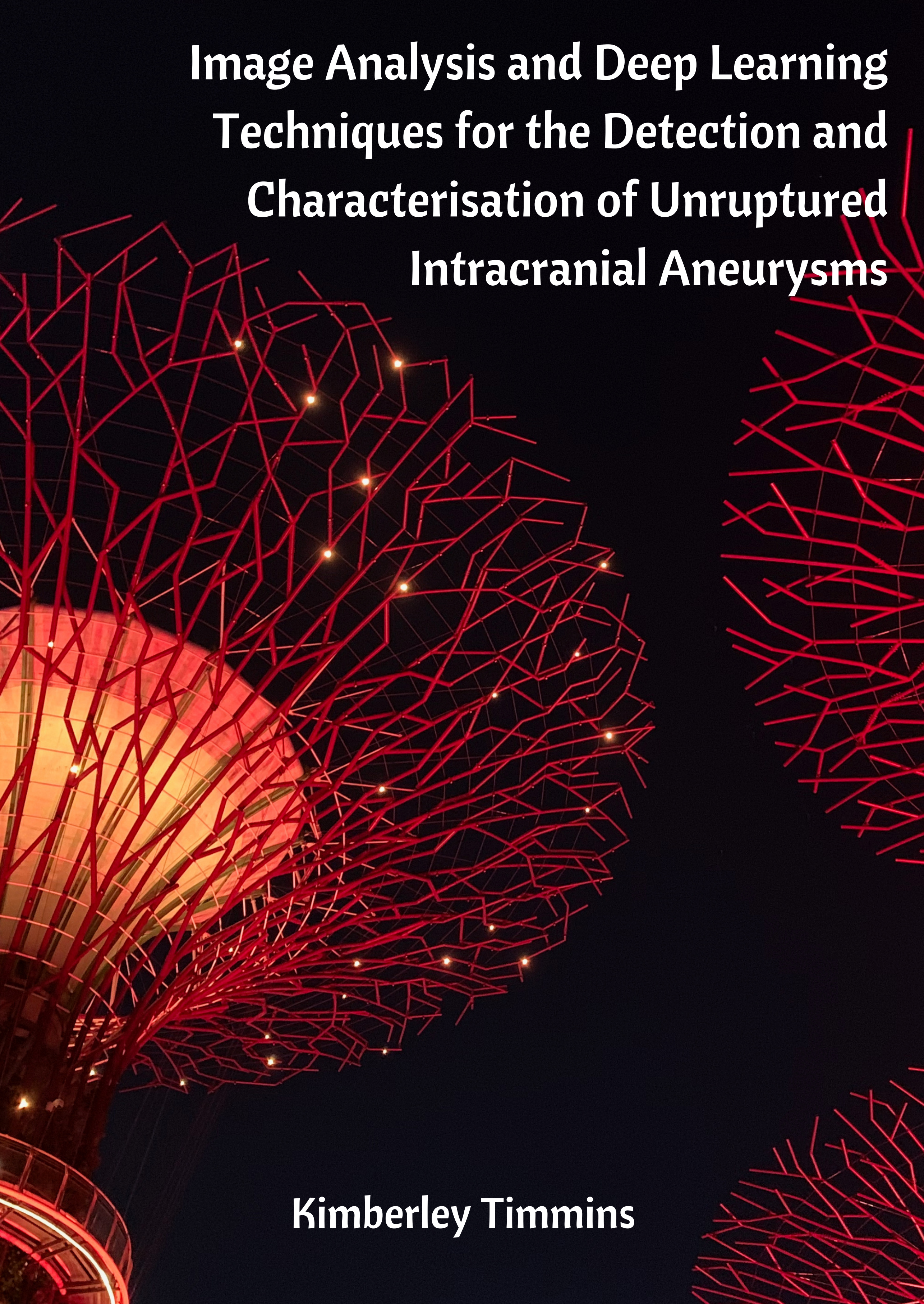
Image Analysis and Deep Learning Techniques for the Detection and Characterisation of Unruptured Intracranial Aneurysms



Kimberley Timmins

Image Analysis and Deep Learning Techniques for the Detection and Characterisation of Unruptured Intracranial Aneurysms

Kimberley Timmins

ISBN: 978-94-6458-854-5

Cover design by Kimberley Timmins. Printed by Ridderprint.

Image Analysis and Deep Learning Techniques for the Detection and Characterisation of Unruptured Intracranial Aneurysms

Beeldanalyse en deep learning technieken voor de detectie en karakterisering van ongebarsten hersenaneurysma's
(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Utrecht
op gezag van de rector magnificus, prof.dr. H.R.B.M. Kummeling ingevolge het
besluit van het college voor promoties
in het openbaar te verdedigen op donderdag 2 februari 2023 des middags te 2.15 uur

door

Kimberley Michelle Timmins

geboren op 10 augustus 1993
te Lancaster, Verenigd Koninkrijk

Promotor:	Prof. dr. B. K. Velthuis
Copromotoren:	Dr. I. C. Schaaf Dr. H. J. Kuijf

The work of this thesis was supported by a grant from the Dutch Heart Foundation (CVON2015-08 ERASE, CVON2018-02 ANEURYSM@RISK).
Financial support by the Dutch Heart Foundation for the publication of this thesis is gratefully acknowledged.

Contents

Chapter 1

Introduction

1.1 Unruptured Intracranial Aneurysms

An unruptured intracranial aneurysm (UIA) is a focal outward bulging in the wall of an intracranial artery. Generally, UIAs occur in the larger blood vessels at the base of the brain, known as the Circle of Willis (CoW). Intracranial aneurysms have a prevalence of approximately 3% in the general adult population [1]. The majority of aneurysms do not rupture during a lifetime, and most patients can have an intracranial aneurysm for a long period of time without any negative consequences. However, if an aneurysm ruptures, it bleeds into the cerebrospinal fluid surrounding the brain, a type of stroke called aneurysmal subarachnoid haemorrhage (aSAH). This can be devastating, resulting in death or severe, long lasting disability [2]. Therefore, it is important that UIAs are detected early to allow treatment decisions to be made. UIAs are often found as incidental findings on CT or MR scans, but screening for unruptured intracranial aneurysms using time-of-flight magnetic resonance angiography (TOF-MRA) is also increasing. This allows us to understand more about risk factors of aneurysm development, for example, in patients with a positive or familial history of aSAH and intracranial aneurysms [3, 4]. When an intracranial aneurysm is first diagnosed in an angiographic scan, a multidisciplinary team of treating physicians will make a rupture risk assessment of the aneurysm, which is of great importance for deciding the best clinical strategy. Aneurysm treatment, by means of neurosurgical clipping or by endovascular approach (coiling, (flow diverting) stents or web-devices) can prevent rupture, but it carries a significant risk of complications that has to be balanced against rupture risk [5]. Therefore, preventative treatment should only be considered in patients with aneurysms that have a high risk of rupture. To make an individual patient treatment decision, it is important to understand the rupture and treatment complication risk factors of the aneurysm, as well as considering patient preference. Known aneurysm rupture risk factors include: the Population of the patient; Hypertension; Age over 70 years; Size of the aneurysm; Earlier aSAH from a previous aneurysm and Site (location) of the aneurysm (PHASES score) [6]. Rupture risk assessment requires accurate and quantitative characterisation of the intracranial aneurysm on an individual level.

1.2 Clinical Aneurysm Assessment

For UIA diagnosis or follow-up, a patient will usually undergo a TOF-MRA (Figure 1.1) or a contrast-enhanced computed tomography angiography scan (CTA). Both of these scans provide brain images where the blood vessels have a higher intensity relative to the rest of the brain scan, allowing the vessels and possible intracranial aneurysms, to be clearly visualised. In TOF-MRAs, the higher intensity in the blood vessels is caused by the magnetic saturation of the blood. In CTA, contrast agent is administered intravenously to enhance the intensity in the blood vessels. TOF-MRAs have preference

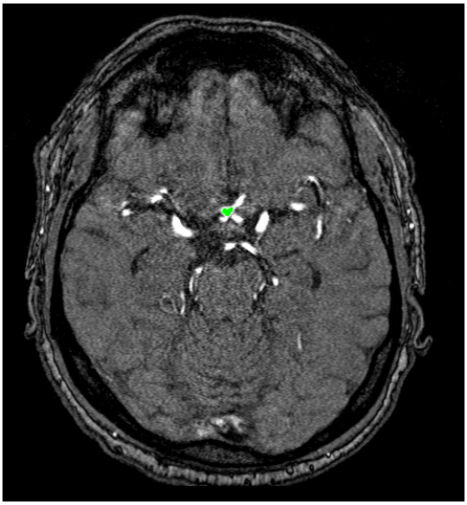


Figure 1.1: An example TOF-MRA of a patient with an unruptured intracranial aneurysm in the anterior communicating artery. Annotation of aneurysm is shown overlaid in green.

over CTAs in most cases as they require neither contrast administration nor radiation exposure. When reviewing scans on the presence of an intracranial aneurysm, a radiologist will first visually inspect the scan carefully and determine the location of the aneurysm. They may do this by scrolling through the 2D slices of the scan and by making a maximum intensity projection (MIP) of the image and/or make a 3D volume render so the vessels can be seen more clearly. Once an aneurysm has been detected, various size measurements of the aneurysm are taken as shown in Figure 1.2. The aneurysm size is a key predictor for risk of rupture of the UIA and is taken into consideration using the PHASES score for rupture risk assessment [6]. If the risk of treatment complication outweighs the rupture risk, the patient will undergo aneurysm follow-up imaging with TOF-MRA or CTA to monitor for signs of aneurysm instability such as aneurysm growth or shape change [7]. In case of aneurysm instability, treatment should be reconsidered.

1.3 Aneurysm Detection and Segmentation

Increasing numbers of patients are being screened, and undergo follow-up, as part of family screening in patients with a ruptured intracranial aneurysm and in patients with autosomal dominant polycystic kidney disease [4]. Automatic detection of intracranial aneurysms would speed up the clinical workflow, however it is important that this does not compromise accuracy. The detection of aneurysms from angiographic scans can be difficult, especially for small intracranial aneurysms. A wide range of sensitivities by visual assessment have been quoted in studies including as

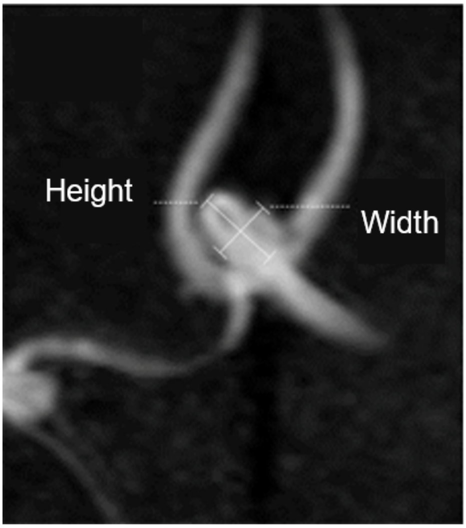


Figure 1.2: Height and width measurements of an unruptured intracranial aneurysm on a TOF-MRA.

low as 28% for small aneurysms to 90% for larger aneurysms [8–10]. After detecting the aneurysm, segmentation (extraction) of the aneurysm from the scan would allow volumetric and morphology measurements to be made. Segmentation of intracranial aneurysms is not easy because there is a large variability in the configuration of the aneurysm relative to parent vessels, making it difficult to define the aneurysm neck. To solve this problem, previous semi-automatic methods defined the neck where the aneurysm is attached to the vessel before segmenting the aneurysm [11]. Several other (semi-) automatic methods for unruptured intracranial aneurysm detection and/or segmentation have already been proposed, including approaches that use blobness filters [12] and shape analysis of vessel surfaces [13]. More recently, various deep learning detection approaches have been developed that prove to have high accuracy [14–16]. Most deep learning approaches use a convolutional neural network (CNN), which is dependent upon the intensities in the image. It is usual to normalise the intensity of MRAs but the scans can still have a large variability in quality and resolution because of differing scanners, protocols and field strengths being used. Deep learning approaches require large annotated datasets for training, which can often be difficult to find [17]. Furthermore, because of the large variance of unruptured intracranial aneurysm configurations, shape and locations as well as difference in scan protocols, the dataset should include as many variations as possible. Prior to this thesis, no such dataset was publicly available.

An alternative to image-based deep learning is geometric deep learning, which can be intensity independent, suggesting it can be applied across different modalities. Geometric deep learning is a relatively new field in medical imaging and includes point clouds [18, 19], graph convolutional neural networks [20, 21] and mesh neural networks [22, 23]. Vessel surface models could be derived from brain MRA and CTA by extracting the higher intensity vessels before fitting meshes to the vessel surface. Aneurysms are identifiable on vessel surface meshes as an outward bulge relative to the smooth, tubular vessel surface. By exploiting this topology of vessel surface meshes, a geometric model for aneurysm detection and or segmentation could be developed. Such a model could be modality independent working across both MRA and CTA, which would be a great aid in clinical practice, where both modalities are used in follow-up imaging.

1.4 Aneurysm Growth and Shape

Once an unruptured intracranial aneurysm has been detected in a scan, measurements of the UIA are made to assess if it has grown. Intracranial aneurysmal growth is difficult to define and measure, since it does not appear to be linear process [24]. The definition of intracranial aneurysmal growth differs between studies and centres around the world, although the generally accepted aneurysm growth definition is an increase in either 2D height or width of at least 1 mm [25]. This growth definition relies on reliable and reproducible height and width measurements of intracranial aneurysms. However, because of the 3D nature of blood vessels and aneurysms, manual 2D measurements can be difficult. Considerable variability between intra- and inter-observer measurements of intracranial aneurysms on CTAs and MRAs has been reported [26, 27]. 3D measures could provide a volume measurement [28, 29], which would be independent of the measurement orientation, thereby removing some observer variability. 3D measurements of UIAs have been performed previously [29, 30], however the reliability of such measurements for growth assessment was not known prior to this thesis. Furthermore, 3D measurements of aneurysms would allow for the inclusion of quantitative shape measurements for aneurysm stability and rupture risk.

Irregular aneurysm shape is a known predictor for growth and subsequently rupture [7, 31]. Currently, shape is included in the clinical growth prediction score, ELAPSS [31], which includes the predictors: Earlier subarachnoid haemorrhage; aneurysm Location; Age; Population; aneurysm Size and Shape. It is known that the shape of an intracranial aneurysm can change, independent of the size of the aneurysm [32]. Up to now, the shape of intracranial aneurysms has often been defined as 'regular' or 'irregular' based on visual assessment by the radiologist, where irregular may be the presence of blebs, wall protrusions or multiple lobes [31]. These visual measurements are observer dependent.

With image analysis techniques, quantitative 3D measures of the shape or morphology of intracranial aneurysms have been introduced [33–35]. Differences have been found in quantified morphology between unruptured and growing or ruptured aneurysms [34] and morphology has also been used as a predictor for the instability of unruptured intracranial aneurysms [36, 37]. There is a variety of different shape/morphology measures used in these studies, which means that no standard definition of intracranial aneurysm shape or morphology has yet been established. The Image Biomarker Standardisation Initiative (IBSI) guidelines were made [38] to standardise radiomics, including morphology measurements, on medical images. By using these standard definitions, a better understanding of change in morphology of aneurysms can be made. Intracranial segmentation approaches of unruptured intracranial aneurysms would allow these volumetric and morphologic measures to be made automatically, speeding up clinical work flow and removing observer bias.

1.5 Outline of the thesis

The objective of this thesis is to develop and investigate image analysis and quantitative techniques for the detection and growth risk assessment of unruptured intracranial aneurysms.

CHAPTER 2 describes the Aneurysm Detection and segMentation (ADAM) Challenge, which we hosted at MICCAI conference 2022. We released a publicly available training set of annotated TOF-MRAs and evaluated method submissions for intracranial aneurysm detection and segmentation.

CHAPTER 3 presents a feasibility study of using variational autoencoders with TOF-MRAs. This could have future use in an anomaly detection method for unruptured intracranial aneurysms from TOF-MRAs.

CHAPTER 4 explores the use of mesh convolutional neural networks for modality independent intracranial aneurysm detection based on vessel surface meshes.

CHAPTER 5 considers how unruptured intracranial aneurysm growth is currently assessed. We performed a reliability and agreement interobserver study for 2D and 3D growth measurements in TOF-MRAs.

CHAPTER 6 investigates quantified morphology changes of unruptured intracranial aneurysms and if these morphology changes relate to UIA growth.

CHAPTER 7 describes the development of an intracranial aneurysm growth prediction model using a mesh convolutional neural network.

Finally, in **CHAPTER 8** the above chapters are summarised and some discussion of the results is provided including limitations and future work that could be implemented.

Chapter 2

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

Published as: K. M. Timmins, I. C. van der Schaaf, E. Bennink, Y. M. Ruigrok, et al. "Comparing methods of detecting and segmenting unruptured intracranial aneurysms on TOF-MRAS: The ADAM challenge," *NeuroImage*, vol. 238 (2021), p. 118216.

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 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-2 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/ .

2.1 Introduction

Approximately 3% of the world general population have an unruptured intracranial aneurysm (UIA) [1]. 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) [3]. 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 [2, 39]. 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 [6, 31]. 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 [40]. This allows informed treatment decisions to be made [41]. 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 [42, 43]. 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 [44–46]. 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 [47].

The detection and measurement of UIAs can be difficult and it has been reported that approximately 10% of all UIAs are missed during screening [10, 26, 27, 39]. 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% [48]. However, detection by radiologists is improving as MRA scan resolution is increasing, especially with higher field strengths [49, 50]. 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 [31–33]. Furthermore, quantified shape measurements of the UIAs may aid in models assessing treatment complication risk [51].

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 [11]. The shape of the aneurysm has been used in some UIA detection techniques, including using blobness filters [12] and shape analysis of the surface of the vessel segmentations [13, 52, 53]. Furthermore, multiple deep learning techniques for UIA detection have been developed with high accuracy [14–16]. However, most methods are developed for CTA or Digital Subtraction Angiography (DSA) 2D images [54, 55] 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 [56]. 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 [57].

2.2 Material and Methods

2.2.1 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 2.2.3) 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 [58] 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 2.2.3) using evaluation code that was made publicly 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 [59, 60] .

2.2.2 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 [57]. 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 [57]. 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.

2.2.3 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 [1]. The dataset was realistic and diverse, reflecting different standard clinical protocols used between MR-scanners and over time.

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. 2.1. 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.

PRE-PROCESSING All images were pre-processed with N4 bias-field correction [62]. The structural image was aligned to the corresponding TOF-MRA using the elastix

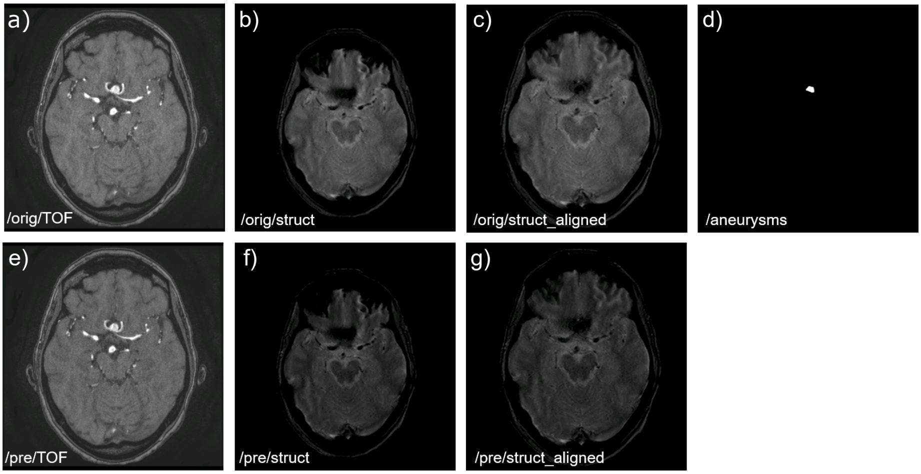


Figure 2.1: An example training case. Top Row: a) Original TOF-MRA, b) Original structural MR image c) original structural MR image aligned to TOF-MRA in a) using registration parameters derived from elastix [61] d) binary aneurysm image derived from annotations as described in Section2.2.3 Bottom Row: e) pre-processed TOF-MRA using n4 bias field correction [62], f) structural MR image pre-processed using n4, g) pre-processed structural MR image aligned to TOF-MRA using pre-determined registration parameters.

toolbox for image registration [61]. The transformation parameters used were provided with the training data. Both original and pre-processed data was provided to the registered participants.

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 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.

2.2.4 Assessment method

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) (95 th percentile) and (3) Volumetric Similarity (VS) [63]. 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.

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 versus test performance.

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 [64] 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.

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.

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.

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 [41]. 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.

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 [65]. 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 intra-subject 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.

Train versus 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 [66], scipy [67], seaborn [68] and pingouin [69] toolboxes with Python 3.7.

2.3 Results

2.3.1 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).

2.3.2 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).

TASK 1 SUBMISSIONS **MiBaumgartner** submitted a 3D neural network based on the Retina U-Net architecture [70]. The decoder was extended to incorporate semantic segmentation information and followed by a Path Aggregation Network [71] to generate the features used for the detection prediction. [72]

Unil_chuv submitted a 3D U-Net [73] which was patch trained using patches selected based on landmark points from a registered vessel atlas [74]. 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. [75]

TASK 2 SUBMISSIONS **IBBM** submitted a 2D convolutional neural network with Tri-Winged-Net architecture based on the BtrflyNet [76]. 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. [77]

Inteneural submitted a method including three 2D neural networks with U-Net architecture based on EfficientNet [78] that were pre-trained using ImageNet [79]. 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 performed using Jerman filter [80]. A loss function including both a generalised dice loss [81] and boundary loss [82] was used. The final prediction was determined as an average of the evaluated models' outputs. [83]

Joker submitted a 3D fully-convolutional neural network based on no new U-Net (nnUNet) [84]. Group Normalisation [85] 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. [86]

JunMa submitted a 3D fully-convolutional neural network based on no new U-Net (nnUNet) [84]. Networks were trained using five-fold cross validation and two different loss functions: Dice loss and cross entropy, and Dice loss with topK loss [87] because the two losses have been proven to be robust on highly imbalanced segmentation tasks [88]. At prediction, the five models with optimum performance were ensembled. [89, 90]

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 [73] style architectures with two paths including contextual information in both the encoding and decoding path [91] trained on different loss functions. The loss functions were the sum of cross entropy, (generalised) Dice loss [81] and boundary loss [82]. [92]

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 [93]. The classification consisted of a 3D convolutional neural network to distinguish between true and false positives. [94]

TUM_IBBM submitted a U-Net based architecture with MRA and aligned structural image as different input channels [95]. Two networks were trained on sagittal and coronal slices and during testing, voxelwise predictions of both models were averaged. [96]

Xlim submitted a hybrid two input neural network: one for 3D patches and the second for the corresponding maximum intensity projection of the patches [97]. 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. [98]

Zelosmediacorp submitted a 3D fully convolutional neural network with a U-Net like architecture [73] 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 [99] was used for both training and inference. [100]

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

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(0.07-0.13)
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(18.47-27.10)	0.43(0.34-0.51)	0.70(0.64-0.76)

Table 2.1: 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 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.

2.3.3 Metrics and rankings

The mean performance of each participating team for task 1 is shown in Table 2.1 and for task 2 is shown in Table 2.2. The 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.2.2 and 2.3 are bar charts and box plots 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 2.2 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.

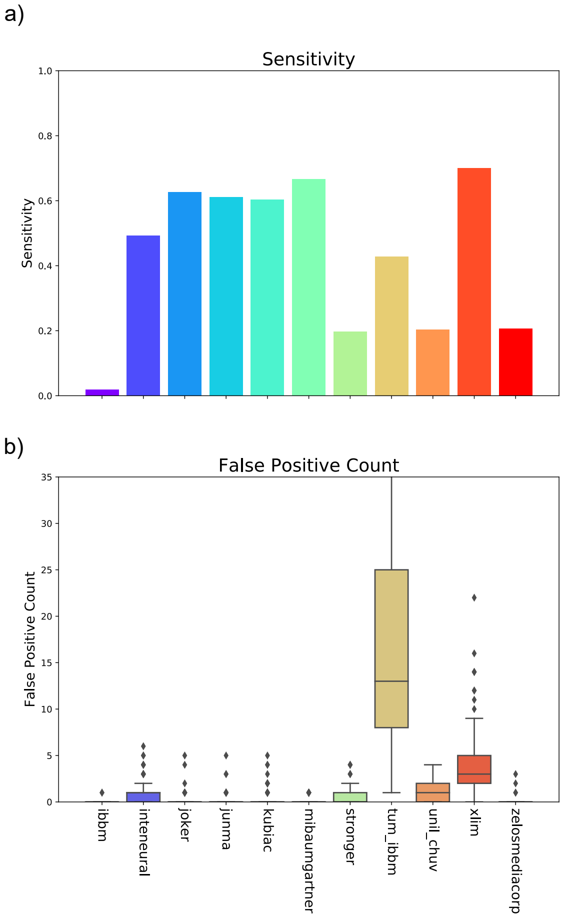


Figure 2.2: Sensitivity and False Positive Count for all teams for all scans in the test set. a) Bar chart of sensitivity of all teams for task 1, taken as an average across all scans in the test set b) Box plot of total false positive count per scan of all teams for all scans in the test set.

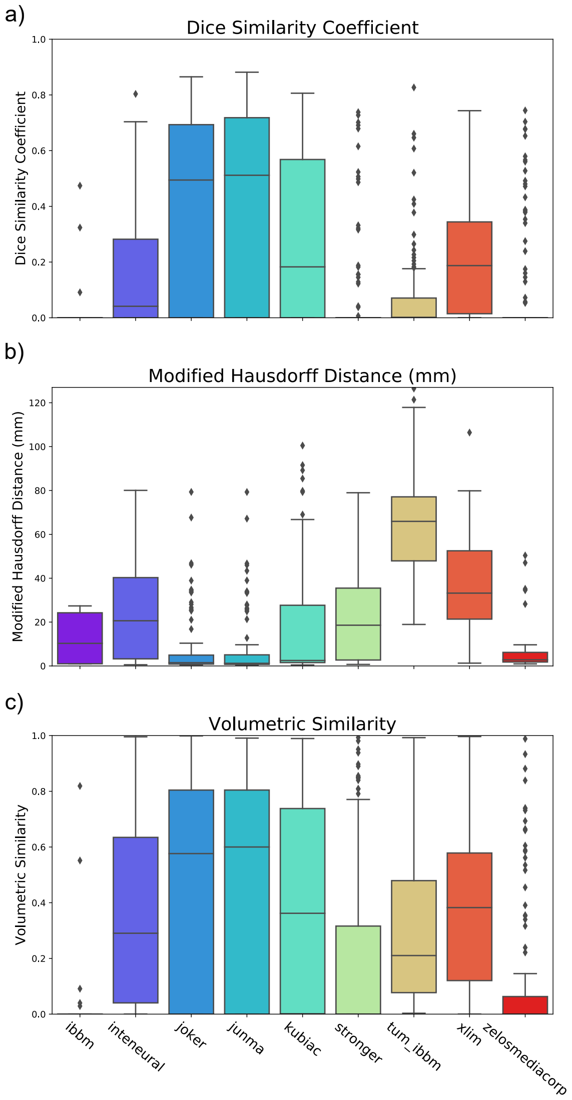


Figure 2.3: Box plots of metrics for all teams for task 2. a) Dice Similarity Coefficient (DSC) b) Modified Hausdorff Distance (MHD) c) Volumetric Similarity for all scans containing a UIA. Each point in the box plots is the metric evaluated on one scan in the test set for each method. The centre line shows the median metric of all scans.

Place	Team	DSC	MHD (mm)	VS	Rank
1	junma	0.41(0.35-0.47)	8.96(5.59-12.71)	0.50(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(0.00-0.09)
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(32.72-41.30)	0.39(0.34-0.44)	0.41(0.35-0.47)
6	zelosmediacorp	0.09(0.06-0.13)	9.79(4.66-15.50)	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(0.00-0.02)	12.77(0.97-25.81)	0.01(0.00-0.03)	0.69(0.67-0.77)
9	TUM_IBBM	0.07(0.05-0.10)	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)
1	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 2.2: Task 2: Average metrics and ranking for each team, and the brackets contain the 95% confidence interval determined using bootstrapping. The 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

2.3.4 Further Analysis

METHOD ANALYSIS 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 2.3 . The U-Net [73] 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-2 ranking segmentation methods used nnU-Net [84] 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 [87]. 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.

SEGMENTATION PERFORMANCE OF TRUE UIAS To evaluate segmentation performance, average segmentation metrics were determined for all teams for only the true UIAs

Team	Task	Place	Architecture	2D/3D	Segmentation loss function	Ensemble	Use of Structural Image	Data Augmentation	Post Processing
mibaumgartner	1	1	Retina U-Net + Path Aggregation Network	3D	Dice and Cross Entropy	Yes	Yes	C,M,R,S	FPS
unil_chuv	1	9	U-Net	3D	Dice	No	No	C,M,R	FPC,L,M
junma	2	1	U-Net (nn-Unet)	3D	Dice and Cross entropy or Top-K	Yes	No	C,M,R,S	
joker	2	2	U-Net (nn-Unet)	3D	Dice ranking	Yes	Yes	C,E,M,R,S	
kubiac	2	3	Multi resolution U-Net style network and CNN classifier	3D	Cross entropy,(generalised) Dice and Boundary	Yes	Yes	T	L
inteneural	2	4	Efficientnet-b1	2D	Generalised Dice and Boundary	Yes	No		FPC, FPS
xlim	2	5	AneurysmNet	2D (MIP), 3D	Dice	No	No		FPS
zelosmediacorp	2	6	U-Net	3D	Dice	Yes	No	M,R,T	FPS
stronger	2	7	U-Net and CNN classifier	3D	Dice and Cross entropt	Yes	No	M,R,T	
IBBM	2	8	TriWingedNet	2.5D	Dice	No	No		
TUM_IBBM	2	9	U-Net	2D	Dice	Yes	Yes	M,R,SH	FPS

Table 2.3: Submitted methods sorted on their final ranking per task, with highest placed ranking first, and information about method design. [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

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(0.45-0.60)	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(0.27-0.50)	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)

Table 2.4: The mean segmentation metrics of each team evaluated only on the detected true UIAs. The arrows and brackets signify the difference between the original task 2 ranking (Table 2.2), 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.

that were detected, as displayed in Table 2.4. 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.

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).

SIZE OF UIAS The detection and segmentation performance improved with the size of the UIA. Fig. 2.4 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. 2.5 a) and b) demonstrate that the segmentation performance also increased with UIA size. In 2.5a) the median DSC over all teams for each UIA was plotted against the individual UIA diameter. In 2.5b), the UIA diameter is again split into four quartiles and the mean DSC over all teams for each UIA was

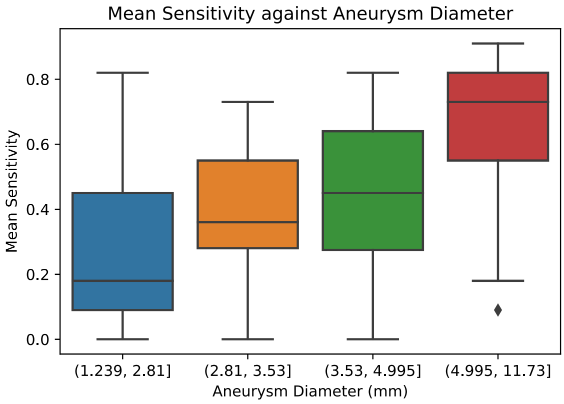


Figure 2.4: Sensitivity of methods for UIA of different sizes. Sensitivity of all teams for each UIA as a function of maximum UIA diameter in mm, when separating UIA diameter into four quartiles. Each point included in the box plot is the mean sensitivity of all teams across each UIA.

included. DSCs for individual teams were plotted in Appendix D, Fig. 2.

