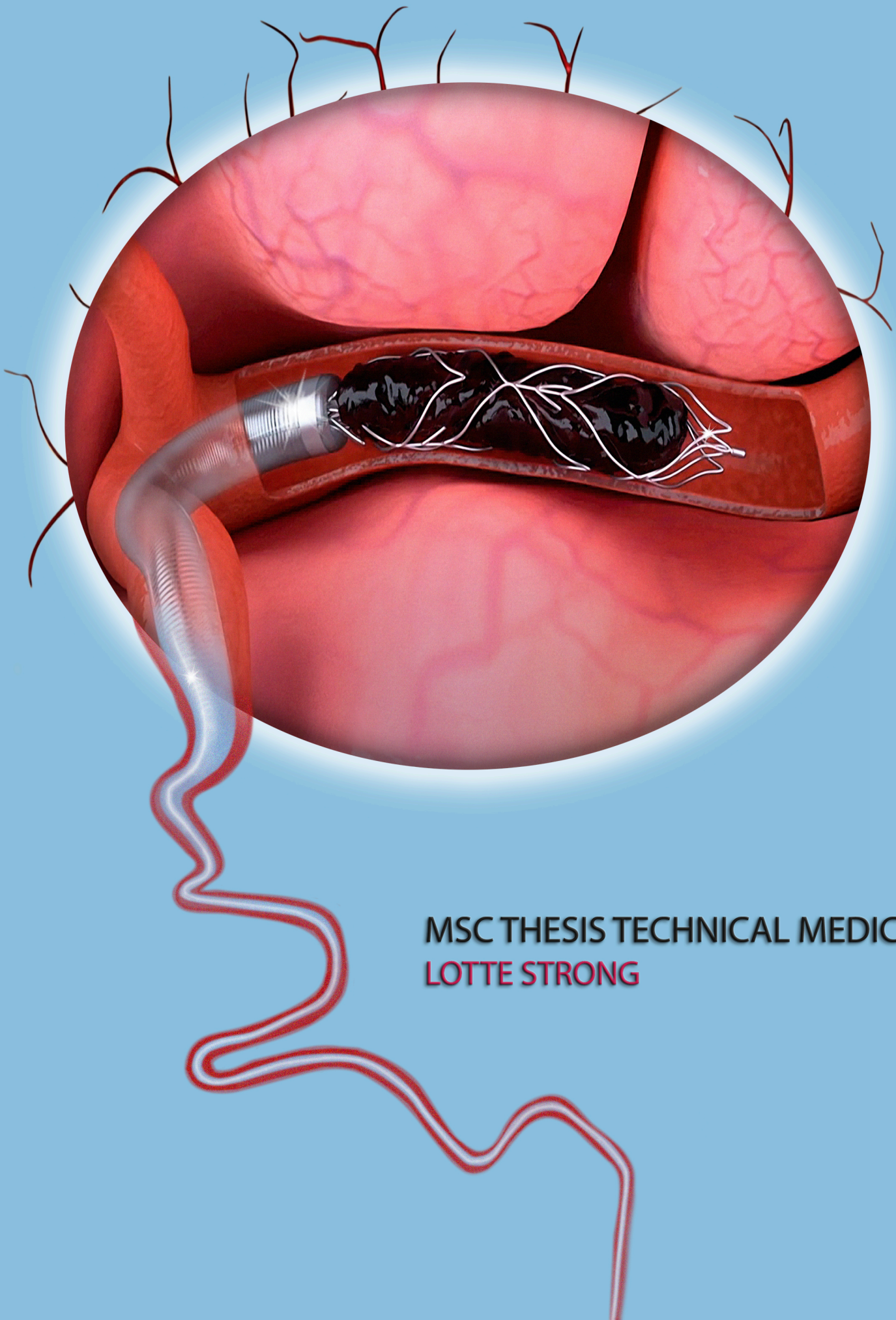
AUTOMATED QUANTIFICATION OF BRAIN REPERFUSION AFTER ACUTE ISCHEMIC STROKE TREATMENT



MSC THESIS TECHNICAL MEDICINE
LOTTE STRONG

Automated quantification of brain reperfusion after acute ischemic stroke treatment

by

Lotte Strong

Student number: 4662024
Date: 27-09-2024

Thesis in partial fulfilment of the requirements for the joint degree of Master of Science in

Technical Medicine

Erasmus MC, Leiden University Medical Center, Delft University of Technology

Master thesis project (TM30004)
Erasmus MC,
Image Guided Interventions and Therapies,
Department of Radiology & Nuclear Medicine;
Philips Medical Systems,
Image Guided Therapy

Technical supervisor: Dr. ir. T. (Theo) van Walsum, Erasmus MC
Medical supervisor: Dr. S.A.P. (Sandra) Cornelissen, Erasmus MC
Independent Thesis Committee member: Dr. A.C.G.M (Adriaan) van Es, Leiden University Medical Center
External Thesis Committee member: Prof. ir. D. (Danny) Ruijters, TU/e

An electronic version of this thesis is available at: `http://repository.tudelft.nl/`

Erasmus MC

TU Delft

Universiteit Leiden

Preface

Before you lies my thesis, which is not only the result of an 8-month research project but also of 7 years of studying Clinical Technology. This report can only capture a fraction of the experience, as these years have been more than just academic; it has been an unforgettable period of personal and professional growth.

My interest in medical imaging and image-guided therapy began in the second year of my bachelor's degree, and from that moment I knew this was the field I wanted to pursue further. Choosing the master's track of 'Imaging and Intervention' was a simple decision, one that has only been reinforced throughout my studies. This dynamic and impactful field, with the opportunity to influence patients' lives, is truly inspiring. Conducting my thesis project in this area has been a valuable opportunity. Now, I can say that I have learned extensively about strokes, registration techniques, and deep learning, while also gaining valuable experience in conducting a clinical study and interviewing clinicians. Managing three projects simultaneously was challenging, but I am proud of the results and the knowledge I have acquired.

I am particularly grateful to have conducted this research in the field of stroke, a topic that holds personal significance for me, as both of my grandparents suffered from a stroke. Stroke affects 15 million people worldwide, touching the lives of many families, and despite many recent advancements, much remains unknown. This has fueled the motivation for myself and the entire research group. I am thankful for the knowledge, enthusiasm, and motivation shared by everyone involved.

I would like to thank my supervisors, Sandra, Theo and Danny, for their guidance and support. Sandra, thank you for your willingness to take the time out of your busy schedule to answer clinically related questions, involve me in fascinating interventional procedures and provide opportunities to witness numerous endovascular thrombectomies, which greatly enriched my understanding of stroke. Theo, thank you for your support and guidance despite your busy schedule. Your expertise, constructive feedback, and kindness made this thesis journey both educational and enjoyable and helped me bring this research to a higher level. I would also like to thank Danny for the opportunity to conduct part of my thesis at Philips Healthcare. Gaining insight into the business perspective was very interesting and valuable. I always enjoyed my days in Best, and I appreciate you involving me in your team meetings and activities, even when I was primarily based at Erasmus MC. I learned a lot while working together on the installation of CloudCast. I look forward to the next step as I start my career as an Image Quality Specialist at Philips Healthcare. Furthermore, I am grateful to the involved interventional radiologists, Sandra, Pieter-Jan, Ad, and Geert, for their participation in the core-laboratory study. I appreciate their patience in working with the executable, which was new to them, and for dedicating their free time to perform the scoring. My gratitude also extends to Frank and Ruisheng, who were consistent in joining the weekly meetings, offering valuable insights and providing feedback. Special thanks to Matthijs, your willingness to discuss the project, share knowledge, and engage in personal conversations made this period even more enjoyable, and your contagious enthusiasm was truly motivating.

Finally, I would like to express my deep appreciation to my family, friends, Ben, Grayson, and roommates, Lucie, Merel and Inge. Thank you for always being there to listen and putting things in perspective. You made my student years unforgettable.

Lotte Strong
Rotterdam, September 2024

List of abbreviations

Abbreviation	Definition
ACA	Anterior cerebral artery
AI	Artificial intelligence
AIS	Acute ischemic stroke
AP	Anteroposterior
CI	Confidence interval
CTA	Computed tomography angiography
DICOM	Digital imaging and communications in medicine
DSA	Digital subtraction angiography
eTICI	Extended thrombolysis in cerebral infarction
EVT	Endovascular thrombectomy
FOV	Field of view
ICA	Internal carotid artery
ICC	Intraclass correlation coefficient
IGT	Image-guided therapy
IQR	Interquartile range
LVO	Large vessel occlusion
MCA	Middle cerebral artery
MinIP	Minimum intensity projection
OIP	Open innovation platform
SUS	System usability scale
TDT	Target downstream territory
TICI	Thrombolysis in cerebral infarction
UI	User interface

Summary

Ischemic stroke, defined as the sudden onset of a focal neurological deficit resulting from hypo-perfusion due to an occluded cerebral artery, is a leading cause of mortality and long-term disability worldwide. The standard treatment for acute ischemic strokes (AIS) is an endovascular thrombectomy (EVT), a minimally invasive procedure for locally removing the occlusion. The effectiveness of an EVT is assessed by estimating the reperfusion status on digital subtraction angiography (DSA) images using a six-category visual grading system termed the extended Thrombolysis In Cerebral Infarction (eTICI) scale. However, eTICI scoring suffers from inter-observer variability, complicating EVT outcome comparisons. To address this, 'autoTICI', a fully automated and objective deep-learning-based TICI scoring method, was developed by our research group. This thesis aimed to improve and evaluate autoTICI in clinical practice.

Chapter 1 provides a general introduction, highlighting the clinical background and relevance of automated TICI scoring.

Chapter 2 presents a pilot study that explored the technical feasibility and clinical implementability of autoTICI within the interventional radiology workflow. The pilot study showed that the performance of autoTICI is currently insufficient for implementation in clinical practice. Nevertheless, with an average System Usability Score (SUS) of 81.3, autoTICI still demonstrated excellent usability within the clinical workflow, indicating its potential for future clinical use. Besides, its potential as a decision-support tool was recognized by multiple interventional radiologists, particularly in complex or high-stress scenarios.

