroke 47, 971–978. doi:10.1161/STROKEAHA.115.012097.

Kamnitsas, K., Bai, W., Ferrante, E., McDonagh, S., Sinclair, M., Pawlowski, N., Rajchl, M., Lee, M., Kainz, B., Rueckert, D., Glocker, B., 2018. Ensembles of multiple models and architectures for robust brain tumour segmentation. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics) 10670 LNCS 450–462. doi:10.1007/978-3-319-75238-9_38.

Keedy, A., 2006. An overview of intracranial aneurysms. McGill J. Med..

Kendall, M.G., 1938. A new measure of rank correlation. Biometrika 30, 81–93. doi:10.1093/biomet/30.1-2.81.

Kervadec, H., Bouchtiba, J., Desrosiers, C., Granger, E., Dolz, J., Ben Ayed, I., 2021. Boundary loss for highly unbalanced segmentation. Med. Image Anal. 67, 1–21. doi:10.1016/j.media.2020.101851.

Kim, H.J., Yoon, D.Y., Kim, E.S., Lee, H.J., Jeon, H.J., Lee, J.Y., Cho, B.M., 2017. Intraobserver and interobserver variability in CT angiography and MR angiography measurements of the size of cerebral aneurysms. Neuroradiology 59, 491–497. doi:10.1007/s00234-017-1826-y.

Klein, S., Staring, M., Murphy, K., Viergever, M.a., Pluim, J., 2010. <emphasis emphasistype="mono">elastix</emphasis>: a toolbox for intensity-based medical image registration. IEEE Trans. Med. Imaging 29, 196–205.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2017. ImageNet classification with deep convolutional neural networks. Commun. ACM 60, 84–90. doi:10.1145/3065386.

Kuijf, H.J., Casamitjana, A., Collins, D.L., Dadar, M., Georgiou, A., Ghafoorian, M., Jin, D., Khademi, A., Knight, J., Li, H., Llado, X., Biesbroek, J.M., Luna, M., Mahmood, Q., McKinley, R., Mehrtash, A., Ourselin, S., Park, B.Y., Park, H., Park, S.H., Pezold, S., Puybareau, E., De Bresser, J., Rittner, L., Sudre, C.H., Valverde, S., Vilaplana, V., Wiest, R., Xu, Y., Xu, Z., Zeng, G., Zhang, J., Zheng, G., Heinen, R., Chen, C., van der Flier, W., Barkhof, F., Viergever, M.A., Biessels, G.J., Andermatt, S., Bento, M., Berseth, M., Belyaev, M., Cardoso, M.J., 2019. Standardized assessment of automatic segmentation of white matter hyperintensities and results of the WMH segmentation challenge. IEEE Trans. Med. Imaging 38, 2556–2568. doi:10.1109/TMI.2019.2905770.

Lane, A., Vivian, P., Coulthard, A., 2015. Magnetic resonance angiography or digital subtraction catheter angiography for follow-up of coiled aneurysms: do we need both? J. Med. Imaging Radiat. Oncol. 59, 163–169. doi:10.1111/1754-9485.12288.

Lawonn, K., Meuschke, M., Wickenhöfer, R., Preim, B., Hildebrandt, K., 2019. A geometric optimization approach for the detection and segmentation of multiple aneurysms. Comput. Graph. Forum 38, 413–425. doi:10.1111/cgf.13699.

Li, H., Jiang, G., Zhang, J., Wang, R., Wang, Z., Zheng, W.S., Menze, B., 2018. Fully convolutional network ensembles for white matter hyperintensities segmentation in MR images. Neuroimage 183, 650–665. doi:10.1016/j.neuroimage.2018.07.005.

Lindgren, A.E., Koivisto, T., Björkman, J., Von Und Zu Fraunberg, M., Helin, K., Jääskeläinen, J.E., Frösen, J., 2016. Irregular shape of intracranial aneurysm indicates rupture risk irrespective of size in a population-based cohort. Stroke 47, 1219–1226. doi:10.1161/STROKEAHA.115.012404.

Liu, S., Qi, L., Qin, H., Shi, J., Jia, J., 2018. Path aggregation network for instance segmentation. Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. 8759–8768. doi:10.1109/CVPR.2018.00913.

Loehr, T., Li, H., Menze, B., 2020. A Multi-View Approach for Automatic Segmentation of Intracranial Aneurysms from Time of Flight MRAs [WWW Document]. URL http://adam.isi.uu.nl/results/results-miccai-2020/participating-teams-miccai-2020/tum_ibbm/ .

Ma, J., 2020. Loss Ensembles for Extremely Imbalanced Segmentation. arxiv: http://arxiv.org/abs/2101.10815

Ma, J., An, X., 2020. Loss Ensembles for Intracranial Aneurysm Segmentation: An Embarrassingly Simple Method [WWW Document]. URL http://adam.isi.uu.nl/results/results-miccai-2020/participating-teams-miccai-2020/junma/ .

Ma, J., Chen, J., Ng, M., Huang, R., Li, Y., Li, C., Yang, X., Martel, A.L., 2021. Loss odyssey in medical image segmentation. Med. Image Anal. 71. doi:10.1016/j.media.2021.102035.