INTRA-SUBJECT ANALYSES Table 2.5 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 and 2. For sensitivity, DSC and volumetric similarity, all methods had p >= 0.05 suggesting that performance was not different between baseline and follow-up subjects.

TRAIN VERSUS 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

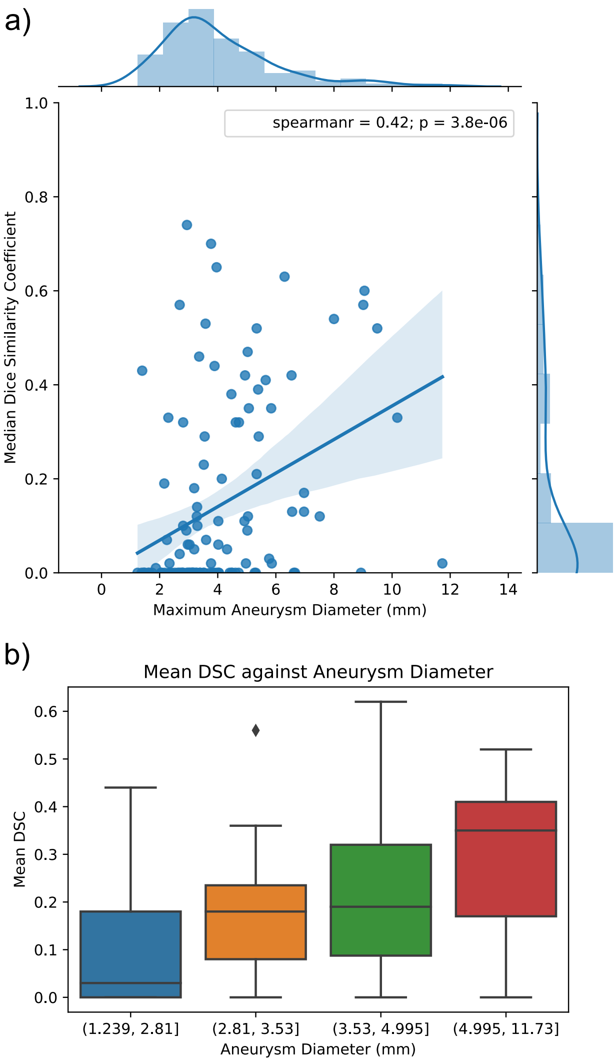


Figure 2.5: Dice Similarity Coefficient (DSC) as a function of aneurysm diameter a) Median Dice Similarity Coefficient (DSC) of all teams for each UIA as a function of maximum UIA diameter in mm, b) Mean DSC of all teams for each UIA as a function of maximum UIA diameter in mm, when separating UIA diameter into four quartiles.

Team	GT Volume Change(µl)	Predicted Volume Change(µl)	Kedall'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)

Table 2.5: Comparison of volume change measurements (median (IQR)) for ground truth (GT) 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 (GT) 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.

correspond to Tables 2.1 and 2.2 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.

2.4 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 [10, 49]. 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.

2.4.1 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 [70], before classifying and regressing anchor boxes using a Path Aggregation Network [71]. 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).[84] 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. 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.

2.4.2 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 [101] and segmentation tasks [59, 102]. 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-3 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.

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 [81], 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 [82]. 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.

2.4.3 Segmentation Performance of true UIAs

The top-3 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.

2.4.4 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.

2.4.5 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 to detect, which is also consistent with studies investigating visual detection of aneurysms. White et. al. [10] 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.

2.4.6 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 intra-subject performance, was the train, test and validation splits between the methods, as many methods did not take baseline-follow-up pairs into account.

2.4.7 Train versus 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.

2.4.8 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 junma's method 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.

2.5 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.

2.6 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).

2.7 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).

2.8 Appendices

All appendices can be found online at: https://www.sciencedirect.com/science/article/pii/S1053811921004936

Chapter 3

Variational Autoencoders with a Structural Similarity Loss in Time of Flight MRAs

Published as: K. Timmins, I. van der Schaaf, Y. Ruigrok, B. Velthuis, and H. J. Kuijf. "Variational Autoencoders with a Structural Similarity Loss in Time of Flight MRAs," *Medical Imaging 2021: Image Processing*, SPIE, 2021, p. 115.

Abstract

Time-of-Flight Magnetic Resonance Angiographs (TOF-MRAs) enable visualisation and analysis of cerebral arteries. This analysis may indicate normal variation of the configuration of the cerebrovascular system or vessel abnormalities, such as aneurysms. A model would be useful to represent normal cerebrovascular structure and variabilities in a healthy population and to differentiate from abnormalities. Current anomaly detection using autoencoding convolutional neural networks usually use a voxelwise mean-error for optimisation. We propose optimising a variational-autoencoder (VAE) with structural similarity loss (SSIM) for TOF-MRA reconstruction.

A patch-trained 2D fully-convolutional VAE was optimised for TOF-MRA reconstruction by comparing vessel segmentations of original and reconstructed MRAs. The method was trained and tested on two datasets: the IXI dataset, and a subset from the ADAM challenge. Both trained networks were tested on a dataset including subjects with aneurysms. We compared VAE optimisation with L2-loss and SSIM-loss. Performance was evaluated between original and reconstructed MRAs using mean square error, mean-SSIM, peak-signal-to-noise-ratio and dice similarity index (DSI) of segmented vessels.

The L2-optimised VAE outperforms SSIM, with improved reconstruction metrics and DSIs for both datasets. Optimisation using SSIM performed best for visual image quality, but with discrepancy in quantitative reconstruction and vascular segmentation. The IXI dataset had overall better performance, potentially due to the larger, more diverse training data. Reconstruction metrics, including SSIM, were lower for MRAs including aneurysms.

A SSIM-optimised VAE improved the visual perceptive image quality of TOF-MRA reconstructions. A L2-optimised VAE performed best for TOF-MRA reconstruction, where the vascular segmentation is important. SSIM is a potential metric for anomaly detection of MRAs.

3.1 Introduction

3.1.1 Background

Time-of-Flight Magnetic Resonance Angiographs (TOF-MRA) allow the visualisation and analysis of the configuration of the cerebral arteries. This analysis can indicate normal variation in the cerebrovascular system or vessel abnormalities such as aneurysms or stenosis. In addition, variation in the geometry and configuration of the Circle of Willis, could indicate patients at risk for development of an aneurysm and cerebral vascular disease [103–105]. To diagnose these differences, a model would be helpful which can represent the normal variation of the cerebral vessels in a healthy population.

Previous work has demonstrated that autoencoders and variational autoencoders (VAEs) [106] can be trained on images of healthy subjects to reconstruct healthy images. When these models are presented with a different image, containing an anomaly, the model will ignore the anomaly and reconstruct the image as if it would be, as a healthy image. The anomaly or variation may be determined by comparing the reconstructed ‘healthy’ image with the original image. This enables autoencoders and VAEs to be used for unsupervised anomaly detection, including pathology variation in brain MRIs [107, 108]. Most previously implemented VAEs for anomaly detection use per-pixel loss functions such as L1-loss or L2-loss. Such loss functions, make the assumption that intensity values of neighbouring pixels are independent. These approaches are less suitable where the anomaly may result in a change in structure, rather than pixel intensity. This is true in the case for example of aneurysms or vessel irregularities, where the irregularity often has the same intensity as the surrounding vasculature in the MRA and can only be defined as an anomaly by its structure. A contextual loss would allow for more structural features in the image to be taken into consideration. The Structural Similarity Index Measure (SSIM) [109] is an image quality measure which can identify difference in structure of images by comparing patterns in intensities which are normalised for luminance and contrast.

3.1.2 Aim

We investigate the use of the Structural Similarity Index as a loss function for optimisation of a VAE for reconstructing normal TOF-MRAs in healthy patients, compared to a voxelwise L2 loss function. The experiments will be trained on a large publicly available dataset, and a smaller in-house dataset. Prediction of the trained methods will also be performed on a third dataset including subjects with aneurysms.

3.2 Materials and Methods

3.2.1 Datasets

The IXI dataset (a) consists of 570 TOF-MRAs of healthy patients collected from three different hospitals in London, United Kingdom, between 2005 and 2006: Hammersmith Hospital (Philips Healthcare, 3T), Guy's Hospital (Philips Healthcare, 3T) and Institute of Psychiatry (GE, 1.5T). This set was randomly split into sets for training (365 MRAs), validation (91 MRAs) and test (114 MRAs). [110] The IXI database contains subjects who have been screened by radiologists and diagnosed to be healthy, and therefore the scans do not contain any diagnosed aneurysms.

The in-house dataset (b) was a subset of the healthy patient data used for the Aneurysm Detection and segMentation (ADAM) challenge for MICCAI 2020 (https://adam.isi.uu.nl/), consisting of MRAs without aneurysms collected from the University Medical Center Utrecht, the Netherlands [111]. This consists of 46 MRAs which were randomly split into sets for training (16 MRAs), validation (3 MRAs) and test (27 MRAs). The scans were diagnosed by radiologists in the clinic as having no present un-ruptured intracranial aneurysms.

A third, additional test dataset (c) of 30 MRAs from the ADAM challenge [111], of subjects diagnosed with un-ruptured intracranial aneurysms, was included for testing of both the trained methods. This dataset contained images using the same protocols as the in-house dataset (b). It was only used for testing and not used for training.

3.2.2 Pre-processing

All MRAs were corrected for bias field in-homogeneities with the N4 bias field correction algorithm [62]. The intensity of the images was normalised between 0 and 1, based on 1 being 95% of the maximum intensity of the original MRA. Otsu thresholding [112] was used to form a crude brain mask. Patches of 32 x 32 voxels were randomly selected from the masked, normalised MRAs in the training set. A patch size of 32 x 32 voxels was chosen, because this allowed the full width of a vessel to be included in a patch. For the IXI training dataset (a) 1,000 patches per MRA were extracted leading to a total of 365,000 patches for training. For the smaller healthy ADAM training dataset (b) 10,000 patches were extracted per MRA, with a total of 160,000 patches for training. More patches were taken from the ADAM dataset per image, as there were less images in the ADAM dataset than the IXI dataset. The same patches were used for all experiments.

3.2.3 Architecture

A fully convolutional VAE was developed using PyTorch to conserve spatial information in the latent space. The latent space was a multidimensional tensor of 32 x 4 x 4, the size of which was optimised by comparing visual quality of the reconstructed output. The reconstruction loss functions optimised were the voxelwise L2-loss function and a differentiable SSIM loss function. The SSIM loss was implemented with a window size of 11 x 11 and given a weighting of 1,000. The Kullback-Leiber divergence loss term was also included to standardise the distribution of latent space as is standard in VAEs. The learning rate and batch size were optimized for memory and performance with a batch size of 100 patches, and learning rate of 0.01 for L2 and 0.001 for SSIM for all experiments. A total of four networks were trained: two networks each optimising L2 loss and SSIM loss for each of (a) the IXI and (b) the healthy ADAM training datasets. The networks were trained until convergence of the validation loss. The trained networks were then used to predict for each of the test sets: (a) the IXI healthy test set, (b) ADAM healthy test set and (c) the ADAM aneurysm test set. Since they were fully convolutional networks, the predictions were made on the full-sized original pre-processed MRAs. These were tested slice per slice before combining to the full 3D TOF-MRA. The intensity of the resulting reconstructed images was re-scaled to the same intensity scale as the original image. Vessel segmentation was performed on the resulting reconstructed images using a previously trained vessel segmentation U-Net [113].

3.2.4 Evaluation

We compared the performance of two different trained networks for each dataset, one optimised with L2 loss function and the other with a SSIM loss function. Prediction was performed on the test sets from each of the same dataset, a) and b) (Table 3.1). Prediction of both trained networks was also performed on the test set (c) of MRAs of subjects with aneurysms (Table 3.2). Reconstruction performance was evaluated between the original and reconstructed images using mean square error (MSE), mean SSIM, peak signal-to-noise-ratio (PSNR). The dice similarity index (DSI) was used to determine overlap of the vessel segmentations of the reconstructed images and the original images.

3.3 Results

Using L2 and SSIM loss functions for both (a) the IXI and (b) the ADAM datasets allowed for sufficient image reconstruction and vessel segmentation to be performed as shown in Table 3.1 and Figure 3.1. Quantitative reconstruction was evaluated using MSE, mean SSIM and PSNR, with MSE close to zero and all mean SSIM >0.7. L2 loss performed on average better than SSIM loss for all quantitative evaluative purposes, with a higher DSI, SSIM and PSNR and lower MSE for both datasets. The networks trained and tested on the IXI dataset using SSIM loss outperformed the healthy ADAM dataset with regard to quantitative reconstruction of the images when assessing with

		DSI	MSE	Mean SSIM	PSNR
a) IXI	SSIM	0.573 (0.199)	0.003 (0.002)	0.851 (0.039)	26.9 (3.16)
	L2	0.604 (0.111)	0.001 (0.000)	0.914 (0.024)	33.5 (1.09)
b) ADAM	SSIM	0.729 (0.183)	0.012 (0.007)	0.706 (0.089)	20.8 (3.36)
	L2	0.837 (0.065)	0.001 (0.001)	0.883 (0.031)	29.3 (2.41)

Table 3.1: Reconstruction and segmentation metrics for the test set for the datasets from a) the IXI dataset and b) the ADAM challenge, trained on the respective training dataset. Values are provided as mean (standard deviation). DSI: Dice similarity index of the resulting vessel segmentation, MSE: mean-square-error, SSIM: structural similarity loss, PSNR: peak-signal-to-noise-ratio, L2: L2-loss.

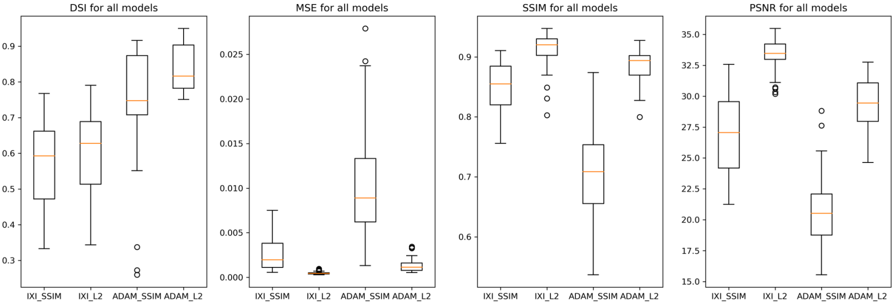


Figure 3.1: Box plots of reconstruction and segmentation metrics for all scans in the test set for the datasets from a) the IXI dataset and b) the ADAM challenge, trained on the respective training dataset. The centre bars correspond to the median value

MSE, mean SSIM and PSNR. However, the MSE, mean SSIM and PSNR of the L2 trained model for ADAM were higher than the SSIM trained IXI model. All DSI scores larger than 0.5 for all reconstructed images as seen in Table 3.1. However, all the DSI scores were lower for the IXI dataset.

Optimization of the network using SSIM loss resulted in reconstructed images with an improved visual perceptual image quality, with more structural details. These structural details were smoother on reconstruction based on the L2 loss VAE, as seen in Figure 3.2.

For (c) the third test set containing aneurysms, reconstruction metrics were poorer for all trained networks as seen in Table 3.2 and Figure 3.3. The IXI trained method had a worse DSI score and lower MSE, SSIM and PSNR than the ADAM trained method for evaluation on (c). The IXI trained VAEs had a larger DSI for (c) the aneurysm dataset compared to (a) the IXI test set.

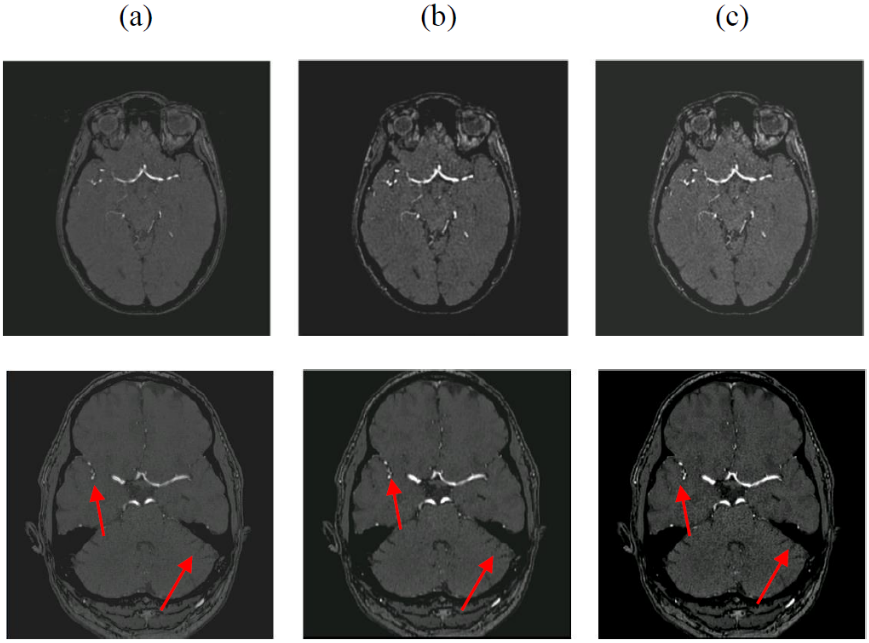


Figure 3.2: Reconstructions of TOF-MRAs using fully convolutional VAE trained with different loss functions a) Original TOF-MRA b) Reconstructed MRA: VAE trained with L2 loss c) Reconstructed MRA: VAE trained with SSIM loss. Top Row: IXI healthy dataset, Bottom Row: ADAM healthy dataset Red arrows indicate areas of more clear structure in both vessel and brain tissue in SSIM reconstructed image compared to original and L2 reconstructed image.

3.4 Discussion and Conclusion

Our results show that using SSIM as a loss function in a VAE provides a better perceptual image quality than with a voxelwise L2 loss for the reconstruction of TOF-MRAs. However, L2 loss performs better for quantitative vessel segmentation and reconstruction metrics.

In the MRAs, there are multiple structures inside the brain which are emphasised when the SSIM loss is optimised (see Figure 3.2). This results in a lower contrast (as demonstrated by the lower PSNR relative to the L2 loss) between the vessels and surrounding brain tissue. This may result in the lower vessel segmentation performance. For vascular segmentation, these structures in the surrounding brain tissue are not of interest, but they can be for the diagnosis of other cerebral disease. For vessel segmentation from the reconstructed VAEs, selection of patches for training containing only vessel, merits further investigation as this narrows the problem. Furthermore, the vessel segmentation method was trained on original TOF-MRAs, potentially leading to a lower performance for reconstructed TOF-MRA with different image qualities

c) Aneurysm Test Set

Network	DSI	MSE	Mean SSIM	PSNR
IXI SSIM	0.602 (0.217)	0.018 (0.008)	0.652 (0.081)	18.0 (2.72)
IXI L2	0.782 (0.133)	0.003 (0.001)	0.845 (0.040)	26.3 (2.31)
ADAM SSIM	0.692 (0.185)	0.012 (0.006)	0.696 (0.074)	20.1 (2.90)
ADAM L2	0.790 (0.014)	0.001 (0.001)	0.880 (0.028)	28.7 (1.95)

Table 3.2: Reconstruction and segmentation metrics for the prediction of the third test set containing aneurysms, for each of the networks trained on a) the IXI dataset and b) the ADAM challenge dataset. Values are provided as mean (standard deviation).

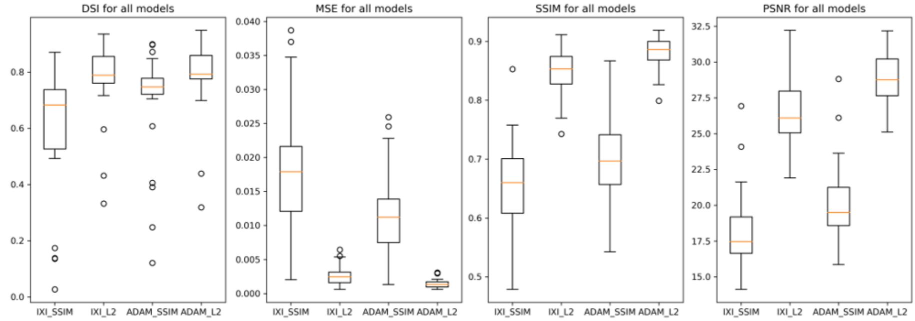


Figure 3.3: Box plots of reconstruction and segmentation metrics for all scans in the third test set containing aneurysms, for each of the networks trained on a) the IXI dataset and b) the ADAM challenge. The centre bars correspond to the median value.

compared to the original MRAs.

The higher reconstruction metrics for the IXI dataset may be caused by the larger quantity and diversity in vascular confirmation and aneurysms in the training data. The lower DSI scores of the IXI set are likely due to the fact that the vessel segmentation network was trained using data from the ADAM challenge. This results in the original vessel segmentation being sub-optimal and consequently the reconstructed vessel segmentation might be even poorer. For more valid assessment of vessel segmentation performance, an alternative vessel segmentation method could be used which performs comparably for both datasets, or the current vessel segmentation network could be re-trained on the IXI dataset. A further limitation of our study, was that validation was performed on a single random split of the dataset and in future studies different validation splits or k-fold validation should be used to ensure a fair distribution of the data.

The networks trained on (a) the IXI dataset performed well on reconstruction of (c) the aneurysm dataset and were not substantially worse than the (b) healthy ADAM

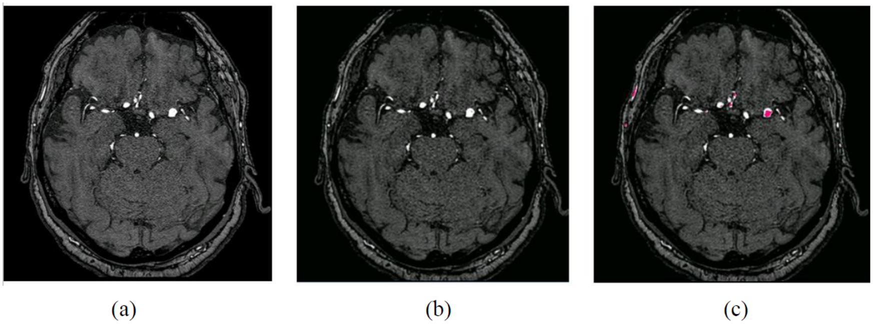


Figure 3.4: a) Original TOF-MRA with aneurysm b) Reconstructed using ADAM SSIM trained network c) Possible anomaly detection method using SSIM values and thresholds. Anomalies in SSIM are shown overlaid in dark pink.

trained networks. The aneurysm dataset was made up of MRAs that had the same protocols and variety in field strength used in the healthy ADAM training set. This suggests that the IXI trained dataset has good performance, even on data with a different protocol from which it was trained, and could potentially work on a variety of different datasets.

Notably, the mean SSIM was lower for all trained networks on the aneurysm test set relative to the healthy test sets. This suggests that an anomaly (such as an aneurysm) may result in a reconstructed image that has a structure more similar to that of a healthy subject, less similar to the original image (lower SSIM). This can be seen in Figure 3.4, where the aneurysm in the reconstructed image has a lower SSIM value. This suggests that VAEs trained using either an L2 or SSIM loss may be useful for anomaly detection when evaluated against the original images using SSIM. Further investigation into this would need to be performed.

In conclusion, our study demonstrates that using SSIM loss to train a VAE does improve the perceptive visual quality of the reconstructed MRA over L2 loss. However, it should be used with caution as it does not necessarily improve the quantitative voxelwise representation of specific features which may be required for future analysis. L2-loss trained VAEs may be used for accurate reconstruction of TOF-MRAs. Furthermore, we suggest that SSIM may be a potential metric for anomaly detection in TOF-MRAs.

3.5 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.

Chapter 4

Geometric Deep Learning using Vascular Surface Meshes for Modality-Independent Unruptured Intracranial Aneurysm Detection

Published as: K. M. Timmins, I. C. van der Schaaf, I. N. Vos, Y. M. Ruigrok, B. K. Velthuis, and H. J. Kuijf. “Geometric Deep Learning using Vascular Surface Meshes for Modality-Independent Unruptured Intracranial Aneurysm Detection,” *Under Review: IEEE Transitions in Medical Imaging* (2022).

Abstract

Early detection of unruptured intracranial aneurysms (UIAs) enables better rupture risk and preventative treatment assessment. UIAs are usually diagnosed on Time-of-Flight Magnetic Resonance Angiographs (TOF-MRA) or contrast-enhanced Computed Tomography Angiographs (CTA). Various automatic voxel-based deep learning UIA detection methods have been developed, but these are limited to a single modality. We propose a modality-independent UIA detection method using a geometric deep learning model with high resolution surface meshes of brain vessels. A mesh convolutional neural network with ResU-Net style architecture was used. UIA detection performance was investigated with different input and pooling mesh resolutions, and including additional edge input features (shape index and curvedness). Both a higher resolution mesh (15,000 edges) and additional curvature edge features improved performance (average sensitivity: 65.6%, false positive count/image (FPC/image): 1.61). UIAs were detected in an independent TOF-MRA test set and a CTA test set with average sensitivity of 52.0% and 48.3% and average FPC/image of 1.04 and 1.05 respectively. We provide modality-independent UIA detection using a deep-learning vascular surface mesh model with comparable performance to state-of-the-art UIA detection methods.

4.1 Introduction

An intracranial aneurysm (IA) is a focal bulging of the vessel wall in the brain. They have a prevalence of approximately 3% in the general population [1]. If an IA ruptures it leads to subarachnoid haemorrhage which can be fatal or lead to long-term disability [39, 114]. It is important that unruptured IAs (UIAs) are detected early to allow for clinicians to make informed rupture and preventative treatment risk assessments [115]. A radiologist usually diagnoses UIAs by visually inspecting Time-Of-Flight Magnetic Resonance Angiographs (TOF-MRA) or contrast-enhanced Computed Tomography Angiographs (CTA) but this inspection can be time-consuming and unreliable. Visual inspection for UIA detection has been found to have a large variability in sensitivity across different studies, varying from as low as 28% for small UIAs to as high as 88% [8–10]. Various (semi-) automatic detection methods for UIAs exist, including those developed for TOF-MRAs as part of the Aneurysm Detection and segMentation (ADAM) Challenge [111]. All methods submitted to this challenge were voxel-based deep learning methods, including nnU-net [84] and nnDetection [116] with a varying range of sensitivity and false positive count (FPC) /image (top 10 methods: sensitivity = 76% - 59%, FPC/image = 0.18 - 9.37). Voxel-based methods are often limited by their sensitivity to modality and scan acquisition parameters. Geometric deep learning methods operating on vessel surface meshes generally would not have these limitations and could be used instead. This paper investigates the use of a mesh convolutional neural network for modality-independent UIA detection.

4.2 Background

4.2.1 Geometric Vessel Surface Models for UIA detection

Vessel surface models could be directly used for UIA detection, since the shape of the vessel surface where an UIA occurs is different from the surrounding vasculature; the UIA bulges out beyond the smooth tubular vessel structure. Until now most UIA detection methods use voxel-based methods which tend to be limited to one modality and often struggle to generalise to different scanning acquisitions. UIAs are often followed up with a different modality and follow-up can occur over a long period of time and/or across different centres with different scanning protocols. Using vessel surface meshes reduces the modality and/or scan protocol dependence of a method, allowing for a generalised detection method.

Geometric surface analysis of vessels in TOF-MRAs can aid in the detection of UIAs [117, 118]. Studies have found that the local shape index and curvedness [119] can be used as a visual aid to detect UIAs, especially UIAs that are smooth and round [117]. Prasetya et al. [118] found shape index to be more effective at identifying UIA location than the mean or Gaussian curvature alone. Shape index combines mean and Gaussian curvature to give a rotation, translation and scale invariant value that is indicative of the local shape. Curvedness provides information about the local scale of the object. The coordinate independence of these measures makes them ideal geometric features to consider in geometric UIA detection and in geometric deep learning.

4.2.2 Geometric Deep-learning Methods

Geometric deep learning using 3D point clouds and meshes has good performance for 3D object classification and segmentation. Both PointNet++ [18] and PointCNN [19] operate on point clouds. Meshes may have a benefit over point clouds, because they include connectivity information about the points on the 3D object surface. Geometric deep learning models using meshes include MeshNet [22] and MeshCNN [23]. Studies have shown that geometric deep learning performs well for UIA segmentation when considering a smaller region-of-interest around an already detected UIA with its parent vessels [120–122].

Bizjak et al. [52] provide an UIA detection method from point clouds representing intracranial vessels extracted from Digital Subtraction Angiograghy (DSA), CTA and TOF-MRA. A PointNet [123] model was trained using smaller parcellated point clouds from DSAs for predicting points as UIA or vessel. The resulting predictions were merged to generate the detection prediction for the full vasculature resulting in a 98.6% sensitivity with 0.2 FPC/image. The majority (92%) of UIAs were larger than 5 mm (median size: 9.22 mm).

Yang et al. [121] investigated multiple geometric deep learning methods to perform UIA segmentation, with optimal performance using SO-Net [124] (Intersection-Over-Union (IOU): 81.4%). All input meshes were a small region-of-interest containing part of the parent vessel and with an UIA (>3 mm). They used MeshCNN [23] with three different resolution input meshes (namely 750, 1,500 and 2,250 edges). The IOU was highest for 2,250 edges (72% (95% CI: 64-79%)) demonstrating that a larger number of input edges improves the UIA segmentation performance. They state that every UIA was segmented at least in part, suggesting MeshCNN could be a good UIA detection method. The same group have recently published a two-step pipeline for UIA segmentation from full brain vasculature [125]. The first step uses Point-Net++ [18] to classify vessel segments with and without UIAs. The second step uses SO-Net [124] to segment the UIA from the vessel (Dice Similarity Coefficient (DSC) : 72%).

Schneider et al. [122] proposed an application of MeshCNN to segment UIAs. They extended the original MeshCNN implementation to use sparse matrices, lowering memory costs and allowing higher resolution input meshes. Input meshes of 19,200 edges were used to segment four classes: inlet, vessel, bifurcation and UIA from a small region-of-interest with good performance (average IOU of 63.24%).

4.2.3 MeshCNN

MeshCNN [23] is a convolutional neural network (CNN) developed for triangular 3D meshes. For the sake of completeness, we provide some description of MeshCNN, but for full details we refer the reader to the original paper [23].