Chapter 3 presents a core-laboratory study assessing the inter-observer agreement and efficiency of eTICI scoring with and without autoTICI feedback, showing a substantial agreement both with and without autoTICI, with weighted kappa values of 0.65 and 0.67, respectively. Additionally, autoTICI had no significant effect on scoring efficiency, suggesting that the supplementary information did not increase the time required for eTICI scoring.

Chapter 4 details the development of an automated deep-learning approach for direct segmentation of vascular territories on cerebral DSA, aimed at improving the robustness of autoTICI by replacing the atlas registration step in its pipeline. A nnUNet model was trained, validated, and benchmarked against the atlas registration method. The proposed model demonstrated excellent performance on a held-out test set, achieving a 100% success rate compared to 52.5% for atlas registration. Additionally, it significantly outperformed the atlas registration method in both segmentation accuracy and computational efficiency.

Chapter 5 provides a general discussion and conclusion, synthesizing the results from the previous chapters and offering recommendations for future research on autoTICI.

In this master's thesis, autoTICI was successfully integrated into the clinical workflow of the interventional radiology. Through the pilot study and inter-observer analysis, we were able to identify autoTICI's intended use and added value within the clinical workflow, along with key issues, challenges, pitfalls, and future opportunities. By addressing one of these identified issues — atlas registration — through automated vascular territory segmentation, the performance of autoTICI was significantly improved. Although autoTICI is not yet ready for clinical implementation, with further optimization needed, the contributions of this master thesis have brought it significantly closer to that goal.

Contents

1

General introduction

1.1. Clinical background

1.1.1. Strokes

Stroke, defined as the rapid development of a focal neurological deficit due to a disrupted blood supply to the brain, is a leading cause of death and long-term disability worldwide [1, 2]. With 15 million people suffering from a stroke annually, it is a disabling condition with a high societal and economic burden [2]. Strokes are classified as either ischemic or hemorrhagic, with ischemic strokes, caused by a perfusion defect due to an occluded cerebral artery, accounting for 87% of cases [2]. The majority of ischemic strokes (80%) involve the anterior circulation, with approximately 75% occurring in the middle cerebral artery (MCA), 20% in the internal carotid artery (ICA), and 5% in the anterior cerebral artery (ACA) (Figure 1.1) [3–6]. The MCA originates directly from the ICA and is divided into a proximal part (involving the M1 and M2 segments) and a distal part (involving the M3 and M4 segments) [7]. The more proximal the occlusion, the greater the restriction of cerebral blood flow, leading to a more extensive infarct (Figure 1.1) [4, 5, 8].

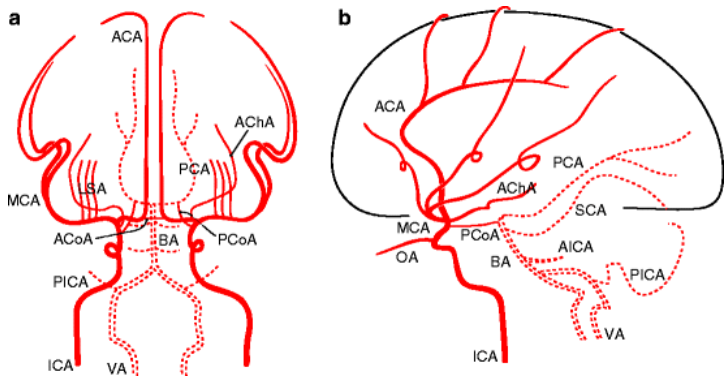


Figure 1.1: Schematic representation of the cerebral circulation, visualized in the anteroposterior (AP) and lateral view [9].

The standard of care for acute ischemic strokes (AIS) with a large vessel occlusion (LVO), an obstruction in one of the large proximal cerebral arteries, is an endovascular thrombectomy (EVT). This minimally invasive mechanical procedure involves navigating endovascular equipment through a groin puncture to the occlusion site to locally remove the blood clot causing the obstruction [10, 11].

1.1.2. Digital subtraction angiography

During EVT procedures, digital subtraction angiography (DSA) and fluoroscopy are imaging modalities used to locate the occlusion and perfusion defect and to navigate to the occlusion site. Fluoroscopy images are two-dimensional (2D) low-dose X-ray projections, while DSA images are created by acquiring a sequence of 2D high-dose X-ray projections while injecting an iodine-containing contrast medium

into the targeted vessels. Images obtained before the arrival of the contrast bolus, referred to as mask images, are subtracted from those taken during contrast injection, referred to as contrast images. This process removes background structures from the contrast images, resulting in a subtraction image exclusively showing the contrasted vessels (Figure 1.2). This enables the visualization of cerebral blood perfusion, providing crucial information about the location of the perfusion defect [12].

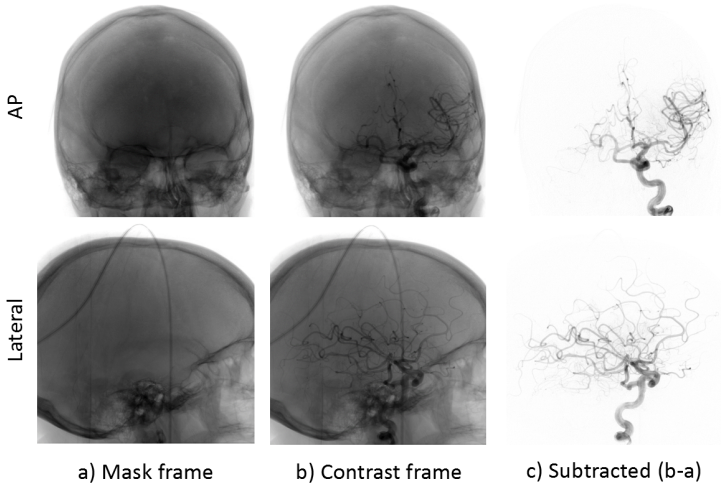


Figure 1.2: Example of a DSA image in the anteroposterior (AP) and lateral view: a) mask frame; b) contrast frame; c) subtracted frame (b-a).

1.1.3. TICI grading

The effectiveness of an EVT can be assessed by grading the reperfusion status on DSA images taken before and after an attempt to remove a blood clot. In EVT terms, reperfusion is defined as the antegrade restoration of a capillary blush [13]. The grading system, termed Thrombolysis In Cerebral Infarction (TICI), originally adapted from a coronary reperfusion grading system, is constructed as a five-category scale [14]. The scoring is performed by visually estimating the percentage of antegrade reperfused territory beyond the previously occluded area, known as the Target Downstream Territory (TDT) [14, 15]. Subtle differences in how each grade is correlated with patient outcomes have led to multiple modifications of the TICI scoring metrics, such as modified TICI (mTICI) [16], extended TICI (eTICI) [17] and expanded TICI [18], of which eTICI is the most widely used [10, 17–19]. eTICI consists of 6 categories:

- **eTICI 0**: no reperfusion
- **eTICI 1**: blood flow past the obstruction, but no reperfusion of brain tissue
- **eTICI 2A**: $<$50% of TDT reperfused
- **eTICI 2B**: 50% - 90% of TDT reperfused
- **eTICI 2C**: 90-99% of TDT reperfused
- **eTICI 3**: complete TDT reperfusion

Despite their widespread use in clinical practice, TICI scoring metrics have several shortcomings. First, the subjective and error-prone nature of visual reperfusion estimation results in moderate to substantial inter-observer variability [20]. This variability may be further compounded by the confusion existing between the concepts of reperfusion and recanalization in clinical practice. Reperfusion refers to the antegrade restoration of a capillary blush, whereas recanalization is the restoration of blood flow, which is necessary but not sufficient for reperfusion [21, 22]. This inter-observer variability poses a significant challenge for structural EVT outcome comparison. Second, the use of various modified TICI scales can reduce the ability to compare treatment success across different clinical trials. Third, the TICI grading system consists of unevenly spaced coarse ordinal scales that confine a wide range of reperfusion percentages per grade. As studies have shown that the functional outcome of EVT is associated with

the increasing degree of reperfusion [23–26], this scoring system could therefore be inadequate for accurately assessing treatment success.

1.1.4. Automated TICI grading

To mitigate these issues, a fully automated deep-learning-based reperfusion assessment method, referred to as 'autoTICI', was developed and investigated by our research group [27, 28]. The proposed autoTICI pipeline, derived from the method of visual eTICI scoring, consists of four key components: phase classification, motion correction, perfusion segmentation, and TICI scoring. First, a DSA acquisition is divided into four vascular phases — non-contrast, arterial, parenchymal, and venous — using a convolutional neural network (CNN). Non-contrast and venous phase frames are subsequently removed from the DSA sequence. Second, affine registration is employed to correct for motion artefacts in the remaining frames. From these aligned frames, a 2D minimum intensity projection (MinIP) is generated and segmented into vessels, reperfused and non-perfused areas, including the background. Then, a brain mask is segmented on the post-EVT MinIP using a method called atlas registration, after which the pre-EVT is aligned with the post-EVT MinIP. For both steps, initial registration is achieved by aligning landmarks detected at the superior part of the ICA. Comparing the reperfused areas in the pre- and post-EVT MinIPs enables calculation of a continuous TICI output, reflecting the reperfusion percentage. Detailed information on the autoTICI pipeline is provided in a study by Su et al. [27]. Recent studies have shown that the proposed autoTICI approach showed a strong correlation with eTICI, with an average area under the curve (AUC) score of 0.81. Moreover, autoTICI demonstrated comparability to eTICI in terms of clinical outcome prediction, achieving a prediction accuracy of 0.66 compared to 0.62 for eTICI [27, 29].

These results suggest that autoTICI could offer a more uniform and reliable method for reporting EVT outcomes and enhance the comparability of treatment success across clinical trials. However, further improvements are essential before autoTICI's implementation can be considered viable:

1. Artificial intelligence (AI) applications often fail to achieve widespread adoption in clinical practice; no matter how effective autoTICI may be, this barrier must be overcome if the method is to succeed [30, 31].
2. Though it is anticipated that implementing autoTICI will decrease inter-observer variation, this has not yet been investigated.
3. AutoTICI's robustness still requires improvement, as in 30% of cases the model prediction is unreliable due to unsuccessful atlas registration [28].

