Zurich University of Applied Sciences

Department Life Sciences and Facility Management

Institute of Computational Life Sciences

MASTER THESIS

Comparative Evaluation of Deep Neural Networks for Intracranial Aneurysm Segmentation

Author:
Ekaterina Golubeva

Supervisors:
Dr. Norman Juchler
Dr. Stefan Glüge
Prof. Dr. Thomas Ott

Submitted on
July 26, 2024

Study program:
Applied Computational Life Sciences, M.Sc.

Imprint

Project:	Master Thesis
Title:	Comparative Evaluation of Deep Neural Networks for Intracranial Aneurysm Segmentation
Author:	Ekaterina Golubeva
Date:	July 26, 2024
Keywords:	deep learning, computer vision, neuroimaging, intracranial aneurysm segmentation
Copyright:	Zurich University of Applied Sciences

Study program:
Applied Computational Life Sciences, M.Sc.
Zurich University of Applied Sciences

Supervisor 1:
Dr. Norman Juchler
Zurich University of Applied Sciences
Email: juch@zhaw.ch

Supervisor 2:
Dr. Stefan Glüge
Zurich University of Applied Sciences
Email: glue@zhaw.ch

Supervisor 3:
Prof. Dr. Thomas Ott
Zurich University of Applied Sciences
Email: ottt@zhaw.ch

Abstract

Background: Intracranial Aneurysms (IAs) are sac-like pathological dilatations of cerebral arteries with a global prevalence of 3-5%. Although they are usually asymptomatic, IAs pose an increased risk of Subarachnoid Hemorrhage (SAH). Detection in routine angiographic imaging is challenging due to the overwhelming amount of information radiologists must process.

Purpose: This thesis assesses the use of Deep Learning (DL) methods for the automatic detection and segmentation of intracranial aneurysms in 3D medical imaging. It discusses the evaluation of the no-new-Net (nnU-Net) segmentation model on Computed Tomography (CT) and Magnetic Resonance (MR) angiographies. The ultimate aim is to assist clinicians in identifying unruptured IAs, enabling future quantitative risk analysis and potentially aiding in the development of a Computer-Assisted Diagnosis (CAD) tool to improve early detection in clinical settings.

Methods: This study used Computed Tomography Angiography (CTA) (n=1186) and Magnetic Resonance Angiography (MRA) (n=284) datasets, including NIfTI images with IAs and corresponding segmentations. The nnU-Net was trained from scratch on both datasets and validated through 5 fold cross-fold validation. The model's performance was evaluated at the voxel and aneurysm level.

Results: The nnU-Net demonstrated superior voxel-wise performance on the CTA dataset, achieving high precision (71.0%), recall (71.0%), and DSC (0.70). On the MRA dataset, the model showed lower performance, with precision (46.0%), recall (33.0%), and DSC (0.38). Aneurysm-wise performance was also higher on the CT data compared to other models, varying depending on the criteria chosen for defining true positives.

Conclusion: The nnU-Net model exhibited superior performance in the segmentation and detection of IAs in larger CT datasets with voxel-wise labels compared to other segmentation models. However, it encountered difficulties when processing MR data, primarily due to the smaller size and the weaker labels of the available training dataset. This study highlights the necessity for larger datasets across different modalities to enhance performance and facilitate comparisons. The creation of benchmark datasets would facilitate easier comparison of models. It is crucial to adopt standardized guidelines to ensure consistent evaluation metrics, including definitions of aneurysm size, choices of metrics, and criteria for true positives. It is recommended that the nnU-Net architecture, particularly the foreground sampling strategy for small instances, be customized. The exploration of alternative models, such as vessel-attention models that leverage anatomical knowledge, may also improve performance.

Zusammenfassung

Hintergrund: Intrakranielle Aneurysmen (IA) sind pathologische, sackförmige Erweiterungen der Hirnarterien mit einer weltweiten Prävalenz von 3-5%. Obwohl sie in der Regel asymptomatisch bleiben, stellen IAen ein erhöhtes Risiko für SAH dar. Die frühzeitige Diagnose von IAen in der angiographischen Routinebildgebung ist aufgrund der überwältigenden Menge an Informationen, die Radiologen verarbeiten müssen, eine Herausforderung.

Zweck: In dieser Arbeit wird der Einsatz von Deep-Learning-Methoden für die automatische Erkennung und Segmentierung von intrakraniellen Aneurysmen in der medizinischen 3D-Bildgebung untersucht. Es wird die Evaluierung des nnU-Net-Segmentierungsmodells auf CT- und MR-Angiographien diskutiert. Ziel ist es, Kliniker bei der Identifizierung nicht rupturierter IA zu unterstützen, eine zukünftige quantitative Risikoanalyse zu ermöglichen und die Entwicklung eines CAD-Tools zur Verbesserung der Früherkennung im klinischen Umfeld zu unterstützen.

Methoden: Für diese Studie wurden CTA- und MRA-Datensätze verwendet, die NIfTI-Bilder mit IAs und entsprechenden Segmentierungen enthalten. Das nnU-Net wurde auf beiden Datensätzen trainiert, durch 5-fache Kreuzvalidierung validiert, und die Modell-Performanz auf Voxel- und Aneurysmenebene bewertet.

Ergebnisse: Das nnU-Net wies auf dem CTA-Datensatz eine sehr gute voxelweise Leistung auf und erreichte eine hohe *precision* (0.71), *recall* (0.71) und DSC (0.70). Für den MRA-Datensatz zeigte das Modell eine geringere Leistung mit *precision* (0.46), recall (0.33) und einem DSC (0.38). Auch bei den CT-Daten war die zielgerichtete Leistung im Vergleich zu den anderen Modellen höher, wobei sie je nach den für die Definition wahrer Positivwerte gewählten Kriterien variierte.

Schlussfolgerung: Das nnU-Net-Modell schnitt bei der Segmentierung und Detektion von IAs in grösseren CT-Datensätzen mit voxelweiser Beschriftung im Vergleich zu anderen Segmentierungsmodellen sehr gut ab. Die Methode stiess bei der Verarbeitung von MR-Daten jedoch auf Schwierigkeiten, was in erster Linie auf die geringeren Grösse und die schwächeren Segmentierungen des Trainingdatensatezs zurückzuführen ist. Diese Studie unterstreicht die Notwendigkeit grösserer Datensätze für verschiedene Modalitäten, um die Leistung zu verbessern und Vergleiche zu erleichtern. Die Erstellung von Benchmark-Datensätzen würde den Vergleich von Modellen erleichtern. Es ist von Bedeutung, standardisierte Richtlinien zu verabschieden, um einheitliche Bewertungsmetriken zu gewährleisten, einschliesslich der Definitionen der Aneurysmengrösse, der Auswahl der Metriken und der Kriterien für echte positive Ergebnisse. Es wird empfohlen, die nnU-Net-Architektur, insbesondere die Vordergrund- Sampling-Strategie für kleine Instanzen, anzupassen. Die Erforschung alternativer Modelle, wie z. B. Gefässauffälligkeitsmodelle, die anatomisches Wissen nutzen, kann die Leistung ebenfalls verbessern.

Acknowledgements

I would like to express my deepest gratitude to my supervisors, Dr. Norman Juchler and Dr. Stefan Glüge, for their continous support and guidance. Their patience and reassurance when I encountered challenges made me feel safe and confident. They taught me scientific rigor and inspired me to strive for excellence in my research. Their presence, availability, and reliability were invaluable throughout this journey. Thank you for continuously pushing me to improve and for being dedicated mentors.

I am also grateful to Fabio Musio for his advice and for generously sharing his expertise with the nnU-Net model.

I extend my thanks to Professor Thomas Ott, who provided crucial guidance during my humble beginnings in image processing with Python.

I am grateful to my classmates, particularly Mengxiao Wang and Preethi Bhavani, for their camaraderie and for the mutual support we shared in overcoming academic challenges. Their friendship made the journey more enjoyable.

Furthermore, I would like to extend my gratitude to the MIC-DKFZ (Division of Medical Image Computing, German Cancer Research Center) GitHub team, whose prompt and helpful responses to my numerous inquiries regarding the model were consistently appreciated.

Finally, I am grateful to my mother and my close friend Syakira Ramli for their support and encouragement during challenging periods.

Contents

Acronyms

CAD Computer-Assisted Diagnosis. iii, iv, 2

CI Confidence Interval. 9, 10, 16

CoW Circle of Willis. 1

CT Computed Tomography. iii, iv, 11, 16, 21

CTA Computed Tomography Angiography. iii, iv, vi, 2, 3

DL Deep Learning. 1

DSC Dice Similarity Coefficient. 11, 16, 20–22

FN False Negative. 17, 19, 20

FP False Positive. 17, 19, 20

HD Hausdorff Distance. 11, 12, 16

IA Intracranial Aneurysm. iii, iv, 1–4, 21, 25

IoU Intersection over Union. 11, 14, 15, 19–23

mm milimeters. 7, 11, 17, 20

MR Magnetic Resonance. iii, iv, 11, 20, 21

MRA Magnetic Resonance Angiography. iii, iv, 2–4

nnU-Net no-new-Net. iii–v, 6–12, 16–18, 20, 21, 23–25

ROI Region of Interest. 22

SAH Subarachnoid Hemorrhage. iii, iv, 1

SGD Stochastic gradient descent. 9

TOF Time-of-Flight. 4

TP True Positive. 17–20

TS TotalSegmentator. 8, 9

Chapter 1

Introduction

1.1 Intracranial Aneurysms

IAs are sac-like pathological dilatations of cerebral arteries resulting from weakened vessel walls. With a global prevalence of approximately 3-5% [1, 2], these anomalies, while often asymptomatic, carry the potential for life-threatening complications, including SAH. Notably, intracranial aneurysms are frequently situated within the circle of Willis (CoW), particularly in cases featuring anatomical anomalies or variations [3, 4].

Aneurysms are frequently encountered incidentally during angiographic imaging of the brain. Professional radiologists can reliably identify aneurysms when deliberately seeking them. However, amidst a information abundance in medical imaging and the immediate clinical priorities of patients, these abnormalities might go unnoticed.

Treatment for an unruptured aneurysm typically necessitates invasive neurosurgical intervention, with the risk of rupture being relatively low [5]. Hence, it is advisable to refrain from intervention in cases where aneurysms are not dangerous. One of the ongoing medical pursuits involves identifying distinguishing criteria to differentiate between dangerous aneurysms and benign ones.