Mesh edge convolutions, pooling and unpooling are implemented in MeshCNN analogous to a CNN operating on voxels in an image. For each edge, five relative geometric edge features (the dihedral angle, two inner angles and two edge-length ratios for each face) are determined as input features for the model. These features are scale, translation and rotation invariant. Convolutions are symmetric operations performed on an edge and its four 1-ring neighbouring edges. Pooling layers consist of collapsing edges; which are prioritised based on the weighting of the edge features, collapsing edges that are less important to the task. The resulting number of edges after pooling is a tunable hyperparameter. The link between the old collapsed edges and the new edges is logged and used in the paired unpooling layer to up-sample the mesh to its original resolution. MeshCNN has been shown to work well for semantic segmentation in large datasets such as the human body segmentation dataset [126] and co-segmentation dataset (COSEG) [127]. In the medical domain, MeshCNN has only been used for a few classification and segmentation problems, including age prediction based on the neonatal white matter cortical surface [128] and UIA segmentation from a parent vessel [122].

The original MeshCNN implementation [23] is limited by memory intensive book-keeping of edge collapses in the pooling layers which reduces the possible resolution of input meshes with large number of edges and parameters. In the original paper, input meshes with 750 edges were used for classification, and 2,250 edges for segmentation. This is too limited for meshes generated from full 3D medical images. For example, taking a full unsimplified vessel surface mesh from a 3D TOF-MRA results in as many as 25,000 edges. Although these edges can be reduced, it is important to have a high resolution to include the full detail and topology of the surface. The original implementation was also limited by edge collapses, which often failed if the collapse would result in a non-manifold mesh.

Modifications can be made to make MeshCNN more suitable for high resolution meshes from medical images. MedMeshCNN [122] introduced sparse matrices for book-keeping of the edge collapses, improving the memory capacity by a factor of 8.5. In another adaption for CAD model surface segmentation [129], sparse matrices were again included in the pooling layers and further changes were made to improve efficiency, such as deleting no longer used tensors and rewriting functions to perform in place. Additionally, edge collapses which result in non-manifold meshes were skipped.

4.2.4 Aim

Mesh neural networks have been used successfully in a few previous medical applications for classification [128] and segmentation [120, 125], and we demonstrated that modality-independent UIA detection using mesh convolutional neural networks was possible with a small dataset [130]. As mesh convolutional neural networks continue to be developed, there is little information on optimal hyperparameter and configuration of mesh convolutional neural networks such as the resolution of input meshes, pooling layers and additional input features. In this paper, we further explore the optimal configuration and hyperparameters of a mesh convolutional neural network for modality-independent UIA detection in a large, heterogeneous dataset.

We performed modality-independent UIA detection using a modified mesh convolutional neural network with full intracranial vessel surface meshes. First, we explored the impact of input mesh resolution on the model performance. Second, the impact of different edge resolutions in the pooling layers by using different pooling schemes was investigated. Third, since shape index and curvedness are known to aid visual UIA detection, we added these in our model. Fourth, the generalisability and modality independence of our model for UIA detection was evaluated against a public dataset of TOF-MRAs (derived from the ADAM challenge) [111] and on an in-house dataset of CTAs.

4.3 Materials and Methods

4.3.1 Dataset

The training data consisted of 93 brain TOF-MRAs with diagnosed UIAs released as part of the ADAM challenge [111]. The dataset included manual annotations for UIAs present in the scans. The UIAs ranged in size, with a median diameter of 3.9 mm and a range from 1.0 – 15.9 mm.

The test TOF-MRA data was the separate hold-out test dataset used for the ADAM Challenge, which consisted of 142 TOF-MRAs, of which 117 contained UIAs and 25 did not. As organisers of the challenge, we had direct access to the test dataset, but this was only used for the final inference. The test data was not used for fine-tuning or in the development of our method, to ensure fair bench-marking against other challenge method submissions.

The test CTA data consisted of 20 CTAs with corresponding non-contrast CTs, which all contained at least one untreated, UIA. The median UIA diameter was 5.1 mm and a range of 2.3 - 16.2 mm. All CT scans were made at the UMC Utrecht and UIAs were labelled by the same radiologist and using the same procedure as the ADAM data [111].

4.3.2 Labelled Mesh Generation

VESSEL SEGMENTATION TOF-MRAs were corrected for bias field inhomogeneities using N4 [62]. The TOF-MRAs and corresponding manual labels were re-sampled to have median voxel size of the dataset (0.357 mm x 0.357 mm x 0.500 mm). Vessel segmentation was performed automatically using an existing 3D U-net [131]. Automatic connected component analysis was used to leave only the main vessels (>= 1000 connected voxels). We removed six TOF-MRAs from the training set at this point because of poor quality vessel segmentation.

All CTAs were re-sampled and the corresponding non-contrast CT was registered to the CTA using elastix [61] with B-Spline registration. A bone mask was generated by thresholding the registered non-contrast CT, which was dilated and then subtracted from the CTA to remove the bone structures. The main vessels were then semi-automatically segmented from the CTA using thresholding and connected component analysis.

VESSEL SURFACE MESH LABELLING Labelled vessel surface meshes were generated entirely automatically, based on the vessel segmentations and the binary image UIA annotations. Using a Neighboring Cells algorithm, developed in MeVisLab (MeVis Solution A.G, Fraunhofer MEVIS) [132], each voxel in the vessel segmentation was scanned to generate a triangular surface mesh representation. For standardisation of the dataset, all meshes were downsampled to 15,000 edges, using a quadratic error metric to decide which edges to collapse first. UIAs on the vessel surface were identified and the corresponding vertices and edges were labelled. Morphological mesh closing was performed to fill any holes in the labelling and to ensure that all meshes were closed and manifold. This pre-processing was performed automatically for all TOF-MRAs and CTAs in both the training and test datasets. Figure 4.1 shows an example of a labelled vessel mesh.

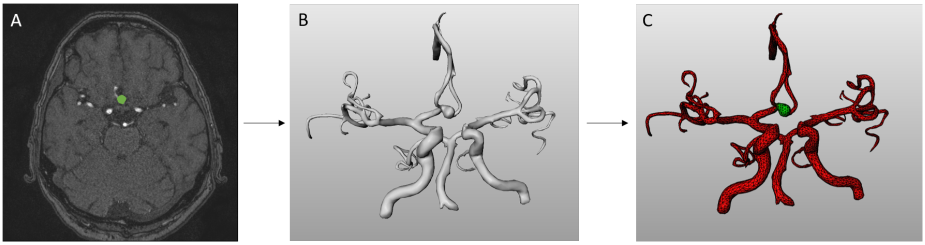


Figure 4.1: Vascular surface mesh generation. A: TOF-MRA with labelled UIA, B: Vessel segmentation from TOF-MRA using pre-trained U-net. C: Labelled vessel surface mesh used to train model.

4.3.3 Model

MODEL ARCHITECTURE For our method, we use a ResU-net style architecture [73, 133] with MeshCNN [23] convolution and pooling layers. Sparse matrices were implemented in the pooling layers, to make the network more memory efficient and enable higher resolution mesh inputs than the original MeshCNN implementation. Furthermore, the pooling layers were modified to not collapse edges that resulted in non-manifold meshes [129]. The model was further modified to allow additional input features for each edge to the original five geometric features. Specifically, two extra local curvature features were added for each edge: shape index and curvedness [119], as these are known to aid in UIA detection [117, 118]. Shape index and curvedness were determined for every vertex on the vessel mesh surface. An edge was then given a curvature feature (shape index or curvedness), as being the average of the values at the corresponding end vertices of the edge. The final model included seven features for each edge: the dihedral angle, two inner angles, two edge-length ratios for each face, shape index and curvedness. A softmax layer was added at the end of the model, to allow saving soft segmentation results for each mesh, which indicate the probability of each edge in the mesh being an UIA.

MODEL TRAINING The remaining 87 TOF-MRAs in the ADAM training set, after the six exclusions, were randomly split into sets for three-fold cross validation for training (58 TOF-MRAs) and validation (29 TOF-MRAs). Paired baseline and follow-up scans were considered when making the training and validation splits. The loss function used was weighted cross entropy with weighting 0.9 to the aneurysm class, 0.1 to the vessel class. A learning rate of 0.001 was used with a learning rate scheduler, AdamW was used as optimizer and the network was trained for 350 epochs. The training was implemented in Python 3.8.5 with Pytorch version 1.8.0 on either a NVIDIA GEOFORCE 2080 Ti GPU (11GB) or NVIDIA TITAN X Pascal (12GB) GPU with CUDA version 11.2.

4.3.4 Experiments

In three different experiments we assessed the influence of: (1) the input mesh resolution, (2) pooling layer resolution of the network and (3) including shape index and curvedness as additional input edge features, on the performance of the model for UIA detection. These experiments were evaluated using three-fold cross validation. The best performing network was trained on the full training set and tested on the TOF-MRA and CTA test sets.

INPUT MESH EDGE RESOLUTION The use of different input mesh resolutions for training the model was investigated. In the original MeshCNN paper, input meshes with 2,250 edges were used for segmentation. With the increased capacity of our model, higher resolution meshes could be included. 15,000, 10,000, 7,500 and 5,000 were chosen as input edge resolutions. 15,000 edges was chosen as the maximum, because this was the full resolution of the smallest mesh in the training data. This resolution allowed for high detail of the surface of the vessel meshes whilst taking into account memory constraints. 5,000 was chosen as the minimum number of edges, because any lower number resulted in severe loss of detail so that any smaller UIAs were no longer clearly defined or that vessel bifurcations were no longer distinct. For this experiment, a fixed pooling scheme was kept with ratios as similar as possible to the original MeshCNN. Only the original five input geometric features per edge were used as input.

POOLING SCHEME Different pooling schemes were chosen to investigate the effect of increasing or reducing the minimum number of edges the meshes are pooled to (see Table 4.2). The number of input edges was chosen based on Experiment 1 and all other hyperparameters were kept the same. As before, only the original five input geometric features per edge were used as input.

INPUT FEATURES The addition of shape index and/or curvedness features per edge was investigated. The baseline model includes the five geometric features from the original MeshCNN. The shape index and curvedness were added as input, first individually and then combined. The number of input edges and pooling scheme were based on Experiments 1 and 2 and all other hyperparameters were kept the same.

TOF-MRA AND CTA TEST SET The best performing method from Experiments 1-3 was trained using the full TOF-MRA training set. The trained network was used for inference on the TOF-MRA and CTA test sets.

4.3.5 Evaluation

The output of the methods was a soft-labelled mesh and a threshold of 0.9 was used to select edges that were predicted to be part of an UIA. This threshold was chosen to balance optimal sensitivity and false positive count, based on empirical results of the development phase on the validation set. The threshold was kept consistent in all experiments. Each connected component of labelled vertices was considered a detection. Connected components that overlapped with treated aneurysms were excluded, similar to the ADAM challenge [111]. A true positive was considered a connected component that overlapped with the true UIA. The bottom ten slices of the images were removed, because false positives often occurred here, owing to the rounded end of the vessels. Moreover these slices do not depict the Circle-of-Willis with its bifurcations where UIAs usually occur. Sensitivity and FPC/image were determined. All experiments were evaluated using three-fold cross validation and the metrics were averaged across all validation splits. For each experiment a similar ranking to the ADAM challenge[111] was used to determine the optimum method, using an average ranking of sensitivity and FPC/image for each experiment. These metrics were chosen as it is important that all UIAs are detected, indicated by a high sensitivity, even if this results in some false positives. However, too high a false positive count would also be disadvantageous to the clinician. FPC/image can be determined even if there are no true UIAs in the image, allowing evaluation of the full dataset.

4.3.6 Analyses

A comparison was made between the sensitivity and FPC/image on the TOF-MRA and the CTA test set using a Mann Whitney U-test with a statistical significance level of p <0.05. Results of the best performing method evaluated on the full TOF-MRA test set were stratified for UIA size and location.

For the TOF-MRA test set, the predicted UIAs were converted back to binary images and detection metrics were determined based on the resulting binary segmentations for direct comparison to the official ADAM challenge ranking.

4.4 Results

4.4.1 Experiments

INPUT MESH EDGE RESOLUTION Table 4.1 shows the performance of Experiment 1 with the top ranking method having an input resolution of 15,000 edges, suggesting that higher edge resolution has better performance. This edge resolution was then used for Experiment 2.

Input Edges	Sensitivity	FPC/Image	Ranking
5,000	50.8%	1.99	0.50
7,500	60.5%	3.31	0.42
10,000	65.2%	3.56	0.50
15,000	61.8%	2.87	0.60

Table 4.1: Experiment 1: Influence of Input Edges. Detection metrics of model with different input edge resolution meshes using three-fold cross-validation and averaged over all validation splits, FPC: False Positive Count

POOLING SCHEME Table 4.2 shows the detection metrics on the validation splits for each of the pooling scheme experiments. Figure 4.2 shows the meshes at the inner pooling layer and at the final output for pooling schemes: 15,000 12,000 6,000 2,000 (A & B) and 15,000 12,000 10,000 8,000 (C & D). The optimal pooling scheme was found to be 15,000 12,000 6,000 2,000.

Input Edges	Sensitivity	FPC/Image	Ranking
15,000 12,000 6,000 2,000	58.8%	1.45	0.65
15,000 12,000 8,000 4,000	61.8%	2.87	0.50
15,000 12,000 9,000 6,000	55.7%	1.10	0.50
15,000 12,000 10,000 8,000	56.5%	1.57	0.43

Table 4.2: Experiment 2: Influence of Pooling Scheme. Detection metrics of model with different pooling schemes with 15,000 input edges using 3-fold cross-validation and averaged over all validation splits. FPC: False Positive Count

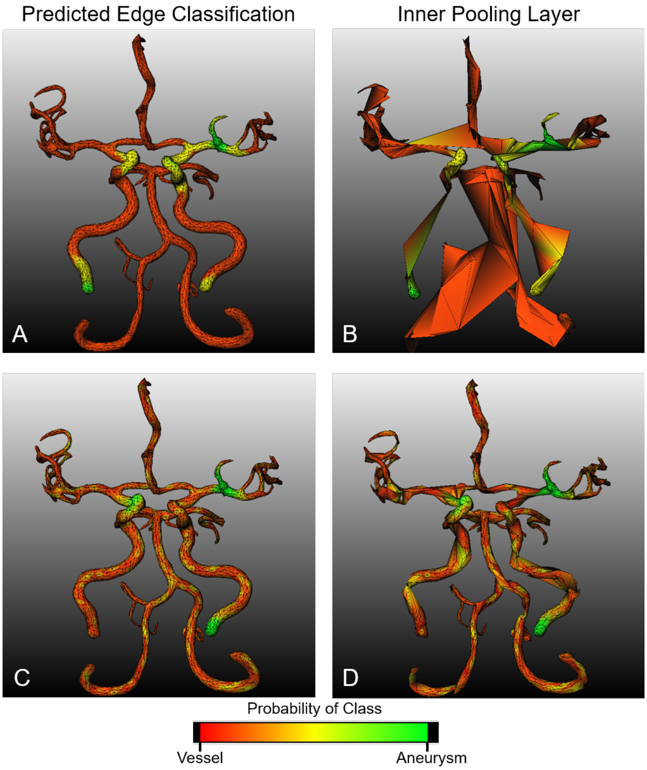


Figure 4.2: Experiment 2: An example of pooling layer resolution and the effect on network output results demonstrated in Experiment 2. Figures A and B: Pooling scheme 15,000 12,000 6,000 2,000. Figures C and D: Pooling scheme 15,000 12,000 10,000 8,000. The first column (Figs A and C) show the output of the trained model after inference on an unlabelled vessel mesh with 15,000 edges. The second column (Figs B and D) is the output of the final predicted mesh, in the centre of the network at its most pooled stage (B: 2,000 edges, D: 8,000 edges). At this point, the most important edges to identify UIAs should be identifiable and not be collapsed and all unimportant edges should be collapsed.

INPUT FEATURES The optimal network configuration (input features: 15,000, pooling layers: 15,000 12,000 6,000 2,000) was trained to include the extra curvature features and the results of inference can be seen in Table 4.3. Figure 4.3 shows the output predictions for the experiments with the additional curvature features per edge. Adding both shape index and curvedness had optimal performance. As seen in Table 4.3, adding shape index increased the sensitivity of the method but also increased the false positive count. By adding curvedness as well, the increased sensitivity was retained but the number of false positives was reduced. Figure 4.4 A.i) and B.i), demonstrate that shape index is higher in certain regions such as in the curved ICAs (shown by the white arrows) which leads to false positives. In Figure 4.4.A.ii) and 4.B.ii), these regions have a lower curvedness and therefore are not considered as false positives. Thus, using both the shape index and curvedness leads to a more optimal performance with higher sensitivity and reduced false positive count.

Input Edges	Sensitivity	FPC/Image	Ranking
Geometric	58.8%	1.45	0.66
Geometric and SI	67.2%	2.97	0.50
Geometric and CV	55.0%	1.71	0.41
Geometric, SI and CV	65.6%	1.61	0.88

Table 4.3: Experiment 3: Influence of Input Features. All models used input number of edges as 15,000 and pooling layer resolution: 15,000 12,000 6,000 2,000. All values averaged across all validation sets, Geometric: Original geometric features in MeshCNN [23], SI: Shape Index, CV: Curvedness, FPC: False Positive Count

TOF-MRA AND CTA TEST SET the model performance on the TOF-MRA test set and the CTA test set (Table 4.4) was not statistically significantly different when assessing both sensitivity and FPC/image with a Mann Whitney U-test with both p >0.05 (p = 0.20, p = 0.50 respectively).

Input Edges	Sensitivity	FPC/Image
TOF-MRA test set	52.0%	1.04
CTA test set	48.3%	1.05

Table 4.4: Detection Performance using Optimal Network. Model with input number of edges as 15,000, pooling layer resolution: 12,000 6,000 2,000 and including shape index and curvedness as extra features. Trained on full TOF-MRA training data. Sensitivity: average of all scans containing UIA. FPC/image: average false positive count per image over all scans including those with and without UIAs.

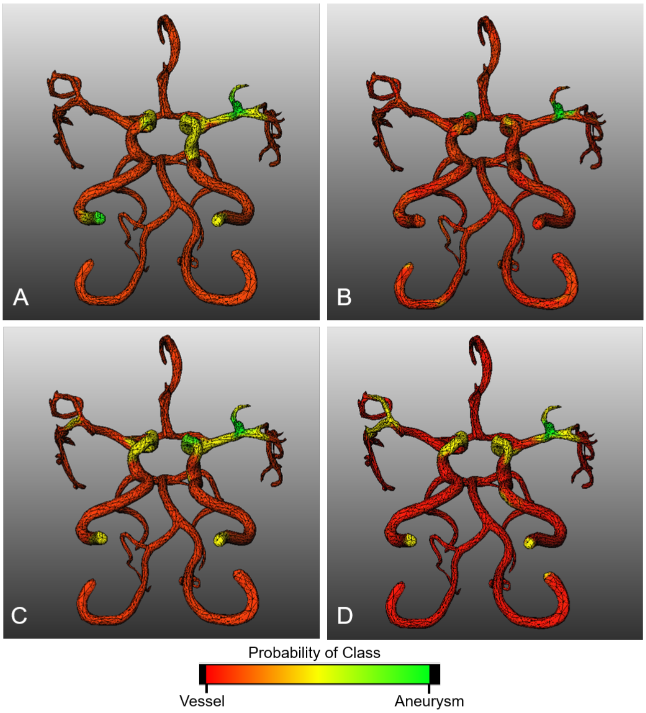


Figure 4.3: Experiment 3: All images show the predicted model output for one example with the model trained using different input features. A: geometric features only, B: geometric features and shape index, C: geometric features and curvedness, D: geometric features, shape index and curvedness

4.4.2 Analyses

Figure 4.5 indicates how the model performs on the TOF-MRA test set with regard to size. UIAs were split into size categories based on diameter; large: ≥ 7 mm (n = 12), medium: > 3mm and < 7mm (n = 78), and small: maximum diameter ≤ 3 mm (n = 58). The average sensitivity to detect UIAs for large, medium and small UIAs were 83.3%, 73.1% and 32.7% respectively, with number of true positives (TP) of 10, 57 and 19. The number of false negatives (FN) (missed aneurysms) for large, medium and small UIAs was 2, 21 and 39. Figure 4.6 shows the detection results stratified based on location of the UIA.

Based on the binary images generated from the meshes in the TOF-MRA test set, the average sensitivity was 59.3% and the FPC/image was 1.08. Our method would rank 8th in the ADAM challenge [111] (assessed in December 2021).

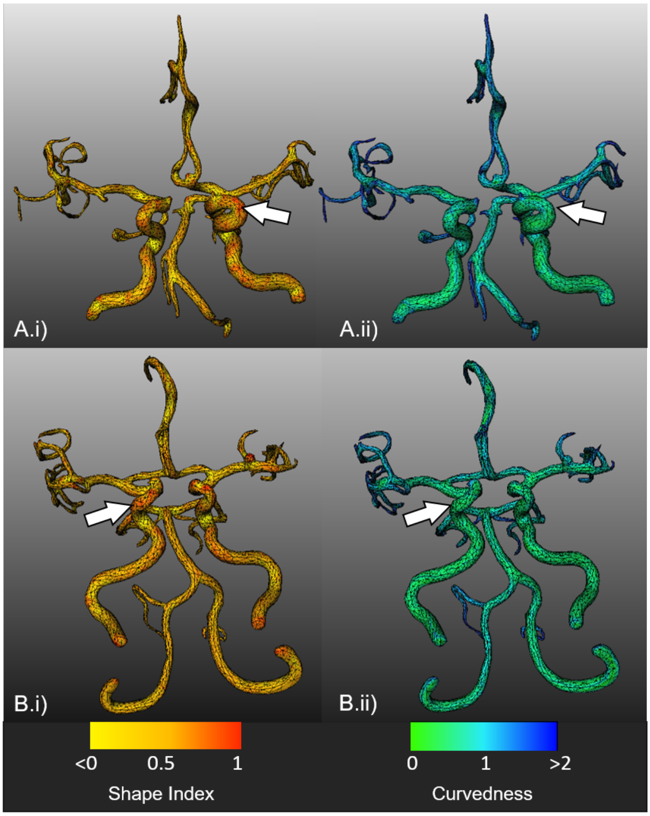


Figure 4.4: Experiment 3: Two vessel meshes used for inference. Left hand column colours indicate the shape index of the surface, the right hand column is coloured based on curvedness. The white arrows indicate areas of high shape index, but low curvedness. Using both measures ensures these areas are not considered false positives.

4.5 Discussion

We demonstrate modality-independent UIA detection with a mesh convolutional neural network using high resolution intracranial vessel surface meshes. The optimal model set-up was found to include input meshes of 15,000 edge resolution, and pooling layers 15,000 12,000, 8,000 and 2,000. Additional input features of shape index and curvedness improved the detection performance of the model. The modality independence of the model was validated on both TOF-MRAs and CTAs with comparable performance. Our model based on vascular surface meshes had UIA detection performance comparable to voxel-wise methods (top ten ranking method in ADAM challenge).

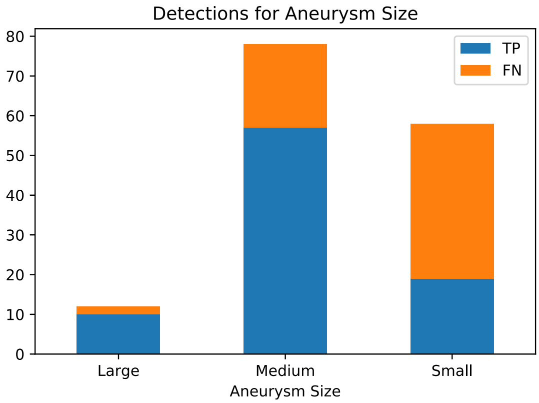


Figure 4.5: The number of true positives (TP) and false negatives (FN) for all UIAs in the ADAM test set evaluated with the best performing method, stratified into groups based on UIA size.

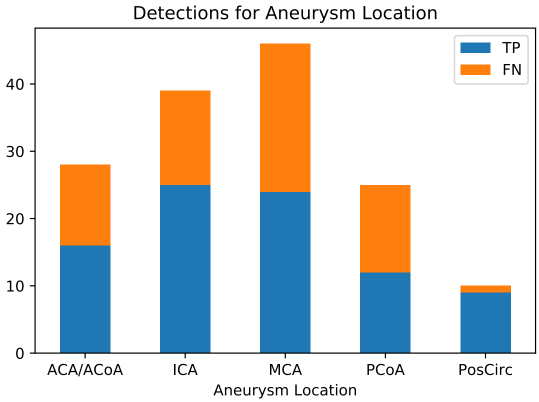


Figure 4.6: The number of true positives (TP) and false negatives (FN) for all UIAs in the ADAM test set evaluated with the best performing method, stratified into UIA location. ACA/ACoA: anterior cerebral or communicating artery, ICA: internal carotid artery, MCA: middle cerebral artery, PCoA: posterior communicating artery, and Pos Circ: posterior circulation.

INPUT MESH EDGE RESOLUTION A larger input mesh resolution (15,000 edges) was determined to have the best UIA detection performance. Higher edge resolution meshes provided detailed topology of the vessel surfaces which became concentrated at the UIA borders during edge collapsing and pooling. The smaller edge resolution input meshes have fewer possible edges to collapse, sometimes resulting in removing edges defining UIAs. Our implementation extends the original MeshCNN by allowing these better performing, high resolution meshes as input. It is worth noting that because of GPU memory constraints, larger input resolution requires smaller batch sizes and hence takes longer to train. Therefore, a balance should be reached based on performance and efficiency requirements for each independent application.

POOLING SCHEME We found the model performance to be sensitive to the pooling layer scheme with the lowest inner pooling resolution (15,000 12,000 6,000 2,000) performing optimally. Figure 4.2 B shows how the larger number of edge collapses results in the vessels having less detail and so more confidently being classified. Furthermore, the concentration of edges around the UIA provides more localised UIA detections. Figure 4.2 D demonstrates that there is still high resolution and detail in the pooling layers. With a large number of edges in the inner pooling layer there is more possibility of false positives. Our experiments highlight that the pooling layer resolutions influence the performance of the method. The correct scheme depends on the shape and structure of the 3D object, so this is a hyperparameter which should be chosen for each specific application.

INPUT FEATURES The addition of shape index as an edge input feature increased the sensitivity of the method by almost 10%. This could be expected because it is known that shape index can aid in the sensitivity of visual UIA detection [117, 118]. However, the addition of shape index also increased the number of false positives, shown by higher regions of shape index in Figure 4.4.A.i) and B.i). Therefore, shape index should be considered for a preferred high sensitivity detection method (i.e. ensuring all possible UIAs are detected) over a precise method. Figure 4.3 shows that the resulting UIA detections appear to be smaller and more localised, with a higher confidence. Curvedness by itself as an additional feature reduced the sensitivity and increased the FPC/image compared to baseline (geometric features only). Visually (Figure 4.3), the addition of curvedness appears to highlight more candidate UIA areas, with low probabilities (yellow areas). A combination of curvedness and shape index increases the sensitivity (increase of 7%) with only a marginal increase in false positives relative to geometric features alone (increase of 0.24/scan). The curvedness appears to correct for the false positives detected as a result of adding the shape index, without reducing the sensitivity of the method. Based on our experiments, we believe it could be useful to include both of these features for other classification/regression problems using MeshCNN.

TOF-MRA AND CTA TEST SET Our method performed comparably for both the TOF-MRA and CTA test sets, demonstrating the modality independence of the model. The sensitivity was slightly higher and FPC/image was slightly lower for the TOF-MRA test set relative to the CTA test set. This may be attributed to the to the better and smoother TOF-MRA vessel segmentations on which the model was trained. The CTA vessel segmentation was not as smooth and was more difficult, due to the extra step to separate the vessel from the skull base. For optimum inter-modality performance vessel segmentations of similar qualities for all modalities should be used. The median UIA diameter in the CTAs was higher than in the TOF-MRAs. This reflects clinical practice, because some UIAs on our TOF-MRAs were found incidentally on scans during familial screening, whereas the CTAs were made specifically for UIA diagnosis or monitoring.

ANALYSES Compared to visual UIA detection from TOF-MRAs by radiologists, the sensitivity of our method is lower than quoted in literature: (87%) [10], however our dataset consisted of a large proportion of small UIAs (39%). The study of Bizjak et al. [52] found a sensitivity of 98.6% and FP count of 0.2. However, the median size of UIA was 9.22 mm, which is much larger than our median size of 3.4 mm. As seen in Figure 4.5, our method performed best for larger UIAs with more true positives than false negatives (missed UIAs) and a sensitivity comparable to literature (83.3%). Medium and large UIAs have a higher clinical relevance since preventive treatment in those UIAs might be considered as they have a higher growth and rupture risk. For small UIAs, the sensitivity of our method (32.7%) was low but found to be comparable to that in literature, where the sensitivity for radiologists was determined to drop as low as 38% [10] for UIAs <3 mm.

The UIA location influenced the performance of the model as seen in Figure 4.6, with the UIAs in the posterior circulation having the least number of false negatives. False positives were frequently found at the branch of the ophthalmic artery on the ICA, which often looked similar to an UIA, and the highly curved carotid siphon (see Figure 4.3). Improvements on the vessel segmentation to include smaller vessels, would reduce false positives and allow detection of more UIAs. This is also true with regard to the CTAs, where the vessel segmentation was not as smooth as the TOF-MRAs, resulting in deformities in the vessels which looked similar to UIAs.

Compared to voxel-wise methods submitted for the ADAM challenge [111], our method ranked 8th on the test data. Our model based on vascular surface meshes removes any image protocol specific dependencies, making this approach ideal for the diverse, heterogeneous dataset. The advantage of our method over the top 7 methods in the ADAM challenge is that it can be used on any modality for UIA detection and not just TOF-MRAs. It is possible that a common framework for CTAs and TOF-MRAs could be used by training separate models for each modality. However, the proposed method was trained using TOF-MRAs only. Separate intensity-based models would require new training data for each modality. Our proposed method also had a lower FPC/image than two of these top methods, which may be due to our method including only the vessels, eliminating any possibility of false positives in the rest of the image. The ADAM test dataset includes a large portion of small UIAs (39%) which also affects the average sensitivity of our method as it is lower for smaller UIAs.

LIMITATIONS AND FUTURE WORK All UIA annotations were made directly on the TOF-MRAs by an experienced radiologist, enabling the use of the full image information and context. Labels for training were made by projecting the image binary segmentations onto the surface meshes. This performed well for a detection method, however, for accurate 3D segmentation and UIA neck definition, the manual labels should be performed directly on the vessel mesh.

The current implementation is relatively memory inefficient due to book-keeping of unpooling layers, even using sparse tensors. Although distributed training could be used, the model still had long training times due to the edge collapsing being performed on CPU. Future implementations should consider the bottlenecks due to the pooling/unpooling layers and how this could be more efficiently performed. As geometric deep-learning libraries such as Pytorch Geometric [134] and Pytorch3D [135] continue to be developed, more efficient mesh convolutional and pooling layers could be implemented which would speed-up and expand the usefulness of the MeshCNN style framework. However, we note that after training, the current implementation has an inference time of approximately 90 seconds per unlabelled vessel mesh of 15,000 edges when evaluated on a Nvidia TITAN X 12GB GPU.

Previous geometric models for UIA segmentation and/or detection have used patch based and region-of-interest approaches [122, 125]. The current implementation makes a large training set of patches unfeasible with long training time and memory constraints. Our implementation has the benefit that it uses the full high resolution brain Circle-of-Willis vessel segmentation for training and inference with no requirement for extra pre/post-processing of patches. It could be investigated if a combination of our detection method followed by a patch-based segmentation approach could be used for UIA segmentation.