1.2. Objectives and overview

Our research group aims to improve autoTICI and quantify its value, with the ultimate goal of its widespread adoption in clinical practice. The three specific objectives of this thesis address the three obstacles observed in section 1.1.4:

1. Assess the clinical implementability of autoTICI using the AI Funnel approach (described in section 2.1) developed by Erasmus MC to achieve the successful integration of AI applications into clinical settings.
2. Demonstrate the effect of autoTICI on inter-observer agreement.
3. Increase the robustness of autoTICI by improving the atlas registration step.

The remainder of this thesis is organized as follows. Chapters 2-4 each cover one objective and are written to stand on their own, to facilitate independent reading. Chapter 2 describes the methods and results of a pilot study on the clinical implementation of autoTICI in the workflow of interventional radiology. Chapter 3 describes the methods and results of a core-laboratory study to assess the inter-observer agreement and efficiency of eTICI scoring with and without autoTICI. Chapter 4 presents an automated approach for directly segmenting the brain mask on cerebral DSAs, bypassing the current atlas registration step. The results are compared against the atlas registration method employed in autoTICI. Chapter 5 unites this research in a general discussion, synthesizing the results from the previous chapters and offering recommendations for future research on autoTICI.

2

Clinical implementability of automated TICI scoring

2.1. Introduction

The application of artificial intelligence (AI) in medical imaging, particularly in radiology, is a rapidly growing field [32, 33]. Despite the promising technical performance and the increasing number of AI applications, many studies have shown that these innovations often fail to be implemented and utilized in clinical practice [30, 31]. This can be attributed to several factors, including unstructured implementation processes, uncertain added value for clinical practice, and significant variance in acceptance and trust among users [30]. To address these challenges, the Erasmus MC has developed a structured approach, termed the 'AI Funnel', for integrating AI applications into clinical settings, which is detailed in a recently published report [34]. Figure 2.1 illustrates the AI Funnel and summarizes the steps necessary to adopt AI technologies in healthcare practices successfully [34].

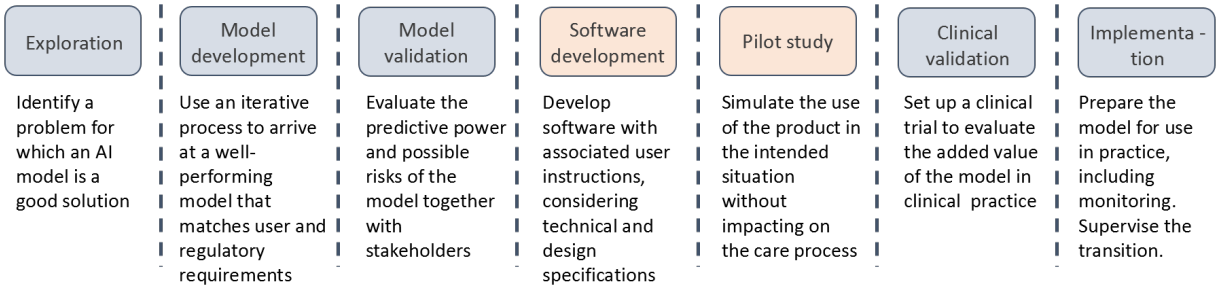


Figure 2.1: Summary of the steps proposed by the Erasmus MC to be followed to ensure the effective integration of AI applications into clinical practice [34]. AutoTICI is currently in the software development and pilot study stages of the AI Funnel, which are depicted in orange.

In two recent studies, autoTICI[1](1.1.4) has undergone model validation [27, 28], demonstrating its readiness for the next phases in the AI Funnel: software development and the initiation of a pilot study (Figure 2.1). The software development phase concentrates on creating a technically feasible and user-friendly infrastructure for deploying autoTICI in the intended clinical environment. The pilot study assesses the use and clinical implementability of autoTICI within the workflow of the intended setting without impacting the patient care process.

2.2. Objectives and overview

This study aims to investigate the technical feasibility and clinical implementability of autoTICI within the workflow of interventional radiology and to provide future recommendations for the next steps regarding

[1]This is one chapter in a larger thesis. See chapter 1 for a full background.

its clinical implementation.

We achieved this by first investigating the workflow of the intended clinical setting and determining autoTICI's role. To integrate autoTICI into this clinical workflow, we implemented autoTICI in a technical infrastructure created by Philips Medical Systems called the Open Innovation Platform (OIP). Besides that, we designed a user-friendly interface for displaying the results of autoTICI within the OIP to ensure smooth adaptation into clinical practice. Finally, we conducted a pilot study, titled 'autoTICI to clinic', to evaluate the use of autoTICI in the clinical workflow of the interventional radiology department and gathered the opinions of interventional radiologists on the challenges, opportunities and potential of autoTICI. Appendix A provides a detailed description of the OIP infrastructure and outlines the steps taken to integrate autoTICI into it. Appendix B describes the process involved in creating the user interface (UI) design for autoTICI, including the list of design requirements and a description of the worked-out designs. Section 2.3 outlines the workflow of the interventional radiology department and describes the pilot study, including the patient inclusion, technical infrastructure, imaging data, and the employed method.

2.3. Materials and methods

2.3.1. Clinical workflow for acute ischemic stroke

The intended clinical setting for the implementation of autoTICI is within the angiography suite of the interventional radiology department, for which Figure 2.2 outlines the identified workflow steps. The clinical workflow is described from the moment a stroke patient is transferred from the emergency room to the angiography suite and concludes upon the patient's transfer to the neurology department following the endovascular thrombectomy (EVT)[2].

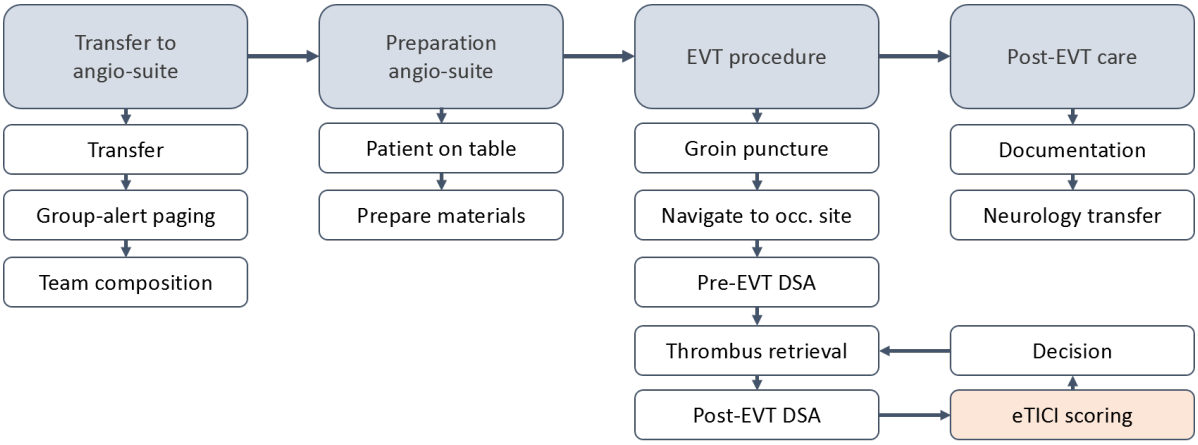


Figure 2.2: Schematic overview of the clinical workflow steps for the EVT treatment of acute ischemic strokes in the interventional radiology department [35]. The orange step depicts where autoTICI could be implemented in this workflow. *EVT = Endovascular Thrombectomy; DSA = Digital Subtraction Angiography; eTICI = extended Thrombolysis In Cerebral Infarction*

Once a patient is diagnosed and EVT is indicated, they are transferred to the angiography suite as quickly as possible. A group-alert paging system for acute stroke ensures that all team members are notified, can prepare the room for the EVT, and are ready to receive the patient. The team typically consists of one or two interventional radiologists who perform the EVT and two technicians who prepare and handle necessary materials and operate the angiography imaging system during the EVT. Additionally, an anesthesiologist is present to administer sedation if needed and to support the interventional team with patient management. The patient is positioned on the angiography table, where a sterile field and all the materials required for the EVT are prepared.

The EVT procedure starts with a groin puncture, with a femoral sheath inserted in the common femoral artery. Under fluoroscopic guidance, an inner catheter is advanced and positioned in the internal carotid artery (ICA). A pre-EVT digital subtraction angiography (DSA) run is then performed to confirm the

[2]Figure 2.2 aims to visualize the distinct steps involved in the workflow, acknowledging that some details are left out for clarity.

location of the occlusion. Subsequently, a microcatheter is navigated to the distal side of the occlusion, where minimal contrast is injected to verify its correct positioning and rule out any perforations. For thrombus removal, a stent retriever is typically employed in combination with negative suction to prevent distal emboli. After retrieving the stent, a post-EVT DSA run is performed, which is used for determining the extended Thrombolysis In Cerebral Infarction (eTICI) score, with a score higher than 2B indicating good reperfusion. Based on this assessment, a decision is made whether to conclude the procedure after successful thrombus removal, make another attempt to achieve a higher eTICI grade, or terminate after an unsuccessful attempt to avoid complications. During the workflow step depicted in orange in Figure 2.2, autoTICI could be implemented as a decision support tool for interventional radiologists by providing an objective score of the measured reperfusion. After the EVT treatment, the patient is transferred to the neurology unit for further care [35–38].

2.3.2. Pilot study workflow

Figure 2.3 shows a schematic overview of the workflow steps of the 'autoTICI to clinic' study. The diagram begins with the clinical workflow steps taken when a stroke patient receives EVT treatment, followed by the procedure for obtaining written deferred consent. The technical workflow of the diagram illustrates the background tasks performed by the OIP: streaming pseudonymized images acquired during the EVT to a cloud storage called the 'IGT cloud'. After obtaining deferred consent, these images are uploaded to the OIP, where data preparation occurs and the autoTICI algorithm is run. In the final phase, an interview is conducted with the interventional radiologist who performed the EVT. During this interview, the radiologist completes the System Usability Scale (SUS) and a questionnaire assessing the clinical implementability of autoTICI, with a comparative analysis against the current eTICI scoring method. The study method is further detailed in the subsections below.