1.2 Problem Statement

This master's thesis investigates the automatic detection and segmentation of intracranial aneurysms in 3D medical imaging using Deep Learning (DL) methods. Automated detection of aneurysms can assist clinicians in identifying unruptured IAs in routine medical imaging. Furthermore, the identification of the location and spatial extent of aneurysms is valuable for the quantitative analysis of risk factors in large patient cohorts.

This study was conducted in the context of the Deep Brain Vessel Profiler (DBVP) research project, a joint venture between the ZHAW and the UZH, and funded by the Digitalization Initiative of the Zurich Higher Education Institutions (DIZH) [6]. The broader goal of the DBVP project is to investigate the extent to which the anatomical variability of the major arteries that form the Circle of Willis (CoW) can have an impact on the prevalence and outcome of cerebrovascular diseases such as IAs.

In this research context, the tool developed herein will facilitate the automatic localization of aneurysms within extensive bio-banks – a prerequisite for large-scale studies utilizing 3D medical imaging. In practice, this research may culminate in the creation of a CAD tool [2], expediting the early detection of aneurysms during routine angiographic imaging by radiologists.

1.3 Related work

This section presents a review of relevant studies conducted between 2018 and 2023, with a focus on detection and segmentation of intracranial aneurysms in 3D CTA or MRA images using AI models.

The following Table 1.1 was originally presented in [7] and has been updated to reflect the current state of the art as of 2023.

Ref.	Modality	Task(s)	#Pats	#IAs	Model	Input	Labels	Use anat.	Multi-site	Key metric
[8]	MRA	Detection	450	508	CNN	2D MIP patches	Voxel-wise	Yes	No	Sens. 94.2%
[9]	MRA	Detection	1271	1477	ResNet	2D patches	Unknown	No	Yes	Sens. 92%
[10]	MRA	Detection	302	336	RCNN	2D MIP patches	Voxel-wise	No	No	Acc.98.8%
[11]	MRA	Detection (via segm.)	85	115	DeepMedic	3D patches	Voxel-wise	Yes	No	Sens. 96%
[12]	CTA	Segm. + CAD	662	358	HeadXNet	3D patches	Voxel-wise	Yes	No	Sens. 95%
[13]	DSA	Detection	281	261	2D CNN	2D DSA images	Boxes	Yes	No	Acc. 93.5%
[14]	DSA	Detection	240	187	2D CNN	2D DSA images	ROI circle	No	No	Sens. 79%
[15]	MRA	Detection	744	761	3D ResNet	3D patches	Voxel-wise	Yes	Yes	Sens. 87.1%
[16]	CTA	Detection + Segm.	1177	1099	3D UNET	3D patches	Voxel-wise	Yes	Yes	Sens. 97.3%
[17]	CTA	Detection	1068	1337	ResNet	3D patches	Unknown	No	Yes	Sens.97.5%
[18]	CTA	Detection	311	352	RCNN	2D NP images	Unknown	No	Yes	Sens. 91.8%
[19]	CTA	Detection + Segm.	1186	1363	GLIA-Net	3D patches	Voxel-wise	Yes	No	Sens. 72.9 %
[20]	DSA	Detection + Segm.	451	485	3D UNET	3D DSA volumes	Voxel-wise	Yes	No	Sens. 98.6%
[21]	MRA	Detection	254	N/A	nnDetection	3D patches	Voxel-wise	No	No	Sens. 64%
[22]	MRA	Detection	284	198	3D UNET	3D patches	Weak	Yes	Yes	Sens. 68%
[23]	MRA	Segm.	104	114	PointNet++, SO-Net	2D patches	Voxel-wise	No	No	Sens. 80%

TABLE 1.1: Summary of papers that use deep learning models to tackle automated brain aneurysm detection/segmentation. Use anatomy: Use anatomical information, whether the method uses some sort of anatomical prior knowledge during training, patch sampling or inference. MRA: Magnetic Resonance Angiography, CTA: Computed Tomography Angiography, DSA: Digital Subtraction Angiography, Pats: Patients [7]

The foundational papers for this study are [19] and [7]. The datasets provided with these papers, along with the models they describe, GLIA-Net and 3D-UNet, were used as primary benchmarks to assess the performance of the state-of-the-art segmentation model, nn-UNet.

It is evident that U-Net-based architectures are the predominant approach for the given task. A recent article employed a novel approach utilizing Vessel Attention (VA-UNet), further enhancing the base model [24]. VA-Unet was designed for vessel segmentation, integrating attention mechanisms to improve feature extraction and the segmentation of intricate and fine structures such as blood vessels.

Chapter 2

Methodology

2.1 Datasets

Two types of data modalities and datasets were utilized in this study: a CTA dataset and an MRA dataset, both of which included intracranial aneurysm NIfTI images and their corresponding segmentations.

2.1.1 CTA Dataset

The CTA dataset, named **Large IA Segmentation dataset**, was sourced from the open repository Zenodo [25]. The current section provides a summarized version of the detailed dataset characteristics described in [19].

It was originally collected from six institutions in China (Guizhou Provincial People's Hospital, Affiliated Hospital of Zunyi Medical University, Tongren Municipal People's Hospital, Xingyi Municipal People's Hospital, The Second People's Hospital of Guiyang, The First People's Hospital of Zunyi).

The images in question contain the head region of the patients, with some also including the neck or heart region. The dataset includes both non-ruptured cases and ruptured cases. All patients were positioned in the same position during the examination. The CTA images were annotated by 5 clinicians and reviewed by two experienced radiologists. The identified IAs were manually segmented on each slice using the open-source annotation software ITK-SNAP [26]. The voxel-wise annotation was used as the ground truth standard both in training and evaluation. However, it should be noted that there might be some bias or noise, given the indistinct boundaries of IAs in CTA images and inter-observer variability. Overall, this may result in inconsistencies in labeling standards.

The entire internal dataset was divided into two subsets: 1,186 cases for training and 152 cases for testing. The internal test set comprises 50 negative cases (no IAs present). Negative cases were excluded from the training set to avoid exacerbating the data imbalance, given that the IAs are small in the brain. Furthermore, it was verified that the positive CTA images were distributed evenly across different institutions, ages, and genders in both the internal training and test sets. The dataset statistics are described in the Table 2.1 and the distribution of aneurysm sizes present in the train and test sets is shown in the Figure 2.1.

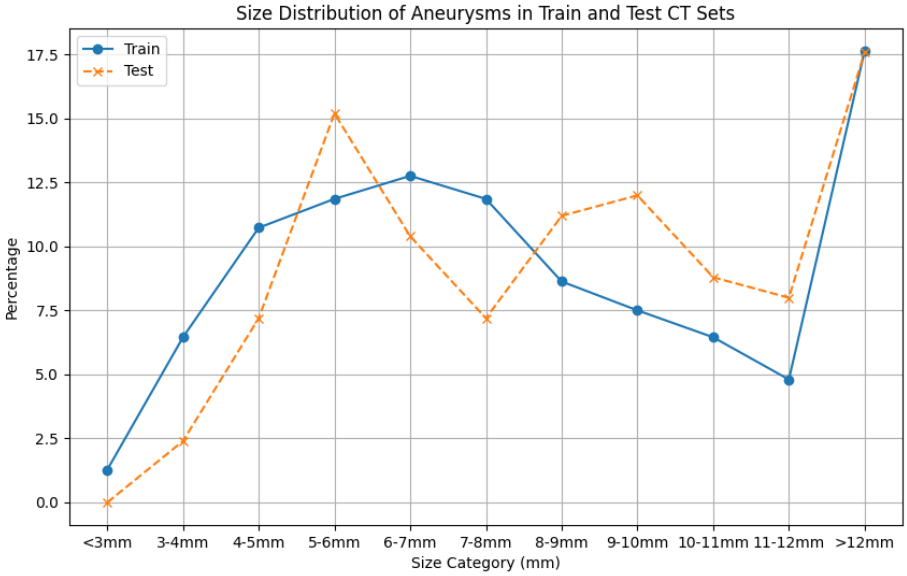


FIGURE 2.1: Distribution of aneurysm sizes in CT dataset. Aneurysm size is defined as twice the maximum distance from the center of mass to any point in the set of coordinates, which effectively represents the diameter of the minimal bounding sphere.

For the purposes of this study, the internal training set of 1,186 images and the internal test set of 152 images were utilized, collectively comprising a total of 1,489 IAs. The training set was employed to train the model, while the test set was utilized solely for the purpose of evaluating its performance. A sample is provided in Figure 2.2.

Dataset	#Pats	#IAs	Ruptured	Non-rupt.	Male	Female	0 IA	1 IA	2 IAs	≥ 3 IAs
Internal training	1,186	1,363	474	712	508	678	0	1,043	119	24
Internal test	152	126	42	60	63	89	50	85	13	4

TABLE 2.1: CTA Dataset statistics in detail [19]

2.1.2 MRA Dataset

The MRA dataset was extracted from the open platform OpenNeuro [27]. The current section provides a summarized version of the detailed dataset characteristics described in [7]. The dataset, named **Lausanne TOF-MRA Aneurysm Cohort**, included patients who underwent Time-of-Flight (TOF)-MRA between 2010 and 2015 and for whom the corresponding radiological reports were available. In total, 284 TOF-MRA subjects were retrieved, with 157 subjects exhibiting one or more IAs and 127 subjects without IAs. The images and their segmentation were obtained from the manual masks folder. The dataset included skull-stripped TOF-MRA volumes for controls and both the skull-stripped volumes and the binary manual masks of the aneurysms for patients. A sample is displayed in Figure 2.3 and the dataset statistics are depicted in Table 2.2. Aneurysm size distribution in the test and train sets is shown in Figure 2.4. Table 2.3 shows locations and sizes of aneurysms [7].