Our modality-independent UIA detection method gives promising results about the use of mesh neural networks in the field of medical image analysis. We believe such an implementation of mesh convolutional neural networks with high resolution meshes could also be useful in other vascular imaging problems such as for abdominal aortic aneurysm or coronary artery segmentation or for 3D lesion classification and regression problems, where the surface topology of a lesion is important for the outcome. Based on our results, we would recommend that future applications of MeshCNN should consider a high resolution input mesh and include both shape index and curvedness as extra input features. The pooling resolution scheme should also be investigated specific to the desired task.

4.6 Conclusion

We demonstrate that a mesh convolution neural network using high resolution brain vessel meshes and additional curvature features, can be used for a modality-independent intracranial aneurysm detection. Our method was validated on TOF-

MRAs in the ADAM challenge test set and a test set of CTAs with comparable performance for both modalities.

Chapter 5

Reliability and Agreement of 2D and 3D Measurements on MRAs for Growth Assessment of Unruptured Intracranial Aneurysms

Published as: K. Timmins, H. Kuijf, M. Vergouwen, M. Otten, Y. Ruigrok, B. Velthuis, and I. van der Schaaf. "Reliability and Agreement of 2D and 3D Measurements on MRAs for Growth Assessment of Unruptured Intracranial Aneurysms," *American Journal of Neuroradiology*, vol. 42 (2021), pp. 1598–1603.

Abstract

Background and Purpose: Reliable and reproducible measurement of unruptured intracranial aneurysm growth is important for unruptured intracranial aneurysm rupture risk assessment. This study aimed to compare the reliability and reproducibility of 2D and 3D growth measurements of unruptured intracranial aneurysms.

Materials and Methods: 2D height, width, and neck and 3D volume measurements of unruptured intracranial aneurysms on baseline and follow-up TOF-MRAs were performed by two observers. The reliability of individual 2D and 3D measurements and of change (growth) between paired scans was assessed (intraclass correlation coefficient) and stratified for aneurysm location. The smallest detectable change on 2D and 3D was determined. Proportions of growing aneurysms were compared, and Bland-Altman plots were created.

Results: Seventy-two patients with 84 unruptured intracranial aneurysms were included. The interobserver reliability was good-to- excellent for individual measurements (intraclass correlation coefficient >0.70), poor for 2D change (intraclass correlation coefficient <0.5), and good for 3D change (intraclass correlation coefficient = 0.76). For both 2D and 3D, the reliability was location-dependent and worse for irregularly shaped aneurysms. The smallest detectable changes for 2D height, width, and neck and 3D volume measurements were 1.5, 2.0, and 1.9 mm and 0.06 mL, respectively. The proportion of growing unruptured intracranial aneurysms decreased from 10% to 2%, depending on the definition of growth (1 mm or the smallest detectable changes for 2D and 3D).

Conclusions: The interobserver reliability of the size measurements of individual 2D and 3D unruptured intracranial aneurysms was good-to-excellent but lower for 2D and 3D growth measurements. For growth assessment, 3D measurements are more reliable than 2D measurements. The smallest detectable change for 2D measurements was larger than 1 mm, the current clinical definition of unruptured intracranial aneurysm growth.

5.1 Introduction

In the adult population, the prevalence of unruptured intracranial aneurysms (UIAs) is around 3% [1]. Intracranial aneurysm rupture leads to SAH with a high case fatality rate. The PHASES (Population, Hypertension, Age, Size, Earlier subarachnoid haemorrhage and Site) study found the 5-year rupture risk of UIAs to be, on average, 3.4% (0.5%–17.8%), depending on patient and aneurysm characteristics [6]. When one makes a treatment decision, the risk of aneurysm rupture is weighed against the complication risk of treatment. Aneurysm size is a key determinant in the prediction models of rupture risk [6, 136]. If a multidisciplinary team decides against preventive aneurysm treatment, the UIA is followed up with repeat TOF-MRA or CTAs to detect potential aneurysm growth. Growth is an additional rupture risk factor [137], and if detected, preventive treatment should be considered. TOF-MRA has been shown to systematically underestimate the size and volume of the aneurysm compared with the criterion standard DSA [138]. However, noninvasive TOF-MRA is the first-choice imaging method for follow-up imaging in clinical practice because neither contrast agent administration nor radiation exposure is required [139, 140].

Assessment of UIAs is performed by taking 2D size measurements of aneurysms on MRA/CTA using electronic calipers. The 3D nature of UIAs makes 2D measurements difficult and dependent on optimal orientation in multi-planar imaging. The 2D measurements by human observers are reported to have comparable mediocre reproducibility on both CTAs and MRAs [26, 27]. These UIA measurements are relevant when comparing aneurysm size in a follow-up scan to assess aneurysm growth. Aneurysm growth is defined as an increase in either 2D height or width of at least 1 mm [25]. A reliable measurement method with good agreement is important for risk assessment. In this context, the reliability depends on the variability of the aneurysm sizes among patients. The agreement describes the interobserver measurement error and is characteristic of the measurement method itself. Without knowledge of reliability and agreement, it is unclear whether a measured change in aneurysm size between baseline and follow-up scans represents real growth or is attributable to observer or scan variations.

In this study, we investigated the reliability and reproducibility of 2D size and 3D volume measurements of UIAs and change in aneurysm size and volume between baseline and follow-up MRAs. For an agreement measure, we calculated the smallest detectable change (SDC) and assessed agreement using Bland-Altman plots.

5.2 Materials and Methods

5.2.1 Study Population

We included 72 patients from a series of patients with UIAs from the University Medical Center Utrecht who met the following inclusion criteria: (1) A TOF-MRA was available at both the baseline admission scan and follow-up, (2) the follow-up scan was per-formed at least 6 months after the baseline scan, and (3) the patient had at least 1 untreated UIA present on both baseline and follow-up MRA. Any treated aneurysm in these subjects was excluded from this study. The most recent follow-up scan in which the UIA remained untreated and unruptured was used. The scans had an in-plane resolution range of 0.175–1.04 mm and a section thickness range of 0.399–1.2 mm. All scans were obtained from 2004 to 2019. Due to the nature of the scans, protocols varied, but all scans were obtained on 1T, 1.5T, or 3T scanners with a median TR of 23 ms and a median TE of 6.4 ms across all scans. This retrospective study required no formal consent from participants. The data that support the findings of this study are available from the corresponding author on reasonable request.

5.2.2 Measurements

2D MEASUREMENTS Manual 2D measurements of the UIAs were performed on the IntelliSpace Portal (Phillips Healthcare). Measurements were obtained using electronic calipers on the TOF-MRAs, which could be rotated in the software. The aneurysm height, width, and neck were measured on the TOF-MRAs on a 0.1 mm scale [31, 40] as shown in parts A and C in Figure 5.1. Aneurysm height was defined as the maximum distance from the aneurysm neck to the dome. Aneurysm width was measured perpendicular to the height along the maximum width of the UIA. The neck was measured as the maximum width of the UIA where it attached to the parent vessel. Observers determined whether the UIA shape was regular or irregular.

All 2D measurements were performed independently by 2 observers. The observers were a neuroradiologist (I.C.v.d.S., with 15 years of experience) and a general radiologist (M.J.O., with 10 years of experience, including cerebral MRA evaluation). Individual measurements were first obtained on the baseline scan, then on the follow-up scan of the same patient. The observers were not blinded to the time order of the scans and had the baseline for comparison, as is standard in clinical practice.

3D MEASUREMENTS For 3D measurement, the UIAs were segmented from the TOF-MRAs using in-house-developed software implemented in MeVisLab (MeVis Medical Solutions). A contour was drawn around the outline of the aneurysm on axial slices, and the parent vessels were not included (Figure 5.1). The UIA volume (in millilitres) was determined on the basis of the voxels contained within the contours and the MRA voxel size. Annotations were performed independently by two observers, first on the

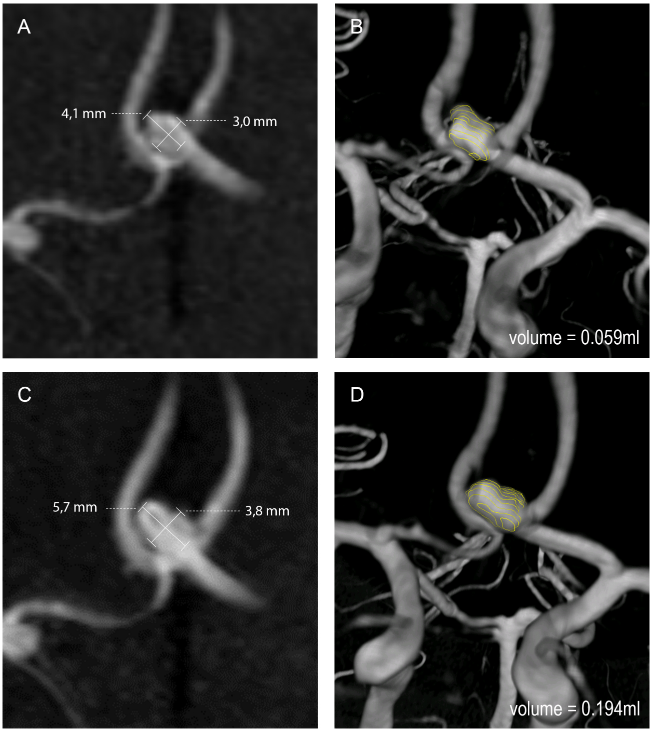


Figure 5.1: Baseline (A and B) and follow-up (C and D) TOF-MRA with an anterior communicating artery aneurysm that shows growth when measured in 2D (A and C) and in 3D (B and D).

baseline scan, followed by the follow-up scan of the same patient. The observers were the neuroradiologist (I.C.v.d.S.) and a trained medical student (D.S.).

STATISTICAL ANALYSIS First, the interobserver reliability of the individual 2D measurements (height, width, neck) and 3D measurements (volume) of the aneurysms was determined. Second, on the basis of the 2D and 3D size measurements, changes in size and volume (growth) between paired baseline and follow-up scans for 2D (difference in height, width, and neck in mm) and 3D (volume difference in millilitres) were calculated. Third, the interobserver reliability of these changes in size (2D) and volume (3D) measurements was assessed by computing the intraclass correlation coefficient (ICC). The ICC was calculated using a single-measurement, absolute-agreement, 2-way random-effects model [141]. An ICC above 0.9 represents excellent reliability; between 0.75 and 0.90, good reliability; between 0.5 and 0.75, moderate reliability; and lower than 0.5, poor reliability [141, 142]. The interobserver reliability for detecting change in 2D and 3D measurements was compared in regular and irregular aneurysms.

The SDC was computed on the basis of the 2D and 3D measurements to assess the interobserver agreement. The SDC represents the minimal change that an aneurysm measurement must show to ensure that the observed change is real and not just due to measurement error. For both 2D size and 3D volume measurements, we calculated the standard error of measurement (SEM) using the ICC previously determined. The SDC was calculated from the standard error, $\mathrm{SEM}_{\mathrm{agreement}}$, $\mathrm{SDC} = 1.96\sqrt{2}\,\mathrm{SEM}_{\mathrm{agreement}}$; where $\mathrm{SEM}_{\mathrm{agreement}} = \mathrm{SD}\sqrt{1 - \mathrm{ICC}_{\mathrm{agreement}}}$; and SD is the standard deviation of all measurements [143].

Bland-Altman plots for the interobserver difference between the change in 2D and 3D measurements between baseline and follow-up scans were created to assess agreement. The difference between each observer and the overall mean of both observers was calculated and plotted. The limits of agreement from the mean (±1.96 SD) were determined. Measurements outside the limits of agreement were considered outliers.

The number of UIAs with change in 2D height and/or width measurements larger than 1 mm, the current clinical definition of aneurysm growth [25], was determined. Next, the number of UIAs with a change in 2D height and/or width and volume larger than the determined 2D and 3D SDCs was determined. The proportion of UIAs showing growth based on the 1 mm clinical definition versus the proportion of UIAs with growth based on the SDCs was compared.

Finally, a sub-analysis was performed stratifying the reliability of change measurements for aneurysm location: anterior cerebral or communicating artery, internal carotid artery, posterior communicating artery, MCA, and posterior circulation.

All data analyses were conducted using Pandas, SciPy, and Pengouin [69] toolboxes with Python 3.7 (https://www.python.org/downloads/release/python-370/).

5.3 Results

We included 72 patients with 84 UIAs. The mean age was 53 years (range, 27–73 years), and 71% were women. Most patients had 1 UIA (n = 63). The median time between baseline and follow-up scans was 4.7 years (range, 0.9–13.1 years). The median aneurysm height was 3.4 mm (range, 0.8–15 mm). 22% of aneurysms were located at the anterior cerebral artery/anterior communicating artery, 27% at the ICA or posterior communicating artery, 38% at the MCA, and 13% in the posterior circulation. Figure 5.1 shows an example of a growing aneurysm measured in 2D and 3D.

The interobserver reliability of the 2D size and 3D volume measurements is summarised in Table 5.1. The ICC of the individual 2D size measurements was excellent for height (0.93), good for width (0.85), and moderate for the neck (0.74). The ICC for the individual 3D volume measurement was excellent (0.98).

The ICCs for the change in measurements (growth) between the paired baseline–follow-up scans for the 2 observers are shown in Table 5.2. The ICC for the change in 2D measurements was poor for height (0.46), width (0.45), and neck (0.26). The ICC for the change in 3D volume measurements was good (0.76). Irregularly shaped aneurysms had a lower reliability for 2D change in height and width (ICCs = 0.23, 0.38) and 3D change in volume (ICC = 0.60) than for regular aneurysms (ICCs = 0.57, 0.47, 0.83, respectively).

On the basis of the standard error of measurement for agreement, between the 2 observers, the SDC for 2D measurements was 1.5 mm for height, 2.0 mm for width, and 1.9 mm for neck. The SDC for 3D volume measurement was 0.062 mL.

The Online Supplemental Data show Bland-Altman plots for the interobserver difference between the change in 2D and 3D measurements between baseline and follow-up scans. The Bland-Altman plots show that there are 4–6 outliers that fall outside the limits of agreement for all change measurements between size and volume. About half of these outliers (55%) were the same for 2D and 3D and were classified as irregularly shaped by the observers. There was no relation between aneurysm size and the outliers.

The number of UIAs with a change in size measurements larger than 1 mm and a change in size and volume larger than the SDCs is shown in the Online Supplemental Data. The proportion of UIAs with growth based on the definition of 1 mm was 10%, compared with 2% when using a 1.5 mm change in height as a cutoff value (SDC for 2D height) or a 2.0 mm change in width as a cutoff value (SDC for 2D width) and a 0.062 mL change (SDC for 3D) as cut-off value.

The Online Supplemental Data indicate the reliability of the change in measurements in different locations. The reliability was found to be location-dependent for both 2D and 3D; however, 3D measurements were more reliable than 2D measurements across all locations (ICC >0.5).

Parameters	Height (mm)	Width (mm)	Neck (mm)	Volume (mL)
Observer A	3.4 (2.4-4.4)	3.4 (2.2-4.5)	2.6 (2.0-3.5)	0.0278 (0.0117-0.0578)
Observer B	3.4 (2.5-4.6)	3.2 (2.2-4.3)	2.0 (1.6-2.8)	-
Observer C	-	-	-	0.0227 (0.0090-0.0470)
Abs. Diff_{obs}	0.40 (0.20-0.60)	0.45 (0.20-0.80)	0.55 (0.30-1.00)	0.0091 (0.0036-0.0206)
$\text{ICC}_{\text{agree}}$(95%)	0.93 (0.90-0.95)	0.85 (0.80-0.89)	0.74 (0.31-0.87)	0.98 (0.97-0.98)

Table 5.1: Interobserver size and volume measurements. 2D and 3D measurements of the aneurysms by observers A and B for 2D and observers A and C for 3D. Total: 168 aneurysms, including both baseline and follow-up scans. Each measurement is provided as a median (quartiles 1–3). Reliability is in the bottom row as an ICC on absolute agreement (95% confidence interval). Note:—– indicates no measurement; $\text{ICC}_{\text{agreement}}$, Intraclass correlation coefficient on absolute agreement between observers' measurements; Abs. Diff_{obs}, absolute difference between observers' measurements.

Parameters	Height (mm)	Width (mm)	Neck (mm)	Volume (mL)
Observer A	0.2 (-0.1-0.7)	0.1 (-0.2-0.4)	0.0 (-0.2-0.4)	0.00001 (-0.0043-0.0011)
Observer B	0.1 (-0.1-0.5)	0.1 (-0.2-0.6)	0.0 (-0.1-0.3)	-
Observer C	-	-	-	0.0015 (-0.0054-0.0106)
Abs. Diff_{obs}	0.40 (0.20-0.70)	0.40 (0.20-0.60)	0.30 (0.10-0.70)	0.0057 (0.0024-0.0135)
$\text{ICC}_{\text{agree}}$(95%)	0.46 (0.27-0.61)	0.45 (0.26-0.60)	0.26 (0.06-0.46)	0.76 (0.65-0.76)

Table 5.2: Interobserver change measurements. Change between baseline and follow-up measurements of the 2D height, width, neck and 3D volume of the aneurysm by observers A and B for 2D and observers A and C for 3D. Total: 84 baseline–follow-up pairs measured by 2 observers. Substantial positive differences between baseline and follow-up may indicate growth of the aneurysm. Each measurement is provided as a median (quartiles 1–3). Reliability of the differences is provided in the bottom row as the ICC on absolute agreement (95% confidence interval).Note:—– indicates no measurement; $\text{ICC}_{\text{agree}}$, Intraclass correlation coefficient on absolute agreement between observers' measurements; Abs. Diff_{obs}, absolute difference between observers' measurements.

5.4 Discussion

In this study, interobserver reliability was better for 3D than 2D measurements of UIAs, both for individual size and detection of change in size (growth). Overall, the interobserver reliability of both 2D and 3D measurements was lower for the detection of change (growth) compared with measurements on individual scans. The SDC between the baseline and follow-up scan for 2D measurements was substantially larger than the current clinical definition (1 mm), and proportions of UIAs showing growth decreased more than three-quarters depending on the growth definition.

Many studies have investigated MRAs for UIA diagnosis [48]. However, few studies have investigated the interobserver reliability of 2D measurements from individual MRAs of patients, and no studies have fully investigated the reliability and agreement of growth measurements between baseline and follow-up MRAs of the same patient. The results of studies for individual 2D height and width measurements are similar to our findings, with the lowest reliability for measuring the neck. Kim et al [27] studied intra- and interobserver individual 2D measurement variability of 33 aneurysms with a mean size of 5.1 mm, finding an ICC of 0.83–0.99 on MRAs with the lowest reliability for the neck measurement (ICC = 0.83-0.86). Mine et al [144] compared the diagnosis and measurements of UIAs between DSAs and MRAs. Three readers assessing 56 aneurysms in MRAs determined an interobserver agreement between individual 2D maximal diameter as moderate-to-substantial (k = 0.53–0.66) and the neck measurement as fair-to-moderate (k = 0.20–0.41). The lower ICC for the neck is likely due to difficulty in defining an aneurysm neck, particularly if there are branching vessels emerging from the neck. This lower measurement reliability for neck measurements may have implications for treatment planning and complication risk assessment [5]. For aneurysm growth assessment, the neck measurement is less important because height and width measurements are commonly used [25, 40].

With ever improving image analysis techniques, 3D measurements of UIAs [30, 34, 145] are more commonly investigated, but little is known of their reliability or reproducibility for individual size and growth measurement of UIAs in TOF-MRAs. D’Argento et al [146] found no significant difference in intra- and interobserver variability of automatic and manual 2D size measurements of UIAs on 3D DSAs and CTAs.

We determined the ICC of absolute agreement to include the systematic error of both observers and random residual errors. A substantially lower ICC for change measurements (growth) between paired baseline–follow-up scans was determined, relative to measurements from individual scans. The ratio of the systematic measurement error compared with the individual aneurysm size is smaller than the ratio of the measurement error compared with the change in aneurysm size. Thus, a small measurement error in individual measurements can have a larger influence on the subsequent change measurements in paired scans.

The interobserver agreement in 2D and 3D measurements was assessed by determining the SDC. The SDC for both 2D and 3D measurements was relatively large, compared with the median aneurysm size (3.4 mm) and median aneurysm volume (0.025 mL). For example, for 2D height, the SDC of 1.5 mm was about half of median aneurysm height. This study has a large proportion of small aneurysms, and the ratio of the SDC to aneurysm size would be better (lower) in larger aneurysms. However, because most patients who undergo follow-up MRAs have small UIAs, our population represents the clinical situation. The SDC for the 2D measurements is larger than the 1 mm used in the current definition of aneurysm growth [25]. The number of UIAs showing growth according to threshold values of the SDC of 2D and 3D measurements decreased by more than three-quarters compared with this 1 mm threshold. This finding shows the influence of the thresholds for growth definition and has potential important clinical consequences for treatment decisions based on aneurysm growth.

The Bland-Altman plots (Online Supplemental Data) show that 3D interobserver differences were more similar than 2D measurements because the measurements were closer together. Most outliers for both the 2D and 3D measurements were irregularly shaped. We also found that irregular aneurysms had a lower inter-observer reliability for detecting change in both 2D size and 3D volume measurements. Irregular aneurysm shape is a risk factor for rupture [32, 40]. 2D measurements and shape assessment of aneurysms are influenced by the selected viewing angle. 3D volume measurements allow a more complete shape of the UIA to be assessed with a single, rotation-invariant measure. Furthermore, 3D segmentation may allow quantitative shape assessment of UIAs, which would be potentially beneficial in risk assessment [31].

We found that aneurysm location affects the reliability of 2D and 3D measurements. We found that the reliability of 3D volume measurements was higher and more consistent for all locations than 2D size measurements.

There were some limitations in our study. One limitation was that the 3D measurements were determined from segmentations based on 2D annotations on axial slices, which is time-consuming and the aneurysm neck definition could be difficult, particularly when the parent vessel did not lie in-plane. Furthermore, the difference in experience of the second observers for 2D (radiologist) and 3D (student) measurements may have introduced bias. If this had influenced our results, it would be toward less agreement for the 3D measurement between the student and the neuroradiologist. However, we found higher agreement in 3D than in 2D.

Second, most scans had small aneurysms with a median diameter of 3.4 mm (range, 0.8–15 mm). The population of patients with small UIAs is, however, representative of patients who undergo follow-up imaging. Because rupture risk increases with aneurysm size, the larger UIAs are more often treated. The protocol and quality of the MRAs between baseline and follow-up differed in some cases, possibly resulting in measurement differences, but they are realistic for clinical practice.

This study investigates TOF-MRAs only because this is the preferred imaging method for follow-up of UIAs [140].

Our findings of a large SDC for 2D size measurements may have implications for the definition of clinical aneurysm growth and growth/rupture models. This subject requires further study because it would have important consequences for rupture and treatment assessment of UIAs. 2D and 3D measurements cannot be directly compared, but instead a standard growth definition should be used for both. The higher reliability of 3D measurements compared with 2D measurements implies that 3D measurements may be important for accurate assessment of aneurysm growth on TOF-MRA. Automatic or semi-automatic 3D UIA segmentation would allow faster and less operator-dependent aneurysm volume measurement for standard 3D growth assessment, alongside quantitative 3D morphological characterisation of UIAs.

5.5 Conclusions

This study found that 3D change measurements are more reliable than 2D with regard to assessing the change in size and volume measurements of UIAs. The SDC for 2D measurements was found to be larger than the current definition for clinical growth, suggesting that more studies into the reliability of 2D measurement on MRA should be performed. This study opens the door for development and incorporation into of automatic and semi-automatic segmentations and volumetric growth assessments of UIAs into clinical practice.

5.6 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 acknowledge Djenghiz Samlal (D. S.), University Medical Center, Utrecht, the Netherlands, for his contributions. We further acknowledge support from the focus area Applied Data Science, Utrecht University.

5.7 Online Supplemental Data

Online Supplemental Data can be found here: http://www.ajnr.org/content/ajnr/suppl/2021/07/01/ajnr.A7186.DC1/1724.pdf

Chapter 6

Relationship between 3D Morphologic Change and 2D and 3D Growth of Unruptured Intracranial Aneurysms

Published as: K. Timmins, H. Kuijf, M. Vergouwen, Y. Ruigrok, B. Velthuis, and I. van der Schaaf. "Relationship between 3D Morphologic Change and 2D and 3D Growth of Unruptured Intracranial Aneurysms," *American Journal of Neuroradiology*, vol. 43 (2022), pp. 416–421.

Abstract

Background and Purpose: Untreated unruptured intracranial aneurysms are usually followed radiologically to detect aneurysm growth, which is associated with increased rupture risk. The ideal aneurysm size cutoff for defining growth remains unclear and also whether change in morphology should be part of the definition. We investigated the relationship between change in aneurysm size and 3D quantified morphologic changes during follow-up.

Materials and Methods: We performed 3D morphology measurements of unruptured intracranial aneurysms on baseline and follow-up TOF-MRAs. Morphology measurements included surface area, compactness, elongation, flatness, sphericity, shape index, and curvedness. We investigated the relation between morphologic change between baseline and follow-up scans and unruptured intracranial aneurysm growth, with 2D and 3D growth defined as a continuous variable (correlation statistics) and a categoric variable (t test statistics). Categoric growth was defined as 1 mm increase in 2D length or width. We assessed unruptured intracranial aneurysms that changed in morphology and the proportion of growing and nongrowing unruptured intracranial aneurysms with statistically significant morphologic change.

Results: We included 113 patients with 127 unruptured intracranial aneurysms. Continuous growth of unruptured intracranial aneurysms was related to an increase in surface area and flatness and a decrease in the shape index and curvedness. In 15 growing unruptured intracranial aneurysms (12%), curvedness changed significantly compared with nongrowing unruptured intracranial aneurysms. Of the 112 nongrowing unruptured intracranial aneurysms, 10 (9%) changed significantly in morphology (flatness, shape index, and curvedness).

Conclusions: Growing unruptured intracranial aneurysms show morphologic change. However, nearly 10% of nongrowing unruptured intracranial aneurysms change in morphology, suggesting that they could be unstable. Future studies should investigate the best growth definition including morphologic change and size to predict aneurysm rupture.

6.1 Introduction

In management decisions on unruptured intracranial aneurysms (UIAs), the risk of rupture needs to be balanced against the risk of treatment complications [6]. UIAs often remain untreated if the risk of treatment complications is higher than the risk of rupture [5, 115]. In that case, UIAs can be monitored with follow-up imaging to detect potential aneurysmal growth, which is associated with an increased risk of rupture [7]. If aneurysmal growth is detected, preventive aneurysm treatment should be reconsidered.

Substantial heterogeneity exists in the definition of UIA growth [40, 147, 148]. Generally, the definition of growth includes a certain increase in aneurysm size and/or any morphologic change [26]. Currently, it remains unclear which definition is most relevant and how morphologic changes relate to any change in aneurysmal size. UIA size and morphology are currently assessed with caliper measurements and visual classification by human observers, which can be prone to measurement errors and poor reproducibility [26, 27]. 3D volumetric segmentations of UIAs enable reproducible and reliable quantification and analysis of UIA volume, morphology, and assessment of changes in morphology [35, 147, 149]. The recent Image Biomarker Standardisation Initiative (IBSI) [38] has been developed to standardise quantitative radiomics extracted from medical imaging, including morphology measurements.

3D quantified morphology measurements of UIAs [34, 145] are more frequently used to better understand growing and unstable UIAs. Previous studies have investigated difference in morphology in growing UIAs [34, 150] and morphology as a predictor of UIA instability [36]. However, no studies have investigated morphologic changes in stable or nongrowing aneurysms. Furthermore, various different morphology measurements are used, making it difficult to make direct comparisons between studies.

More investigation is warranted into both growing and non-growing (stable) UIAs to understand the relationship between growth and morphologic change of UIAs using standardised morphology definitions.

This study aimed to investigate the relationship between UIA growth and morphologic change by considering continuous and categoric (dichotomous) 2D and 3D growth of growing and non-growing UIAs.

6.2 Materials and Methods

6.2.1 Study Population

From the UIA data base of the University Medical Center Utrecht, the Netherlands, we included consecutive patients of >18 years of age who adhered to the following inclusion criteria: (1) at least 1 saccular UIA; (2) a 3D TOF-MRA available both at the baseline admission scan and at follow-up in the period 2004–2020; and (3) the interval between the baseline scan and follow-up scan was at least 6 months. Exclusion criteria were the following: (1) fusiform or arteriovenous malformation–related aneurysm; and (2) aneurysm rupture or preventive treatment between baseline and the first follow-up scan. For each patient, we assessed both a baseline and the most recent follow-up TOF-MRA scan for the analysis. All scans were obtained between 2004 and 2020. Due to the time period, protocols varied, but either a 1 T, 1.5 T, or 3 T scanner was used with a median TR of 23 ms and a median TE of 4 ms across all scans. The scans had a median in-plane resolution range of 0.357 mm and a median section thickness range of 0.5 mm. All scans were pre-processed and re-sampled to the same voxel size (0.357 x 0.357 x 0.500 mm) to account for scan protocol differences. The institutional review board of the University Medical Center Utrecht waived individual patient consent and formal ethics approval for this study because data available from routine patient care were used.

6.2.2 Measurements

2D MEASUREMENTS 2D measurements of the UIAs in all scans were performed manually on the IntelliSpace Portal (Philips Healthcare) by an experienced neuroradiologist (I.C.v.d.S.). The UIA length and width were measured on the TOF-MRAs on a 0.1 mm scale using electronic calipers [31, 40]. UIA length was defined as the maximum distance from the UIA neck to the UIA dome. UIA width was measured perpendicular to the measured length along the maximum width of the UIA. Individual length and width measurements were made on both the baseline and follow-up scans. 2D length and width changes were determined as the difference in the 2D length and width measurements between the follow-up and baseline scans of the same UIA of the same patient.

3D MEASUREMENTS To make 3D quantified morphology measurements of the UIAs, we manually segmented the UIAs from the original TOF-MRAs using in-house developed software implemented in MeVisLab (MeVis Medical Solutions). All annotations were made by drawing a contour around the UIA on axial slices of the original TOF-MRA by the neuroradiologist who made the 2D measurements. The annotation did not include the parent vessels. Annotations were first made on the baseline scan, followed by the follow-up scan of the same patient. The annotations were converted to binary masks in which voxels that were located. 50% inside the contour were labelled as UIAs. The images and annotations were all re-sampled to the median voxel size of 0.357 x 0.357 x 0.500 mm. Using a marching cubes algorithm [151], we automatically fitted a mesh to the outside of the segmented UIA. The volume and surface area of the UIA were determined on the basis of the mesh around the segmented UIA. 3D volume change was determined as the difference in volume between the follow-up and baseline scans. The size of the UIA was determined by performing principal component analysis on the voxels within the segmented UIA and calculating the major, minor, and least extent. From these values, various morphology measurements were calculated on the basis of definitions in accordance with the IBSI guidelines [38], including compactness 1, compactness 2, elongation, flatness, and sphericity. Compactness 1 and 2 and sphericity are different measures that all quantify how similar the morphology of the UIA is to a sphere. Elongation describes the eccentricity of the UIA by describing how long it is relative to its width. Flatness quantifies the amount the UIA is flat relative to the length. Next, on the basis of the generated 3D mesh, the mean and Gaussian curvature of the surface of the UIA was determined, allowing the principal curvatures k1 and k2 to be calculated. By means of these principal curvatures, it was possible to determine the shape index and curvedness (Fig 6.1) [119]. Shape index and curvedness were calculated for every point on the mesh, and a median over the whole mesh of the UIA was determined. The shape index is a descriptor of the local shape of the surface of an object and is scale-invariant. The curvedness is a positive value, which describes the local curvature of the surface and is dependent on the local scale of the object. These values are rotation and translation invariant, and Fig 6.1 depicts examples of how these values vary.