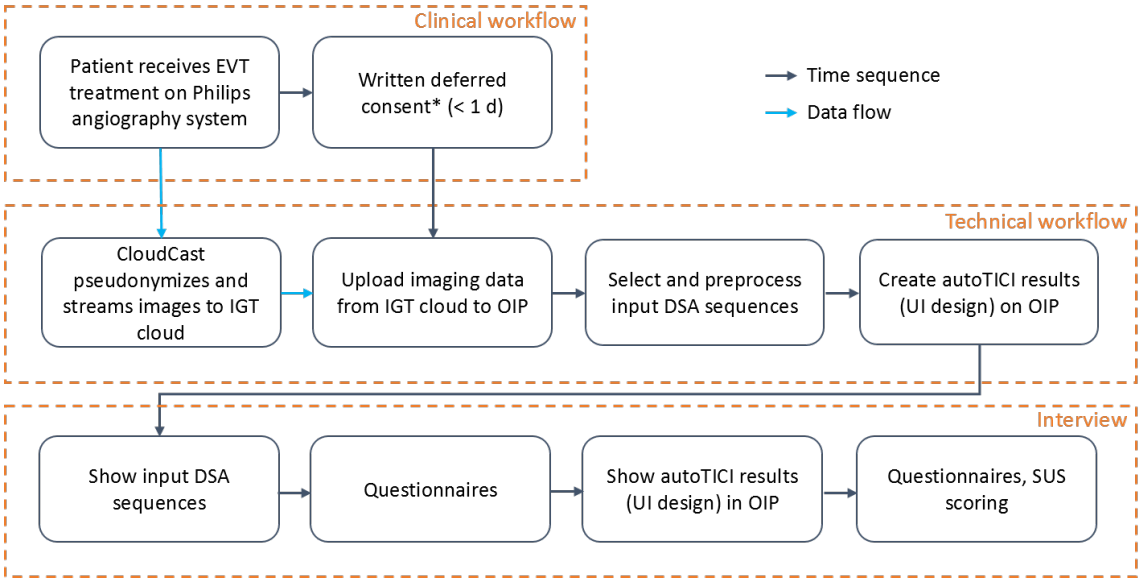


Figure 2.3: Schematic overview of the workflow steps of the 'autoTICI to clinic' study. The diagram consists of a clinical and technical workflow phase, followed by an interview. Data flow and time steps are depicted by light and dark blue arrows, respectively. *As 'autoTICI to Clinic' is part of the DIVINE study, deferred consent was obtained for participation in the DIVINE study [39]

2.3.3. Patient inclusion

'AutoTICI to clinic' was a single-centre, prospective cohort study conducted in the Erasmus MC. The study was established as a part of a broader study, the DIVINE registry, which obtained approval from our non-WMO institutional ethics committee [39].

Patients were included in the 'autoTICI to clinic study' if (1) they underwent EVT treatment for acute ischemic stroke (AIS) in the anterior circulation, and (2) they were treated using the Philips Xper FD20/10 angiography system, which was connected to the OIP. Patients were excluded if the interventional team encountered failed intracranial access. After transfer to the neurology department, a deferred consent

form was presented to the patient or their representative(s), which authorized the use of all imaging data acquired during the EVT. Patients were given one hour to consider their participation, considering the rapid transfers of stroke patients to other hospitals that frequently occur. Eligible subjects were identified using PACS Vue and the electronic health record. Once the written deferred consent was obtained, the interventional radiologists who performed the EVT for the included patient were approached and asked for their permission to participate in an interview.

2.3.4. Technical feasibility

To integrate autoTICI into the intended clinical workflow of the interventional radiology department, we implemented autoTICI in the OIP, an infrastructure hosted by Philips Medical Systems (Best, the Netherlands). The OIP functions by securely streaming data from the imaging system via 'CloudCast' hardware to a cloud storage called the 'IGT cloud'. After consent is obtained, imaging data is uploaded to the 'cloud', where images are received, necessary computations are performed for the execution of applications, and results are sent back to the CloudCast. A detailed description of the OIP and the autoTICI integration can be found in Appendix A.

During 'autoTICI to clinic', we investigated the technical feasibility of running autoTICI using the OIP by determining the reliability and calculation time for each included patient. We evaluated the reliability of the OIP by comparing the number of DICOM images and frames per patient received in the cloud with the imaging data in PACS Vue. The calculation time was defined as the time required for running autoTICI on the OIP, using an NVIDIA A10G GPU with CUDA version 12.2. This encompassed the total computational time, as well as the time for each component of the autoTICI pipeline: DSA preprocessing, landmark detection, atlas registration, pre-post registration, and phase selection [27]. Additionally, a log file was maintained to record any issues encountered during the clinical implementation. This log was regularly communicated to Philips to facilitate further improvements to the CloudCast system and the OIP.

2.3.5. Data preparation

For each included patient, the data collected for this study encompassed all imaging data acquired during the EVT procedure by the Philips angiography system located at the Erasmus MC (Philips Allura Xper FD20/10). This data included fluoroscopy images and DSA sequences in the AP and lateral views, which were pseudonymized and securely streamed to the IGT cloud, where they were temporarily stored for one week. After consent was obtained, the imaging data was uploaded to the cloud. Within the cloud, scripts were executed to filter out the fluoroscopy images, remove the unsubtracted from the subtracted DSA images, and compile the streamed single-frame DICOM files into multi-frame DSA sequences (Appendix A). Subsequently, four DSA sequences were manually selected to serve as input for autoTICI: the pre- and post-EVT DSA sequences in the AP and lateral view. The pre-EVT DSA was defined as the run before the first thrombus removal attempt, while the post-EVT DSA was the run after the final attempt. Finally, the autoTICI pipeline was executed on the OIP, with the results being produced according to the configuration of the selected UI design (Appendix B).

2.3.6. Clinical implementability interview

To investigate autoTICI's clinical implementability, we conducted interviews with interventional radiologists after they performed the EVT on the included patients. These interviews aimed to assess the usability and value of autoTICI within the existing EVT workflow (Figure 2.2), the radiologists' confidence and reasoning when scoring eTICI with and without the addition of autoTICI, and whether autoTICI feedback might influence the number of thrombus removal attempts. To ensure accurate recollection of procedural details and decision-making, interviews were conducted within one week post-EVT.

The interview consisted of two phases. In the first phase, the DSA sequences were presented, and interventional radiologists completed a 0-5 Likert scale assessing their confidence in visual eTICI scoring, along with a questionnaire to identify factors contributing to potential doubt. In the second phase, autoTICI results were displayed alongside the DSA sequences, and the radiologists repeated the confidence assessment. Additionally, they completed the System Usability Scale (SUS) to evaluate the clinical usability of autoTICI. The SUS is a straightforward ten-item Likert scale that addresses various usability aspects, including the need for support, confidence, training, and system complexity [40]. Furthermore, a questionnaire assessed the added value of autoTICI and whether the radiologists would

consider adjusting the number of attempts during EVT. An overview of the interview questions can be found in Appendix C.

2.3.7. Evaluation

The outcomes regarding technical feasibility, reliability and calculation time, were reported as numerical variables. Reliability was expressed as the total amount of streamed images and a percentage, while computational time was reported using the mean and standard deviations. Regarding clinical implementability, the following guidelines were used for interpreting the SUS scores determined for autoTICI:

SUS Score	Grade	Adjective Rating
> 80.3	A	Excellent
68 – 80.3	B	Good
68	C	Okay
51 – 68	D	Poor
< 51	E	Very poor

Table 2.1: System Usability Scale (SUS) interpretation guidelines [41]

2.4. Results

This section presents the results of the 'autoTICI to clinic' study, discussing patient inclusion, outcomes related to technical feasibility, and findings from the interviews on clinical implementability, organized into the following themes: clinical usability, confidence in eTICI scoring, and added value of autoTICI.

2.4.1. Patient inclusion

We included four stroke patients (median age 76 years, 50% female) who underwent EVT on the Allura Xper FD20/10 system. Table 2.2 details the characteristics of each included patient. All EVT procedures were successful. ATC004 was the only patient requiring general anaesthesia due to an epileptic insult during the procedure. For patient ATC004, the occlusion had already resolved with intravenous thrombolysis (IVT) as identified on the pre-EVT DSA, achieving 100% reperfusion. Therefore, no further attempts were made to remove the thrombus, as depicted by 0 passes in Table 2.2.

Table 2.2: Summary of patient characteristics (n = 4).

	ATC001	ATC002	ATC003	ATC004
Gender (female/male)	Female	Male	Male	Female
Age (years)	88	72	80	30
Occlusion location	M1	M2	M2	M1
Anesthesia Type	Local	Local	Local	General
No. passes during EVT	2	1	1	0*
Post-EVT eTICI grade (reperfusion (%))	2C (90-99)	3 (100)	3 (100)	3 (100)
Mean reperfusion percentage autoTICI (AP, lateral) (%)	96 (93, 99)	58 (59, 56)	54 (57, 52)	100 (100, 100)**

*The occlusion resolved with intravenous thrombolysis, achieving 100% reperfusion pre-EVT, so no further attempts were made to remove the thrombus (0 passes). **As autoTICI requires both pre-and post-EVT DSAs as input, the pre-EVT DSA was used twice, resulting in a reperfusion difference of 0%. However, as the pre-EVT eTICI score was 100%, this value is tabulated.

AP: anteroposterior; EVT = Endovascular Thrombectomy

2.4.2. AutoTICI performance

Table 2.2 presents the reperfusion percentages calculated by autoTICI for the included patients and Figure 2.4 shows the corresponding segmentations. The DSA sequence of patient ATC001 exhibited significant motion artefacts, which initially resulted in unusable autoTICI predictions. This issue was resolved by manually removing the motion frames from the DSA. Subsequently, the autoTICI-derived reperfusion percentage of 96% aligned with the reported eTICI grade of 2C (Table 2.2). As illustrated in Figure 2.4, the brain masks for ATC001 were correctly delineated by autoTICI. However, in the

AP view, pixels at the superior border of the brain mask were mistakenly classified as non-perfused. For patients ATC002 and ATC003, the lateral brain masks were undersized, missing part of the MCA vascular territory, and the AP brain mask for ATC002 did not account for slight head rotation in the image. For ATC002 and ATC003, the distal parts of the brain masks were incorrectly segmented as non-perfused, leading to an underestimation of the actual reperfusion percentage as reported in Table 2.2. For ATC004, as no post-EVT DSA was performed, the pre-EVT DSA was shown during the interview. Figure 2.4 shows that the lateral mask was too large for ATC004, extending in the skull, which led to incorrect segmentation of the Target Downstream Territory (TDT).