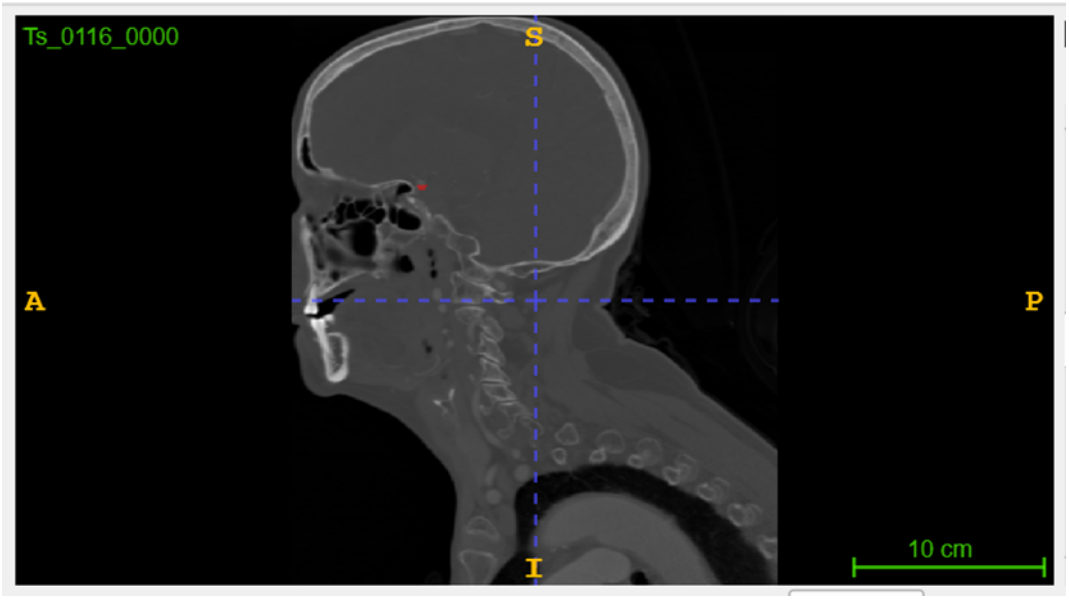


FIGURE 2.2: A sample of an original CT image with the ground truth label of an intracranial aneurysm

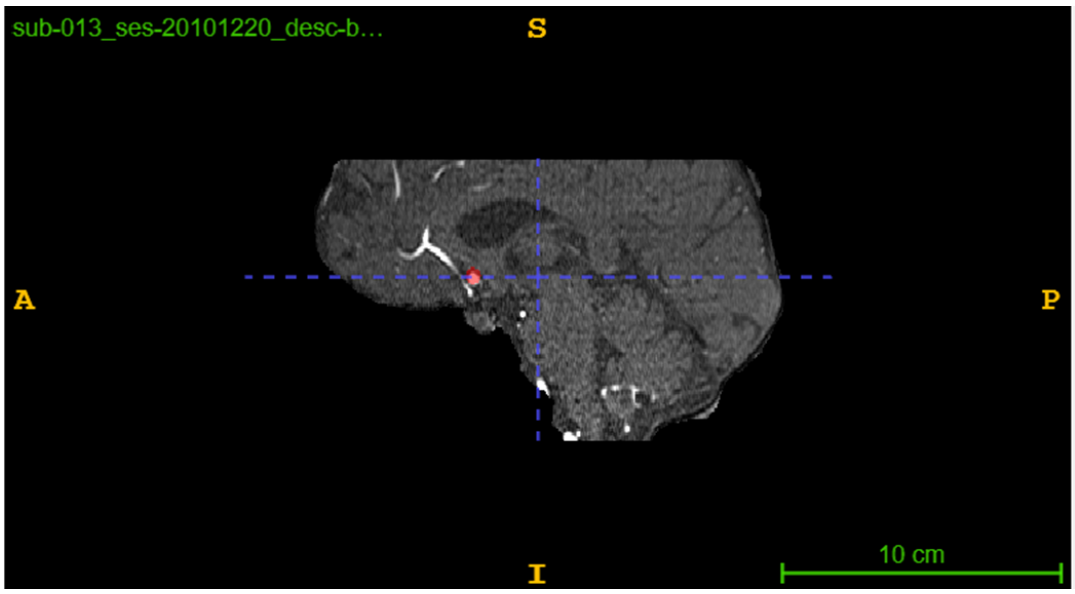


FIGURE 2.3: A sample of an original MR image with the ground truth label of an intracranial aneurysm. The image is originally cropped to the region of interest.

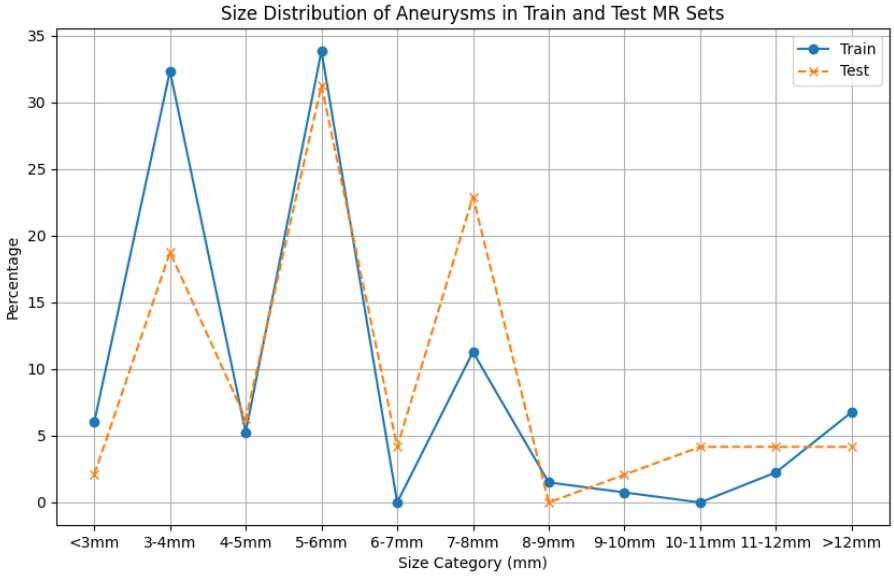


FIGURE 2.4: Distribution of aneurysm sizes in MR dataset. Aneurysm size is defined as twice the maximum distance from the center of mass to any point in the set of coordinates, which effectively represents the diameter of the minimal bounding sphere.

	Patients	**Controls**	**Whole Sample**
N	157	127	284
Age (y)	56 ± 14	46 ± 17	51 ± 16
Sex	53 M, 104 F	61 M, 66 F	114 M, 170 F
# UIA	198	0	198

TABLE 2.2: Demographics of the MR dataset. Patients = subjects with aneurysm(s). Controls = subjects without aneurysms. Age calculated in years and presented as mean $\pm$ standard deviation. *N* number of samples, *M* males, *F* females, *UIA* Unruptured Intracranial Aneurysms[7].

2.2 Segmentation Model

This thesis encompasses the training and validation of deep learning models for aneurysm segmentation, specifically employing the state-of-the-art segmentation method, nnU-Net [28].

The U-Net architecture is a convolutional neural network designed for biomedical image segmentation, characterized by a symmetric U-shaped structure comprising a contracting encoder path and an expansive decoder path. The encoder reduces spatial dimensions while increasing feature complexity, and the decoder combines upsampled features with high-resolution encoder features to produce precise segmentation maps [29].

	Count	%
Location		
ICA	59	29.8 (59/198)
MCA	57	28.8 (57/198)
ACA/Pcom/Posterior	82	41.4 (82/198)
Size		
$d \leq 7$ milimeters (mm)	180	91.0 (180/198)
7 - 9, 9 mm	7	3.5 (7/198)
10 - 19, 9 mm	10	5.0 (10/198)
$d \geq 20$ mm	1	0.5 (1/198)

TABLE 2.3: Locations and sizes of aneurysms for the MR dataset. *ICA* Internal Carotid Artery, *MCA* Middle Cerebral Artery, *ACA* Anterior Cerebral Arteries, *Pcom* Posterior communicating artery, *Posterior* posterior circulation, *d* maximum diameter [7]

The nnU-Net semantic segmentation method is designed to adapt to a given dataset. It analyzes the provided training cases and automatically configures a matching U-Net-based segmentation pipeline. Upon its release, nnU-Net was evaluated on 23 datasets belonging to biomedical domain competitions [30]. This study employed the nnunetv2 framework, version 2.2.

As the nnU-Net architecture is built upon the U-Net model, it also comprises an encoder integrating multiple blocks of convolutional layers, instance normalization, leaky ReLU activation functions, a decoder containing convolutional transpose layers, and skip connections joining corresponding layers in the encoder and decoder. This sophisticated structure comprises approximately 30 million parameters (more precisely: 31,195,594 parameters for CT datasets and 30,785,994 for MR dataset).

Given a new dataset, nnU-Net systematically analyzes the provided training cases to create a "dataset fingerprint". This analysis leads to the generation of several U-Net configurations tailored to the dataset's characteristics: a 2D U-Net for both 2D and 3D datasets, a high-resolution 3D U-Net for 3D datasets, and a 3D U-Net cascade for large 3D datasets, where a low-resolution 3D U-Net is refined by a subsequent high-resolution model. nnU-Net's segmentation pipelines are configured through a three-step process:

- fixed parameters based on a robust pre-identified configuration (see Table 2.5)
- rule-based parameters, which adapt the segmentation pipeline based on heuristic rules derived from the dataset fingerprint (see Figure 2.5)
- empirical parameters, which involve trial-and-error methods to select the best U-Net configuration and optimize post-processing strategies [30]

2.3 Preprocessing

The preprocessing in the fully automated nnU-Net segmentation pipeline operates without user intervention. To begin, nnU-Net computes a dataset fingerprint (See Table 2.4) through the preprocessing command. This dataset fingerprint provides

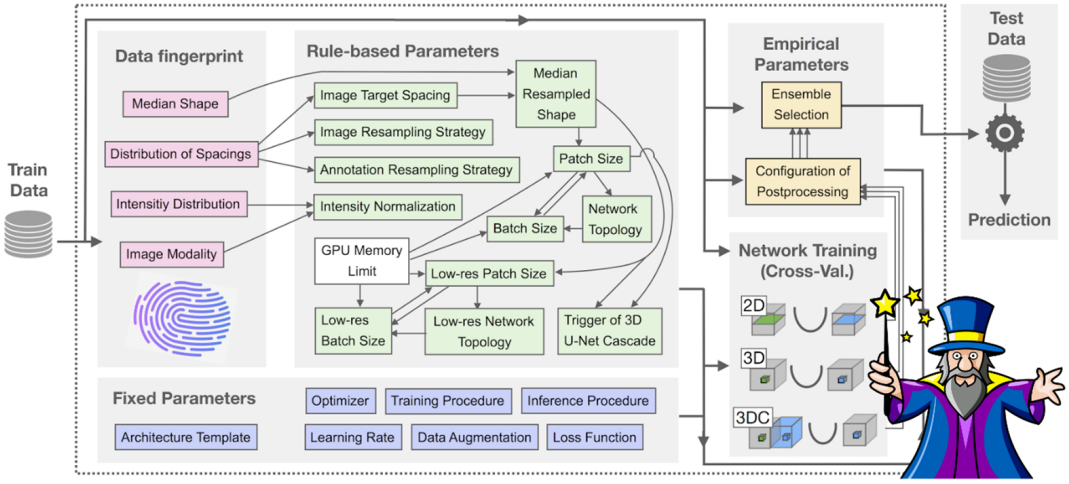


FIGURE 2.5: Overview of nnU-Net's architecture [31]

a comprehensive overview of the pixel intensity distributions within the dataset, which are necessary for the subsequent preprocessing steps.