All measurements and segmentations were performed on anonymised data sets by a neuroradiologist (I.C.v.d.S., with 15 years of experience). 2D measurements and the 3D segmentations were performed in a different order and several months apart to prevent bias. The observer was not blinded to the time order of the scans because this reflects the clinical setting. Morphology measurements were made on both the follow-up and baseline scans. Morphologic change was considered the difference between each morphology measurement at follow-up compared with baseline.

6.2.3 Statistical Analysis

MORPHOLOGIC CHANGES IN RELATION TO CONTINUOUS UIA 2D AND 3D GROWTH. The relation between morphologic change and UIA growth was investigated by assessing growth as a continuous 2D and 3D outcome measurement (2D size: length and width in millimetres and 3D volume in cubic millimetres). Correlations were assessed using the Pearson or Spearman correlation coefficient, in which normality was tested using a Shapiro-Wilk test. The threshold for statistical significance was $P < 0.05$.

MORPHOLOGIC CHANGES IN GROWING AND NONGROWING UIAS. The whole study population was categorised into 2 groups, including either growing or nongrowing UIAs. Growing UIAs were defined according to the clinical definition of 2D growth of ≥ 1 mm increase in any direction between the baseline and follow-up scan [25]. All other UIAs were categorised as nongrowing. The difference in morphology measurements between baseline and follow-up scans was determined for each UIA (morpho-

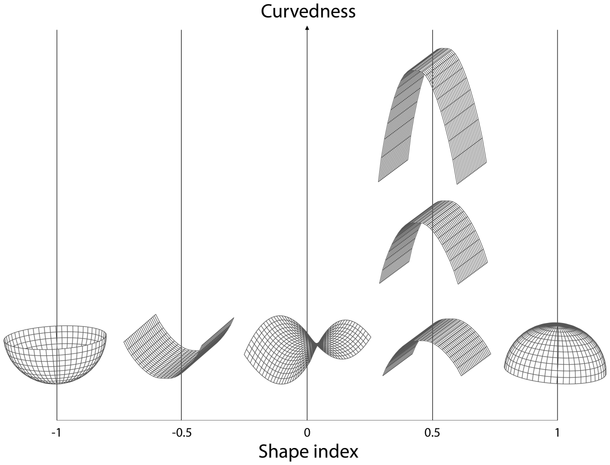


Figure 6.1: Shape index and curvedness. Shape index and curvedness values vary in 3D shapes. The shape index is a descriptor of the local shape of the surface of an object and is scale-invariant. Shape index values range from –1 (concave "cup") through 0 (saddle point) to 1 (convex "dome"). The curvedness is a positive value to the local curvature of the surface, which usually lies between 0 and 1 and is dependent on the local scale of the object. These values are rotation- and translation-invariant.

logic change) and compared between the populations of growing and nongrowing UIAs. An unpaired Student t test was used for normally distributed data, and a Mann-Whitney U test, for non-normally distributed data, for which normality was tested using a Shapiro-Wilk test. The threshold for statistical significance was P<0.05.

MORPHOLOGIC CHANGE BASED ON MODIFIED Z SCORES. We determined the modified z score of the morphologic change, to identify UIAs with morphologic changes that significantly differed from those in most of the study population. This allowed us to differentiate those UIAs that can be considered to change in morphology more than could be expected on the basis of the trend of morphologic change in our population. The modified z score (M_i) for each morphology measurement for each UIA was determined as $M_i = \frac{0.675(x_i - \tilde{x})}{MAD}$, ,where MAD is the median absolute deviation $MAD = median(|x_i - \tilde{x}|)$, and x_i was each morphology measurement for each (ith) UIA. For this subanalysis, we selected the morphologic parameters that were statistically significantly related to growth (either as continuous or categoric variables) on the basis of the first analyses (parameters: flatness, shape index, and curvedness). Statistically significant morphologic change was defined as any change in morphology measurement that had a modified z score >3.5 [152]. Finally, we determined the proportion of UIAs with statistically significant morphologic change in the 2 groups of growing and nongrowing UIAs.

Characteristic	
No. of patients	113
No. of aneurysms	127 (102 patient with 1 UIA, 8 with 2 and 3 with 3 UIAs)
Sex (% women)	81 women, 32 men (72% women)
Age at baseline (mean) (yr)	55 (range, 27-77)
Time between baseline and follow-up scan (median) (yr)	4.1 (range, 0.9–13.1)
Location of Aneurysm	
Anterior cerebral or communicating artery	25 (20%)
ICA or posterior communicating artery	33 (26%)
MCA	53 (42%)
Posterior circulation	16 (13%)

Table 6.1: Patient Characteristics

6.3 Results

6.3.1 Study Population

We included 113 patients with 127 UIAs who met the inclusion criteria (Table 6.1). After a median follow-up time of 4.1 years (range, 0.9–13.1 years), aneurysm growth was observed in 15/127 (12%) UIAs. There was no statistically significant difference in follow-up time between the groups of growing and nongrowing aneurysms (Mann-Whitney U test, P = 0.48). A morphologic change that differed statistically significantly from that in most of the study population was found in 18/127 (14%) UIAs.

6.3.2 Morphologic Change in Relation to Continuous UIA 2D and 3D Growth

The correlation between UIA morphologic change and continuous UIA growth (2D size and 3D volume) is shown in Table 6.2. An increase in volume and surface area showed a statistically significant correlation with 2D growth. An increase in surface area and flatness and a decrease in the shape index and curvedness showed statistically significant correlation with continuous 3D volume growth. Shape index and curvedness were also seen to decrease with increasing continuous 2D length and width measurements, but not enough to be considered statistically significant.

		Correlation Coefficient (P value)		
Change in	Median (IQR)	2D Growth, Length	2D Growth, Width	3D Growth, Volume
Volume (mm^3)	1.60 (-3.70-10.80)	0.29 (<0.01)*	0.28 (<0.01)*	-
Area (mm^2)	2.30 (-6.10-11.40)	0.25 (<0.01)*	0.34 (<0.01)*	0.90 (<0.1)*
Compactness1	0.50 (-1.70-2.60)[a]	0.10 (0.28)	-0.01 (0.89)	0.15 (0.09)
Compactness2	0.01 (-0.04-0.07)	0.09 (0.30)	-0.02 (0.86)	0.15 (0.10)
Elongation	0.01 (-0.03-0.04)	0.01 (0.89)	0.02 (0.79)	0.09 (0.33)
Flatness	0.00 (-0.03-0.04)	0.00 (0.97)	0.05 (0.58)	0.19 (0.03)*
Sphericity	0.01 (-0.02-0.04)	0.10 (0.27)	-0.01 (0.94)	0.15 (0.09)
Shape Index	0.00 (-0.03-0.01)	-0.09 (0.32)	-0.17 (0.06)	-0.33 (<0.01)*
Curvedness	-0.01 (-0.18-0.07)	-0.12 (0.16)	-0.15 (0.09)	-0.33 (<0.01)*

Table 6.2: Change in UIA morphology measurements in relation to continuous UIA 2D and 3D growth. Note:—IQR indicates interquartile range; -, perfect correlation (same input variable). The correlation coefficient was calculated with the Pearson or Spearman correlation based on normality of morphologic change. * P values are statistically significant. [a] All values x10^3.

Change in	Growing	Nongrowing	P value
Volume (mm^3)	21.92 (4.80-33.23)*	1.42 (-4.26-9.76)*	0.01*
Area (mm^2)	28.09 (-4.23-35.37)*	2.07 (-6.47-9.89)*	<0.01*
Compactness1	0.40 (-1.35-3.85)^a	0.00 (-1.65-2.33)^a	0.21
Compactness2	0.01 (-0.03-0.11)	0.01 (-0.04-0.06)	0.22
Elongation	-0.02 (-0.05-0.02)	0.02 (-0.03-0.04)	0.42
Flatness	0.01 (-0.04-0.04)	0.00 (-0.03-0.04)	0.37
Sphericity	0.01 (-0.02-0.05)	0.01 (-0.02-0.03)	0.20
Shape Index	0.00 (-0.13-0.00)	0.00 (-0.03-0.01)	0.06
Curvedness	-0.14 (-0.37-0.01)*	-0.01 (-0.16-0.09)*	0.03*

Table 6.3: Comparing change in 3D quantified morphology of stable and growing UIAs. Values are written as median (IQR). Growth was defined as an increase of at least 1 mm in either width or length of the UIA. P values refer to the relation between parameters of the growing and stable UIAs using a t test or Mann-Whitney U test. * P values are statistically significant. ^{a} All values x10^3.

6.3.3 Morphologic Change in Growing and Nongrowing UIA

Morphologic changes in growing and nongrowing UIAs are shown in Table 6.3. There were 15 growing UIAs (12% of all 127 UIAs). Growing UIAs had a higher increase in volume and surface area and a larger decrease in curvedness compared with nongrowing UIAs ($p<0.05$).

6.3.4 Morphologic Change Based on Modified z Scores

For the parameters flatness, shape index, and curvedness, we determined the proportion of UIAs with morphologic changes that statistically significantly differed from most of the full study population. In total, 18 UIAs (14%) changed statistically significantly in ≥ 1 of the morphology parameters compared with most of the population. Eight of the 15 growing UIAs (53%) and 10 of the 112 nongrowing UIAs (9%) showed a statistically significant morphologic change (Fig 6.2).

6.4 Discussion

This study showed a correlation between UIA 3D quantified morphologic changes and UIA growth, as both continuous and categoric variables. Increase in surface area and flatness and decrease in shape index and curvedness were correlated with continuous 3D volume growth. Surface area and curvedness remained statistically significant for growth as a categoric variable. In addition, nearly 1 of 10 nongrowing UIAs also showed morphologic change, suggesting that UIAs can change in morphology even if they are considered nongrowing.

Several previous studies investigated 3D quantified morphology of UIAs, in relation to UIA growth and as a predictor for UIA rupture. In one study, 56 growing UIAs

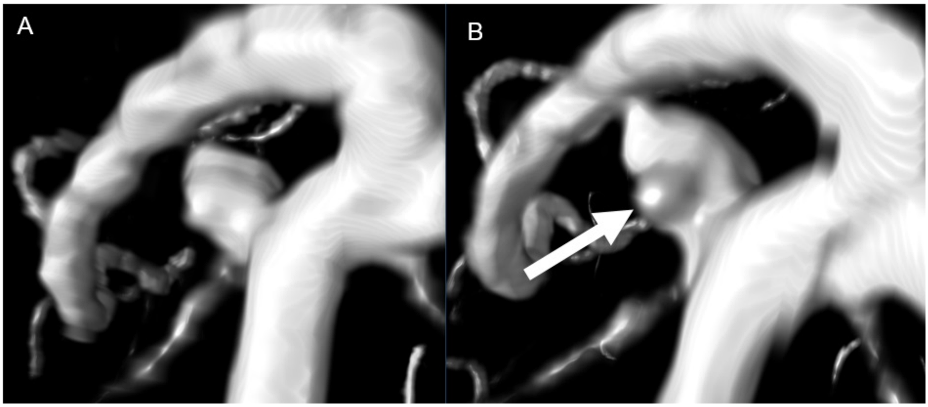


Figure 6.2: Nongrowing UIA with statistically significant change in morphology. An example of a ROI around a UIA taken from baseline and follow-up TOF-MRAs made on a Philips 1.5T scanner (Intera/Achieva). The measured UIA shows statistically significant changes in morphology but was considered to be nongrowing (<1 mm change in length or width). The bulge (arrow)that becomes visible on follow-up results in more saddle points on the surface of the UIA, whereby the shape index decreases. The bulge also increases the curvature of the surface of the UIA, resulting in a slightly increased curvedness value.

and 81 nongrowing UIAs were included [34]. UIA growth was defined as an increase of at least 0.5 mm in any direction or a visual change in shape. Only baseline scans of nongrowing UIAs were assessed. At baseline, no statistically significant morphologic differences were observed between non-growing UIAs and UIAs with future growth. Another study included 38 growing UIAs [150]. Growth was defined as 1 mm growth in 1 direction, 0.5 mm growth in 2 directions, or a significant visual change in shape. Similar to our findings, morphology of the UIA (bottleneck factor and ellipticity index) after growth was statistically significantly different from baseline morphology. A third study included 420 UIAs and investigated whether 12 different morphologic measurements of UIAs predicted UIA stability [36], which was defined as rupture within 1-month, clinically defined growth at radiologic follow-up or symptomatic UIAs with adjacent structure compressive symptoms. They found that flatness was the most important morphologic measurement to predict UIA stability.

A direct comparison with previous studies is difficult because consistent methodology and morphology measurements have not been used. Because the field of quantitative medical image analysis is developing rapidly, the IBSI guidelines provide a standardisation of radiomics and morphology measurements across all medical images [38]. Thus, this study incorporated the morphology measurements as defined in IBSI to assess the growth of UIAs on TOF-MRAs.

Our study differs from previous studies by investigating changes in morphologic measurements between baseline and follow-up of both growing UIAs and UIAs that were considered to be nongrowing. By this method, we were able to show that 9%

of nongrowing UIAs also showed statistically significant morphologic changes. This raises the question of whether UIA growth and stability should be defined only by size measurements and suggests that (quantified) standard morphologic measurements could also be considered when assessing the stability of UIAs with regard to growth and potential subsequent rupture.

New in our study, compared with previous studies, was the definition of growth both as a continuous as well as a categoric variable. We found more differences that are statistically significant using continuous outcome measures for UIA growth compared with categoric outcomes. By assessing growth as a continuous measure, we consider all UIAs with any change in size or volume, without the use of a cutoff value. By dichotomising growth into nongrowing and growing categories, precision is lost, reducing the statistical power to find relationships between growth and morphology measurements, which is especially important in smaller data sets. Growth measurements that are close to the 1 mm cutoff, for example 0.9 and 1.1 mm, can be very similar, but by means of a dichotomous measure, they are categorised as completely different. Because many of our UIA growth measurements are in this range, around the clinical definition of growth, a continuous outcome has much larger power [153]. Despite a larger statistical power of continuous measurements, in the clinical setting a definition of dichotomised growth is important because it allows better interpretation of UIA growth and facilitates clinical decision-making. However, the growth definition of 1 mm is rather arbitrary, and studies suggest that the interobserver variability in growth measurements could be larger than this [147]. Future studies are needed to investigate the best cutoff values for size and morphologic change and a growth definition for predicting aneurysm rupture. Change in size and morphology could aid in rupture prediction modelling, and how this may affect treatment decisions of UIAs should be studied.

A limitation in our study was that the 3D measurements were determined from segmentations based on annotations on axial slices. This is time-consuming, and the definition of the UIA neck was difficult in some UIAs. An alternative and reproducible automatic aneurysm segmentation method could be used. [111] There was variation in the time period between baseline and follow-up because the most recent follow-up MRA was always performed to ensure the longest follow-up time and potential largest proportion of growth and morphologic change. In some cases, the aneurysm was treated or ruptured after the first standard follow-up at 1 year, meaning that the time until follow-up was relatively short. The growing and nongrowing aneurysms did not have statistically significantly different follow-up times. Future studies could assess the longitudinal growth and change in morphology across time. Next, because the MRA scans were performed during a long time period, the scan protocol, scanner field strength, and scan quality differed in some patients between baseline and follow-up scans for both growing and nongrowing UIAs. This difference is realistic in the clinical setting, and we did re-sample all images to median voxel spacing. This step would have influenced both growing and nongrowing UIAs; therefore, we do not think it has biased our results.

6.5 Conclusions

Our study suggests that both aneurysm size and morphologic changes should be taken into account when assessing UIA growth during radiologic follow-up. However, more studies should be undertaken to develop a complete growth definition based on size and standard 3D-quantified morphology measurements.

6.6 Acknowledgements

This work was supported by the European Commission; Horizon 2020 Framework Program, H2020; Research and Innovation Program; European Research Council, 852173; and Hartstichting, 2018T076, CVON2015-08, CVON2018-02.

Chapter 7

Future Unruptured Intracranial Aneurysm Growth Prediction using Mesh Convolutional Neural Networks

Published as: K. Timmins, M. Kamphuis, I. Vos, B. Velthuis, I. van der Schaaf, and H. Kuijf. "Future Unruptured Intracranial Aneurysm Growth Prediction using Mesh Convolutional Neural Networks," *MICCAI Conference Proceedings 2022: tda4biomedicalimaging workshop (in press) *joint last author,*

Abstract

The growth of unruptured intracranial aneurysms (UIAs) is a predictor of rupture. Therefore, for further imaging surveillance and treatment planning, it is important to be able to predict if an UIA is likely to grow based on an initial baseline Time-of-Flight MRA (TOF-MRA). It is known that the size and shape of UIAs are predictors of aneurysm growth and/or rupture. We perform a feasibility study of using a mesh convolutional neural network for future UIA growth prediction from baseline TOF-MRAs. We include 151 TOF-MRAs, with 169 UIAs where 49 UIAs were classified as growing and 120 as stable, based on the clinical definition of growth (>1 mm increase in size in follow-up scan). UIAs were segmented from TOF-MRAs and meshes were automatically generated. We investigate the input of both UIA mesh only and region-of-interest (ROI) meshes including UIA and surrounding parent vessels. We develop a classification model to predict UIAs that will grow or remain stable. The model consisted of a mesh convolutional neural network including additional novel input edge features of shape index and curvedness which describe the surface topology. It was investigated if input edge mid-point co-ordinates influenced the model performance. The model with highest AUC (63.8%) for growth prediction was using UIA meshes with input edge mid-point co-ordinate features (average F1 score = 62.3%, accuracy = 66.9%, sensitivity = 57.3%, specificity = 70.8%). We present a future UIA growth prediction model based on a mesh convolutional neural network with promising results.

7.1 Introduction

Approximately 3% of the general population has a unruptured intracranial aneurysm (UIAs) [6]. If an UIA ruptures, it leads to subarachnoid haemorrhage with a high mortality and morbidity rate. Neurosurgical or endovascular treatment can prevent UIAs from rupture, but carry a considerable risk. Therefore a balanced decision based on the rupture and treatment complication risk must be made [5]. UIA growth is an important rupture risk factor [7], and if detected, preventative treatment should be considered. Most UIAs are monitored, using Time-of-Flight Magnetic Resonance Angiographs (TOF-MRAs) or Computed Tomography Angiographs (CTAs). Currently, 2D size measurements of the UIAs are made by a radiologists and changes in size (>1 mm) would be considered aneurysmal growth [25]. Shape and topology of UIAs is also known to be different in aneurysms that grow [31] and is often visually assessed. The ELAPSS score [31] is a clinical score for UIA growth prediction based on patient and aneurysm characteristics. The predictors are: Earlier subarachnoid haemorrhage, aneurysm Location, Age, Population, aneurysm Size and Shape. Shape is assessed visually as 'Regular' or 'Irregular'.

As computer aided radiology tools continue to be developed, there is the possibility to measure UIAs in 3D, including their shape [147]. Quantitative shape/morphology measures of UIAs could be used, including to distinguish between growing and stable aneurysms [34, 150, 154]. Based on such morphological parameters, as well as classical parameters, UIA rupture risk prediction models have been developed [155, 156]. More recently, some prediction models for aneurysmal stability and growth have been proposed [36, 37].

Liu et al. [36] investigated the feasibility of predicting aneurysm stability using machine learning regression models and 12 morphology radiomics features. The dataset included 420 aneurysms, between 4 and 8 mm in size. Instability was defined as ruptured within a month, growth or adjacent structure compressive symptoms. They determined flatness to be the most important morphology predictor of aneurysm stability. Bizjak et al. [37] found using point clouds with PointNet++ for future UIA growth prediction had a higher accuracy than other machine learning (random forest and multi-layer perceptron) models based on classical shape parameters. The method was performed using only 44 UIAs, where 25 were considered to be growing and 19 to be stable. Growing UIAs were defined by visual inspection in 3D of the UIAs and their configuration.

Various different morphology measurements and definitions of growth or stability have been used in these studies, making it difficult to make direct comparisons. However, it is clear that UIA shape and surface topology is an important predictor of future UIA growth and that deep learning methods may have an advantage over using predefined morphology parameters. Geometric deep learning methods are well suited to this problem, as they accurately describe the shape and topology of a surface by using point clouds or meshes [157]. Meshes may have a preference over point clouds as they include connectivity information, providing more information about the topology of the surfaces. Meshes could be used of the UIA itself as we already know UIA shape is a growth predictor growth. Alternatively, parent vessels in a Region-of-Interest (ROI) around the UIA could be included which allows UIA-vessel configuration to also be considered and exact UIA segmentation is not required.

MeshCNN [23] is a convolutional neural network (CNN) developed for classification and segmentation problems using 3D triangular meshes. Convolutions and pooling are performed on edges of the meshes, based on an edge neighbourhood. Five relative scale, translation and rotation invariant geometric edge are determined for each edge as input features for the model. These five geometric features are: the dihedral angle, two inner angles and two edge-length ratios. MeshCNN has only been used for a few medical imaging classification and segmentation problems, including age prediction based on the neonatal white matter cortical surface [128] and UIA segmentation from a parent vessel [158]. In our previous work, we proposed a modified version of MeshCNN for UIA detection based on brain vessel surface meshes [130]. We use our MeshCNN framework in this study.

In this paper, we propose a prediction model for future UIA growth from baseline TOF-MRAs using a mesh convolutional neural network. We investigate the use of meshes of UIAs alone, and region-of-interest (ROI) meshes including the UIA with parent vessels as input for these models and their performance for future UIA growth prediction. We also investigate the addition of edge mid-point co-ordinate input features of the meshes and the impact on the model performance.

7.2 Materials and Methods

7.2.1 Dataset

The dataset consisted of 151 baseline Time-of-Flight MRAs (TOF-MRAs) taken from routine clinical scans. We included patients with UIAs who met the following inclusion criteria: 1) A TOF-MRA or CTA was available at baseline and follow-up, 2) the follow-up scan was performed at least 6 months after the baseline scan, and 3) the patient had at least 1 untreated UIA present on both baseline and follow-up imaging. The most recent follow-up scan in which the UIA remained untreated and unruptured was used for growth assessment. Fusiform and ruptured aneurysms were excluded. All scans were made from the University Medical Center Utrecht between 2006 and 2020. The average time between baseline and follow-up scans was 5.2 ± 3.3 years (range: 1 - 16 years) The mean baseline aneurysm size was 5.0 ± 2.2 mm with a range of 1.3 – 14.7 mm. Manual 2D length and width UIA measurements were performed in IntelliSpace Portal (Philips Healthcare) by an experienced neuroradiologist (I.C.v.d.S.) and a trained PhD-student (M.J.K.) according to standard clinical protocol. Growth was defined as a ≥ 1.0 mm increase in any direction between the baseline and follow-up scan [25]. Based on this definition, UIAs were categorised as either ‘growing’ (30%, n= 49) or ‘stable’ (70%, n = 120).

7.2.2 Methods

INPUT MESH GENERATION All baseline TOF-MRAs were pre-processed using an N4 bias field correction algorithm and z-score normalised before being resampled to have voxel size 0.357 mm x 0.357 mm x 0.500 mm (median of the dataset). All UIA and ROI selection, mesh generation and processing was performed completely automatically based on UIA annotations.

UIA mesh generation UIA meshes were generated and pre-processed automatically based on the TOF-MRAs and UIA annotations. UIAs were manually segmented from the TOF-MRAs using annotations drawn on axial slices in in-house-developed software implemented in MeVisLab (MeVis Medical Solutions) (performed by I.C.v.d.S. and M.J.K). A triangular mesh was automatically fitted to the outside of the UIA surface using a Marching Cubes algorithm [151]. All UIA meshes were down-sampled to 1000 edges and included just the UIA and no other vessels.

Region-of-Interest (ROI) mesh generation ROI meshes were automatically generated from the TOF-MRAs using the UIA segmentations. An existing 3D U-net was used to automatically perform full vessel segmentation from the scans [113]. Based on the UIA segmentation, a region-of-interest (ROI) including only the UIA and parent vessels was made. The centre-of-mass of the UIA segmentation was determined and the ROI included all connected vessels (and UIA) within a 20 mm cube around the centre-of-mass. A mesh was automatically fitted to the outside of the UIA and parent vessel surface using a Marching Cubes algorithm [151]. All ROI meshes were down-sampled to 2000 edges.

Input edge features Based on the generated UIA and ROI meshes, new input edge features were automatically determined per edge. These were shape index, curvedness and edge mid-point co-ordinates. These further edge features (shape index, curvedness and mid-point co-ordinates) could then be included as input to the network, in addition to the original five geometric edge features.

Shape index and curvedness are rotation and translation invariant measures which describe the topology of the UIA surface. The invariant nature of these novel input edge features ideal for use in MeshCNN. It is known from our work that the addition of both shape index and curvedness as input edge features improve the performance of the original MeshCNN [159]. Shape descriptor values; shape index and curvedness, were calculated for each vertex on the mesh surface using the standard formulae [119].

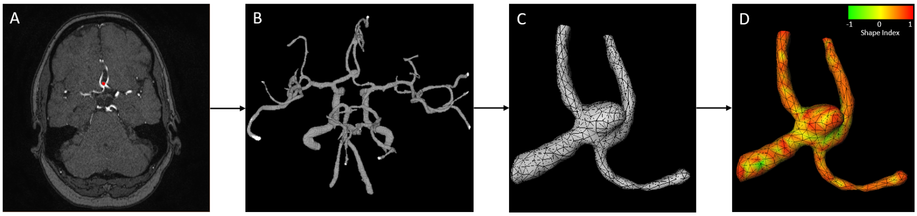


Figure 7.1: Example generation of input region-of-interest (ROI) mesh including parent vessels and UIA. A: TOF-MRA with annotated UIA shown overlaid in red. B: Vessel segmentation performed using 3D U-net [113]. C: ROI selection including UIA and parent vessels, followed by mesh generation. D: Shape index determination for each edge, to be used as an additional input feature alongside curvedness and edge coordinates.

An edge was then given a shape descriptor value (shape index or curvedness), as being the average of the values at the corresponding end vertices of the edge.

The addition of edge mid-point co-ordinate values was suggested in the original MeshCNN paper [23]. We experiment with including these co-ordinates in our models as we know location is important as an aneurysm growth predictor [31]. Edge mid-point co-ordinates (x,y,z) were determined as the average of the world co-ordinates of the corresponding end vertices of the edge.

Figure 7.1 shows an example generation of a ROI mesh including shape index values determined for each edge.

MODEL IMPLEMENTATION A ConvNet style network was set up based on our modified MeshCNN framework [130] including four convolutional layers and four pooling layers. Four different model configurations were investigated. The first model (uia_model) had UIA meshes only as input, with 1000 edges. Pooling layer configuration for the UIA model was: 750, 600, 500, 400. The second model (roi_model) had ROI meshes including UIA and parent vessels as input. The pooling layer configuration for the ROI model was: 1500, 1200, 1000, 800. All models were made to include shape index and curvedness as additional input features to the original five edge geometric features of MeshCNN. This meant that there were seven input edge features as standard. For each different input, two models were trained. The first with the seven input edge features (uia_model_1, roi_model_1), and the second including edge mid-point co-ordinates (x,y,z) as further additional input features (uia_model_2, roi_model_2), meaning there were ten input edge features. No augmentation was used.

For all models, all other hyper parameters were kept the same, and as similar to the original paper as possible [23]. Both a weighted data sampler and weighted cross-entropy loss function were used, based on the class distribution of growing and stable UIAs (0.7 to growing, 0.3 to stable). Batch normalisation was used with a batch size of 50 meshes and a learning rate of 0.0002. The classification model was trained to predict future growth of the UIA as defined by the clinical definition, whereby output was one of the two classes: growing or stable. All experiments were performed using five-fold cross-validation where the validation splits were made randomly and kept the same for each experiment. The models were trained for a maximum of 200 epochs with validation every 5 epochs and the model with the highest average F1 score for each split was selected. The model was implemented in Python 3.8.5 with Pytorch version 1.8.0 on a NVIDIA TITAN X Pascal (12GB) GPU with CUDA version 11.2.

For final model assessment, we determined the classification accuracy, growth prediction sensitivity and specificity, where the metrics were averaged across all validation splits. A true positive was considered a correctly identified growing UIA, a true negative was a correctly identified stable UIA. Sensitivity and Specificity were determined using these definitions, therefore high sensitivity suggests the model is good at detecting growing UIAs and high specificity suggests the model is good at detecting stable UIAs. We plotted the mean ROC curve and calculated the mean area under the curve (AUC) for each model, as the average over all validation splits for each model.

7.3 Results

Results of the growth prediction models averaged across all validation splits are summarised in Table 7.1. Figure 7.2 shows ROC curves for all of the models. Roi_model_1, using ROI meshes and no edge mid-point co-ordinates, had the highest accuracy (0.761), F1 score (0.681) and specificity (0.883) suggesting it performs optimally for stable aneurysm detection. Uia_model_2, using UIA meshes and including edge mid-point co-ordinates, had the highest sensitivity for growth detection. Overall, both the second models including edge mid-point co-ordinates had higher AUC and sensitivity values but slightly lower accuracy and F1 scores.

Model	Accuracy	F1 score	Sensitivity	Specificity	AUC
uia_model$_1$	0.704 (0.077)	0.617 (0.062)	0.389 (0.048)	0.833 (0.121)	0.620 (0.119)
uia_model$_2$	0.669 (0.061)	0.623 (0.073)	0.573 (0.155)	0.708 (0.075)	0.638 (0.116)
roi_model$_1$	0.761 (0.017)	0.681 (0.021)	0.458 (0.038)	0.883 (0.030)	0.606 (0.056)
roi_model$_2$	0.713 (0.075)	0.650 (0.077)	0.498 (0.090)	0.781 (0.110)	0.622 (0.064)

Table 7.1: Classification metrics for each model. F1 score is the average of F1 score for each class (growing and stable). A true positive was considered a correctly identified growing UIA, a true negative was a correctly identified stable UIA. Sensitivity and Specificity were determined using these definitions. AUC is the area under the mean ROC curve in Figure 7.2. Values are provided as mean (standard deviation) across all validation splits (standard deviation)

7.4 Discussion

In this paper, we demonstrate that a future UIA growth prediction model could be developed using a mesh convolutional neural network, which considers the topology of

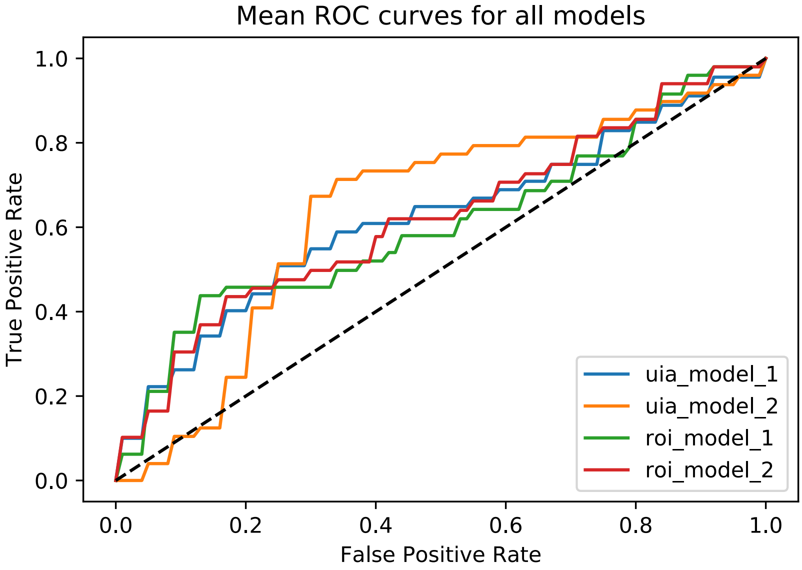


Figure 7.2: ROC curves of all trained models for growth prediction classification. Each line is the mean of the performance across all cross validation splits for each model. The black dotted line indicates a classifier which would give random choice.