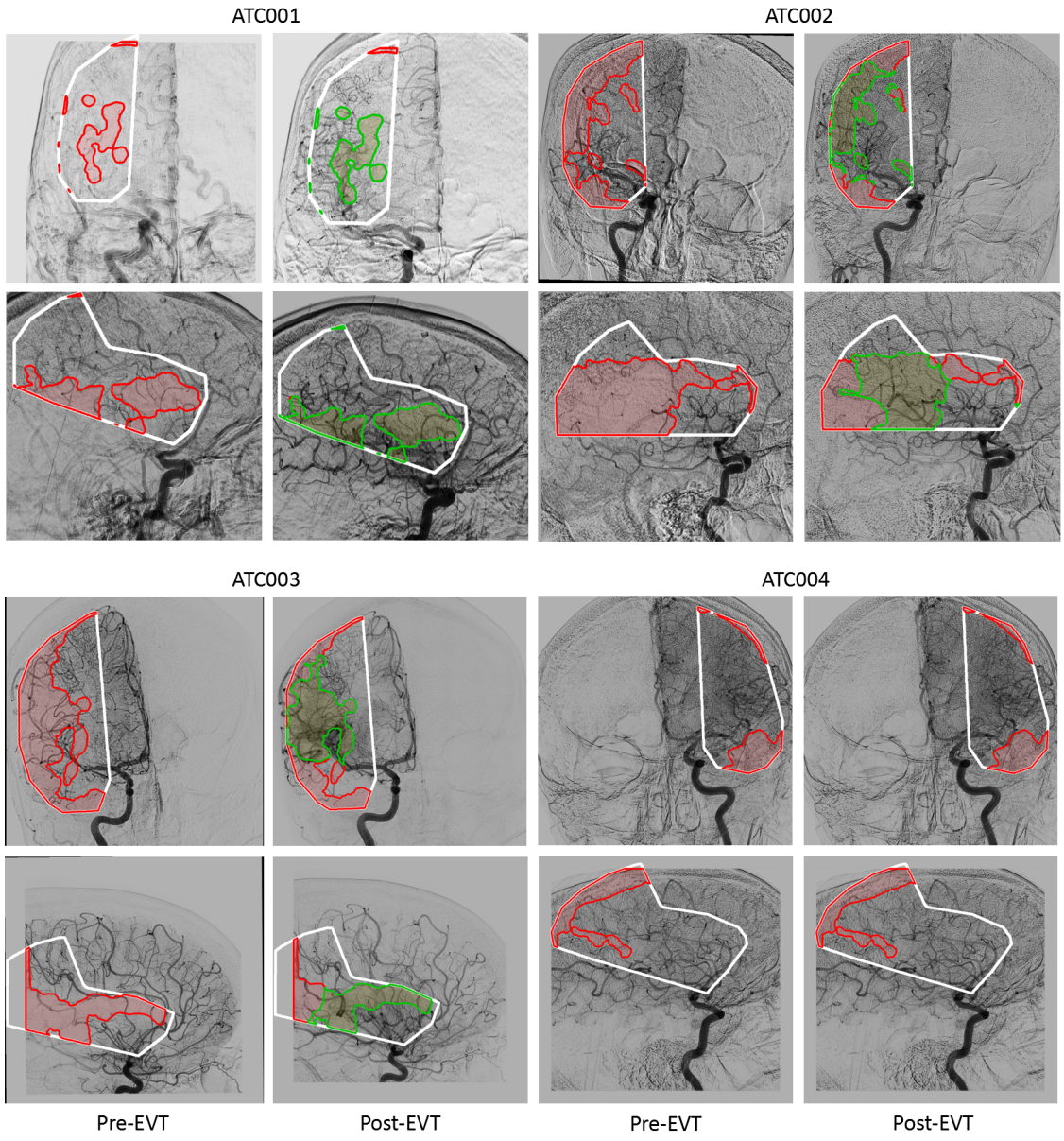


Figure 2.4: AutoTICI predictions for the four included patients. Reperfused and non-perfused segmentations are depicted in green and red, respectively.

2.4.3. Technical feasibility

Due to technical difficulties with connecting the CloudCast to both channels of the biplane angiography system (Allura Xper FD20/10, Philips), only the frontal channel could be installed, allowing only AP images to be streamed to the OIP. Lateral DSA sequences were retrieved from PACS Vue and uploaded

manually to the OIP. All data processing was done by LS.

Due to hardware issues, CloudCast was non-operational for the last two patients, limiting the evaluation of technical feasibility to patients ATC001 and ATC002, as shown in Table 2.3. Regarding reliability, the number of DSA runs and frames per run received in the OIP matched 100% with the imaging data in PACS Vue. The average total computational time for running autoTICI on the OIP was within a 5-minute interval, with the AP view requiring approximately one minute longer than the lateral view. The atlas registration step was the most time-consuming in the autoTICI pipeline, followed by the landmark registration step, while phase selection, pre-processing, and pre-post registration were relatively quick, taking only a few seconds for each view.

Table 2.3: Streaming reliability and computational times for running autoTICI on the OIP for each patient, with separate times reported for the AP and lateral views.

	ATC001	ATC002	ATC003	ATC004
Reliability (%)	100	100	-	-
-No. frames	86	69	-	-
-Acquisition rate (fps)	3	3	-	-
Mean computational time (AP, Lateral) (s)	280 (171, 109)	284 (175, 109)	283 (174, 109)	284 (172, 112)
-Pre-processing	8 (4, 4)	4 (1, 3)	5 (2, 3)	5 (1, 4)
-Landmark prediction	63 (32, 31)	64 (32, 32)	74 (38, 36)	76 (39, 37)
-Atlas registration	187 (124, 63)	189 (127, 62)	185 (125, 60)	186 (126, 60)
-Pre-post registration	6 (3, 3)	6 (3, 3)	6 (3, 3)	6 (3, 3)
-Phase selection	16 (8, 8)	11 (6, 5)	14 (7, 7)	10 (5, 5)

OIP = Open Innovation Platform; fps = frames per second; AP = Anteroposterior

2.4.4. Clinical implementability

Interviews

We conducted interviews for three of the four included patients: the interview for ATC003 could not be scheduled within a week post-EVT and was therefore not conducted according to protocol. The EVTs of ATC002 and ATC004 were performed by the same clinician (obs 3). For ATC001, two interventional radiologists (obs 1 and obs 2) were present during the EVT, and both were interviewed, resulting in a total of four interviews with three distinct clinicians. All interviews were conducted by LS. Obs 1 and 2 were specialized in body procedures, with 1 and 2 years of experience in independently performing EVTs, respectively, and obs 3 was a neuro-interventional radiologist with extensive experience (9 years) in performing EVT procedures. Table D.1 (Appendix D) describes the characteristics of the included clinicians and summarizes their responses given during the interview.

Clinical usability

Figure 2.5 presents the SUS scores assigned to autoTICI during each interview. Table D.2 (Appendix D) displays the corresponding Likert scales awarded for each usability aspect. The average SUS score for autoTICI was 81.3, indicating excellent usability within the workflow of interventional radiology [41]. Higher SUS scores were assigned to ATC001 compared to ATC002 and ATC004, with a notably lower score in the second interview of observation 3, particularly concerning the questions on consistency (Q6) and confidence (Q9) (Table D.1, Appendix D). Obs 3 attributed this decline to the inaccurate autoTICI prediction for patients ATC002 and ATC004 (Figure 2.4). AutoTICI received maximum SUS scores from all clinicians for questions regarding simplicity, ease of use, and intuitiveness (Q2, Q3, Q10). All clinicians found the user interface easy to understand and expected that autoTICI would be quick to learn when implemented (Q7, Q8). In the first three interviews, the clinicians stated they would want to use autoTICI frequently in the angiography suite (Q1). During both interviews conducted for ATC001, autoTICI was rated highly consistent, with an average score of 4.5 out of 5 (Q6). In contrast, for ATC002 and ATC004, the consistency scores dropped to 2 and 1, respectively, due to the inaccurate predictions of autoTICI (Figure 2.4). The question about the need for technical support when using autoTICI in the clinic (Q4) was challenging to answer since ‘autoTICI was not yet integrated in the clinic’ (obs 1). Obs 2 mentioned that ‘if autoTICI were automatically integrated, minimal support would be needed, but if

manual steps are required, assistance with implementation would be necessary'.

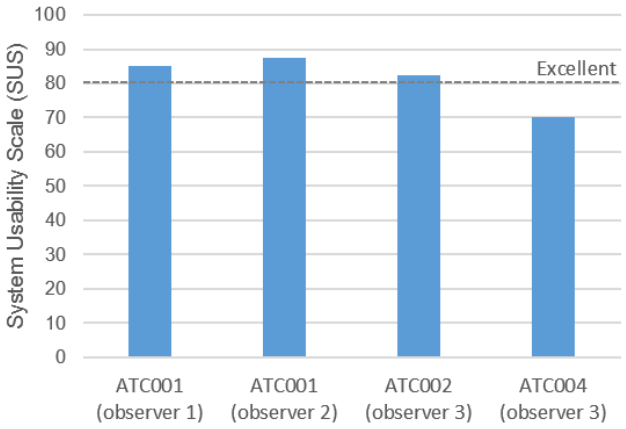


Figure 2.5: System Usability Scale (SUS) scores assigned to autoTICI during each interview.