Maier-Hein, L., Reinke, A., Kozubek, M., Martel, A.L., Arbel, T., Eisenmann, M., Hanbury, A., Jannin, P., Müller, H., Onogur, S., Saez-Rodriguez, J., van Ginneken, B., Kopp-Schneider, A., Landman, B.A., 2020. BIAS: Transparent reporting of biomedical image analysis challenges. Med. Image Anal. 66, 101796. doi:10.1016/j.media.2020.101796.

McKinney, W., 2010. Data Structures for Statistical Computing in Python. pp. 56–61. https://doi.org/10.25080/Majora-92bf1922-00a

Mendrik, A.M., Vincken, K.L., Kuijf, H.J., Breeuwer, M., Bouvy, W.H., De Bresser, J., Alansary, A., De Bruijne, M., Carass, A., El-Baz, A., Jog, A., Katyal, R., Khan, A.R., Van Der Lijn, F., Mahmood, Q., Mukherjee, R., Van Opbroek, A., Paneri, S., Pereira, S., Persson, M., Rajchl, M., Sarikaya, D., Smedby, Ö., Silva, C.A., Vrooman, H.A., Vyas, S., Wang, C., Zhao, L., Biessels, G.J., Viergever, M.A., 2015. MRBrainS challenge: online evaluation framework for brain image segmentation in 3T MRI scans. Comput. Intell. Neurosci. 2015. doi:10.1155/2015/813696.

Merkel, D., 2014. Docker: lightweight Linux containers for consistent development and deployment. Linux J 2014. doi:10.5555/2600239.2600241.

Michael Waskom and the seaborn development team, 2020. mwaskom/seaborn. https://doi.org/10.5281/zenodo.592845

Mouches, P., Forkert, N.D., 2019. A statistical atlas of cerebral arteries generated using multi-center MRA datasets from healthy subjects. Sci. Data 6, 29. doi:10.1038/s41597-019-0034-5.

Nakagawa, D., Nagahama, Y., Policeni, B.A., Raghavan, M.L., Dillard, S.I., Schumacher, A.L., Sarathy, S., Dlouhy, B.J., Wilson, S., Allan, L., Woo, H.H., Huston, J., Cloft, H.J., Wintermark, M., Torner, J.C., Brown, R.D., Hasan, D.M., 2019. Accuracy of detecting enlargement of aneurysms using different MRI modalities and measurement protocols. J. Neurosurg. 130, 559–565. doi:10.3171/2017.9.JNS171811.

Nakao, T., Hanaoka, S., Nomura, Y., Sato, I., Nemoto, M., Miki, S., Maeda, E., Yoshikawa, T., Hayashi, N., Abe, O., 2018. Deep neural network-based computer-assisted detection of cerebral aneurysms in MR angiography. J. Magn. Reson. Imaging 47, 948–953. doi:10.1002/jmri.25842.

Nieuwkamp, D.J., Setz, L.E., Algra, A., Linn, F.H., de Rooij, N.K., Rinkel, G.J., 2009. Changes in case fatality of aneurysmal subarachnoid haemorrhage over time, according to age, sex, and region: a meta-analysis. Lancet Neurol 8, 635–642. doi:10.1016/S1474-4422(09)70126-7.

Nishimori, M., Fukumoto, S., Ueda, D., Shimahara, Y., Shimazaki, A., Choppin, A., Miki, Y., Katayama, Y., Doishita, S., Shimono, T., Yamamoto, A., 2018. Deep Learning for MR Angiography: Automated Detection of Cerebral Aneurysms. Radiology 290, 187–194. doi:10.1148/radiol.2018180901.

Park, A., Chute, C., Rajpurkar, P., Lou, J., Ball, R.L., Shpanskaya, K., Jabarkheel, R., Kim, L.H., McKenna, E., Tseng, J., Ni, J., Wishah, F., Wittber, F., Hong, D.S., Wilson, T.J., Halabi, S., Basu, S., Patel, B.N., Lungren, M.P., Ng, A.Y., Yeom, K.W., 2019. Deep learning–assisted diagnosis of cerebral aneurysms using the HeadXNet Model. JAMA Netw. Open 2, e195600. doi:10.1001/jamanetworkopen.2019.5600.

Raghavan, M.L., Ma, B., Harbaugh, R.E., 2009. Quantified aneurysm shape and rupture risk. J. Neurosurg. 102, 355–362. doi:10.3171/jns.2005.102.2.0355.

Rjiba, S., Urruty, T., Bourdon, P., Maloigne-Fernandez, C., Delepaul, R., Guillevin, R., 2020. AneurysmNet: Deep Neural Network-based Segmentation of Aneurysms in 3D-MR Angiography [WWW Document]. URL http://adam.isi.uu.nl/results/results-miccai-2020/participating-teams-miccai-2020/xlim/ .

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics) 9351, 234–241. doi:10.1007/978-3-319-24574-4_28.