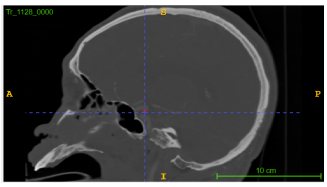
The preprocessing steps include cropping all data to the region of nonzero values to reduce the computational burden, and resampling to standardize voxel spacings, with image data resampled to the median voxel spacing of the dataset. Normalization processes vary by modality: CT images are clipped and z-score normalized based on dataset statistics, while MRI images undergo individual z-score normalization [28].

The nnU-Net plans (See Table 2.6) generated during preprocessing serve to prepare the input data and configure the model for training. Parameters such as batch size, patch size, median image size, and voxel spacing ensure that the data is processed consistently. The specified architecture, with its defined number of features, convolutional layers, pooling operations, and kernel sizes, is optimized to handle the specific characteristics of the dataset.

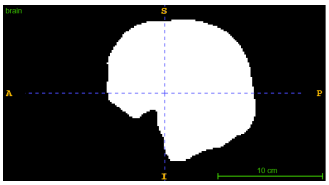
Channel	Min	Max	Mean	Median	Std	Percentile 0.5	Percentile 99.5
CT	-67.01	2395.00	205.99	207.99	132.73	-0.016	579.99
MR	0.00	2564.45	289.16	210.00	275.05	0.00	1521.70

TABLE 2.4: Dataset fingerprint: Foreground intensity properties per channel

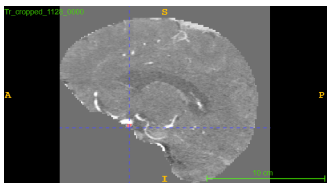
Prior to feeding the images into the model, an additional preprocessing step was conducted using the TotalSegmentator (TS) model to extract the brain [32]. The original images were then cropped to the brain region only, which allowed for a reduction in training time and an improvement in predictions by avoiding the generation of false positives outside of the brain area. The Figure 2.6 illustrates the workflow from the original CT image to the cropped version.



(A) Original CT image



(B) Brain mask segmentation by TS



(C) Original image cropped to the brain region

FIGURE 2.6: Preprocessing using TS

2.4 Training

As a result of the preprocessing, nnU-Net automatically generates a JSON file to capture the characteristics of the datasets (number of cases, modalities, labels). These characteristics were needed to create the dataset fingerprint. The original file names were then renamed to the nnU-Net-compatible format as required [33]. The following commands from the nnU-Net repository were executed: plan and process, verify data integrity, preprocess, train 3d fullres, find best configuration, predict, postprocess and evaluate. All the commands were executed on a workstation running Nvidia A100 80GB Tensor Core GPUs.

The nnU-Net model was trained using a 5-fold cross-validation technique to evaluate its performance comprehensively. This approach involved splitting the dataset into 5 subsets (folds). The model was trained 5 times, each time using a different subset as the validation set and the remaining subsets as the training set. This ensured that all data points were used for both training and validation at least once, providing an assessment of the model's generalizability and robustness. The networks are trained using a combination of dice and cross-entropy loss. A Stochastic gradient descent (SGD) optimizer with Nesterov momentum (=0.99) is used [34], and an epoch is defined as 250 training batches. The learning rate is defined by the polynomial learning rate scheduler, with the exponent parameter 0.9 influencing the rate of decay. The default parameters are summarized in Table 2.5. These default nnU-Net parameters were chosen based on extensive experimentation and were found to be generally robust for various biomedical image segmentation tasks [28].

The training process employed fixed default parameters, which had been optimized and validated in previous research to ensure robust performance across various datasets and tasks [28]. This approach simplified the training process by leveraging pre-established best practices in hyperparameter settings.

The results from all 5 folds were aggregated to compute the mean and 95% Confidence Interval (CI) for the evaluation metrics, allowing a statistical assessment of the model's performance.

Parameter	Description
Learning rate	PolyLR Schedule (Init: 0.01)
Loss function	Dice + Cross-Entropy
Architecture template	Encoder-decoder with skip-conn. ("U-Net-like") and: Instance norm., Leaky ReLU, deep super-vision (topology adapted in Inferred Parameters)
Optimizer	SGD with Nesterov Momentum (=0.99)
Data augmentation	Rotations, scaling, Gaussian noise, Gaussian blur, brightness, contrast, simulation of low resolution, Gamma and mirroring
Training procedure	1000 epochs x 250 minibatches, foreground oversampling
Inference procedure	Sliding window with half patch size overlap, Gaussian patch center weighting

TABLE 2.5: Default nnU-Net training parameters [35]

Parameter	Value
Batch Size	2
Patch Size	[64, 224, 160]
Median Image Size in Voxels	[129.0, 411.0, 334.0]
Spacing	[0.70, 0.41, 0.41]
Normalization Schemes	ZScoreNormalization
UNet Class Name	PlainConvUNet
UNet Base Num Features	32
Num Conv per Stage (Encoder)	[2, 2, 2, 2, 2, 2]
Num Conv per Stage (Decoder)	[2, 2, 2, 2, 2]
Num Pool per Axis	[4, 5, 5]
Pool Op Kernel Sizes	[[1, 1, 1], [2, 2, 2], [2, 2, 2], [2, 2, 2], [2, 2, 2], [1, 2, 2]]
Conv Kernel Sizes	[[3, 3, 3], [3, 3, 3], [3, 3, 3], [3, 3, 3], [3, 3, 3], [3, 3, 3]]
UNet Max Num Features	320

TABLE 2.6: nnU-Net Plans for 3d_fullres MR dataset

All results present the mean of each metric and a 95% CI over all 5 folds. The formula used to compute the 95% CI over all folds is the Eq. 2.1.

$$\text{CI} = \mu \pm 1.96 \times \left(\frac{\sigma}{\sqrt{n}} \right) \tag{2.1}$$

where:

- μ is the average of the metric over all folds,
- σ is the standard deviation of the metric over all folds,
- n is the number of folds.

2.5 Post-processing

The default post-processing of nnU-Net involves performing connected component analysis on the training data's ground truth segmentation labels, interpreting classes that consistently lie within a single connected component as a dataset property, and consequently removing all but the largest connected component for those classes in predicted images. However, despite nnU-Net's default post-processing, more than 10% of all predicted aneurysms were smaller than 1mm. The detection of aneurysms smaller than 1 mm on CT or MR scans is impractical due to the limitations of imaging technology and the human eye. This has led us to the decision to filter out predictions smaller than 1 mm as noise. High-resolution imaging typically resolves down to 0.6 mm – 0.7 mm, but noise, limited contrast, and human visual limitations make reliable detection of such small aneurysms unfeasible [36]. This approach allows us to focus the validation on clinically relevant sizes and to eliminate obvious false positive cases.

2.6 Evaluation

Following the training and post-processing phases, the model's performance was assessed utilizing standard metrics, including the Dice Similarity Coefficient (DSC), Intersection over Union (IoU), recall, precision, Hausdorff Distance (HD) and other pertinent measures of segmentation accuracy. This evaluation ensured that the performance of the model was both consistent and comparable with other models in the context of aneurysm detection and segmentation. The selected metrics were deliberately chosen to align with those used in previous studies, facilitating a direct comparison.

The DSC is used to gauge the similarity between two sets. It is particularly useful in image segmentation tasks for comparing the overlap between the predicted segmentation and the ground truth segmentation. In this work, the DSC was computed as in 2.3. In [19] the DSC was computed using the Eq. 2.2, which is equivalent to Eq. 2.3, used in this work when ϵ approaches 0.

$$\mathrm{DSC} = \frac{2 \cdot (\mathrm{TP} + \epsilon)}{2 \cdot (\mathrm{TP} + \epsilon) + (\mathrm{FP} + \epsilon) + (\mathrm{FN} + \epsilon)} \tag{2.2}$$

$$\mathrm{DSC} = \begin{cases} \frac{2 \times \mathrm{TP}}{2 \times \mathrm{TP} + \mathrm{FP} + \mathrm{FN}} & \text{if } (2 \times \mathrm{TP} + \mathrm{FP} + \mathrm{FN}) \neq 0 \\ 0 & \text{otherwise} \end{cases} \tag{2.3}$$

The IoU measures the overlap between the predicted segmentation and the ground truth segmentation relative to their union (Eq. 2.4).

$$\mathrm{IoU} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \tag{2.4}$$

Recall measures the proportion of actual positives (ground truth) that are correctly identified by the model (Eq. 2.5.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2.5}$$

Precision measures the proportion of predicted positives that are correctly identified as true positives (Eq. 2.6.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2.6}$$

The 95% HD, denoted as $d_H^{95\%}$, measures the 95th percentile of the maximum Euclidean distance between the predicted and ground truth segmentations, thereby providing a robust metric less sensitive to outliers (Eq. 2.7). The function used in this work is described in the MedPy library [37].

$$d_H^{95\%}(A, B) = \max \left\{ P_{\substack{95\% \\ a \in A}} \left(\inf_{b \in B} d(a, b) \right), P_{\substack{95\% \\ b \in B}} \left(\inf_{a \in A} d(b, a) \right) \right\} \tag{2.7}$$

where $P_{95\%}$ denotes the 95th percentile.

The performance evaluation was conducted at two distinct levels: voxel-wise and aneurysm-wise. A summary of the metrics employed for each dataset is presented in Figure 2.7.