Confidence in eTICI scoring

Confidence in eTICI scoring decreased from 4 to 3.75 with the addition of autoTICI feedback (Figure 2.6). For ATC002, obs 3 experienced doubt in visual eTICI scoring due to a 'small borderline thrombus that made it difficult to distinguish between 2B and 2C scores'. For ATC004, doubt arose because 'IVT had resolved the occlusion, but a small perfusion defect was still present'. AutoTICI did not improve confidence for ATC002 and even decreased it for ATC004 due to repeated inaccurate reperfusion predictions. For ATC001, there was no doubt regarding visual eTICI scoring, as the 'case was straightforward with no complications' (obs 1). This is reflected in Table D.1 (Appendix D), where the clinicians chose not to change their eTICI scores or perform another attempt based on autoTICI feedback. However, obs 1 also mentioned that 'during challenging situations — like performing EVT at night, when fatigued, or under stress - maintaining a consistent score can be difficult' and obs 2 added that 'eTICI scoring can be inconsistent due to conceptual confusion regarding the TDT'. Regarding autoTICI, they both expressed doubts about the predicted reperfusion percentage in the AP view, due to a segmentation error at the distal part of the brain mask. Nevertheless, they still suggested that in such cases, autoTICI could increase confidence, as reflected by their Likert scores either remaining at 4 (obs 1) or increasing from 4 to 5 (obs 2).

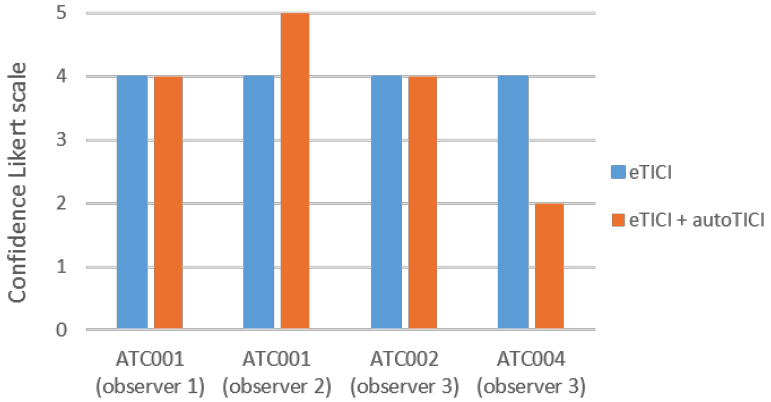


Figure 2.6: Confidence in eTICI scoring, measured on a Likert scale of 0-5, with scores indicated in orange for assessments with autoTICI feedback and in blue for assessments without it.

Added value of autoTICI

All clinicians recognized the intended use of autoTICI as a decision-support tool in the angiography suite. Obs 1 noted that 'autoTICI could be valuable as decision support when I am alone, stressed, or tired, or in difficult cases'. Obs 1 suggested a 5-minute time window for autoTICI processing (representing the waiting time during suction), while obs 2 recommended 1 minute. Obs 3 noted that autoTICI could add value not only as a clinical tool but also as a research tool to provide a consistent and more objective metric for comparing stroke-related studies. In clinical settings, autoTICI would be particularly supportive in cases of uncertainty (e.g., distinguishing between eTICI grades 2B and 2C or in exceptional cases like ATC004). However, autoTICI currently lacks the necessary robustness and accuracy, as reflected by decreased confidence and increased inconsistency. Therefore, 'before clinical implementation, autoTICI must become more robust and accurate' (obs 3).

2.5. Discussion

2.5.1. Interpretation of results

Regarding technical feasibility, the CloudCast hardware reliably streamed all DSA frames to the OIP, acquired at a frame rate of 3 fps, which is the rate typically used in clinical practice. AutoTICI was processed on the OIP with an average computational time of 283 ± 1.90 seconds, with the AP view taking about one minute longer than the lateral view. This discrepancy is attributed to the atlas registration step, which took longer for the AP view due to the higher number of AP atlases involved. The total computational time fits the 5-minute time window mentioned by obs 1 during the clinical implementation interview. However, this interval was based on the waiting time during suction. Since the post-EVT DSA, required as input for autoTICI, is acquired after thrombus retrieval, this time window does not apply to autoTICI. Obs 2 suggested a maximum waiting time of 1 minute for autoTICI in a clinical setting, highlighting the need to reduce its computational time. Most time was consumed by atlas registration (65.7%) and landmark prediction (24.5%), making these steps primary targets for improving speed.

AutoTICI received an average SUS score of 81.3, indicating excellent usability within the intended workflow, particularly for simplicity, intuitiveness, and ease of use. However, we identified issues which resulted in low SUS scores for confidence and consistency, hindering clinical implementation. These problems were mainly due to unreliable atlas registration, resulting in incorrect brain masks in five out of eight images across all patients (Figure 2.4). Another identified problem was the mislabeling of pixels at the mask borders as non-perfused, which occurred in three out of four patients. This issue stems from the lower contrast intensity in the distal brain regions, which is due to the reduced tissue density and the smaller amount of contrast agent that reaches these areas. Consequently, these regions fail to exceed the segmentation threshold, leading to an underestimation of the reperfusion percentage. Despite these issues, clinicians expressed a desire to use autoTICI frequently and recognized its value as a decision support tool during EVT, particularly during night shifts, in stressful or tiring situations, or for difficult cases, such as distinguishing between 2B and 2C scores. AutoTICI could also help resolve conceptual confusion about defining the TDT area. However, it may not provide significant benefits in straightforward EVT cases. Furthermore, it is important to note that the clinicians' interest in autoTICI presupposed a future, more robust version, without the limitations and errors experienced during the pilot study. Section 2.5.2 explores the improvements required to achieve the required accuracy and efficiency.

2.5.2. Further improvements for clinical implementation

First, the atlas registration step requires improvement, as it caused the most issues during the 'autoTICI to clinic' study and had the longest processing time. Atlas registration is a challenging task because it requires inter-patient alignment, which is particularly difficult for DSA images due to variations in vascular anatomy and the lack of anatomical information, as DSA only visualizes the vessels. This complexity is further increased in the context of stroke, where additional variations in perfusion levels are caused by the occlusion [42–44]. Therefore, we recommend bypassing the atlas registration step in autoTICI by directly segmenting the brain masks on the MinIPs, which is further discussed in Chapter 4. Eliminating atlas registration would also remove the need for landmark detection, which currently serves as its initial registration step, further reducing computational time. To further minimize computational time, we suggest exploring the removal of landmark detection in the pre-post registration step. Alternative

registration techniques, such as SIFT, which have shown promising results in similar registration tasks in other studies, could offer a potential solution [44]. To potentially address the issue identified in the perfusion segmentation step, we suggest exploring the use of adjustable thresholds based on depth or developing region-specific thresholds for areas with varying contrast intensities.
Second, further automation of autoTICI is necessary for clinical implementation. The current version requires users to manually specify input sequences, view, hemisphere, occlusion location, and head orientation. Automating the detection of these parameters would reduce manual inputs, making autoTICI more suitable for the clinical workflow. To accommodate case-specific variations, we recommend keeping the selection of input DSA sequences a manual step.
Third, we recommend continuing to use the OIP for the clinical implementation of autoTICI, given its potential to enable reliable real-time image data streaming in the angiography suite. However, since the pilot study was not conducted during the procedure, the final steps of the OIP pipeline, such as sending results from the OIP back to CloudCast and displaying them on the screen in the room, were not done. Investigating these steps in a clinical setting would be valuable for future research.

2.5.3. Limitations

The results of this study should be interpreted considering several limitations. The first limitation was the small sample size (n = 4), leading to a descriptive case analysis rather than a population-based evaluation. A larger sample size would have provided greater clinical variability (different eTICI grades and occlusion locations) and a broader range of scenarios to better assess the added value of autoTICI. Furthermore, conducting interviews with more clinicians with varied specialities and experience levels could offer further insights and opinions valuable for improving autoTICI and enhancing its clinical implementation.
Second, due to issues with CloudCast, we could only assess the technical feasibility for 2 patients, which is insufficient to draw definitive conclusions about streaming reliability. Further investigation is needed to determine if reliability remains at 100% when streaming a larger volume of images or images acquired at a higher frame rate.
The third limitation concerns the SUS scoring method for investigating autoTICI's usability. Since autoTICI is not yet implemented, physicians found it challenging to answer certain questions, such as the need for technical support when using autoTICI in the clinic (Q4). Although we provided a detailed explanation of the potential future application, this is not equivalent to the experience of actually using it in the angiography suite. Nevertheless, the SUS still provided a good indication of the strengths and weaknesses of autoTICI in terms of usability.

A strength of this research is that the interviews were conducted in close proximity to a real clinical setting, providing valuable insights into the intended use of autoTICI within its actual clinical context. Furthermore, by involving clinicians early in the process, both in the design phase and through interviews, they became familiar with autoTICI, which is advantageous for future implementation.

2.6. Conclusions

In conclusion, autoTICI was successfully integrated into the workflow of the interventional radiology practice. The pilot study showed that the performance of autoTICI is currently insufficient for implementation in clinical practice. Nevertheless, with an average SUS score of 81.3, autoTICI still demonstrated excellent usability within the clinical workflow, indicating its potential for future clinical use. Besides that, multiple interventional radiologists recognized the added value of autoTICI as a clinical decision-support tool in the angiography suite, particularly in challenging situations such as night shifts, high-stress conditions, or complex cases.

Future research should prioritize enhancing the reliability of autoTICI by improving the atlas registration and perfusion segmentation steps. Additionally, further automation of autoTICI is required for clinical implementation. Once these advancements are made, we recommend continuing the 'autoTICI to clinic' study to collect more data, utilizing the improved autoTICI method for a more accurate evaluation of its clinical implementability. Following these improvements, autoTICI will be ready for the next phase in the AI funnel, involving a clinical trial to establish its role as a decision support tool during EVT.

3

Inter-observer agreement in eTICI scoring: a comparative analysis with and without autoTICI

3.1. Introduction

3.1.1. Inter-observer variation in eTICI scoring

Extended Thrombolysis In Cerebral Infarction (eTICI) is a six-category grading scale used to assess the reperfusion status and thus the success of an endovascular thrombectomy (EVT) after acute ischemic stroke (AIS). Despite its widespread use in clinical practice, eTICI scoring suffers from moderate to substantial inter-observer variability due to its dependence on the clinician's visual assessment of digital subtractive angiography (DSA) images [20–22]. To address this and other drawbacks of eTICI scoring, our research group has developed a fully automated deep-learning-based variant, referred to as 'autoTICI'[1]. Due to its objective and standardized nature, autoTICI is expected to increase inter-observer agreement of eTICI scoring.