Sekuboyina, A., Rempfler, M., Kukačka, J., Tetteh, G., Valentinitsch, A., Kirschke, J.S., Menze, B.H., 2018. Btrfly net: vertebrae labelling with energy-based adversarial learning of local spine prior. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics) 11073 LNCS 649–657. doi:10.1007/978-3-030-00937-3_74.

Shit, S., Shah, D., Fava Sanches, A., Menze, B.H., 2020. 2D TriWingedNet for 3D Intracranial Aneurysm Segmentation [WWW Document]. URL http://adam.isi.uu.nl/results/results-miccai-2020/participating-teams-miccai-2020/ibbm/ .

Sudre, C.H., Li, W., Vercauteren, T., Ourselin, S., Jorge Cardoso, M., 2017. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics) 10553 LNCS 240–248. doi:10.1007/978-3-319-67558-9_28.

Sulayman, N., Al-Mawaldi, M., Kanafani, Q., 2016. Semi-automatic detection and segmentation algorithm of saccular aneurysms in 2D cerebral DSA images. Egypt. J. Radiol. Nucl. Med. doi:10.1016/j.ejrnm.2016.03.016.

Taha, A.A., Hanbury, A., 2015. Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool. BMC Med. Imaging 15. doi:10.1186/s12880-015-0068-x.

Tan, M., Le, Q.V., 2019. EfficientNet: rethinking model scaling for convolutional neural networks. In: 36th Int. Conf. Mach. Learn. ICML 2019 2019-June, pp. 10691–10700.

Timmins, K., Bennink, E., Schaaf, I. van der, Velthuis, B., Ruigrok, Y., Kuijf, H., 2020. Intracranial Aneurysm Detection and Segmentation Challenge. https://doi.org/10.5281/ZENODO.3715848

Tustison, N.J., Cook, P.A., Gee, J.C., 2011. N4Itk 29, 1310–1320. https://doi.org/10.1109/TMI.2010.2046908.N4ITK

Vallat, R., 2018. Pingouin: statistics in Python. J. Open Source Softw. 3, 1026. doi:10.21105/joss.01026.

Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S.J., Brett, M., Wilson, J., Millman, K.J., Mayorov, N., Nelson, A.R.J., Jones, E., Kern, R., Larson, E., Carey, C.J., Polat, İ., Feng, Y., Moore, E.W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E.A., Harris, C.R., Archibald, A.M., Ribeiro, A.H., Pedregosa, F., van Mulbregt, P., 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat. Methods 17, 261–272. doi:10.1038/s41592-019-0686-2.

Vlak, M.H.M., Algra, A., Brandenburg, R., Rinkel, G.J.E., 2011. Prevalence of unruptured intracranial aneurysms, with emphasis on sex, age, comorbidity, country, and time period: a systematic review and meta-analysis. Lancet Neurol 10, 626–636. doi:10.1016/S1474-4422(11)70109-0.

Walińska, U., Klimont, M., Kraft, M., Pieczyński, D., Mikołajczak, M., Pawlak, M., 2020. Inteneural at Aneurysm Detection And segMentation Challenge [WWW Document]. URL http://adam.isi.uu.nl/results/results-miccai-2020/participating-teams-miccai-2020/inteneural/ .

Wang, S., Manning, C., 2013. Fast Dropout Training, in: ICML.

Wardlaw, J.M., White, P.M., 2000. The detection and management of unruptured intracranial aneurysms. Brain 123, 205–221. doi:10.1093/brain/123.2.205.

Warfield, S.K., Zou, K.H., Wells, W.M., 2004. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. IEEE Trans. Med. Imaging 23, 903–921. doi:10.1109/TMI.2004.828354.

White, P.M., Teasdale, E.M., Wardlaw, J.M., Easton, V., 2001. Intracranial aneurysms: CT angiography and MR angiography for detection - prospective blinded comparison in a large patient cohort. Radiology 219, 739–749. doi:10.1148/radiology.219.3.r01ma16739.

White, P.M., Wardlaw, J.M., Easton, V., 2000. Can noninvasive imaging accurately depict intracranial aneurysms? A systematic review. Radiology 217, 361–370. doi:10.1148/radiology.217.2.r00nv06361.

Wrede, K.H., Matsushige, T., Goericke, S.L., Chen, B., Umutlu, L., Quick, H.H., Ladd, M.E., Johst, S., Forsting, M., Sure, U., Schlamann, M., 2017. Non-enhanced magnetic resonance imaging of unruptured intracranial aneurysms at 7 Tesla: Comparison with digital subtraction angiography. Eur. Radiol. 27, 354–364. doi:10.1007/s00330-016-4323-5.

Wu, Y., He, K., 2018. Group Normalization.

Yang, Y., Lin, Y., Li, Y., Wei, D., Ma, K., Yang, X., Zheng, Y., 2020. Automatic Aneurysm Segmenattion via 3D U-Net Ensemble [WWW Document]. URL http://adam.isi.uu.nl/results/results-miccai-2020/participating-teams-miccai-2020/joker/ .

Zhou, Z., Sodha, V., Siddiquee, M.M.R., Feng, R., Tajbakhsh, N., Gotway, M.B., Liang, J., 2019. Models Genesis: Generic Autodidactic Models for 3D Medical Image Analysis.