UIAs and their parent vasculature. We found that adding edge mid-point co-ordinates as input features to the network increases the AUC and sensitivity of growth prediction but reduces the overall accuracy of the model (uia_model_2, roi_model_2). Using ROI meshes as opposed to UIA meshes alone, improved the accuracy and F1 score of the model but has a decreased AUC for growth prediction. A sensitive growth prediction model should consider using UIA meshes as input and including edge mid-point co-ordinates as input features (uia_model_2).

We found using UIA meshes alone (uia_model_2) improved the AUC relative to using the ROI including parent vessels (roi_model_2). This suggests that it is the topology of the aneurysm surface itself which is high indicative of growth or stability as opposed to UIA configuration relative to parent vessels. This result is also on par with previous studies, where measurements of just the UIA were used for distinguishing between growing and stable UIAs [34, 150, 154]. However, it is also worth noting that using a ROI as opposed to the UIA mesh does not greatly reduce the performance of the method. A ROI mesh, is easier for a clinician to achieve in the clinic as it requires only a simple click of a centre point from which to select the ROI. Whereas, currently the UIA meshes require accurate manual segmentation of the UIA. Therefore, a ROI model may be more useful in the clinic and we prove it could still have adequate performance for UIA growth prediction.

The inclusion of input edge mid-point co-ordinate features increased the AUC and growth prediction sensitivity. We believe this to be because the co-ordinate provides information of the location of the aneurysm to the network. Location is a known predictor of growth [31]. In the original MeshCNN paper [23] it was commented that adding in edge co-ordinates reduced the model performances. They suggest this may be because the additional features, removes the rotation, translation and uniform scaling in-variance of the usual relative geometric input edge features. However, in real-life applications, such as in medical images, the co-ordinates give important information about the location of lesions. Therefore, the addition of these features only appears to improve the performance in this scenario. Further studies could be performed to investigate the use of relative position input features, which would ensure the in-variance to rotation, translation and scaling is kept. Another possibility could be to include position/location, and potentially other known growth predictors, as global features in the final layers of the network.

The models all had a relatively high specificity, suggesting they perform well for detecting stable UIAs. This may be useful in clinic to identify those UIAs which are stable and do not need further investigation. In our study, we had a relatively large class imbalance of only 30% growing UIAs to 70% stable UIAs. Although weighted loss functions and samplers were used, this does not eliminate the class imbalance. In the future, a more balanced dataset, including more growing UIAs could be used. The validation results displayed a large range in sensitivity, and all models performed particularly badly for one validation split. It was also clear, that the model tended to over-fit relatively quickly to the training set. This is due, in part, to the heterogeneous nature of the UIAs and configurations leading to the validation sets being quite different to the training data. This could be improved by including more training and validation data. Furthermore, a larger dataset would allow for independent evaluation on a separate test set.

The ELAPSS growth prediction score was determined to have a c-statistic (AUC) of 0.69 in an external validation study [160]. Our model performed only slightly inferior to this (AUC = 0.64), suggesting that our model has comparable performance to current clinical prediction models. Future studies should consider combining the patient characteristics used in the ELAPSS score, with the aneurysm characteristics used in our model.

Our proposed method did not perform as well as the method using PointNet++ put forward by Bizjak [37] (accuracy = 82%). This may be for a variety of reasons. Firstly, our dataset was imbalanced (30% growing, 70% stable) compared to the dataset they used which included more growing than stable aneurysms. Secondly, our model predicts clinically defined growth, assessed directly on the TOF-MRAs by radiologists. In Bizjak et al. they assess growth visually on the pre-processed 3D meshes. Instead, we propose a model, which can provide prediction for growth as is currently clinically assessed and accepted in the clinic. Future studies should investigate different definitions of growth, and as computer aided tools for UIA diagnosis and assessment continue to be developed and improve, a definition for volumetric growth should be considered [147, 161]. It is difficult to make a comparison of our model performance to the study by Liu et al. [36] as they are predicting aneurysm stability, which included rupture and not just growth. Furthermore, they also include aneurysms all larger than 4 mm and have a much larger dataset. However, future studies could investigate if our mesh based model could also predict rupture/aneurysm instability as well as growth.

In our previous paper [130], we demonstrated the mesh convolutional neural networks could be used for a modality independent UIA detection method. Based on these results, we believe that our growth prediction method could also be modality independent. This would be helpful in the clinic, where UIAs are often assessed or followed-up with different modalities such as CTA or DSA.

7.5 Conclusion

We present a future UIA growth prediction model using a mesh convolutional neural network. We demonstrate that both UIA and ROI meshes can be used as input for such a prediction model, and that edge mid-point co-ordinates improve the growth prediction sensitivity. This model may have potential clinical use as an aid for radiologists assessing potential future UIA growth.

7.6 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.

Chapter 8

Summary and discussion

8.1 Summary

This thesis presents and investigates image analysis and quantitative techniques for the detection and growth risk assessment of unruptured intracranial aneurysms (UIAs). Using such methods, robust and reliable growth assessment of UIAs can be made to aid in treatment decision making. Chapters 2 to 4 in this thesis consider methods for automatic UIA detection and segmentation from angiographic brain scans; specifically Time-of-Flight Magnetic Resonance and Computed Tomography Angiography scans (TOF-MRAs and CTAs). Chapters 5 and 6 investigate UIA volume and morphology and their relationship to UIA growth assessment. Finally in Chapter 7, a geometric deep learning prediction model for UIA growth based on a mesh convolutional neural network is presented.

CHAPTER 2 describes the organisation of an international biomedical image analysis challenge: the Aneurysm Detection And segMentation (ADAM) challenge, as part of MICCAI 2020. This included the release of 113 annotated TOF-MRAs. Teams submitted automatic UIA detection and segmentation methods, which were evaluated on a secret held-out test set. The winning detection method of the ADAM challenge has since been developed into an open-source, self-configuring framework for medical image detection (nnDetection). The challenge remains open as an important benchmark for UIA detection and segmentation methods.

CHAPTER 3 describes a novel feasibility study of an anomaly detection method using a variational autoencoder (VAE) trained on healthy TOF-MRAs. Reconstructed TOF-MRAs with diagnosed aneurysms had a lower Structural Similarity Index Measure (SSIM), than TOF-MRAs of subjects with no aneurysms. SSIM could be a potential metric for anomaly/aneurysm detection. Importantly, the results identified that structure and shape within the scans, and not just intensity, is important for UIA detection.

The UIA detection method in **CHAPTER 4** exploits the fact that the vessel surface of an UIA is different from the surrounding tubular-shaped brain vessels. Vessels were segmented from TOF-MRAs and meshes were fitted to the surface. A mesh convolutional neural network was trained using the labelled vessel meshes, to detect UIAs on the vessel surface. The method is modality-independent, with comparable performance for both TOF-MRAs and CTAs and to voxel-wise detection methods.

Automatic detection and segmentation of UIAs from TOF-MRAs as described in Chapters 2 to 4 allow 3D volume and morphology UIA measurements to be made automatically. Such measures could be used for reliable growth and rupture risk assessment.

The reliability and agreement of UIA growth assessment using both 2D size and 3D volume measurements was studied in **CHAPTER 5**. 3D growth assessment was more reliable than 2D, with smaller interobserver differences, and was more consistent across all UIA locations. The smallest detectable change for 2D growth (1.5 mm) was larger than the current accepted growth definition of 1 mm. This might lead to ambiguity in the current 2D growth definition.

3D UIA quantitative morphology measures, such as flatness, shape index and curvedness were introduced in **CHAPTER 6** and their relationship with UIA growth was investigated. Continuous UIA growth was related to an increase in surface area and flatness, and a decrease in shape index and curvedness. Morphology also change in non-growing (stable) aneurysms, suggesting that non-growing aneurysms could still be unstable. Quantified morphologic change should be considered when assessing UIA growth and rupture risk.

Finally, in **CHAPTER 7**, UIA growth prediction from baseline TOF-MRAs was investigated. Combining the concepts learnt from previous chapters on UIA detection, morphology, and growth, an UIA growth prediction model based on a vessel surface mesh convolutional neural network was developed. The model had comparable prediction performance to patient demographic growth prediction models (ELAPSS).

In conclusion, this thesis provides a complete description of UIA characterisation from TOF-MRAs using computer-aided techniques. The automatic detection and segmentation of UIAs from TOF-MRAs and CTAs allow UIA measurements to be made automatically and reliably. 3D volume and morphology measurements aid in UIA growth assessment and formal UIA growth definitions including these measures should be investigated. As the accuracy of automatic UIA segmentation methods and growth prediction models increase, these will become more commonplace in clinical workflows. This could result in a fully automatic UIA characterisation tool, which determines UIA volume, morphology and growth prediction scores. This would allow complete assessment and prediction of UIA growth, aiding the clinician in the treatment decision and improve patient outcome.

8.2 General Discussion

This thesis presents image analysis techniques for the detection and characterisation of UIAs, as well as assessment and prediction of growth. This could aid medical specialists in making UIA rupture risk assessments to allow informed treatment decisions to be made. First, methods for automatic detection and segmentation of UIAs from angiographic scans (TOF-MRAs and CTAs) were investigated. Second, the use of 3D volumetric and morphology UIA quantification was explored and compared to current growth assessment of UIAs. Finally, a geometric deep learning prediction model for UIA growth was developed based on a mesh convolutional neural network. This thesis covers developments in (geometric) deep learning methods for medical imaging detection and prediction models, and their application in UIA characterisation and growth assessment.

8.2.1 Detection and Segmentation of UIAs

Automatic detection and segmentation of UIAs from angiographic scans (e.g. TOF-MRAs or CTAs) would allow for fully automatic, observer independent quantification of UIAs. Volume and morphology of the UIA can be derived from a UIA segmentation. Up until now and including Chapters 5 and 6, UIA segmentation has been performed by manual annotations. These annotations are time-consuming and made on each slice, as the radiologist scrolls through the scan. The image orientation often makes it difficult to define the UIA neck relative to the parent vessel to segment the UIA. Automatic detection and segmentation techniques (such as those described in Chapters 2 to 4), could remove this manual step and any observer bias and dependence. It would also allow 3D measures such as volume and morphology to be determined automatically and relatively fast. Of course, automatic measures are not without flaws, such as bias from training data and inability to deal with abnormal data. For an automatic technique to be beneficial over manual segmentation, it should have performance comparable to, if not better than, human observers and it should speed up the segmentation procedure. Automatic segmentations will likely have more power when combined with quality assessment by a human observer, and in this way, they would be more likely to be accepted for use by clinicians. Further clinical research studies of automatic UIA segmentation techniques need to be performed before it would be considered suitable for clinical use.

DEEP LEARNING METHODS Supervised deep learning models are commonly used for medical image detection and segmentation problems. They require a large amount of labelled data to train the models. Prior to this thesis, no such labelled database of TOF-MRAs including UIAs was publicly available. In 2020, we organised the Aneurysm Detection And segMentation (ADAM) Challenge as part of the MICCAI conference as described in Chapter 2. As part of this challenge, a large dataset of 113 TOF-MRAs including manually annotated UIAs were released. Such biomedical image analysis challenges are important, not only to find a solution to one particular application (here, UIA detection and segmentation), but also for the development of techniques, frameworks and open-source resources that could be used and applied to different or similar tasks. Furthermore, it provides a benchmark for the field, which allows direct comparison of methods by evaluation on the same dataset with the same evaluation method. All participants that entered the ADAM challenge submitted a deep learning based method and the top method for the segmentation task was based on the nnU-Net framework [84]. The nnU-Net has been developed since the start of my PhD (2018), and is an "out-of-the box tool for state-of-the-art segmentation". It is an open-source, self-configuring deep learning segmentation pipeline, which dynamically adapts to the input, selecting optimal parameters for the network based on the input data. The architecture itself is the same for all applications; the U-net architecture that is known for its optimal performance in medical image segmentation problems [162]. In December 2019, nnU-Net had top-ranking performance on 19 different biomedical image analysis challenges. Now by 2022, nnU-Net has become the benchmarking standard for medical image segmentation methods. Their results suggest that perhaps a 'brute-force', 'one-fits-all' architecture and workflow may be possible and suitable for multiple medical imaging problems. As it develops and continually outperforms other methods, nnU-Net could change the field of image analysis and segmentation problems. However, with the advance of transformer networks [163], CNN-based architectures such as the nnU-Net could soon be out-performed. A transformer model uses the self-attention mechanism that allows the entire image data to be considered and not just filters on image patches as in a CNN. Regardless, the concept of a self-configuring framework is likely to remain and is a good way to make image analysis accessible to everyone. For segmentation tasks, manual experimentation of different architectures, data augmentations, hyper-parameter optimisation and configurations are no longer required. Perhaps in the future, we will see nnU-Net developed into a full application that can be used by someone with any background and no coding ability but just some annotated data. Nevertheless, it is no surprise that nnU-Net was the top-performer for the segmentation task of the ADAM challenge. Leading on from nnU-Net is nnDetection [116], which was the top ranking method for the detection task of the ADAM challenge and comes from the same centre as nnU-Net. nnDetection has since been developed into a full open source framework, similar to that of nnU-Net, but for lesion detection tasks. nnDetection uses a similar self-configuring method that outputs bounding boxes with confidence scores for detection. It has been great to see the use of the ADAM challenge as validation for the nnDetection framework as it develops to become the benchmark medical image detection tool.

As discussed, both implementations of nnU-Net and nnDetection, are anatomy independent and there were few methods submitted to the challenge that used prior anatomical information. Only two methods performed vessel segmentation to aid in their method. This is surprising, as the aneurysms are very small relative to the full 3D TOF-MRAs, and by concentrating on regions around the vessels, it would reduce the imbalance problem. Furthermore, the vessel surface of the aneurysms are likely to be very different from the rest of the brain vessels. This motivated the proposed method in Chapter 4. There is a relatively similar configuration of the brain and brain vessels between subjects. The similarity of healthy control scans, and the imbalanced problem, suggests that an anomaly (or out-of-distribution) detection method could potentially be suitable here, but no such method was submitted to the ADAM challenge. Such an anomaly detection method was what motivated Chapter 3 of this thesis.

In Chapter 3, an anomaly detection method for UIA detection by considering image reconstruction using a variational autoencoder (VAE) was investigated. Anomaly detection is an unsupervised method and only requires normal data for training without annotated lesions. Instead, a model is trained on healthy control data to learn on scans that represent subjects without the disease, or here, subjects without aneurysms. This means that large publicly available datasets can be used, such as the IXI dataset [110], and no time-consuming annotations from radiologists are required. Anomaly detection requires a large training dataset to ensure that all variabilities of possible healthy configurations are included in the training data. The first stage of this process would be to ensure that the model, when trained on healthy data, can reproduce the same image. If subsequently a 'diseased' patient, i.e. a patient with an UIA, is input to the model, reconstruction metrics will be worse in the output 'diseased' image relative to a healthy output image. This is, indeed, what was determined in Chapter 3, using a VAE, where the SSIM (Structural Similarity Index Measure) was found to be lower in reconstructed scans of patients with UIAs, relative to healthy control subject scans. Our study was merely the beginning of an anomaly detection method for UIA detection, and future work would be needed to explore if using SSIM and VAEs could directly detect UIAs. However, an important result was that using a structural loss (SSIM) performed better at identifying patients with UIAs than an intensity-based loss (L2). This confirms that that the shape and structure of the image is more important than the image intensity values to identify aneurysms. This further motivated the work of Chapter 4, assessing only the shape of vessels in the scans using geometric deep learning and vessel surface meshes.

GEOMETRIC DEEP LEARNING In Chapter 4, geometric deep learning with vessel surface meshes for UIA detection was proposed. Voxel-wise methods such as convolutional neural networks (CNNs), like those described in Chapters 2 and 3, operate on Euclidean grid data, such as 2D and 3D images. These methods use the full image and voxel information, thus the intensities in the image is the most important variable. As a result, for medical images, the application of such voxel-wise methods is usually limited to a single imaging modality, the modality of scans on which the model has been trained. Furthermore, they will often struggle to generalise to different scanning acquisitions and protocols, which could have different intensity distributions. Geometric deep learning is a rapidly advancing field of deep learning methods for non-Euclidean data such as graphs, meshes or point clouds [157]. These techniques do not require any resampling of 3D objects to a structured 2D or 3D space, since generally convolutions and pooling layers are performed on the non-Euclidean surface of the 3D object itself. As a result, geometric deep learning techniques depend only on the 3D object surface and can be intensity independent. Meshes and graphs also include connectivity of points on the surface of the object. In medical imaging, geometric methods would allow for modality-independent models and reduce the scan protocol dependence compared to a voxel-wise deep learning model. Geometric deep learning has demonstrated good performance for 3D object classification and segmentation, for example PointNet++ [18] using 3D point clouds and MeshCNN [23] for 3D meshes.

The original MeshCNN includes five relative geometric input edge features that are rotation, translation and scale invariant [23]. In Chapters 4 and 7, it was proposed to include additional input edge features to MeshCNN, namely: shape index, curvedness [119] and the edge midpoint coordinates. These features were specifically chosen for the tasks: UIA detection in Chapter 4, and UIA growth prediction in Chapter 7. Shape index and curvedness are already known to aid in UIA detection [118], and their invariance made them ideal features for MeshCNN. Edge midpoint coordinates were included for the growth prediction model, since location is a known indicator of UIA growth prediction [31]. Our results indicate that model performance can be improved by including specific, predetermined input edge features for a particular purpose. However, input edge features should be carefully chosen for the specific application and the translation, rotation and scale invariance of the resulting model should be considered. Relative features solve the issues with lack of invariance, and in the future, relative features such as a relative distance to an atlas or a template space could be considered.

The original MeshCNN was developed for low-resolution 3D objects, since high resolution medical 3D images result in memory constraints. A large amount of training in the original MeshCNN is restricted to the CPU (central processing unit) because this is where the edge collapsing and bookkeeping of the MeshPool layers is performed. Distributed training and performing all computations on GPU (graphics processing unit) would speed up training times and greatly increase the usability of MeshCNN, allowing for the use of higher resolution meshes and larger batch sizes, which is important in medical imaging problems. As geometric deep learning is increasing in popularity, libraries such as Pytorch Geometric and Pytorch3D are being expanded. These libraries provide potential for the implementation of more efficient MeshCNN pooling and convolutional layers that could reduce the memory associated problems.

The results of Chapters 4 and 7 give promising results for the use of MeshCNN, and more generally geometric deep learning techniques, in the field of medical image analysis. The use of MeshCNN could be extended for other modality-independent vascular imaging problems, such as abdominal aortic aneurysm or coronary artery segmentation. However, there can also be 3D lesion classification and regression problems, where the surface topology of a lesion is important for the outcome. In these situations, I believe that MeshCNN or geometric deep learning could play a large role in developing models based on the surface topology, including information that is perhaps missed in intensity-based voxel-wise deep learning models. In the future, a combination of both intensity-based and geometric-based deep learning models would allow for complete inclusion of surface topology and image voxel information in a model.

8.2.2 UIA Measurement

2D VERSUS 3D MEASUREMENTS Currently, UIAs are assessed in the clinic using 2D length and width measurements made by radiologists on TOF-MRAs or CTAs using digital calipers. The observer selects the orientation of the length and width measurements and adjusts the contrast setting of the scan. This results in a relatively large heterogeneity in length and width measurements between observers, as was described in Chapter 5. However, since length and width measurements are easily performed and interpreted by clinicians, currently UIA growth is still assessed using these 2D measurements. UIA segmentation using the previously described automatic methods (Chapters 2 to 4) or using manual annotations (Chapters 5 and 6) allow the volume of UIA to be determined. Volumetric measurements are independent of orientation, and incorporate the full shape of the UIA in a single measure. In Chapter 5, it was found that manual 3D volume measurements were more reliable when assessing interobserver measurements then 2D measurements. However, volume measurements are currently not used in clinical settings since these require time-consuming manual annotations. As automatic segmentation methods improve, automatic volume measurements could be made. This will lead to more studies using and understanding UIA volume and eventually these volume measurements can become applicable in the clinic.

MORPHOLOGY MEASUREMENTS UIA segmentation also allows morphological measurements of the UIAs to be made, which describe the shape of the aneurysm. Morphological characterisation of lesions in medical images is a well-studied field, and recently the IBSI guidelines [38] have standardised radiomics, including morphology parameters, across medical imaging and radiology. This allows direct comparison between studies and formal, standard morphology definitions, which can be used in clinical research studies, and eventually in the clinic. For this reason, in Chapter 6, the IBSI morphology parameters were used and I believe that future studies should continue to do so. Quantifying UIA shape using morphology measurements is important because irregular UIA shape is a risk factor for growth and rupture [31, 32].

GROWTH ASSESSMENT After diagnosis, UIAs are followed up over time, and considered to grow significantly if either the measured length or width of the UIA increases by more than 1 mm [25]. Aneurysm growth is considered a proxy for aneurysm rupture. If the UIA is determined to be growing, then it is likely to be treated. An important result from Chapter 5 was that the smallest detectable change (SDC) of the interobserver 2D length measurements was determined to be 1.5 mm. This is larger than the currently accepted definition of growth of 1 mm [25], suggesting that that UIAs could be incorrectly identified as growing or non-growing based on interobserver measurement error. This questions the robustness and reliability of the 1 mm definition for 2D growth assessment. Alternatively, a volumetric definition of growth could be considered and in Chapter 5, it was found that volumetric measurements to have a lower interobserver error. Future studies to understand more about volumetric measures and UIA progression would need to be performed before a growth definition including a cut-off for UIA volume increase can be determined.

Volume measurement includes more information than size alone, and in Chapter 6 it was determined that changes in morphological measures of UIAs could also be indicative of aneurysmal instability. It was found that aneurysms considered to be non-growing (stable), could still change in shape. Although these aneurysms are not growing, their change in shape may indicate that they are becoming more unstable, and possibly more likely to rupture. These patients undergo be follow-up, to fully understand the implications of this change in morphology and UIA outcome. (Chapters 5 and 6) indicate that a definition for UIA growth should include morphology measurements as well as volume and/or size measurements.

In Chapter 6, growth was assessed both as a continuous measure (e.g. volume in ml, size in mm) and as a categorical value (growing or non-growing). By assessing growth as a continuous value, all measurements could be included as outcome, which increases the statistical power. Growth measurements that are close to the 1 mm cut-off, for example 0.9 and 1.1 mm, can be very similar, but by means of a dichotomous measure, they are categorised as opposite outcomes. However, a dichotomised definition of growth is important for clearer interpretation of aneurysmal growth and clinical decision-making. Therefore, future studies and clinical research trials should investigate a 3D volume growth definition including morphology changes, for full characterisation of UIAs. A main limitation in Chapters 5 and 6 was that scans were assessed only at two time points, and the time between the baseline and follow-up scans varied. In the future, longitudinal studies with longer follow-up time, and more scan time points should be made to assess the longitudinal growth and change of morphology over time. Furthermore, a percentage volume change could be considered relative to the initial baseline volume to consider relative growth of the UIA.

8.2.3 UIA Growth Prediction Model

Based on morphological parameters, UIA stability and rupture risk prediction models have been developed [36, 155, 156]. Models using handpicked morphological parameters have the benefit that it is known which parameters influence the outcome of the prediction model. However, it also limits the model to only using the specifically chosen parameters to make the outcome decision. The shape of UIAs and their configuration relative to the parent vessels can be diverse. An ideal prediction model, should include all possible UIA shapes and configurations in the most complete sense. After the success with mesh convolutional networks in Chapter 4, a UIA growth prediction model using MeshCNN [23] was developed in Chapter 7 with vessel surface meshes extracted from baseline TOF-MRAs. Comparable performance of the mesh based model to clinical prediction models based on patient demographics and aneurysm characteristics (ELAPSS) was found [31, 160]. For a complete prediction model, both models could be used together or patient demographics should be included in the mesh model. This gives a probability score of growth that could aid clinicians when making a treatment decision of a patient with an UIA.

One limitation of the mesh model is the 'black box' nature of such a prediction model, where it is not clear why the model has chosen such an outcome. As such, explainable AI methods could be used to give understanding of the model's decisions. This could be in the form of highlighting areas on the surface mesh of the UIA and vessel, which had higher activations, and were more important in the decision-making. Using explainable AI, it could be possible to find new growth risk markers. These features could be previously unidentified markers and are common between all the growing UIA meshes or vessel surface meshes. An explainable model with determined risk factors would be more understandable to clinicians and therefore, more trusted and applicable in the clinic. The explainability of AI prediction models is a rapidly developing field [164], and understanding and defining learned features by networks would make prediction models more appealing for clinical use.

8.2.4 Future perspectives

The development of automatic UIA detection methods could speed up the workflow and remove any observer bias in UIA detection. This will be especially important to aid clinicians in the future, especially since preventative screening may become more commonplace [4]. However, before such methods could be used in the clinic, they should have a detection sensitivity greater than clinicians – realistically this should be larger than 90% [10] for larger aneurysms. The current methods require development and improvement before this sensitivity would be achieved. False positives could be easily removed by a radiologist, as long as all UIAs are detected, thus a highly sensitive, less precise method may be beneficial. Small aneurysms (<3 mm) will always be difficult to detect, even for radiologists, due to their similarities to vessel irregularities or infundibulums. However, the sensitivity of a method for detecting larger UIAs is the most important and can still be of clinical value, as these are the UIAs that may have treatment implications.

For growth assessment, more studies into the usefulness and applicability of these morphology measurements in different patient populations using the same morphological parameters should be made. The main limiting factor is the use of manual, time-consuming annotations for 3D segmentation. Automatic segmentations would greatly improve this, removing the observer dependence of the measurements. However, current automatic segmentation methods, have an even lower performance than the interobserver error. Automatic segmentation methods need to be greatly improved, or semi-automatic methods could also be investigated to overcome this hurdle. Such detection and segmentation algorithms followed by morphology measurements could eventually be implemented in radiology software, allowing for standard UIA measurement techniques throughout all institutions. Currently, there is no 3D definition of growth so future studies should consider a 3D volume growth definition including morphology changes, before automatic quantification can have clinical use.

For growth prediction from baseline scans, future studies should investigate more time points during the UIA life cycle. More time-points would allow for a better understanding of the non-linear growth process of UIAs and the pathophysiology of UIA instability. A model could then be developed to predict growth per year, or over a longitudinal time period. Geometric deep learning methods appear to work well for growth prediction, but the current model has not been validated on a large dataset. Explainability of such a growth prediction model could be investigated using explainable AI techniques. Models using handpicked morphology measurements should also be compared. Eventually, a growth prediction model including morphology and/or UIA meshes and patient demographics could be developed.

Growth prediction is considered a proxy for rupture prediction, and this is something that was not studied in this thesis. Future models could consider UIA outcome, including rupture, however, data is limited as few patients have aneurysm rupture during follow-up. Collaborations with other centres studying 3D quantified UIA morphology and aneurysm rupture could aid in developing a 3D volume definition for aneurysm rupture and models for aneurysm instability including growth and rupture.

All of the methods and studies presented in this thesis, from detection to growth assessment and prediction, would be improved by including more patients and also patients from other institutions and populations. In the future, a large collaboration, such as was performed for ELAPSS or PHASES, should be performed including the TOF-MRAs/CTAs and UIA morphological measurements.

8.2.5 Conclusion

This thesis presented methods using image analysis techniques to detect, quantify and characterise UIAs for use in growth assessment. Although, 3D volumetric and morphological measures are not currently accepted for clinical use, the work of this thesis clearly demonstrates their usefulness in aiding reliable growth assessments and prediction alongside known growth risk factors. By assessment in further clinical studies, the added value of UIA volume and morphology for clinical UIA growth and rupture risk assessment can be determined. The growth prediction model presented in this thesis was a feasibility study with promising results. It is important that automatic UIA segmentation methods and growth prediction models have a higher accuracy than intra- or interobserver errors in current clinical assessment. Such a high performance would be required for these models to become trusted and applicable for use in clinical research and workflows. This could result in a fully automatic UIA characterisation tool, which determines volume and morphology of the UIA and provides potential growth prediction scores. This information could aid the treating clinicians greatly in the final treatment decision of the patient and overall improve patient outcome.

Nederlandse samenvatting

Dit proefschrift presenteert beeldanalysetechnieken voor het opsporen en kwantificeren van niet-gebarsten hersenaneurysma's. Met behulp van dergelijke technieken kunnen er betrouwbare metingen van groei worden gemaakt, die worden meegewogen in de beslissing om al dan niet over te gaan op behandeling. Hoofdstukken 2-4 van dit proefschrift bevatten verschillende methoden voor het automatisch detecteren en segmenteren van aneurysma's op basis van hersenscans, specifiek op Time-of-Flight Magnetische Resonantie Angiografie (TOF-MRA) en Computed Tomography Angiografie (CTA) scans. Hoofdstukken 5 en 6 van dit proefschrift behandelen de relatie tussen het volume en de morfologie van aneurysma's en de beoordeling van aneurysmagroei. Tot slot presenteert Hoofdstuk 7 een predictie model voor groei van hersen aneurysma's op basis van TOF-MRA's.

HOOFDSTUK 2 beschrijft de opzet van een internationale wedstrijd voor biomedische beeldanalysetechnieken: de zogeheten Aneurysm Detection en segMentation (ADAM) challenge. Deze wedstrijd werd georganiseerd als onderdeel van het MICCAI 2020 congres. Ten behoeve hiervan zijn er meer dan honderd geannoteerde TOF-MRA's vrijgegeven. Deelnemende teams hebben hun methode voor automatische aneurysmasegmentatie ingediend, welke werd beoordeeld aan de hand van een achtergehouden, geheime dataset. De methode die verkozen is als winnaar van de wedstrijd is sindsdien verder ontwikkeld tot een vrij te gebruiken (open-source), zelfstandig model voor detectie van objecten in medische beelden (nnDetectie). De wedstrijd blijft geopend als een belangrijk ijkpunt voor detectie- en segmentatiemethoden van aneurysma's.