3.1.2. Goals and objectives

This study primarily aims to determine whether the implementation of autoTICI enhances inter-observer agreement in eTICI scoring, thereby potentially improving the consistency of reperfusion grading. The secondary objective is to evaluate whether autoTICI improves observer efficiency, measured as the time required for eTICI scoring of a patient's DSA sequences. Both objectives were assessed in a core-laboratory setting by comparing both inter-observer agreement and observer efficiency with and without autoTICI feedback.

3.2. Methods

This study took place in an imaging core-laboratory where four experienced interventional radiologists evaluated patient DSA sequences in two individual scoring sessions, one using the eTICI scale alone and one using the eTICI scale supplemented by autoTICI results. Their assessments were compared using standard statistical outcome measures to determine what effect autoTICI had on the inter-observer agreement and scoring efficiency. The rest of this section describes the study's technical setup; results are discussed in section 3.3.

3.2.1. Patient inclusion

Patients were selected from the MR CLEAN Registry (parts 1 and 2), a prospective multi-centre registry of consecutive patients treated with EVT in the Netherlands between March 2014 and November 2017 [4, 15]. Inclusion criteria required the presence of a large vessel occlusion on DSA within the internal

[1] This is one chapter in a larger thesis. See chapter 1 for a full background.

carotid artery (ICA), ICA-top, or the M1 and M2 segments of the middle cerebral artery (MCA). Patients were excluded if (1) their DSA sequences contained only unsubtracted images, (2) substantial motion artefacts were present, (3) the DSAs were of poor image quality, (4) there was a limited field of view (FoV) of the target downstream territory (TDT), (5) pre- and post-EVT DSA sequences were poorly registered by autoTICI, (6) atlas registration in autoTICI resulted in an unusable brain mask, or (7) no venous frames were classified by autoTICI.
For each included patient, a total of four DSA sequences, acquired before and after the EVT in AP and lateral views, were selected for the core-laboratory study.

3.2.2. Core-laboratory study

The four participating radiologists (observers 1 - 4) were blinded to all clinical data. The study, following a fully crossed design, consisted of two individual scoring sessions separated by two months to minimize recall bias [45, 46]. To further reduce bias, the DSA sequences were presented in randomized, different orders in each session.

3.2.3. eTICI scoring

For each scoring session, observers were provided with a local executable developed for the study, which the observers ran on their local computers. The executable displayed a DICOM viewer showing the four DSA sequences for each patient alongside a dropdown menu with the eTICI grades to select from. During the second session, the autoTICI results were displayed alongside the DSA sequences. The executable saved the grades and recorded the time needed to score each patient. Appendix E shows an example of the executable for both sessions.

3.2.4. Training session

To mitigate bias caused by observers' unfamiliarity with the executable and with autoTICI, as well as potentially differing definitions for reperfusion and the TDT, we conducted a training session prior to the study. The training session detailed the study design, provided precise eTICI scoring definitions, and explained the autoTICI model output. The observers also practiced on a training set of 10 patients to become familiar with the executable. We provided the following eTICI scoring definitions to the observers: reperfusion was defined as the antegrade restoration of capillary blush in the TDT. The TDT was defined as the occluded brain region that was supplied via antegrade blood flow prior to stroke onset.

3.2.5. Outcome measures

Inter-observer agreement

Inter-observer agreement for eTICI with and without autoTICI was quantified using the two-sided weighted kappa and two-way mixed intraclass correlation coefficient (ICC) [46]. The two-sided weighted kappa was used to assess absolute differences in eTICI scores between observers, facilitating direct comparison with other studies found in literature [47–51]. Agreement levels for the weighted kappa were categorized using Landis and Koch's guidelines: less than 0 indicates no agreement, 0–0.2 slight agreement, 0.2–0.4 fair agreement, 0.4–0.6 moderate agreement, 0.6–0.8 substantial agreement, and 0.8–1 almost perfect agreement [52]. A two-way mixed ICC was calculated for both absolute agreement, assessing differences between exact values, and consistency, evaluating how closely different raters provided similar rank orders [46]. This dual approach allowed for a comprehensive assessment of inter-rater reliability. ICC values below 0.5 indicate poor reliability, 0.5–0.75 moderate reliability, 0.75–0.9 good reliability, and above 0.9 indicate excellent reliability [53].

The two-sided weighed kappa was also determined for the dichotomized eTICI score, which is a categorization used during EVT to divide perfusion levels into failure (eTICI grades ≤ 2A) and success (eTICI grades ≥ 2B). This analysis aimed to provide insights into the potential impact of autoTICI on decision-making during EVT, such as whether to attempt further thrombus removal.

The change in eTICI scores between the two sessions was assessed by calculating the eTICI scores for each observer and across all observers. Scores were reported as either mean or median values, along with 95% Confidence Intervals (CIs) or Interquartile Ranges (IQRs), depending on the normality of the data. Statistical analysis was conducted using a two-sided paired Student's t-test for normally

distributed data or a two-tailed Mann-Whitney U-test for non-normally distributed results.

Additionally, the percentage of observer consensus for each eTICI grade was calculated by dividing the number of times all observers assigned the same grade by the total number of times that grade was assigned across all observers.

Time to score

The time required to complete the eTICI scoring, with and without the addition of autoTICI feedback, was reported using means and 95% CIs or medians and IQRs, depending on the normality of the results. Comparisons were made using the two-sided paired Student's t-test or the two-tailed Mann-Whitney U-test. Additionally, scoring times were calculated for each eTICI grade.

3.2.6. Sample size calculation

For the inter-observer agreement calculation, it was expected that the two-sided weighted kappa would increase by one category, moving from 'moderate to substantial' (κ = 0.60) to 'substantial to almost perfect' (κ = 0.80) [20, 51]. Considering a significance level (α) of 0.05 (95% confidence), a power (1-β) of 80%, and a proportion of positive ratings of 0.5, the sample size calculation indicated that a minimum of 126 patients was required to detect a significantly different kappa between eTICI with and without autoTICI feedback [54]. For the dichotomized eTICI score, a higher agreement was expected, resulting in a lower required sample size [54].

To calculate the number of patients needed to capture a significant change in scoring time between eTICI and eTICI combined with autoTICI, a minimum of 34 images is required. This calculation considered the same significance level (α) and power (1-β) as previously mentioned, with an expected scoring time of 60 seconds, a difference of 30 seconds and a standard deviation of 60 seconds.

3.3. Results

3.3.1. Patient inclusion

A total of 126 patients were selected from the MR CLEAN Registry (parts 1 and 2) using randomized stratification to reflect the distribution of occlusion locations and eTICI grades within the registry, as detailed in Table 3.1 [4]. During the core-laboratory study, the raters observed an additional occlusion in the anterior cerebral artery (ACA) for 4 patients, which caused confusion. Therefore, we decided to remove these patients from the analysis, resulting in a total of 122 patients (Table 3.1).

Table 3.1: Summary of patient characteristics (n = 122). AutoTICI reperfusion percentages are reported as median and Inter Quartile Ranges [IQRs].

No. patients	122
No. occlusion locations	
- ICA (%)	24 (20)
- M1 (%)	70 (57)
- M2 (%)	28 (23)
No. of eTICI grades	
- 0 (%)	22 (18)
- 1 (%)	4 (3)
- 2A (%)	23 (19)
- 2B (%)	27 (22)
- 2C (%)	11 (9)
- 3 (%)	35 (29)
Reperfusion autoTICI (%)	
- Overall	68 [34, 88]
- AP	79 [49, 93]
- Lateral	57 [28, 79]

ICA = Internal Carotid Artery; M1 = first segment of Middle Cerebral Artery (MCA); M2 = second segment of MCA; eTICI = extended Thrombolysis In Cerebral Infarction; AP = Anteroposterior

3.3.2. AutoTICI performance

Table 3.1 presents the medians and IQRs of the reperfusion percentages calculated by autoTICI for the included patients, separately for AP and lateral views, and the overall median. Notably, the reperfusion percentage was higher in the AP view than in the lateral view.

Figure 3.1 shows four examples of autoTICI segmentations on included patient data. AutoTICI demonstrated good performance in 67 patients (55%), as illustrated in the first example of Figure 3.1, where it accurately predicted an average reperfusion percentage of 53% (AP: 57%, lateral: 48%), corresponding to an eTICI grade of 2B. For this patient, inter-observer agreement was low in the session without autoTICI, with scores ranging from 2A to 2C, but with autoTICI feedback, all observers consistently predicted 2B. The brain registration step was generally accurate, due to the exclusion of cases with failed atlas registration. However, in 8 cases (6.5%), the brain mask was still too small, leading to an underestimation of the reperfusion percentage, as shown in the second example in Figure 3.1. Additional issues were noted in the perfusion segmentation step. In 9 cases (7.5%), vascular overprojection led to an undersized TDT, as depicted in the third example in Figure 3.1, which shows overprojection of the ACA territory. Another issue in the perfusion segmentation step was the mislabelling of pixels as non-perfused at the mask borders, as illustrated in the fourth example of Figure 3.1. This issue was consistently observed in 38 patients (31%), leading to slightly underestimated reperfusion percentages.