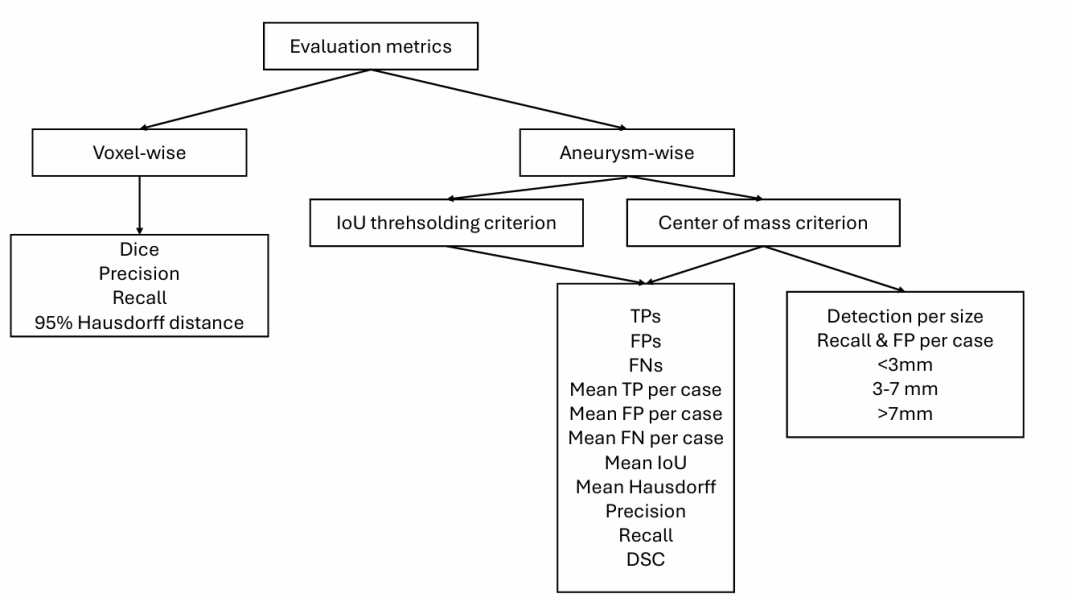


FIGURE 2.7: Evaluation metrics

2.6.1 Vowel-wise Performance

Vowel-wise performance results are automatically calculated using nnU-Net's evaluate command, providing results on a per-voxel and per-image basis. However, since nnU-Net does not compute all the above metrics for comparison, a custom evaluation script was developed to calculate correctly classified voxels and additional metrics, including Dice score, precision, recall, and 95% Hausdorff distance.

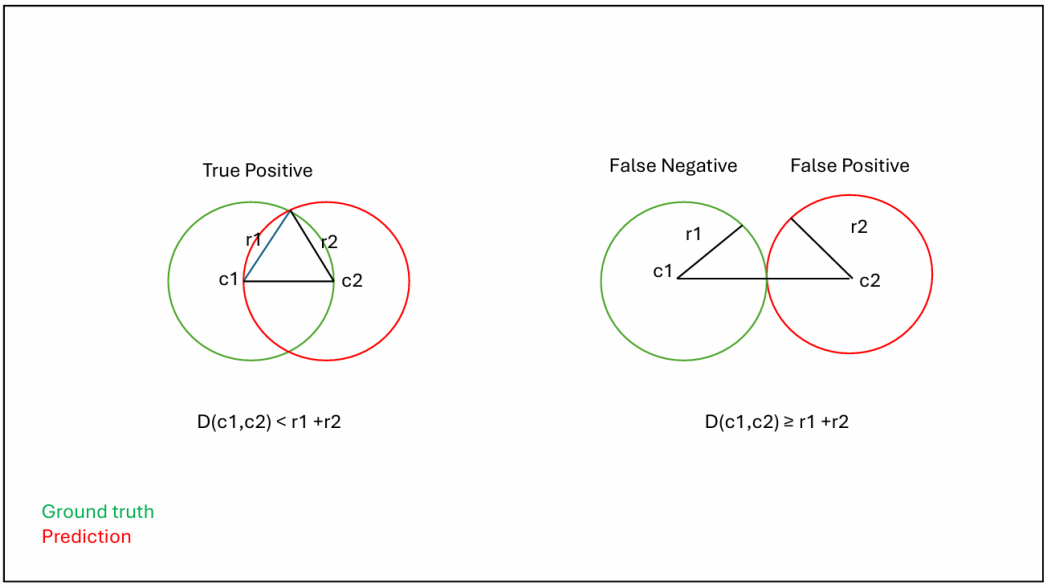


FIGURE 2.8: Illustration of the center of mass distance criterion

2.6.2 Target-wise Performance

A more refined and practically relevant evaluation was performed using target-based calculations, where the metrics were calculated at the aneurysm level instead of voxels.

Two different evaluation approaches were used to comprehensively and comparatively evaluate the model's *detection* performance against the other models used in the papers accompanying the datasets [7, 19].

In the first approach, the Euclidean distance between centers of mass was used to decide if an aneurysm was predicted correctly:

$$d(\mathbf{c}_{\text{pred}}, \mathbf{c}_{\text{gt}}) < r_{\text{pred}} + r_{\text{gt}}. \tag{2.8}$$

In accordance with the metrics presented in [19], the criterion employed was the distance between the centers of mass of predicted and ground truth aneurysms ($\mathbf{c}$), with the stipulation that this distance be less than the sum of their respective radii r. It is important to note that this method has its limitations, as aneurysms can have various shapes that are not necessarily spherical. A more accurate approach would take into account the actual shape of the aneurysm. However, for the sake of comparison, the same metrics as in the original paper were used in this work. The criterion is illustrated in Figure 2.8.

In [19], the size dependency of the detection performance was examined by stratifying the aneurysms into three size groups: smaller than 3 mm, between 3 mm and 7 mm, and greater than 7 mm. Unfortunately, the paper does not explicitly state the definition of aneurysm size. Our definition measures aneurysm size as the diameter of the minimum bounding sphere of an aneurysm. Using this definition, there are no aneurysms in the first size group (<3mm). This implies that the authors of the [19] used a different definition of aneurysm size. To compensate for this discrepancy, we

FIGURE 2.9: Illustration of the IoU threshold criterion

have used our own size groups: <5 mm, 5-10 mm and >10 mm, as opposed to those in the reference paper.

The second approach to assess the detection performance makes use of IoU thresholding.

$$IoU \geq 30\% \tag{2.9}$$

The IoU thresholding method was employed to assess the accuracy of the predictions. Specifically, a prediction was considered a true positive if the IoU between the predicted and ground truth aneurysms was equal to or greater than 30%. The reference papers did not employ the same methodology, thus the threshold was selected without conducting a direct comparison. Nonetheless, this metric is reported to provide an additional perspective on model performance, facilitating future research and comparative evaluations. The idea is illustrated in 2.9.

A sensitivity analysis was performed to determine a suitable IoU threshold by re-evaluating the prediction performance on the test data for different choices of the threshold (c.f.Figure 2.10).

Increasing values of the IoU threshold (horizontal axis) increase the FP rate and lower the TP rate. The optimal threshold is identified at the point of intersection between the two curves, which occurs at approximately 0.5 in this case. The true positive and false positive rates showed minimal variation between IoU thresholds of 10% and 50% (cf. Fig. 2.10). An IoU threshold of 30% was selected as optimal compromise for classifying TPs. A higher IoU threshold would be too restrictive, potentially leading to the omission of aneurysms with smaller intersection areas, thus increasing the FN rate. Conversely, a lower IoU threshold would be too lenient, thereby increasing the FP rate.

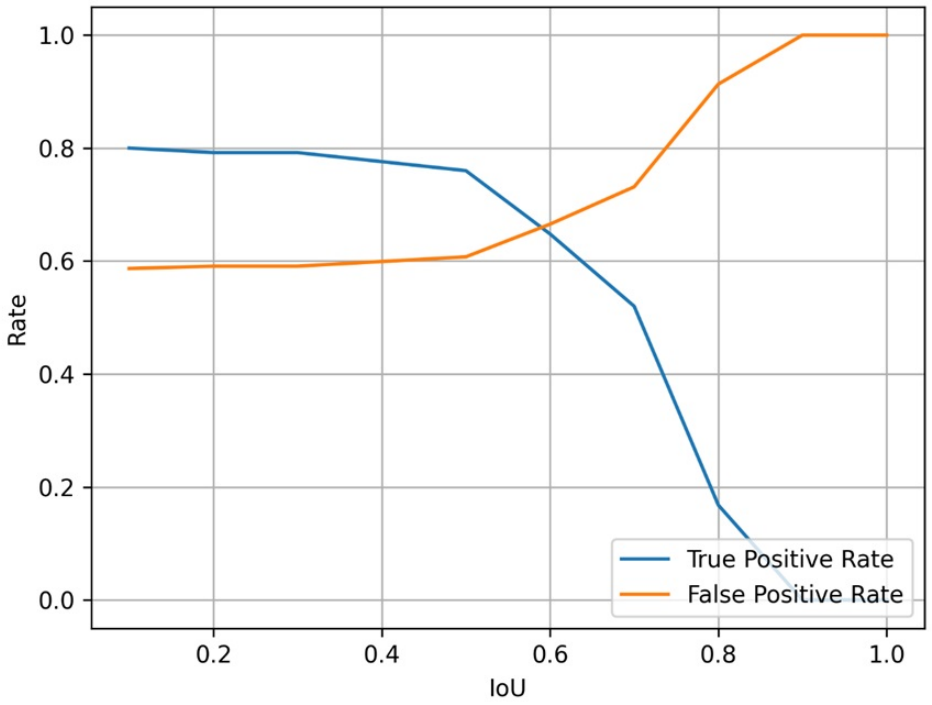


FIGURE 2.10: Sensitivity analysis for the IoU thresholding.

Chapter 3

Results

3.1 Performance on CT dataset

This section presents the evaluation results of the nnU-Net model on the CT dataset and compares them to the results reported in the paper [19] for U-Net, HeadXNet, and GLIA-Net. Additionally, it provides the evaluation results without comparison as a baseline for future research.

Voxel-wise performance on CT dataset

The Table 3.1 provides voxel-wise segmentation performance metrics for four models: U-Net, HeadXNet, GLIA-Net, and nnU-Net. The metrics include Precision, Recall, DSC, and 95% HD, along with their 95% CIs. The performance results for the first three models were sourced from [19] for comparative purposes, while the results for nnU-Net were computed as part of this study.

Furthermore, in Table 3.2 all other results computed on the CT dataset are presented, despite the absence of a benchmark for comparison.

Dataset	Model	Precision (%) ↑ (95% CI)	Recall(%) ↑ (95% CI)	DSC ↑ (95% CI)
Internal test (n= 151)	U-Net	14.0 (11.9–16.2)	71.3 (63.9–78.7)	23.2 (20.5–25.9)
	HeadXNet	16.2 (13.1–19.2)	55.6 (33.0–78.2)	23.2 (20.6–25.9)
	GLIA-Net	48.8 (44.5–53.0)	72.9 (66.9–78.9)	57.9 (56.4–59.5)
	nnU-Net	71.0 (69.0–73.0)	70.0 (68.0–72.0)	70.0 (69.0–72.0)

TABLE 3.1: Voxel-wise segmentation performance on the CT dataset in comparison

Metric	Mean (95% CI)
Mean TP per case	1234.20 (1199.08 - 1269.31)
Mean FP per case	508.91 (454.27 - 563.55)
Mean FN per case	528.85 (487.81 - 569.88)
Mean IoU	0.61 (0.59 - 0.63)

TABLE 3.2: Voxel-wise segmentation performance on the CT dataset

Target-wise performance on CT dataset

The target-wise detection performance was evaluated based on IoU thresholding criteria and center of mass distance criteria. The results are summarized in Table 3.3 and Table 3.4, respectively.