HOOFDSTUK 3 beschrijft een nieuwe studie over de uitvoerbaarheid van een methode voor de detectie van onregelmatigheden in de hersenen. Deze methode maakt gebruik van een kunstmatige intelligentie techniek (Variational Autoencoder) en is getraind op een dataset met TOF-MRA afbeeldingen van gezonde hersenen. De gereconstrueerde TOF-MRA's van patiënten met bevestigde aneurysma's hadden een lagere Structural Similarity Index Measure (SSIM) dan TOF-MRA's van hersenen zonder aneurysma. SSIM is mogelijk te gebruiken als maatstaf voor onregelmatigheden/aneurysma detectie. Een interessante uitkomst van het onderzoek was dat, naast de verschillen in structuur en vorm binnen de scans, ook de verschillen in intensiteit belangrijk zijn voor detectie van aneurysma's.

De methode voor detectie van hersenaneurysma's in **HOOFDSTUK 4**, maakt gebruik van het verschil tussen het vaatoppervlak van een aneurysma en de omliggende bloedvaten. De hersenvaten zijn gesegmenteerd uit de TOF-MRA's en van deze segmentaties zijn driedimensionale (3D) oppervlakte-modellen gemaakt. Een zogeheten mesh neuraal netwerk, specifiek voor deze oppervlakte-modellen, is getraind op de gesegmenteerde hersenvaten om hersenaneurysma's te herkennen op basis van afwijkingen in het vaatoppervlak ten opzichte van gezonde vaten. Deze detectiemethode is onafhankelijk van de beeldvormingsmodaliteit, met vergelijkbare prestaties bij TOF-MRA's en CTA's. Tevens zijn de prestaties vergelijkbaar met detectiemethodes waarbij voxels worden gebruikt als input.

Automatische detectie en segmentatie van hersenaneurysma's uit TOF-MRA's, zoals beschreven in Hoofdstukken 2-4, dragen bij aan het automatisch meten van allerlei eigenschappen van een aneurysma. Deze metingen zijn te gebruiken voor betrouwbare beoordeling van aneurysmagroei en voor risicomodellen voor het barsten van een aneurysma.Deze metingen zijn te gebruiken voor betrouwbare beoordeling van aneurysmagroei en risico op barsten.

De betrouwbaarheid van de beoordeling van groei van een aneurysma met behulp van zowel tweedimensionale (2D) oppervlaktegroei als 3D volumemetingen is onderzocht in **HOOFDSTUK 5**. 3D beoordeling van aneurysmagroei bleek betrouwbaarder dan beoordeling aan de hand van 2D metingen, met kleinere verschillen tussen beoordelaars en consistentere resultaten over alle aneurysma locaties. De kleinst detecteerbare verandering in 2D groei (1,5 mm) is echter groter dan de huidige klinisch geaccepteerde definitie van groei: 1 mm. Dit kan leiden tot onduidelijkheid tussen de kleinst meetbare waarde en de huidige standaard voor 2D groei.

Kwantitatieve morfologische metingen van 3D hersenaneurysma's - zoals vlakheid, vorm en kromming - zijn geïntroduceerd in **HOOFDSTUK 6** en hun relatie tot aneurysmagroei is onderzocht. Continue groei van hersenaneurysma's is gerelateerd aan een toename in oppervlak en vlakheid, en een afname in vorm en kromming. Zelfs in niet-groeiende (stabiele) aneurysma's veranderde de morfologie, wat de suggestie wekt dat niet-groeiende aneurysma's mogelijk ook instabiel kunnen zijn. De kwantificatie van morfologische verandering zou in acht genomen kunnen worden bij het beoordelen van de groei van hersenaneurysma's en risico op barsten.

Tot slot is in **HOOFDSTUK 7** de groeivoorspelling van hersenaneurysma's op basis van TOF-MRA's onderzocht. De technische concepten over hersenaneurysma detectie, morfologie en groei die in de voorgaande hoofdstukken zijn geïntroduceerd, zijn gecombineerd tot een groeipredictie model voor hersenaneurysma's. Het model is gebaseerd op een mesh neuraal netwerk. Het model levert vergelijkbare prestaties ten opzichte van bestaande patiënt-demografische en anerysma gerelateerde modellen voor groeipredictie (ELAPPS). In de toekomst kunnen mesh en/of morfologische modellen gecombineerd worden met bovenbeschreven patient en aneurysma gerelateerde modellen, om zo een meer compleet hersenaneurysma groeipredictie model te creëren.

In conclusie, dit proefschrift levert een compleet overzicht van het opsporen en kwantificeren van hersenaneurysma's uit TOF-MRA's, door middel van computergestuurde technieken. Automatische detectie en segmentatie van hersenaneurysma's uit

TOF-MRA's en CTA's geeft de mogelijkheid voor automatisering van betrouwbare hersenaneurysma metingen. 3D volume en morfologische metingen komen van pas in de beoordeling van groei en vormverandering van hersenaneurysma's. Klinische geaccepteerde definities van groei en aneurysma instabiliteit, gebaseerd op 3D volume en morfologische veranderingen, evenals de waarde van groei predictiemodellen op basis van 3D gekwantificeerde aneurysma morfologie moet worden onderzocht. Deze voorspellingsmodellen krijgen een grotere rol in de klinische praktijk naarmate de betrouwbaarheid en nauwkeurigheid toeneemt. Dit kan resulteren in een volledig automatisch hulpmiddel voor de bepaling van het volume, de morfologie en de groeipredictie van hersenaneurysma's. Dit zou bijdragen aan een complete beoordeling en voorspelling van de groei van hersenaneurysma's, wat behandelaars kan bijstaan in het maken van behandelkeuzes en daarmee het verbeteren van patiëntenzorg.

Bibliography

1. M. H. Vlak, A. Algra, R. Brandenburg, and G. J. Rinkel. “Prevalence of unruptured intracranial aneurysms, with emphasis on sex, age, comorbidity, country, and time period: A systematic review and meta-analysis,” *The Lancet Neurology*, vol. 10 (2011), pp. 626–636 (cited on pp. 10, 19, 22, 57, 77).

2. D. J. Nieuwkamp, L. E. Setz, A. Algra, F. H. Linn, N. K. de Rooij, and G. J. Rinkel. “Changes in case fatality of aneurysmal subarachnoid haemorrhage over time, according to age, sex, and region: a meta-analysis,” *The Lancet Neurology*, vol. 8 (2009), pp. 635–642 (cited on pp. 10, 19).

3. A. S. E. Bor, G. J. Rinkel, J. van Norden, and M. J. Wermer. “Long-term, serial screening for intracranial aneurysms in individuals with a family history of aneurysmal subarachnoid haemorrhage: A cohort study,” *The Lancet Neurology*, vol. 13 (2014), pp. 385–392 (cited on pp. 10, 19).

4. G. J. Rinkel and Y. M. Ruigrok. “Preventive screening for intracranial aneurysms,” *International Journal of Stroke*, vol. 17 (2022), pp. 30–36 (cited on pp. 10, 11, 122).

5. A. M. Algra, A. Lindgren, M. D. Vergouwen, J. P. Greving, I. C. Van Der Schaaf, T. P. Van Doormaal, and G. J. Rinkel. “Procedural Clinical Complications, Case-Fatality Risks, and Risk Factors in Endovascular and Neurosurgical Treatment of Unruptured Intracranial Aneurysms: A Systematic Review and Meta-analysis,” *JAMA Neurology*, vol. 76 (2019), pp. 282–293 (cited on pp. 10, 83, 89, 103).

6. J. P. Greving, M. J. Wermer, R. D. Brown, A. Morita, et al. “Development of the PHASES score for prediction of risk of rupture of intracranial aneurysms: A pooled analysis of six prospective cohort studies,” *The Lancet Neurology*, vol. 13 (2014), pp. 59–66 (cited on pp. 10, 11, 19, 77, 89, 103).

7. L. T. van der Kamp, G. J. E. Rinkel, D. Verbaan, R. van den Berg, et al. “Risk of Rupture After Intracranial Aneurysm Growth,” *JAMA Neurology* (2021) (cited on pp. 11, 13, 89, 103).

8. R. S. Bechan, S. B. van Rooij, M. E. Sprengers, J. P. Peluso, M. Sluzewski, C. B. Majoie, and W. J. van Rooij. “CT angiography versus 3D rotational angiography in patients with subarachnoid hemorrhage,” *Neuroradiology*, vol. 57 (2015), pp. 1239–1246 (cited on pp. 12, 57).

9. Z. L. Yang, Q. Q. Ni, U. J. Schoepf, C. N. De Cecco, et al. “Small Intracranial Aneurysms: Diagnostic Accuracy of CT Angiography,” *Radiology*, vol. 285 (2017), pp. 941–952 (cited on pp. 12, 57).

10. P. M. White, J. M. Wardlaw, and V. Easton. “Can noninvasive imaging accurately depict intracranial aneurysms? A systematic review,” *Radiology*, vol. 217 (2000), pp. 361–370 (cited on pp. 12, 19, 37, 40, 57, 71, 122).

11. R. Cardenes, J. M. Pozo, H. Bogunovic, I. Larrabide, and A. F. Frangi. “Automatic aneurysm neck detection using surface voronoi diagrams,” *IEEE Transactions on Medical Imaging*, vol. 30 (2011), pp. 1863–1876 (cited on pp. 12, 20).

12. C. M. Hentschke, O. Beuing, R. Nickl, and K. D. Tönnies. “Automatic cerebral aneurysm detection in multimodal angiographic images,” *IEEE Nuclear Science Symposium Conference Record* (2012), pp. 3116–3120 (cited on pp. 12, 20).

13. H. Arimura, Q. Li, Y. Korogi, T. Hirai, et al. “Computerized detection of intracranial aneurysms for three-dimensional MR angiography: Feature extraction of small protrusions based on a shape-based difference image technique,” *Medical Physics*, vol. 33 (2006), pp. 394–401 (cited on pp. 12, 20).

14. A. Faron, R. Sijben, N. Teichert, J. Freiherr, M. Wiesmann, and T. Sichtermann. “Deep learning-based detection of intracranial aneurysms in 3D TOF-MRA,” *American Journal of Neuroradiology*, vol. 40 (2019), pp. 25–32 (cited on pp. 12, 20).

15. M. Nishimori, S. Fukumoto, D. Ueda, Y. Shimahara, et al. “Deep Learning for MR Angiography: Automated Detection of Cerebral Aneurysms,” *Radiology*, vol. 290 (2018), pp. 187–194 (cited on pp. 12, 20).

16. A. Park, C. Chute, P. Rajpurkar, J. Lou, et al. “Deep Learning–Assisted Diagnosis of Cerebral Aneurysms Using the HeadXNet Model,” *JAMA Network Open*, vol. 2 (2019), p. e195600 (cited on pp. 12, 20).

17. G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, et al. “A survey on deep learning in medical image analysis.” *Medical image analysis*, vol. 42 (2017), pp. 60–88 (cited on p. 12).

18. C. R. Qi, L. Yi, H. Su, and L. J. Guibas. “PointNet++: Deep hierarchical feature learning on point sets in a metric space,” *Advances in Neural Information Processing Systems*, vol. 2017-Decem (2017), pp. 5100–5109 (cited on pp. 13, 58, 119).

19. Y. Li, R. Bu, M. Sun, W. Wu, X. Di, and B. Chen. “PointCNN: Convolution on X-transformed points,” *Advances in Neural Information Processing Systems*, vol. 2018-Decem (2018), pp. 820–830 (cited on pp. 13, 58).

20. T. N. Kipf and M. Welling. “Semi-Supervised Classification with Graph Convolutional Networks,” (2016) (cited on p. 13).

21. C. Morris, M. Ritzert, M. Fey, W. L. Hamilton, J. E. Lenssen, G. Rattan, and M. Grohe. “Weisfeiler and Leman Go Neural: Higher-Order Graph Neural Networks,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33 (2019), pp. 4602–4609 (cited on p. 13).

22. Y. Feng, Y. Feng, H. You, X. Zhao, and Y. Gao. “MeshNet: Mesh neural network for 3D shape representation,” *33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019* (2019), pp. 8279–8286 (cited on pp. 13, 58).

23. R. Hanocka. “MeshCNN: A network with an edge,” *ACM Transactions on Graphics*, vol. 38 (2019) (cited on pp. 13, 58, 59, 62, 66, 104, 106, 109, 119, 122).

24. H. Koffijberg, E. Buskens, A. Algra, M. J. H. Wermer, and G. J. E. Rinkel. “Growth rates of intracranial aneurysms: exploring constancy,” *Journal of Neurosurgery*, vol. 109 (2008), pp. 176–185 (cited on p. 13).

25. K. A. Hackenberg, A. Algra, R. A. S. Salman, J. Frösen, et al. “Definition and Prioritization of Data Elements for Cohort Studies and Clinical Trials on Patients with Unruptured Intracranial Aneurysms: Proposal of a Multidisciplinary Research Group,” *Neurocritical Care*, vol. 30 (2019), pp. 87–101 (cited on pp. 13, 77, 80, 83, 84, 91, 103, 105, 120, 121).

26. G. Forbes, A. J. Fox, J. Huston, D. O. Wiebers, and J. Torner. “Interobserver variability in angiographic measurement and morphologic characterization of intracranial aneurysms: A report from the International Study of Unruptured Intracranial Aneurysms,” *American Journal of Neuroradiology*, vol. 17 (1996), pp. 1407–1415 (cited on pp. 13, 19, 77, 89).

27. H. J. Kim, D. Y. Yoon, E. S. Kim, H. J. Lee, H. J. Jeon, J. Y. Lee, and B. M. Cho. “Intraobserver and interobserver variability in CT angiography and MR angiography measurements of the size of cerebral aneurysms,” *Neuroradiology*, vol. 59 (2017), pp. 491–497 (cited on pp. 13, 19, 77, 83, 89).

28. J. W. Van Keulen, J. Van Prehn, M. Prokop, F. L. Moll, and J. A. Van Herwaarden. “Potential value of aneurysm sac volume measurements in addition to diameter measurements after endovascular aneurysm repair,” *Journal of Endovascular Therapy*, vol. 16 (2009), pp. 506–513 (cited on p. 13).

29. S.-H. V. Chan, K.-S. A. Wong, Y.-M. P. Woo, K.-Y. Chan, and K.-M. Leung. “Volume Measurement of the Intracranial Aneurysm: A Discussion and Comparison of the Alternatives to Manual Segmentation,” *Journal of Cerebrovascular and Endovascular Neurosurgery*, vol. 16 (2015), p. 358 (cited on p. 13).

30. M. Piotin, P. Gailloud, L. Bidaut, S. Mandai, M. Muster, J. Moret, and D. A. Rüfenacht. “CT angiography, MR angiography and rotational digital subtraction angiography for volumetric assessment of intracranial aneurysms. An experimental study,” *Neuroradiology*, vol. 45 (2003), pp. 404–409 (cited on pp. 13, 83).

31. D. Backes, G. Rinkel, J. Greving, B. K. Velthuis, et al. “ELAPSS score for prediction of risk of growth of unruptured intracranial aneurysms,” *Neurology*, vol. 88 (2017), pp. 1600–1606 (cited on pp. 13, 19, 20, 78, 84, 90, 103, 106, 109, 119, 120, 122).

32. A. E. Lindgren, T. Koivisto, J. Björkman, M. Von Und Zu Fraunberg, K. Helin, J. E. Jääskeläinen, and J. Frösen. "Irregular Shape of Intracranial Aneurysm Indicates Rupture Risk Irrespective of Size in a Population-Based Cohort," *Stroke*, vol. 47 (2016), pp. 1219–1226 (cited on pp. 13, 20, 84, 120).

33. M. L. Raghavan, B. Ma, and R. E. Harbaugh. "Quantified aneurysm shape and rupture risk," *Journal of Neurosurgery*, vol. 102 (2009), pp. 355–362 (cited on pp. 13, 20).

34. E. L. Leemans, B. M. Cornelissen, C. H. Slump, C. B. Majoie, J. R. Cebral, and H. A. Marquering. "Comparing Morphology and Hemodynamics of Stable-versus-Growing and Grown Intracranial Aneurysms," *American Journal of Neuroradiology*, vol. 40 (2019), pp. 2102–2110 (cited on pp. 13, 14, 83, 89, 96, 103, 108).

35. R. D. Millán, L. Dempere-Marco, J. M. Pozo, J. R. Cebral, and A. F. Frangi. "Morphological characterization of intracranial aneurysms using 3-D moment invariants," *IEEE Transactions on Medical Imaging*, vol. 26 (2007), pp. 1270–1282 (cited on pp. 13, 89).

36. Q. Liu, P. Jiang, Y. Jiang, H. Ge, S. Li, H. Jin, and Y. Li. "Prediction of Aneurysm Stability Using a Machine Learning Model Based on PyRadiomics-Derived Morphological Features," *Stroke*, vol. 50 (2019), pp. 2314–2321 (cited on pp. 14, 89, 96, 103, 110, 121).

37. Z. Bizjak, F. Pernus, and Z. Spiclin. "Deep Shape Features for Predicting Future Intracranial Aneurysm Growth," *Frontiers in Physiology*, vol. 12 (2021), pp. 1–10 (cited on pp. 14, 103, 109).

38. A. Zwanenburg, M. Vallières, M. A. Abdalah, H. J. W. L. Aerts, et al. "The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping," *Radiology*, vol. 295 (2020), pp. 328–338 (cited on pp. 14, 89, 91, 96, 120).

39. A. Keedy. "An overview of intracranial aneurysms," *McGill Journal of Medicine*, vol. 9 (2006), pp. 141–146 (cited on pp. 19, 57).

40. D. Backes, M. D. Vergouwen, A. T. Tiel Groenestege, A. S. E. Bor, et al. "PHASES Score for Prediction of Intracranial Aneurysm Growth," *Stroke*, vol. 46 (2015), pp. 1221–1226 (cited on pp. 19, 78, 83, 84, 89, 90).

41. J. M. Wardlaw and P. M. White. "The detection and management of unruptured intracranial aneurysms," *Brain*, vol. 123 (2000), pp. 205–221 (cited on pp. 19, 26).

42. R. D. Brown and J. P. Broderick. "Unruptured intracranial aneurysms: Epidemiology, natural history, management options, and familial screening," *The Lancet Neurology*, vol. 13 (2014), pp. 393–404 (cited on p. 19).

43. D. Nakagawa, Y. Nagahama, B. A. Policeni, M. L. Raghavan, et al. "Accuracy of detecting enlargement of aneurysms using different MRI modalities and measurement protocols," *Journal of Neurosurgery*, vol. 130 (2019), pp. 559–565 (cited on p. 19).

44. A. S. E. Bor, H. Koffijberg, M. J. H. Wermer, and G. J. E. Rinkel. "Optimal screening strategy for familial intracranial aneurysms: A cost-effectiveness analysis," *Neurology*, vol. 74 (2010), pp. 1671–1679 (cited on p. 19).

45. A. Flahault, D. Trystram, F. Nataf, M. Fouchard, B. Knebelmann, J.-P. Grünfeld, and D. Joly. "Screening for intracranial aneurysms in autosomal dominant polycystic kidney disease is cost-effective," *Kidney International*, vol. 93 (2018), pp. 716–726 (cited on p. 19).

46. E. M. Hopmans, Y. M. Ruigrok, A. S. Bor, G. J. Rinkel, and H. Koffijberg. "A cost-effectiveness analysis of screening for intracranial aneurysms in persons with one first-degree relative with subarachnoid haemorrhage," *European Stroke Journal*, vol. 1 (2016), pp. 320–329 (cited on p. 19).

47. A. Lane, P. Vivian, and A. Coulthard. "Magnetic resonance angiography or digital subtraction catheter angiography for follow-up of coiled aneurysms: Do we need both?" *Journal of Medical Imaging and Radiation Oncology*, vol. 59 (2015), pp. 163–169 (cited on p. 19).

48. P. M. White, E. M. Teasdale, J. M. Wardlaw, and V. Easton. "Intracranial aneurysms: CT angiography and MR angiography for detection - Prospective blinded comparison in a large patient cohort," *Radiology*, vol. 219 (2001), pp. 739–749 (cited on pp. 19, 83).

49. L. HaiFeng, X. YongSheng, X. YangQin, D. Yu, W. ShuaiWen, L. XingRu, and L. JunQiang. "Diagnostic value of 3D time-of-flight magnetic resonance angiography for detecting intracranial aneurysm: a meta-analysis," *Neuroradiology*, vol. 59 (2017), pp. 1083–1092 (cited on pp. 19, 37).

50. K. H. Wrede, T. Matsushige, S. L. Goericke, B. Chen, et al. "Non-enhanced magnetic resonance imaging of unruptured intracranial aneurysms at 7 Tesla: Comparison with digital subtraction angiography," *European Radiology*, vol. 27 (2017), pp. 354–364 (cited on p. 19).

51. W. Ji, A. Liu, X. Lv, H. Kang, et al. "Risk score for neurological complications after endovascular treatment of unruptured intracranial aneurysms," *Stroke*, vol. 47 (2016), pp. 971–978 (cited on p. 20).

52. Z. Bizjak, B. Likar, F. Pernuš, and Z. Špiclin. "Modality agnostic intracranial aneurysm detection through supervised vascular surface classification," *Medical Imaging 2021: Com-puter-Aided Diagnosis*, edited by K. Drukker and M. A. Mazurowski. SPIE, 2021, p. 21 (cited on pp. 20, 58, 71).

53. K. Lawonn, M. Meuschke, R. Wickenhöfer, B. Preim, and K. Hildebrandt. "A geometric optimization approach for the detection and segmentation of multiple aneurysms," *Computer Graphics Forum*, vol. 38 (2019), pp. 413–425 (cited on p. 20).

54. H. Duan, Y. Huang, L. Liu, H. Dai, L. Chen, and L. Zhou. "Automatic detection on intracranial aneurysm from digital subtraction angiography with cascade convolutional neural networks," *Biomedical engineering online*, vol. 18 (2019), p. 110 (cited on p. 20).

55. N. Sulayman, M. Al-Mawaldi, and Q. Kanafani. "Semi-automatic detection and segmentation algorithm of saccular aneurysms in 2D cerebral DSA images," *Egyptian Journal of Radiology and Nuclear Medicine* (2016) (cited on p. 20).

56. K. Timmins, E. Bennink, I. v. d. Schaaf, B. Velthuis, Y. Ruigrok, and H. Kuijf. "Intracranial Aneurysm Detection and Segmentation Challenge," (2020) (cited on p. 20).

57. L. Maier-Hein, A. Reinke, M. Kozubek, A. L. Martel, et al. “BIAS: Transparent reporting of biomedical image analysis challenges,” *Medical Image Analysis*, vol. 66 (2020), p. 101796 (cited on pp. 20, 21).

58. D. Merkel. “Docker: lightweight Linux containers for consistent development and deployment,” *Linux J.*, vol. 2014 (2014) (cited on p. 20).

59. H. J. Kuijf, A. Casamitjana, D. L. Collins, M. Dadar, et al. “Standardized Assessment of Automatic Segmentation of White Matter Hyperintensities and Results of the WMH Segmentation Challenge,” *IEEE transactions on medical imaging*, vol. 38 (2019), pp. 2556–2568 (cited on pp. 21, 38).

60. A. M. Mendrik, K. L. Vincken, H. J. Kuijf, M. Breeuwer, et al. “MRBrainS Challenge: Online Evaluation Framework for Brain Image Segmentation in 3T MRI Scans,” *Computational Intelligence and Neuroscience*, vol. 2015 (2015) (cited on p. 21).

61. S. Klein, M. Staring, K. Murphy, M. Viergever, and J. Pluim. “elastix: A Toolbox for Intensity-Based Medical Image Registration,” *IEEE Transactions on Medical Imaging*, vol. 29 (2010), pp. 196–205 (cited on pp. 23, 61).

62. N. J. Tustison, P. A. Cook, and J. C. Gee. “N4Itk,” vol. 29 (2011), pp. 1310–1320 (cited on pp. 22, 23, 48, 61).

63. A. A. Taha and A. Hanbury. “Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool,” *BMC Medical Imaging*, vol. 15 (2015) (cited on p. 24).

64. S. K. Warfield, K. H. Zou, and W. M. Wells. “Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation,” *IEEE Transactions on Medical Imaging*, vol. 23 (2004), pp. 903–921 (cited on p. 25).

65. M. G. KENDALL. “A NEW MEASURE OF RANK CORRELATION,” *Biometrika*, vol. 30 (1938), pp. 81–93 (cited on p. 26).

66. W. McKinney. “Data Structures for Statistical Computing in Python,” 2010, pp. 56–61 (cited on p. 26).

67. P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, et al. “SciPy 1.0: fundamental algorithms for scientific computing in Python,” *Nature Methods*, vol. 17 (2020), pp. 261–272 (cited on p. 26).

68. Michael Waskom and the seaborn development team. “mwaskom/seaborn,” 2020. DOI 10.5281/zenodo.592845 (cited on p. 26).

69. R. Vallat. “Pingouin: statistics in Python,” *Journal of Open Source Software*, vol. 3 (2018), p. 1026 (cited on pp. 26, 80).

70. P. F. Jaeger, S. A. A. Kohl, S. Bickelhaupt, F. Isensee, T. A. Kuder, H.-P. Schlemmer, and K. H. Maier-Hein. “Retina U-Net: Embarrassingly Simple Exploitation of Segmentation Supervision for Medical Object Detection,” (2018) (cited on pp. 27, 38).

71. S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia. “Path Aggregation Network for Instance Segmentation,” *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2018), pp. 8759–8768 (cited on pp. 27, 38).

72. M. Baumgartner, P. F. Jaeger, F. Isensee, and K. H. Maier-Hein. "Retina U-Net for Aneurysm Detection in MR Images," 2020 (cited on p. 27).

73. O. Ronneberger, P. Fischer, and T. Brox. "U-net: Convolutional networks for biomedical image segmentation," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9351 (2015), pp. 234–241 (cited on pp. 27, 28, 32, 62).

74. P. Mouches and N. D. Forkert. "A statistical atlas of cerebral arteries generated using multi-center MRA datasets from healthy subjects," *Scientific data*, vol. 6 (2019), p. 29 (cited on p. 27).

75. T. Di Noto, G. Marie, S. Tourbier, Y. Alemán-Gómez, et al. "Weak labels and anatomical knowledge: making deep learning practical for intracranial aneurysm detection in TOF-MRA," (2021) (cited on p. 27).

76. A. Sekuboyina, M. Rempfler, J. Kukačka, G. Tetteh, A. Valentinitsch, J. S. Kirschke, and B. H. Menze. "Btrfly Net: Vertebrae Labelling with Energy-Based Adversarial Learning of Local Spine Prior," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11073 LNCS (2018), pp. 649–657 (cited on p. 27).

77. S. Shit, D. Shah, A. Fava Sanches, and B. H. Menze. "2D TriWingedNet for 3D Intracranial Aneurysm Segmentation," 2020 (cited on p. 27).

78. M. Tan and Q. V. Le. "EfficientNet: Rethinking model scaling for convolutional neural networks," *36th International Conference on Machine Learning, ICML 2019*, vol. 2019-June (2019), pp. 10691–10700 (cited on p. 27).

79. L. Fei-Fei, J. Deng, and K. Li. "ImageNet: Constructing a large-scale image database," *Journal of Vision*, vol. 9 (2010), pp. 1037–1037 (cited on p. 27).

80. T. Jerman, F. Pernuš, B. Likar, and Ž. Špiclin. "Beyond Frangi: an improved multiscale vesselness filter," edited by S. Ourselin and M. A. Styner. 2015, 94132A (cited on p. 28).

81. C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. Jorge Cardoso. "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10553 LNCS (2017), pp. 240–248 (cited on pp. 28, 39).

82. H. Kervadec, J. Bouchtiba, C. Desrosiers, E. Granger, J. Dolz, and I. Ben Ayed. "Boundary loss for highly unbalanced segmentation," *Medical Image Analysis*, vol. 67 (2021), pp. 1–21 (cited on pp. 28, 39).

83. U. Walińska, M. Klimont, M. Kraft, D. Pieczyński, M. Mikołajczak, and M. Pawlak. "Inteneural at Aneurysm Detection And segMentation Challenge," 2020 (cited on p. 28).

84. F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein. "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18 (2021), pp. 203–211 (cited on pp. 28, 32, 38, 57, 116).

85. Y. Wu and K. He. “Group Normalization,” (2018) (cited on p. 28).

86. Y. Yang, Y. Lin, Y. Li, D. Wei, K. Ma, X. Yang, and Y. Zheng. “Automatic Aneurysm Segmenattion via 3D U-Net Ensemble,” 2020 (cited on p. 28).

87. L. Berrada, A. Zisserman, and M. P. Kumar. “Smooth Loss Functions for Deep Top-k Classification,” (2018) (cited on pp. 28, 32).

88. J. Ma. “Loss Ensembles for Extremely Imbalanced Segmentation,” (2020) (cited on p. 28).

89. J. Ma and X. An. “Loss Ensembles for Intracranial Aneurysm Segmentation: An Embarrassingly Simple Method,” 2020 (cited on p. 28).

90. J. Ma, J. Chen, M. Ng, R. Huang, et al. “Loss odyssey in medical image segmentation,” *Medical Image Analysis*, vol. 71 (2021) (cited on p. 28).

91. A. Hilbert, V. I. Madai, E. M. Akay, O. U. Aydin, et al. “BRAVE-NET: Fully Automated Arterial Brain Vessel Segmentation in Patients With Cerebrovascular Disease,” *Frontiers in Artificial Intelligence*, vol. 3 (2020) (cited on p. 28).

92. R. De Feo, J. Kaiponen, and J. Tohka. “Aneurysm Segmentation in the ADAM Challenge: KUBIAC team,” 2020 (cited on p. 28).

93. Z. Zhou, V. Sodha, M. M. R. Siddiquee, R. Feng, N. Tajbakhsh, M. B. Gotway, and J. Liang. “Models Genesis: Generic Autodidactic Models for 3D Medical Image Analysis,” (2019) (cited on p. 28).

94. M. Hu, X. Feng, and L. Yin. “Unruptured Intracranial Aneurysm Segmentation from TOF-MRA Images Using Cascaded 3D Convolutional Neural Networks,” 2020 (cited on p. 28).

95. H. Li, G. Jiang, J. Zhang, R. Wang, Z. Wang, W. S. Zheng, and B. Menze. “Fully convolutional network ensembles for white matter hyperintensities segmentation in MR images,” *NeuroImage*, vol. 183 (2018), pp. 650–665 (cited on p. 28).

96. T. Loehr, H. Li, and B. Menze. “A Multi-View Approach for Automatic Segmentation of Intracranial Aneurysms from Time of Flight MRAs,” 2020 (cited on p. 28).

97. T. Nakao, S. Hanaoka, Y. Nomura, I. Sato, et al. “Deep neural network-based computer-assisted detection of cerebral aneurysms in MR angiography,” *Journal of Magnetic Resonance Imaging*, vol. 47 (2018), pp. 948–953 (cited on p. 28).

98. S. Rjiba, T. Urruty, P. Bourdon, C. Maloigne-Fernandez, R. Delepaul, and R. Guillevin. “AneurysmNet: Deep Neural Network-based Segmentation of Aneurysms in 3D-MR Angiography,” 2020 (cited on p. 28).

99. S. Wang and C. Manning. “Fast Dropout Training,” *ICML*, 2013 (cited on p. 28).

100. C. Giroud and F. Dubost. “Probabilistic Segmentation and Detection of Aneurysm from brain MRA with an Ensemble of 3D Convolutional Neural Networks and Monte Carlo Dropout,” 2020 (cited on p. 28).