Metric	Mean (95% CI)
True Positives (TPs)	98.60 (98.12 - 99.08)
False Positives (FPs)	136.20 (108.15 - 164.25)
False Negatives (FNs)	125.00 (125.00 - 125.00)
Mean TP per case	0.66 (0.66 - 0.66)
Mean FP per case	0.92 (0.72 - 1.11)
Mean FN per case	0.84 (0.83 - 0.85)
Mean IoU	0.71 (0.70 - 0.71)
Mean Hausdorff	2.62 (2.29 - 2.95)
Precision (%)	43.0 (38.0 - 47.0)
Recall (%)	44.0 (44.0 - 44.0)
DSC	0.43 (0.41 - 0.46)

TABLE 3.3: Target-wise detection performance on CT data (iou thresholding criterion)

Dataset	Model	TPs ↑	FPs ↓	FNs ↓	Recall (%) ↑	FPs per case ↓
Internal test (n = 151)	U-Net	92.4 (89.0–95.8)	4.68k (3.72k–5.64k)	33.6 (30.2–37.0)	73.3 (70.6–76.0)	30.8 (24.5–37.1)
	HeadXNet	69.2 (45.1–93.3)	2.41k (1.21k–3.62k)	56.8 (32.7–80.1)	54.9 (35.8–74.1)	15.9 (7.96–23.8)
	GLIA-Net	103 (98.5–108)	666 (443–889)	22.6 (17.7–27.5)	82.1 (78.2–86.0)	4.38 (2.91–5.85)
	nn-UNet	108.2 (105.43–110.97)	127.6 (97.39–157.81)	19.8 (17.52–22.08)	85.0 (83.0–86.0)	0.85 (0.64–1.05)

TABLE 3.4: Target-wise detection performance on CT dataset (center of mass distance criterion)

Detection per aneurysm size

Table 3.5 presents the detection performance for aneurysms of different sizes on the CT internal test dataset for nnU-Net. The metrics evaluated include recall and FPs per case for different aneurysm size categories (<5 mm, 5–10 mm, >10 mm).

Size Category	Metric	Value
<5mm (n=12)	Recall(%) ↑	43.34 (38.71–47.96)
	FPs per case ↓	0.65 (0.43–0.88)
5-10mm (n=70)	Recall(%) ↑	85.23 (81.97–88.48)
	FPs per case ↓	0.17 (0.12–0.21)
>10mm (n=43)	Recall(%) ↑	94.29 (92.74–95.84)
	FPs per case ↓	0.03 (0.01–0.04)

TABLE 3.5: Detection performance for aneurysms of different sizes on CT dataset.

3.2 Performance on MR dataset

This section presents the evaluation results of the nnU-Net model on the MR dataset and compares them to the results reported in the paper [7] for 3D-UNet. Additionally, it provides the evaluation results without comparison as a baseline for future research.

Voxel-wise performance on MR dataset

Table 3.6 presents the voxel-wise segmentation performance on the MR dataset, providing detailed information on various performance metrics along with their associated confidence intervals. In the absence of a benchmark comparison, these values can nevertheless serve as a baseline for future studies and improvements in segmentation algorithms.

Metric	Mean (95% CI)
Mean TP per case	820.85 (737.54 - 904.15)
Mean FP per case	1024.95 (658.38 - 1391.53)
Mean FN per case	1661.74 (1578.11 - 1745.37)
Mean IoU	0.36 (0.33 - 0.39)
Precision (%)	46.0 (39.0 - 53.0)
Recall (%)	33.0 (30.0 - 36.0)
DSC	0.38 (0.35 - 0.41)

TABLE 3.6: Voxel-wise segmentation performance on MR dataset

Target-wise performance on MR dataset

The Table 3.7 compares the target-wise detection performance of 3D-UNet and nnUNet on the MR dataset. The TP was defined based on center of mass distance criterion. Table 3.8 presents the target-wise detection performance on the MR dataset using the IoU thresholding method and the center of mass distance criteria side-by-side for comparison.

Model	Recall	Avg. FP rate
3D-UNet [7]	68%	2.50
nn-UNet	68%	0.24

TABLE 3.7: Target-wise detection performance on the MR dataset in comparison

Detection per aneurysm size

Table 3.9 presents the detection performance of the nnU-Net model for aneurysms of different sizes on the MR dataset.

Metric	IoU Thresholding	Center of Mass Distance
TPs	20.60 (18.07 - 23.13)	32.80 (29.10 - 36.50)
FPs	25.80 (18.43 - 33.17)	13.60 (6.40 - 20.80)
FNs	49.00 (49.00 - 49.00)	15.40 (11.89 - 18.91)
Mean TP per case	0.36 (0.32 - 0.41)	0.58 (0.51 - 0.64)
Mean FP per case	0.45 (0.32 - 0.58)	0.24 (0.11 - 0.36)
Mean FN per case	0.86 (0.86 - 0.86)	0.27 (0.21 - 0.33)
Mean IoU	0.54 (0.51 - 0.56)	0.39 (0.36 - 0.42)
Mean Hausdorff	2.87 (2.75 - 2.99)	5.40 (5.02 - 5.78)
Precision(%)	46.0 (38.0 - 53.0)	72.0 (63.0 - 81.0)
Recall (%)	30.0 (27.0 - 32.0)	68.0 (61.0 - 75.0)
DSC	0.36 (0.33 - 0.38)	0.69 (0.66 - 0.73)

TABLE 3.8: Target-wise detection on MR dataset using IoU thresholding and center of mass distance criteria

Size Category	Metric	Value
<5mm (n=13)	Recall(%) ↑	29.23 (18.77–39.69)
	FPs per case ↓	0.15 (0.02–0.28)
5-10mm (n=29)	Recall(%) ↑	78.62 (67.96–89.28)
	FPs per case ↓	0.02 (-0.01–0.06)
>10mm (n=6)	Recall(%) ↑	100.0
	FPs per case ↓	0.07 (0.02–0.11)

TABLE 3.9: Detection performance for aneurysms of different sizes on MR dataset.

Chapter 4

Discussion

4.1 NN-Unet performance

On CT dataset

nnU-Net significantly outperforms GLIA-Net in terms of voxel-wise precision (71.0% vs. 48.8%) indicating that nnU-Net is more accurate in identifying positive instances with fewer false positives. nnU-Net has a higher DSC compared to GLIA-Net (0.70 vs. 0.58), indicating better overlap between predicted and actual segmentations (See Table 3.1). Both models have comparable recall values, with GLIA-Net having a slight edge (72.9% vs 70.0%). This indicates that both models are similarly effective in identifying relevant instances, but GLIA-Net might capture slightly more true positives.

On the aneurysm level, the nnU-Net model outperforms the other models in terms of TPs and recall, indicating its higher sensitivity in detecting targets. Also, nnU-Net's FPs per case and the number of FN are the lowest. However, it also has a higher number of FPs compared to GLIA-Net, which is likely due to its advanced post-processing procedures(See Table 3.4).

The nnU-Net model demonstrates high recall rates of 94.29% for larger aneurysms (>10 mm). However, for smaller aneurysms (<5 mm), the nnU-Net shows a smaller recall of 43.34%, indicating a limitation in detecting smaller targets. Additionally, for all size categories nnU-Net's FP rate is lower than 1. Due to discrepancies in the definitions of aneurysm size and the selection of different size categories, a comparison with other models could not be conducted.

On MR dataset

The evaluation of voxel-wise segmentation performance metrics for the nnU-Net model on the MR dataset is presented in Table 3.6. The mean IoU is 0.36 (0.33-0.39). The precision, recall, and DSC are 46.0% (39.0%–53.0%), 33% (30%–36.0%), and 0.38 (0.35–0.41), respectively.

On the aneurysm level, precision, recall, and DSC values are relatively low at 46.0%, 30.0%, and 0.36 respectively. These metrics suggest that while the nnU-Net can identify some true positives, there is a significant number of false positives and false negatives, indicating potential challenges in accuracy and reliability (Table 3.8).

A comparison of the 3D-UNet and nnU-Net models in terms of recall and average

false positive rate (see Table 3.7) reveals that both models exhibit an identical recall rate of 68%. However, the nnU-Net significantly outperforms the 3D-UNet in terms of the average false positive rate, achieving a rate of 0.24 compared to 2.50 for the 3D-UNet. As previously stated, an additional post-processing step may be a contributing factor to the enhanced performance observed.

As illustrated in Table 3.8, the center of mass method demonstrates enhanced performance compared to the IoU thresholding method. The precision, recall, and DSC metrics exhibit notable improvements, reaching 72.0%, 68.0%, and 0.69, respectively. These outcomes indicate that the nnU-Net exhibits enhanced overall performance and robustness under this criterion.

A similar trend is observed in the MR dataset with respect to target-wise performance across different aneurysm sizes. The model has difficulty to identify IAs measuring less than 5 mm in diameter (29.23% recall). However, it demonstrates moderate to high recall rates for aneurysms in the 5–10 mm and a perfect recall in >10 mm size categories, with respective recall rates of 78.62% and 100% (see Table 3.9).

In terms of false positives per case, the model maintains a relatively low rate across all size categories. The lowest false positive rate is noted in the 5-10 mm category of aneurysms (0.02 FPs per case), whereas the rates for the <5mm (0.15 per case) and > 10 mm (0.07 per case) categories are slightly higher.

CT and MR comparison

The model demonstrates considerably superior performance on the CT dataset compared to the MR dataset. The comparative analysis indicates that the performance of the nnU-Net on the MR dataset is approximately half that observed on the CT dataset. This discrepancy can be attributed to the differing sizes of the training sets. Larger training sets, such as the 1159 samples in the CT test set, allow the model to learn more robust and generalizable features, leading to improved performance. Conversely, the smaller MR test set, with only 225 samples, provides less data for the model to learn from, which can result in reduced performance. The limited data in the smaller training set can hinder the model's ability to generalize well to new, unseen data, thereby affecting overall accuracy and effectiveness.

The nnU-Net demonstrates superior performance on CT data in most evaluations. However, a direct comparison between the CT and MR datasets in our study is not feasible due to significant differences in dataset sizes (1186 vs. 284 images) and label quality (voxel-wise labels vs. weak labels). The larger size of the CT dataset likely contributed to more effective model training, enabling better aneurysm detection. Additionally, voxel-wise segmentations provided precise boundary information, enhancing segmentation accuracy. CT scans typically offer higher resolution, whereas MR scans are cropped to regions of interest (i.e., brain areas prone to aneurysms). MR images are generally smaller, and their annotations are less precise (coarse or oversized labels). These factors likely account for the nnU-Net's reduced performance on the MR dataset.

Target vs Voxel level comparison

It is observed that voxel-wise results are generally higher than target-wise results.

For instance, the DSC at the voxel level is 70.0% (69.0%–72.0%), whereas at the aneurysm level, it is 43.0% (41.0%-46.0%). For MR, the voxel-level Dice score is 0.38 (0.35-0.41) compared to the target-wise score of 0.36 (0.33-0.38). This discrepancy may be attributed to the greater challenge of target-wise detection compared to voxel-wise detection, or possibly due to the more restrictive criteria for defining true positive targets. The differences between voxel-wise and target-wise evaluations are described in Table 4.1.

	Target-wise	**Voxel-wise**
Complexity of Detection	Involves identifying and localizing entire structures (aneurysms), which can be inherently more complex due to variations in size, shape, and location. Misclassifying or missing parts of these structures can significantly impact the performance metrics. Criteria are often more stringent. Requires a substantial overlap between the detected target and the ground truth target, or specific distance metrics. Small discrepancies can lead to false negatives.	Involves classifying individual voxels within an image. Since the task is to label each voxel independently, there is often less impact from missing small parts. Allows for partial detections to still contribute positively to performance metrics.
Statistical Differences	Metrics are evaluated on a smaller number of instances (number of aneurysms), leading to higher variability and less robust statistics.	Metrics are evaluated over a large number of voxels, providing more stable and reliable statistical measures due to the larger sample size.
Clinical Relevance	Aligns more closely with aneurysm detection task, where identifying the entire aneurysm structure is crucial for assessing the risk of rupture.	Might be more suited for preliminary image segmentation tasks, where the focus is on voxel-level accuracy rather than the holistic identification of structures (e.g. tissue classification, volume quantification, identifying Region of Interests (ROIs)).

TABLE 4.1: Comparison between target-wise and voxel-wise detection [38]

Metrics criteria comparison

In assessing the efficacy of aneurysm segmentation and detection methodologies, two principal criteria have been taken into account: the center of mass distance criterion and the IoU threshold method [2]. Each criterion possesses distinctive strengths and weaknesses, as detailed in Table 4.2.

The center of mass criterion assumes that aneurysms are spherical; however, this assumption is frequently invalid in practice due to the complex morphologies of aneurysms [39]. This limitation can result in inaccurate estimations of the aneurysm's true shape and extent.

In contrast, the IoU thresholding method offers a more comprehensive evaluation

	Center of Mass Criterion	IoU Thresholding Criterion
Strengths	• Useful for comparison with studies using the same metric	• Accounts for actual shape and size of aneurysms • Flexible and adaptable threshold
Weaknesses	• Assumes aneurysms are spherical, which may not capture true shape and extent of irregular aneurysms	• Selection of threshold can be arbitrary • High threshold may miss true positives • Low threshold may lead to false positives

TABLE 4.2: Comparison of IoU thresholding and center of mass criteria [38]

by taking into account the actual shape and size of aneurysms. Nevertheless, the selection of an appropriate IoU threshold can be somewhat arbitrary. This issue can be addressed through the implementation of a sensitivity analysis and empirical testing, with the objective of establishing an optimal threshold.

The results of this study indicate that the IoU thresholding criterion, with a chosen threshold of 30%, is slightly more restrictive. This is evidenced by the fact that the results for IoU are less high than for the results based on the center of mass criterion across most tables.

The flexibility of the IoU method suggests that it may be a superior choice for clinical applications involving the segmentation and detection of aneurysms. Therefore, it is recommended that the IoU thresholding method be used for such purposes, while also reporting the center of mass distance metric to facilitate comparisons with existing studies.

Furthermore, the center of mass distance criterion can be generalized by adding an arbitrary weight to the sum of radii. This is demonstrated by the equation

$$d = r_1 + r_2 \tag{4.1}$$

which is equivalent to IoU = 0, indicating no intersection. In this setting, the two criteria are equivalent.

In summary, the nnU-Net demonstrates varying performance based on the criterion used for true positive detection. This underscores the necessity of selecting appropriate criteria tailored to specific applications in medical imaging. Engaging in discussions with radiologists, clinicians, and medical doctors is essential to identify the most relevant metrics and criteria that align with clinical needs and practices.

4.2 Study Limitations

One of the limitations of this comparative evaluation study is that the literature review cannot be conducted in an exhaustive manner due to the disparate metrics employed by various researchers.

One other challenge is the variety of methods used to define aneurysm size. In the results reported in [19], an evaluation of the detection method was conducted across different size categories: smaller than 3 mm, between 3 mm and 7 mm, and greater than 7 mm. However, the definition of aneurysm size in this study was the diameter of the minimal bounding sphere. This approach results in a lack of aneurysms smaller than 3 mm in the dataset. Consequently, the study's comparison with other results was not presented due to the inconsistent definition of aneurysm size. In order to facilitate a more rigorous comparison, it would be beneficial for future studies to standardize the definition of aneurysm size, the criteria for true positive identification, and the metrics used for evaluation.

Finally, it is challenging to assess the differences in aneurysm segmentation and detection performance for CT and MR images when using the two available datasets. This is due to the discrepancy in their size (thousands vs. hundreds of images) and label quality (voxel-wise vs. weak labels).

4.3 Further work

To enhance model performance and explore alternative solutions to the research question, several strategies are recommended. Increasing dataset sizes is crucial, and establishing collaborations with hospitals to access larger cohorts can significantly improve model training.

Extending the training epochs has also shown promising results; in-house experiments indicated that training the model for 2000 epochs, as opposed to 1000, can yield a 1-3% improvement in performance. If computational resources allow, this extended training duration should be pursued. Additionally, customizing the nnU-Net model architecture to target specific performance metrics, particularly by modifying the foreground sampling strategy for small instances, can lead to better outcomes.

Exploring alternative models, such as those incorporating attention mechanisms for vascular structures [24], may also provide performance enhancements for specific tasks. Introduction of vessel proximity metrics and anatomically informed models might leverage the knowledge of aneurysm locations and increase segmentation performance.

Another potential avenue for investigation is the combination of CT and MR datasets to create a more robust and general model. Alternatively, the nnU-Net-trained model on the CT dataset could be fine-tuned on the MR dataset.

Finally, to conduct a more robust comparative evaluation, it is essential to adopt standardized guidelines and benchmarks for performance assessment and datasets. This standardization would facilitate coherent comparisons and ensure that new models are trained and evaluated on large benchmark datasets using consistent metrics and criteria.

Chapter 5

Conclusion

The findings from this study demonstrates the superior performance of the nnU-Net model in the segmentation and detection of IAs in large CT datasets, particularly when voxel-wise labels are available. The nnU-Net model consistently outperformed GLIA-Net in terms of voxel-wise precision and DSC, indicating its higher accuracy in identifying and segmenting aneurysms with fewer FPs. However, the model's performance on MR datasets was significantly lower, largely due to the smaller size and weaker labels of the available training data. This performance discrepancy highlights the importance of dataset size and label quality in training robust models. The comparative analysis between CT and MR datasets illustrates the impact of dataset characteristics on model performance.

Additionally, the study highlighted the importance of selecting appropriate evaluation criteria. The center of mass criterion, despite its assumption of spherical aneurysms, provided a practical approach for comparison with existing studies. However, the IoU thresholding method, with its flexibility in accounting for the actual shape and size of aneurysms, proved to be a more comprehensive evaluation metric.

To address the limitations identified in this study, future work should focus on increasing dataset sizes through collaborations with hospitals to access larger, more diverse datasets, and creating benchmark datasets for robust model training and evaluation. Customizing the nnU-Net architecture to better handle small instances and exploring alternative models with attention mechanisms can further enhance performance. Standardizing evaluation metrics and criteria in collaboration with clinical experts will ensure models are clinically relevant and comparable across studies. These strategies will enhance model performance and contribute to a deeper understanding of how different evaluation metrics impact the assessment of aneurysm segmentation and detection.

In conclusion, while the nnU-Net shows promise in the segmentation and detection of IAs, particularly in CT datasets, there is a clear need for further research to improve its performance on MR datasets and smaller aneurysms. The study's findings provide a foundation for future work aimed at developing more robust and accurate models for medical imaging applications.

DECLARATION OF ORIGINALITY

Master's Thesis for the School of Life Sciences and Facility Management

By submitting this Master's thesis, the student attests of the fact that all the work included in the assignment is their own and was written without the help of a third party.

The student declares that all sources in the text (including Internet pages) and appendices have been correctly disclosed. This means that there has been no plagiarism, i.e. no sections have been partially or wholly taken from other texts and represented as the student's own work or included without being correctly referenced.

Any misconduct will be dealt with according to paragraphs 39 and 40 of the General Academic Regulations for Bachelor's and Master's Degree Courses at the Zurich University of Applied Sciences (dated 29 Januar 2008) and subject to the provisions for disciplinary action stipulated in the University regulations (Rahmenprüfungsordnung ZHAW (RPO)).

Town/City, Date: Signature:

Zürich, 26/07/24 ..... Golubva .....

The original signed and dated document (no copies) must be included in the appendix of the ZHAW version of all Master's theses submitted.

A.1 Use of generative AI

Finally, it should be mentioned that generative AI systems and tools were utilized as sources of inspiration and for initial brainstorming in this work. Specifically, the critical dialogic interaction with ChatGPT 4o and its contents enriched my work and enhanced the quality of my scientific output. Using the OpenAI model not only enabled me to generate ideas but also helped me recognize the limitations and possible distortions in the generated content, particularly in the aspect of literature review. This fostered a deeper understanding of language technologies and raised my awareness regarding the critical reflection upon their use in my work. Additionally, AI tools were used for various purposes including summarizing literature, understanding concepts, formulating texts, and ensuring academic style. The specific AI tools employed are listed below :

- Translation of the abstract [40]
- Refining academic style [41]
- Generation of text proposals, data processing, and brainstorming [38]

Bibliography

[1] M.H. Vlak et al. "Prevalence of unruptured intracranial aneurysms, with emphasis on sex, age, comorbidity, country, and time period: a systematic review and meta-analysis". In: *Lancet Neurology* 10.7 (2011), pp. 626–636. DOI: 10.1016/S1474-4422(11)70109-0.