101. A. Krizhevsky, I. Sutskever, and G. E. Hinton. "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60 (2017), pp. 84–90 (cited on p. 38).

102. K. Kamnitsas, W. Bai, E. Ferrante, S. McDonagh, et al. "Ensembles of multiple models and architectures for robust brain tumour segmentation," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10670 LNCS (2018), pp. 450–462 (cited on p. 38).

103. A. S. E. Bor, B. K. Velthuis, C. B. Majoie, and G. J. Rinkel. "Configuration of intracranial arteries and development of aneurysms: a follow-up study." *Neurology*, vol. 70 (2008), pp. 700–705 (cited on p. 47).

104. K. N. Kayembe, M. Sasahara, and F. Hazama. "Cerebral aneurysms and variations in the circle of Willis." *Stroke*, vol. 15 (1984), pp. 846–850 (cited on p. 47).

105. N. G. Campeau and J. Huston. "Vascular Disorders-Magnetic Resonance Angiography: Brain Vessels," *Neuroimaging Clinics of North America*, vol. 22 (2012), pp. 207–233 (cited on p. 47).

106. D. P. Kingma and M. Welling. "Auto-Encoding Variational Bayes," (2013), pp. 1–14 (cited on p. 47).

107. C. Baur, B. Wiestler, S. Albarqouni, and N. Navab. "Deep autoencoding models for unsupervised anomaly segmentation in brain MR images," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11383 LNCS (2019), pp. 161–169 (cited on p. 47).

108. D. Zimmerer, F. Isensee, J. Petersen, S. Kohl, and K. Maier-Hein. "Unsupervised Anomaly Localization Using Variational Auto-Encoders," *Informatik aktuell*, 2019, pp. 289–297 (cited on p. 47).

109. Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli. "Image Quality Assessment: From Error Visibility to Structural Similarity," *IEEE Transactions on Image Processing*, vol. 13 (2004), pp. 600–612 (cited on p. 47).

110. E. GR/S21533/02. "IXI Dataset – Information eXtraction from Images," (cited on pp. 48, 118).

111. K. M. Timmins, I. C. van der Schaaf, E. Bennink, Y. M. Ruigrok, et al. "Comparing methods of detecting and segmenting unruptured intracranial aneurysms on TOF-MRAS: The ADAM challenge," *NeuroImage*, vol. 238 (2021), p. 118216 (cited on pp. 48, 57, 60, 63, 67, 71, 97).

112. N. Otsu. "A Threshold Selection Method from Gray-Level Histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9 (1979), pp. 62–66 (cited on p. 48).

113. V. de Vos, K. Timmins, I. van der Schaaf, Y. Ruigrok, B. Velthuis, and H. J. Kuijf. "Automatic Cerebral Vessel Extraction in TOF-MRA using Deep Learning," (2021), p. 83 (cited on pp. 49, 105, 106).

114. K. A. Hackenberg, D. Hänggi, and N. Etminan. "Unruptured Intracranial Aneurysms," *Stroke*, vol. 49 (2018), pp. 2268–2275 (cited on p. 57).

115. N. Etminan and G. J. Rinkel. "Unruptured intracranial aneurysms: development, rupture and preventive management," *Nature Reviews Neurology*, vol. 12 (2016), pp. 699–713 (cited on pp. 57, 89).

116. M. Baumgartner, P. F. Jäger, F. Isensee, and K. H. Maier-Hein. "nnDetection: A Self-configuring Method for Medical Object Detection," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12905 LNCS of (2021), pp. 530–539 (cited on pp. 57, 117).

117. N. Hayashi, Y. Masutani, T. Masumoto, H. Mori, et al. "Feasibility of a Curvature-based Enhanced Display System for Detecting Cerebral Aneurysms in MR Angiography," *Magnetic Resonance in Medical Sciences*, vol. 2 (2005), pp. 29–36 (cited on pp. 57, 62, 70).

118. H. Prasetya, T. L. Mengko, O. S. Santoso, and H. Zakaria. "Detection method of cerebral aneurysm based on curvature analysis from 3D medical images," *Proceedings - International Conference on Instrumentation, Communication, Information Technology and Biomedical Engineering 2011, ICICI-BME 2011* (2011), pp. 141–144 (cited on pp. 57, 62, 70, 119).

119. J. Koenderink and A. Doorn. "Surface shape and curvature scales," *Image and Vision Computing*, vol. 10 (1992), pp. 557–564 (cited on pp. 57, 62, 91, 105, 119).

120. Z. Bizjak, B. Likar, F. Pernus, and Z. Spiclin. "Vascular Surface Segmentation for Intracranial Aneurysm Isolation and Quantification," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12266 LNCS (2020), pp. 128–137 (cited on pp. 58, 60).

121. X. Yang, D. Xia, T. Kin, and T. Igarashi. "INTRA: 3D intracranial aneurysm dataset for deep learning," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2020), pp. 2653–2663 (cited on p. 58).

122. L. Schneider, A. Niemann, O. Beuing, B. Preim, and S. Saalfeld. "MedmeshCNN - Enabling meshcnn for medical surface models," *Computer Methods and Programs in Biomedicine*, vol. 210 (2021), pp. 1–7 (cited on pp. 58, 59, 72).

123. C. R. Qi, H. Su, K. Mo, and L. J. Guibas. "PointNet: Deep learning on point sets for 3D classification and segmentation," *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, vol. 2017-Janua (2017), pp. 77–85 (cited on p. 58).

124. J. Li, B. M. Chen, and G. H. Lee. "SO-Net: Self-Organizing Network for Point Cloud Analysis," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, 2018, pp. 9397–9406 (cited on p. 58).

125. X. Yang, D. Xia, T. Kin, and T. Igarashi. "A Two-step Surface-based 3D Deep Learning Pipeline for Segmentation of Intracranial Aneurysms," vol. 1 (2020) (cited on pp. 58, 60, 72).

126. H. Maron, M. Galun, N. Aigerman, M. Trope, et al. "Convolutional neural networks on surfaces via seamless toric covers," *ACM Transactions on Graphics*, vol. 36 (2017), pp. 1–10 (cited on p. 59).

127. Y. Wang, S. Asafi, O. van Kaick, H. Zhang, D. Cohen-Or, and B. Chen. "Active co-analysis of a set of shapes," *ACM Transactions on Graphics*, vol. 31 (2012), pp. 1–10 (cited on p. 59).

128. V. Vosylius, A. Wang, C. Waters, A. Zakharov, et al. "Geometric Deep Learning for Post-Menstrual Age Prediction Based on the Neonatal White Matter Cortical Surface," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12443 LNCS (2020), pp. 174–186 (cited on pp. 59, 60, 104).

129. A. Mandado. "Surface class segmentation in CAD models with MeshCNN," (cited on pp. 59, 62).

130. K. M. Timmins, I. C. Schaaf, I. N. Vos, Y. M. Ruigrok, B. K. Velthuis, and H. J. Kuijf. "Deep Learning with Vessel Surface Meshes for Intracranial Aneurysm Detection," *Medical Imaging 2022: Computer-Aided Diagnosis*, SPIE, 2022, p. 110 (cited on pp. 60, 104, 106, 110).

131. V. de Vos, K. Timmins, I. van der Schaaf, Y. Ruigrok, B. Velthuis, and H. J. Kuijf. "Automatic Cerebral Vessel Extraction in TOF-MRA using Deep Learning," *Medical Imaging 2021: Image Processing*, SPIE, 2021, p. 83 (cited on p. 61).

132. F. Ritter, T. Boskamp, A. Homeyer, H. Laue, M. Schwier, F. Link, and H.-O. Peitgen. "Medical Image Analysis," *IEEE Pulse*, vol. 2 (2011), pp. 60–70 (cited on p. 61).

133. Z. Zhang, Q. Liu, and Y. Wang. "Road Extraction by Deep Residual U-Net," *IEEE Geoscience and Remote Sensing Letters*, vol. 15 (2018), pp. 749–753 (cited on p. 62).

134. M. Fey and J. E. Lenssen. "Fast Graph Representation Learning with PyTorch Geometric," (2019) (cited on p. 72).

135. J. Johnson, N. Ravi, J. Reizenstein, D. Novotny, S. Tulsiani, C. Lassner, and S. Branson. "Accelerating 3D deep learning with PyTorch3D," *SIGGRAPH Asia 2020 Courses*, New York, NY, USA: ACM, 2020, pp. 1–1 (cited on p. 72).

136. S. Tominari, A. Morita, T. Ishibashi, T. Yamazaki, et al. "Prediction model for 3-year rupture risk of unruptured cerebral aneurysms in Japanese patients," *Annals of Neurology*, vol. 77 (2015), pp. 1050–1059 (cited on p. 77).

137. J. P. Villablanca, G. R. Duckwiler, R. Jahan, S. Tateshima, et al. "Natural history of asymptomatic unruptured cerebral aneurysms evaluated at CT angiography: Growth and rupture incidence and correlation with epidemiologic risk factors," *Radiology*, vol. 269 (2013), pp. 258–265 (cited on p. 77).

138. H. Takao, Y. Murayama, T. Ishibashi, T. Saguchi, et al. "Comparing Accuracy of Cerebral Aneurysm Size Measurements From Three Routine Investigations: Computed Tomography, Magnetic Resonance Imaging, and Digital Subtraction Angiography," *Neurologia medico-chirurgica*, vol. 50 (2010), pp. 893–899 (cited on p. 77).

139. B. G. Thompson, R. D. Brown, S. Amin-Hanjani, J. P. Broderick, et al. "Guidelines for the Management of Patients With Unruptured Intracranial Aneurysms: A Guideline for Healthcare Professionals From the American Heart Association/American Stroke Association," vol. 46 (2015), pp. 2368–2400 (cited on p. 77).

140. A. Malhotra, X. Wu, D. Gandhi, P. Sanelli, and C. C. Matouk. "Management of Small, Unruptured Intracranial Aneurysms," *World Neurosurgery*, vol. 135 (2020), pp. 379–380 (cited on pp. 77, 85).

141. T. K. Koo and M. Y. Li. "A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research," *Journal of Chiropractic Medicine*, vol. 15 (2016), pp. 155–163 (cited on p. 80).

142. J. Kottner, S. Brorson, A. Donner, and B. J. Gajewski. "Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed . J Clin Epidemiol Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed," vol. 64 (2015), pp. 96–106 (cited on p. 80).

143. H. C. de Vet, C. B. Terwee, D. L. Knol, and L. M. Bouter. "When to use agreement versus reliability measures," *Journal of Clinical Epidemiology*, vol. 59 (2006), pp. 1033–1039 (cited on p. 80).

144. B. Mine, M. Pezzullo, G. Roque, P. David, T. Metens, and B. Lubicz. "Detection and characterization of unruptured intracranial aneurysms: Comparison of 3T MRA and DSA," *Journal of Neuroradiology*, vol. 42 (2015), pp. 162–168 (cited on p. 83).

145. S.-H. V. Chan, K.-S. A. Wong, Y.-M. P. Woo, K.-Y. Chan, and K.-M. Leung. "Volume Measurement of the Intracranial Aneurysm: A Discussion and Comparison of the Alternatives to Manual Segmentation," *Journal of Cerebrovascular and Endovascular Neurosurgery*, vol. 16 (2014), p. 358 (cited on pp. 83, 89).

146. F. D'Argento, A. Pedicelli, C. Ciardi, E. Leone, et al. "Intra- and inter-observer variability in intracranial aneurysm segmentation: comparison between CT angiography (semi-automated segmentation software stroke VCAR) and digital subtraction angiography (3D rotational angiography)," *Radiologia Medica* (2020) (cited on p. 83).

147. K. Timmins, H. Kuijf, M. Vergouwen, M. Otten, Y. Ruigrok, B. Velthuis, and I. van der Schaaf. "Reliability and Agreement of 2D and 3D Measurements on MRAs for Growth Assessment of Unruptured Intracranial Aneurysms," *American Journal of Neuroradiology*, vol. 42 (2021), pp. 1598–1603 (cited on pp. 89, 97, 103, 110).

148. A. S. E. Bor, A. T. Groenestege, K. G. Terbrugge, R. Agid, B. K. Velthuis, G. J. Rinkel, and M. J. Wermer. "Clinical, radiological, and flow-related risk factors for growth of untreated, unruptured intracranial aneurysms," *Stroke*, vol. 46 (2015), pp. 42–48 (cited on p. 89).

149. B. Ma, R. E. Harbaugh, and M. L. Raghavan. "Three-dimensional geometrical characterization of cerebral aneurysms," *Annals of Biomedical Engineering*, vol. 32 (2004), pp. 264–273 (cited on p. 89).

150. E. L. Leemans, B. M. W. Cornelissen, M. Said, R. van den Berg, C. H. Slump, H. A. Marquering, and C. B. L. M. Majoie. “Intracranial aneurysm growth: consistency of morphological changes,” *Neurosurgical Focus*, vol. 47 (2019), E5 (cited on pp. 89, 96, 103, 108).

151. W. E. Lorensen and H. E. Cline. “Marching Cubes: A High Resolution 3D Surface Construction Algorithm,” *ACM siggraph computer graphics*, vol. 21 (1987), pp. 163–169 (cited on pp. 90, 105).

152. B. Iglewicz and D. C. Hoaglin. “How to Detect and Handle Outliers: Vol 16,” edited by E. F. Mykytka. Vol. 16 (The ASQC Basic References in Quality Control: Statistical Techniques, 1993), p. 12 (cited on p. 92).

153. D. G. Altman and P. Royston. “The cost of dichotomising continuous variables,” *BMJ*, vol. 332 (2006), p. 1080.1 (cited on p. 97).

154. K. Timmins, H. Kuijf, M. Vergouwen, Y. Ruigrok, B. Velthuis, and I. van der Schaaf. “Relationship between 3D Morphologic Change and 2D and 3D Growth of Unruptured Intracranial Aneurysms,” *American Journal of Neuroradiology*, vol. 43 (2022), pp. 416–421 (cited on pp. 103, 108).

155. H. C. Kim, J. K. Rhim, J. H. Ahn, J. J. Park, et al. “Machine Learning Application for Rupture Risk Assessment in Small-Sized Intracranial Aneurysm,” *Journal of Clinical Medicine*, vol. 8 (2019), p. 683 (cited on pp. 103, 121).

156. J. Liu, Y. Chen, L. Lan, B. Lin, et al. “Prediction of rupture risk in anterior communicating artery aneurysms with a feed-forward artificial neural network,” *European Radiology*, vol. 28 (2018), pp. 3268–3275 (cited on pp. 103, 121).

157. W. Cao, Z. Yan, Z. He, and Z. He. “A Comprehensive Survey on Geometric Deep Learning,” *IEEE Access*, vol. 8 (2020), pp. 35929–35949 (cited on pp. 104, 118).

158. L. Schneider, A. Niemann, O. Beuing, B. Preim, and S. Saalfeld. “MedmeshCNN - Enabling meshcnn for medical surface models,” *Computer Methods and Programs in Biomedicine*, vol. 210 (2021), p. 106372 (cited on p. 104).

159. K. M. Timmins, I. C. van der Schaaf, I. N. Vos, Y. M. Ruigrok, B. K. Velthuis, and H. J. Kuijf. “Geometric Deep Learning using Vascular Surface Meshes for Modality-Independent Unruptured Intracranial Aneurysm Detection,” *Under Review: IEEE Transitions in Medical Imaging* (2022) (cited on p. 105).

160. M. Sánchez van Kammen, J. P. Greving, S. Kuroda, D. Kashiwazaki, et al. “External Validation of the ELAPSS Score for Prediction of Unruptured Intracranial Aneurysm Growth Risk,” *Journal of Stroke*, vol. 21 (2019), pp. 340–346 (cited on pp. 109, 122).

161. X. Liu, H. Haraldsson, Y. Wang, E. Kao, et al. “A Volumetric Metric for Monitoring Intracranial Aneurysms: Repeatability and Growth Criteria in a Longitudinal MR Imaging Study,” *American Journal of Neuroradiology* (2021) (cited on p. 110).

162. F. Isensee, P. F. Jäger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein. “Automated Design of Deep Learning Methods for Biomedical Image Segmentation,” (2019) (cited on p. 117).

163. F. Shamshad, S. Khan, S. W. Zamir, M. H. Khan, M. Hayat, F. S. Khan, and H. Fu. “Transformers in Medical Imaging: A Survey,” (2022) (cited on p. 117).

164. B. H. van der Velden, H. J. Kuijf, K. G. Gilhuijs, and M. A. Viergever. “Explainable artificial intelligence (XAI) in deep learning-based medical image analysis,” *Medical Image Analysis*, vol. 79 (2022), p. 102470 (cited on p. 122).

Acknowledgments

Although the front of this thesis has my name on it, it is definitely not a work that I made entirely on my own. So many people have played a large part in my PhD, and in this section I would like to extend my thanks to you all for supporting me along the way. When I started this journey, I was an insecure, wee Scottish physicist, in a foreign country and a bit confused about my place, career or what I was really doing. I would like to say that after the last four years, that has changed but perhaps I have just become more comfortable in accepting who I am, and that you can never really know everything!!

Firstly and foremost, I would like to thank my promotor and supervisors who have supported me throughout my PhD.

Birgitta, I was very happy to have been able to complete my PhD under your knowledgeable and experienced guidance. Even with your busy daily life, you managed to always find time for me and provide helpful input and viewpoints. Thanks for all the dinners and being so easy to talk to, including our amiable discussions about British tea and Scottish dancing!

Irene, I really valued your input and supervision throughout my PhD trajectory. You made me focus my research to be more clinically applicable, keeping me grounded and without that it would not be the thesis that it now is. I enjoyed the challenges of trying to explain deep learning and AI to you, or describing shapes using apples and pears. You were always patient with me as I got caught up in technicalities, and made me consider things from other perspectives.

Hugo, I'm not sure one paragraph is enough to explain how much I am grateful for your support, this thesis would not have been possible without you! From day 1, you believed in me and gradually I began to believe in myself. Your passion for image analysis is contagious, igniting my enthusiasm and making me strive to know and explore more. I miss our weekly 9 am Wednesday meetings, where our conversations included everything from experiment results and journal papers to discussions about cycling and travelling. A lot changed during the trajectory of my PhD, including of course, a global pandemic and you having kids, but still you always found time for me both professionally and personally - even coming to my wedding in Scotland, and finally getting to have one conference together in Singapore right at the end of my PhD! I feel lucky to have had a supervisor I could be so honest and open with and our conversations are something I truly valued. I am eternally grateful for your continued support, and hope that we continue to stay in touch as we both embark on new journeys.

Members of the committee, Prof. dr. J Hendrikse, Prof. dr. G.J.E Rinkel, Prof. dr. J.P.W. Pluim, Prof. dr. W.J.Niessen, Prof. dr. C.H.Slump, thank you for your time and efforts in reading and evaluating this thesis. Prof. dr. Max Viergever, thanks for the support during my PhD within ISI and for agreeing to be Rector for my defence. I am grateful that you are able to be part of my PhD story.

In my journey before my PhD, I was lucky to have had many fantastic academic opportunities and be inspired by many great scientists. Thanks to my masters supervisor Dr Ewan Eadie, for introducing me to the field of medical physics and supporting me in applying for a PhD. Thanks as well to Dr Bruce Sinclair, who ignited my real enthusiasm for physics and science in St Andrews, and inspired me to use those skills to make a difference.

I would like thank everyone at ISI for being so accommodating for the last four years and being my home from home. Wilbert, I am incredibly grateful that you 'adopted' me in ISI from day 1, where I felt a true family member and I am grateful for all of the opportunities and resources that ISI provided me with. Gerard, I cannot thank you enough for your patience with me during many IT failures and computer lock-outs and your guidance and support with the servers and the ADAM challenge set-up. Edwin, thank-you for your knowledge and support with organising the ADAM challenge, it would not have been possible without you. Jaco, thanks for all the support with the 7T scanner and always managing to fix things when they went wrong! Maria, Anna and Renee, your endless help during my PhD was invaluable. Especially Maria, you never failed to make me laugh or smile on a given day. Thanks for making Q, such a happy and cheerful place to work.

Of course, my time at ISI was made the best by the past and present colleagues that I was lucky to be able to spend time with, have fun with and learn from throughout my PhD. Special shout out to these guys: Bas: You never fail to make me laugh and take life a little less seriously. Thanks for all the fun!, Erik: I guess now I never will catch you on your Estafette run leg time!, Ishaan: I loved our technical discussions, and you always left me with inspiration and ideas. I have learnt from our trip to Singapore, that there is some food that is even too spicy for you!, Julia: Your crazy, infectious personality made the days fun at ISI and I really missed you when you left. I'm glad we've managed to stay in touch and I know that we will continue to make memories!, Marielle: Even at the end of my PhD, I still can't say "ui" properly...thanks for the fun chats and being such a good officemate! Mark: I will miss being able just to drop in your office for a chat and all the Dutch knowledge I learnt from you!, Mateusz: Tush, you are one of the smartest guys I know, and I learnt so much from you. Your endless patience and energy (for deep learning or dancing) and is something I truly respect! Max MD: Your sense of humour led to there never being a dull day on the OIO steeg, Ruurd: I always thought of you as my crazy brother, keep on adventuring! Sam: You taught me a lot at the start of my PhD, from GIT to Mendeley tips, all which aided me to get to this point - Thanks for being there! Saeed: In the time I have known you, you have become a confident researcher and person, and I am proud of you!, Sanne: You inspire me as a strong woman, and I'm grateful for our honest chats, Steffen: Thanks for all the honest and genuine conversations. Your work-life balance helped me learn to keep mine too!

There are so many people, both past and present, from ISI and the UMCU that had a positive influence on my time during my PhD. Alberto, Cyrano, Fenghua, Giulia, Hui Shan, Hui, Jiggy, Jorg, Kees, Mathijs, Matthijs, Majd, Mike, Myrthe, Nick, Nils, Rens, Rick B, Ryanne and Tessa thanks for all the fun memories, coffee breaks and sharing your knowledge and experience. Special thanks to Tessa for helping with the Dutch translation of the summary of this thesis!

Rick, we spent many hours at the 7T scanner together, mainly trying to figure out what went wrong. Thanks for the PhD chats and companionship! Maarten, thanks for always being so interested and I enjoyed explaining things to you and our discussions together. I know that you are capable to doing some really great things!

Iris, my dear paranymph, you started midway through my PhD during COVID but somehow, even through teams calls, you just really understood me as a person and I felt a bond with you from your day 1. I really appreciated all the honest and open conversations, from which I always come away positive and inspired. The overlap of our personalities and interests continues to surprise me! I am incredibly grateful to have been able to get to know you better and to be able to call you a good friend. My only regret is that I wish you started earlier and that we could have written more papers and gone to conferences together! But I leave you, safe in the knowledge you will complete an excellent PhD!

Bea and Nadieh, I know I should have only two paranymphs but deciding between you was like choosing a favourite child – impossible. You've both been there for me since the very start, thanks for all the supportive and inspirational chats, dinners, dancing and memories! I feel incredibly lucky to have made such good friends as you during my PhD. Bea, my dear fellow physicist. Your optimism and enthusiasm are contagious, you genuinely light up the room when you come in. Thank you for always listening and being there, with no judgement and your supportive ways. Nadieh, my dear Dutchie. Our deep chats are something I truly appreciate and enjoy, I would come to your office at the end of the day and leave a long time later with new viewpoints and thoughts. I admire your ambition and strength and I only hope to be half as brave as you one day!

There were a lot of other people without whose support, I could not have completed this PhD. My friends and family, who kept me grounded and calm (most of the time) throughout my journey.

Julia and Ollie, I was so lucky to have my best friend move to the same country during my PhD! You were a wee piece of home and family in the Netherlands, especially during those tough COVID months. Thanks for the support, and for keeping my weekends full of many happy times, adventures and cycles.

Scott and Elspeth, our never ending silly messages kept morale high throughout my PhD journey. I hope that I have done you proud, even if I never quite made it to Prof. Kim.

All my other friends from home and uni, you know who you are: you were never more than a message or call away to support me and I am incredibly grateful for all of your friendships. Special shout-out to Kate, Lauren and Louisa for our crazy online circuit classes that motivated me through each week in COVID times!

My in-laws, Trish and Raymond, and the entire Milne/Stewart family, thanks for all your support. Raymond, thanks always for your interest in my work and wanting to learn more about it. Natalie, your regular calls with Lachlan made COVID life that little bit brighter!

My sisters, Danielle, Natasha and Davina, you all continually inspire me in different ways to lead a happy life doing something you love. I am grateful for all our conversations and proud to have such strong women as sisters and role models.

Mum and Dad, thank you always for your support in my crazy ambitions and ideas. I'm not sure you ever anticipated that I would give up a professional career to go back and do a PhD in a different country. But I'm sure that now, you probably agree that it was one of the best decisions I made. Since I started enjoying Physics at school, you have always encouraged me. I remember when Dad and I took apart the washing machine to see how it worked (we couldn't get it back together again!) or when Mum took me to a windfarm so I could see how they made electricity. Dad, special thanks for putting up with me arguing with you about how best to solve maths problems. Thank you both for the encouragement and support that inspired me to question the world around us, and has allowed me to pursue and complete this PhD.

Finally, I save the biggest thank-you to last and that would be to my incredibly patient, supportive husband and best COVID officemate, Cameron. This book reflects only a small part of my PhD, but I know that you saw the full rollercoaster journey that took to producing it. You believed in me, when I didn't believe in myself. You picked me up on those tough days, and you celebrated with me on the good days. You listened to my never-ending rants and my interesting self-discussions about concepts I'm not sure you ever quite fully understood. You always tried to help me whilst I mulled over problems, even deriving your own amazing Cameron algorithm (which already existed). Your patience with me, my self-doubt and ability to calm me down and see clearly, brings out the best in me and allowed me to be able to produce this thesis. It really would not have been possible without you and your continued support. You keep me sane and happy through all the craziness that life has thrown at us! And perhaps the most important point, that I will now be the Dr. Milne between us. I can't wait to get our first letter through the post addressed as that!!

Publications

Journal publications

K. M. Timmins, I. C. van der Schaaf, E. Bennink, Y. M. Ruigrok, et al. "Comparing methods of detecting and segmenting unruptured intracranial aneurysms on TOF-MRAS: The ADAM challenge," *NeuroImage*, vol. 238 (2021), p. 118216.

K. Timmins, H. Kuijf, M. Vergouwen, M. Otten, Y. Ruigrok, B. Velthuis, and I. van der Schaaf. "Reliability and Agreement of 2D and 3D Measurements on MRAs for Growth Assessment of Unruptured Intracranial Aneurysms," *American Journal of Neuroradiology*, vol. 42 (2021), pp. 1598–1603.

K. Timmins, H. Kuijf, M. Vergouwen, Y. Ruigrok, B. Velthuis, and I. van der Schaaf. "Relationship between 3D Morphologic Change and 2D and 3D Growth of Unruptured Intracranial Aneurysms," *American Journal of Neuroradiology*, vol. 43 (2022), pp. 416–421.

K. M. Timmins, I. C. van der Schaaf, I. N. Vos, Y. M. Ruigrok, B. K. Velthuis, and H. J. Kuijf. "Geometric Deep Learning using Vascular Surface Meshes for Modality-Independent Unruptured Intracranial Aneurysm Detection," *Under Review: IEEE Transitions in Medical Imaging* (2022).

R. Tuijl, **K. Timmins**, H. Kuijf, Y. Ruigrok, B. Velthuis, J. Zwanenburg, and I. van der Schaaf. "Relating Hemodynamic and Morphologic Measurements in Unruptured Intracranial Aneurysms using 7T 4D Flow MRI," *In preparation* (2022).

R. Tuijl, **K. Timmins**, Y. Ruigrok, B. Velthuis, P. Van Ooij, J. Zwanenburg, and I. van der Schaaf. "Wall shear stress and velocity pulsatility in the parent artery of an unruptured intracranial aneurysm - a 7T 4D flow MRI study," *In preparation* (2022).

M. Kamphuis, **K. Timmins**, H. Kuijf, G. Rinkel, M. Vergouwen, and I. van der Schaaf. "3D Morphologic Change of Intracranial Aneurysms Before and Around Rupture," *In preparation* (2022).

I. Vos, **K. Timmins**, Y. Ruigrok, B. Velthuis, and H. Kuijf. "Improving Automated Intracranial Artery Labeling using Atlas-based Features in Graph Convolutional Nets," *In preparation* (2022).

Conference proceedings

K. Timmins, I. van der Schaaf, Y. Ruigrok, B. Velthuis, and H. J. Kuijf. "Variational Autoencoders with a Structural Similarity Loss in Time of Flight MRAs," *Medical Imaging 2021: Image Processing*, SPIE, 2021, p. 115.

V. de Vos, **K. Timmins**, I. van der Schaaf, Y. Ruigrok, B. Velthuis, and H. J. Kuijf. "Automatic Cerebral Vessel Extraction in TOF-MRA using Deep Learning," *Medical Imaging 2021: Image Processing*, SPIE, 2021, p. 83.

K. M. Timmins, I. C. Schaaf, I. N. Vos, Y. M. Ruigrok, B. K. Velthuis, and H. J. Kuijf. "Deep Learning with Vessel Surface Meshes for Intracranial Aneurysm Detection," *Medical Imaging 2022: Computer-Aided Diagnosis*, SPIE, 2022, p. 110.

I. N. Vos, Y. M. Ruigrok, **K. M. Timmins**, B. K. Velthuis, and H. J. Kuijf. "Improving Automated Intracranial Artery Labeling using Atlas-based Features in Graph Convolutional Nets," *Medical Imaging 2022: Image Processing*, SPIE, 2022, p. 45.

K. Timmins, M. Kamphuis, I. Vos, B. Velthuis, I. van der Schaaf, and H. Kuijf. "Future Unruptured Intracranial Aneurysm Growth Prediction using Mesh Convolutional Neural Networks," *MICCAI Conference Proceedings 2022: tda4biomedicalimaging workshop (in press) *joint last author*,

Conference abstracts

R. Tuijl, Y. Ruigrok, **K. Timmins**, B. Velthuis, J. Zwanenburg, and I. van der Schaaf. "Wall shear stress and velocity pulsatility in the parent artery of an unruptured intracranial aneurysm - a 7T 4D flow MRI study," *ISMRM 2022*,

Biography

Kimberley Timmins was born on 10th August 1993 in Lancaster, England, before moving to Scotland with her family at the age of 4. She undertook an MPhys (Hons) Physics degree at the University of St Andrews, Scotland. During her studies she undertook a variety of diverse internships including a research project at Fermilab, Illinois working with the South Telescope team. She performed her final year masters project in collaboration with Ninewells hospital, Dundee characterising light meters for use in daylight photodynamic therapy. After her studies, she worked for two years as a trainee Patent Attorney in Electronics at HGF Limited, Glasgow. Realising that the legal lingo and life was not for her, she made a tough (but wise!) decision to leave the professional world and go back to research and academia. This led her to the Netherlands, to start a PhD in Medical Imaging at the Image Sciences Institute, UMC Utrecht. Her PhD project involved using quantitative image analysis and deep learning to aid in the understanding of the development, diagnosis and outcome of intracranial aneurysms and cerebrovascular diseases. The results of which are shown in this thesis. Kimberley now works as an Research Engineer for Canon Medical Research Europe in Edinburgh, Scotland.