[2] H. Arimura 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". In: *Medical Physics* 33.2 (2006), pp. 394–401. DOI: 10.1118/1.2163389.

[3] B.M.W. Cornelissen et al. "Aneurysmal Parent Artery–Specific Inflow Conditions for Complete and Incomplete Circle of Willis Configurations". In: *American Journal of Neuroradiology* 39 (2018), pp. 910–915. DOI: 10.3174/ajnr.A5602.

[4] N.T. Anh et al. "The Features of the Circle of Willis and Cerebral Aneurysm in Patients with Cerebral Aneurysms through Films of Multi-slice Computed Tomography". In: *Journal of Medical and Pharmaceutical Sciences* 36 (2020). DOI: 10.25073/2588-1132/vnumps.4224.

[5] C. Rutledge et al. "Small Aneurysms with Low PHASES Scores Account for a Majority of Subarachnoid Hemorrhage Cases". In: *World Neurosurgery* (2020). DOI: 10.1016/j.wneu.2020.04.074.

[6] S. Hirsch et al. *Enhancing Brain Angiograms for Personalized Stroke Management*. Running time: 2021-2023. 2022. URL: https://www.dizh.uzh.ch/en/2022/01/03/deep-brain-vessel-profiler-2/.

[7] T. Di Noto, G. Marie, S. Tourbier, et al. "Towards Automated Brain Aneurysm Detection in TOF-MRA: Open Data, Weak Labels, and Anatomical Knowledge". In: *Neuroinformatics* 21 (2023), pp. 21–34. DOI: 10.1007/s12021-022-09597-0.

[8] T. Nakao et al. "Deep neural network-based computer-assisted detection of cerebral aneurysms in MR angiography". In: *Journal of Magnetic Resonance Imaging* 47.4 (2018), pp. 948–953. DOI: 10.1002/jmri.25842. URL: https://doi.org/10.1002/jmri.25842.

[9] D. Ueda, S. Doishita, and A. Choppin. "Deep learning for MR angiography: automated detection of cerebral aneurysms". In: *Radiology* (2019). DOI: 10.1148/radiol.2018180901. URL: https://doi.org/10.1148/radiol.2018180901.

[10] J.N. Stember et al. "Convolutional neural networks for the detection and measurement of cerebral aneurysms on magnetic resonance angiography". In: *Journal of Digital Imaging* 32.5 (2019), pp. 808–815. DOI: 10.1007/s10278-018-0162-z. URL: https://doi.org/10.1007/s10278-018-0162-z.

[11] T. Sichtermann et al. "Deep learning–based detection of intracranial aneurysms in 3D TOF-MRA". In: *American Journal of Neuroradiology* (2019). DOI: 10.3174/ajnr.A5911. URL: https://doi.org/10.3174/ajnr.A5911.

[12] A. Park et al. "Deep learning-assisted diagnosis of cerebral aneurysms using the HeadXNet model". In: *JAMA Network Open* 2.6 (2019), e195600. DOI: 10.1001/jamanetworkopen.2019.5600. URL: https://doi.org/10.1001/jamanetworkopen.2019.5600.

[13] H. Duan et al. "Automatic detection on intracranial aneurysm from digital subtraction angiography with cascade convolutional neural networks". In: *BioMedical Engineering Online* 18.1 (2019). DOI: 10.1186/s12938-019-0726-2. URL: https://doi.org/10.1186/s12938-019-0726-2.

[14] N. Hainc et al. "Deep learning based detection of intracranial aneurysms on digital subtraction angiography: A feasibility study". In: *Neuroradiology Journal* 33.4 (2020), pp. 311–317. DOI: 10.1177/1971400920937647. URL: https://doi.org/10.1177/1971400920937647.

[15] B. Joo et al. "A deep learning algorithm may automate intracranial aneurysm detection on MR angiography with high diagnostic performance". In: *European Radiology* 30.11 (2020), pp. 5785–5793. DOI: 10.1007/s00330-020-06966-8. URL: https://doi.org/10.1007/s00330-020-06966-8.

[16] Z. Shi et al. "A clinically applicable deep-learning model for detecting intracranial aneurysm in computed tomography angiography images". In: *Nature Communications* (2020). DOI: 10.1038/s41467-020-19527-w. URL: https://doi.org/10.1038/s41467-020-19527-w.

[17] J. Yang et al. "Deep learning for detecting cerebral aneurysms with CT angiography". In: *Radiology* 298.1 (2020), pp. 155–163. DOI: 10.1148/RADIOL.2020192154. URL: https://doi.org/10.1148/RADIOL.2020192154.

[18] X. Dai et al. "Deep learning for automated cerebral aneurysm detection on computed tomography images". In: *International Journal of Computer Assisted Radiology and Surgery* 15.4 (2020), pp. 715–723. DOI: 10.1007/s11548-020-02121-2. URL: https://doi.org/10.1007/s11548-020-02121-2.

[19] Z.-H. Bo et al. "Toward human intervention-free clinical diagnosis of intracranial aneurysm via deep neural network". In: *Patterns* 2.2 (2021), p. 100197. DOI: https://doi.org/10.1016/j.patter.2020.100197. URL: https://www.sciencedirect.com/science/article/pii/S2666389920302671.

[20] X. Liu et al. "Deep neural network-based detection and segmentation of intracranial aneurysms on 3D rotational DSA". In: *Interventional Neuroradiology* (2021). DOI: 10.1177/15910199211000956. URL: https://doi.org/10.1177/15910199211000956.

[21] M. Baumgartner et al. "nnDetection: a self-configuring method for medical object detection". In: *MICCAI*. Accessed July 2021. 2021. URL: https://link.springer.com/chapter/10.1007/978-3-030-87240-3_51.

[22] T. Di Noto et al. "Towards Automated Brain Aneurysm Detection in TOF-MRA: Open Data, Weak Labels, and Anatomical Knowledge". In: *Neuroinformatics* 21.1 (2023), pp. 21–34. ISSN: 1559-0089. DOI: 10.1007/s12021-022-09597-0. URL: https://doi.org/10.1007/s12021-022-09597-0.

[23] X. Yang, D. Xia, T. Kin, et al. "A two-step surface-based 3D deep learning pipeline for segmentation of intracranial aneurysms". In: *Comp. Visual Media* 9 (2023), pp. 57–69. DOI: 10.1007/s41095-022-0270-z.

[24] W. You et al. "Diagnosis of intracranial aneurysms by computed tomography angiography using deep learning-based detection and segmentation". In: *Journal of NeuroInterventional Surgery* (2024). ISSN: 1759-8478. DOI: 10.1136/jnis-2023-021022. eprint: https://jnis.bmj.com/content/early/2024/01/17/jnis-2023-021022.full.pdf. URL: https://jnis.bmj.com/content/early/2024/01/17/jnis-2023-021022.

[25] Z.-H. Bo. *Large IA Segmentation dataset*. Version 1. This dataset is for non-commercial purposes. Feb. 2021. DOI: 10.5281/zenodo.6801398. URL: https://zenodo.org/record/6801398.

[26] Paul A. Yushkevich et al. "User-Guided Segmentation of Multi-modality Medical Imaging Datasets with ITK-SNAP". In: *Neuroinformatics* 17.1 (2019), pp. 83–102. ISSN: 1559-0089. DOI: 10.1007/s12021-018-9385-x. URL: https://doi.org/10.1007/s12021-018-9385-x.

[27] T. Di Noto et al. *Lausanne_TOF-MRA_Aneurysm_Cohort*. Version 1.0.1. 2022. DOI: 10.18112/openneuro.ds003949.v1.0.1. URL: https://openneuro.org/datasets/ds003949/versions/1.0.1.

[28] F. Isensee et al. "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation". In: *Nature Methods* 18.2 (2021), pp. 203–211.

[29] O. Ronneberger, P. Fischer, and T. Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation". In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Ed. by N. Navab et al. Cham: Springer International Publishing, 2015, pp. 234–241. ISBN: 978-3-319-24574-4.

[30] F. Isensee et al. *nnU-Net: Self-adapting Framework for U-Net-Based Medical Image Segmentation*. Accessed: 2024-06-12. 2024. URL: https://github.com/MIC-DKFZ/nnUNet/blob/master/readme.md.

[31] F. Isensee et al. *nnU-Net Overview Image*. Accessed: 2024-07-19. 2024. URL: https://github.com/MIC-DKFZ/nnUNet/raw/master/documentation/assets/nnU-Net_overview.png.

[32] J. Wasserthal. *TotalSegmentator*. Tool for robust segmentation of over 100 important anatomical structures in CT and MR images. 2023. URL: https://github.com/wasserth/TotalSegmentator.

[33] F. Isensee et al. *nnU-Net: Dataset Format*. Accessed: 2024-06-12. 2024. URL: https://github.com/MIC-DKFZ/nnUNet/blob/master/documentation/dataset_format.md.

[34] PyTorch Contributors. *torch.optim.SGD*. Accessed: 2024-07-09. 2023. URL: https://pytorch.org/docs/stable/generated/torch.optim.SGD.html.

[35] F. Isensee. *nnU-Net: Self-adapting Framework for U-Net Based Medical Image Segmentation*. Accessed: 2024-06-24. 2020. URL: https://youtu.be/K3qRY9Q-BFo?si=_vey6rHWVp1eaDLx.

[36] N.K. Yoon et al. "Imaging of cerebral aneurysms: a clinical perspective". In: *Neurovascular Imaging* 2.1 (Mar. 2016), p. 6. ISSN: 2055-5792. DOI: 10.1186/s40809-016-0016-3. URL: https://doi.org/10.1186/s40809-016-0016-3.

[37] Alex Rothberg Oskar M. and contributors. *MedPy: Medical Image Processing in Python*. https://github.com/loli/medpy/blob/master/medpy/metric/binary.py. Accessed: 2024-07-24. 2024. URL: https://github.com/loli/medpy/blob/master/medpy/metric/binary.py.

[38] ChatGPT. *Generation of text proposals, data processing, and brainstorming*. https://www.openai.com/chatgpt. Accessed: 2024-07-24.

[39] N. Juchler. "Shape-based analysis of intracranial aneurysms". Published as part of the ZHAW project: AneuX. Doctoral thesis. Zurich, Switzerland: University of Zurich, 2020. DOI: 10.21256/zhaw-22031. URL: https://digitalcollection.zhaw.ch/handle/11475/22031.

[40] DeepL Translator. *Translation of the abstract*. https://www.deepl.com. Accessed: 2024-07-24.

[41] DeepL Write. *Refining academic style*. https://www.deepl.com/write. Accessed: 2024-07-24.