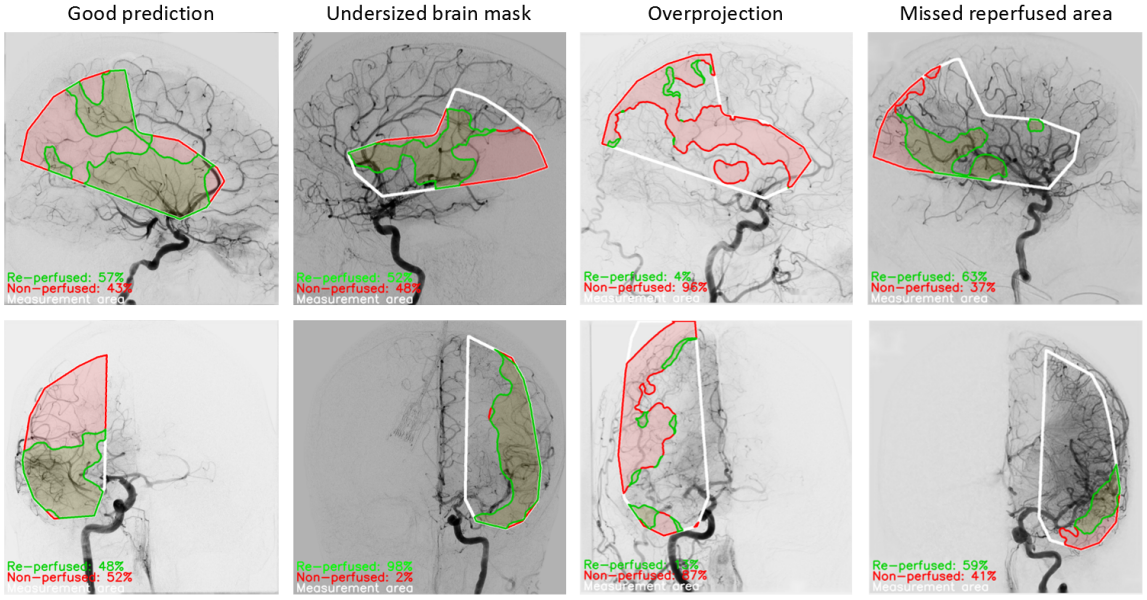


Figure 3.1: Four examples of autoTICI predictions; with a good prediction, undersized brain mask, overprojection leading to an undersized TDT and missed reperfused area at the mask borders. Reperfused and non-perfused segmentations are depicted in green and red, respectively.

3.3.3. Inter-observer agreement

Table 3.2 summarizes the inter-observer agreement analysis for eTICI scoring with and without autoTICI feedback. The two-sided weighted kappa test indicated substantial agreement for visual eTICI scoring alone (κ = 0.65) and with autoTICI feedback (κ = 0.67) (p = 0.99). Similar agreement levels were observed for the dichotomized scores, with a more pronounced increase in the weighted kappa when autoTICI feedback was included (p = 0.47). The two-way mixed ICC analysis showed 'good' inter-rater reliability for absolute measurements, which improved to 'excellent' for the consistency ICC. Across both scoring sessions, observer agreement was highest for eTICI grades 0 and 3, with lower agreements for the intermediate grades. The addition of autoTICI generally improved observer agreement rates, despite the minor reductions in agreement for grade 1 (from 15% to 13%) and grade 2C (from 16% to 15%).

Table 3.3 provides the medians and IQRs of the eTICI scores assigned by each observer when scoring with and without autoTICI feedback, along with the overall eTICI grade across all observers. The

Table 3.2: Results of the inter-observer agreement analysis for eTICI scoring without (session 1) and with (session 2) autoTICI feedback.

	Session 1: eTICI	Session 2: eTICI + autoTICI
Two-sided weighted kappa		
- eTICI score	0.65	0.67
- Dichotomized score	0.68	0.76
Intraclass Correlation Coefficient (ICC)		
- Consensus	0.91	0.92
- Absolute agreement	0.81	0.84
Observer agreement (%)		
- Overall	50	55
- eTICI 0	60	65
- eTICI 1	15	13
- eTICI 2A	36	42
- eTICI 2B	18	24
- eTICI 2C	16	15
- eTICI 3	42	47

eTICI = extended Thrombolysis In Cerebral Infarction

median eTICI grade across both scoring sessions was 2B [2A, 3], aligning with the median autoTICI reperfusion percentage of 68 [34, 88] (Table 3.1), which corresponds to an eTICI grade of 2B [2A, 2B]. In the second scoring session, the median score for observer 2 showed a slight decrease toward the overall median. However, the addition of autoTICI did not result in significant differences in the eTICI grades, as indicated by the Mann-Whitney U-test ($p = 0.24$, $p = 0.06$, $p = 0.63$, and $p = 0.45$ for observers 1 - 4, respectively).

Table 3.3: eTICI grades assigned by each observer when scoring without (session 1) and with (session 2) autoTICI feedback, along with the overall average eTICI grade. Values in the table represent the medians and Inter Quartile Ranges [IQRs].

	Grade eTICI	Grade eTICI + autoTICI
Observer 1	2B [2A, 2C]	2B [2A, 2C]
Observer 2	2C [2A, 3]	2B [2A, 2C]
Observer 3	2C [2A, 3]	2C [2A, 3]
Observer 4	2B [2A, 3]	2B [2A, 3]
Overall	2B [2A, 3]	2B [2A, 3]

eTICI = extended Thrombolysis In Cerebral Infarction

3.3.4. Time to score

Figure 3.2 visualizes the scoring times and IQRs of each observer in boxplots. Four outliers were excluded from the analysis due to scoring times exceeding 600 seconds. The addition of autoTICI feedback did not result in a significant increase in overall median scoring time, as indicated by the Mann-Whitney U-test ($p = 0.11$). However, for individual observers 1, 2 and 3, the scoring time did increase significantly ($p = 0.0001$, $p = 0.040$, and $p = 0.011$, respectively), whereas the time for observer 4 significantly decreased ($p = 0.031$). Figure 3.2 shows considerable variation in scoring times recorded per observer, particularly for observers 2 and 3, with this variation being more pronounced during the session with autoTICI feedback.

Table 3.4 presents the median scoring times for each assigned eTICI grade, the overall median scoring time across all observers, and the percentual difference between the two sessions. For most eTICI grades, the inclusion of autoTICI feedback led to increased scoring times, except for grades 1 and 3, where scoring times were reduced.

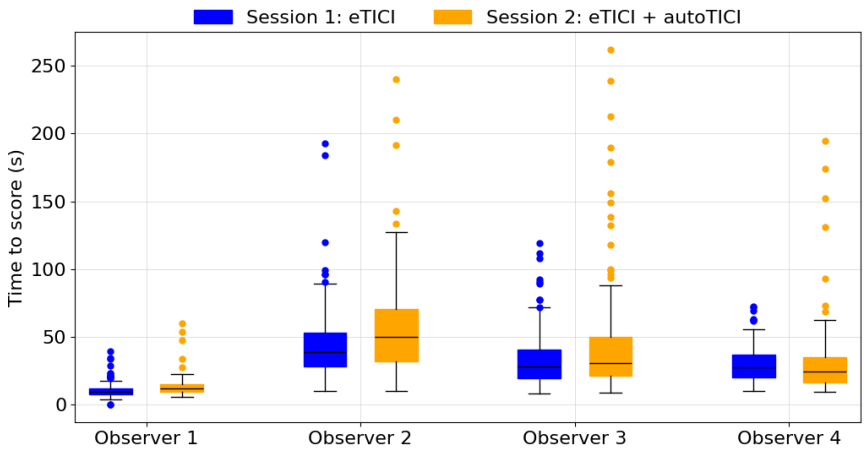


Figure 3.2: Time to score measured for all observers when performing eTICI grading without (session 1) and with (session 2) autoTICI feedback.

Table 3.4: Scoring time for each assigned eTICI grade, measured when scoring without (session 1) and with (session 2) autoTICI feedback, along with the percentual difference between the two sessions. Values in the table represent the median and Inter Quartile Ranges [IQRs].

eTICI grades	Scoring time eTICI (s)	Scoring time eTICI + autoTICI (s)	Difference in scoring time (%)
0	22 [12, 33]	24 [13, 40]	+9.1
1	31 [10, 35]	23 [15, 69]	-26
2A	21 [10, 35]	22 [13, 42]	+4.8
2B	31 [14, 41]	36 [22, 55]	+16
2C	30 [15, 45]	31 [16, 61]	+3.3
3	26 [18, 41]	22 [15, 34]	-15
Overall	25 [14, 39]	26 [15, 46]	+4.0

eTICI = extended Thrombolysis In Cerebral Infarction

3.4. Discussion

Due to the subjective nature of visual reperfusion estimation, eTICI scoring suffers from fair to moderate inter-observer agreement [20, 47, 50, 51]. This study aimed to assess whether the implementation of autoTICI could improve inter-observer agreement in eTICI scoring.

The study demonstrated substantial inter-observer agreement for eTICI scoring both alone and with autoTICI feedback, with weighted kappa values of 0.65 and 0.67, respectively. The addition of autoTICI feedback did not lead to a statistically significant increase in agreement. Dichotomization resulted in a higher kappa value for the session with autoTICI feedback, reflecting the increased observer agreement observed across nearly all eTICI grades (Table 3.2). However, the overall agreement remained substantial across the two scoring sessions, and the difference in weighted kappa was not statistically significant. Moreover, the ICC analysis showed that interobserver reliability improved from 'good' for absolute agreement to 'excellent' for consistency ICC. This suggests that while the observers maintained similar rank orders for eTICI grades, discrepancies existed in the exact assigned grades. The addition of autoTICI feedback slightly reduced these discrepancies by aligning individual observers' eTICI scores more closely with the overall median, although this improvement was not statistically significant (Table 3.3).

The lack of statistical significance could be explained by the fact that the observers were all experienced neuro-interventional radiologists, who were already highly skilled in eTICI scoring and therefore might benefit less from autoTICI feedback. Furthermore, the training session emphasized standardized definitions for eTICI scoring, potentially reducing the variability associated with conceptual differences that

typically contribute to inter-observer variability in clinical settings. While this approach was intended to improve result comparability, it may have limited the potential impact of autoTICI in reducing variability. Future research could include clinicians with varying levels of experience to better assess the influence of autoTICI in a broader clinical context. Moreover, the lack of significance may also be attributed to issues identified in autoTICI. Three issues were observed (3.3.2): undersized brain masks (6.5%), vascular overprojection (7.5%) and incorrect segmentation of non-perfused pixels at the mask borders (31%), also noted in studies by Su et al. and van der Sluijs et al. [27, 28]. These issues likely reduced observer confidence in autoTICI during the second scoring session, leading to continued reliance on visual eTICI scoring. Addressing these problems should be a priority in future research. Exploring different thresholds based on the level of contrast intensity could be a potential approach to improving perfusion segmentation accuracy.

It was observed that the median autoTICI reperfusion percentage was higher in the AP view than in the lateral view (Table 3.1). This difference is attributed to the greater overprojection of contrast intensity from the lateral hemisphere onto the AP view compared to the overprojection from the AP hemisphere onto the lateral view.

Regarding scoring efficiency, there was no significant increase in the time to perform eTICI scoring without and with the addition of autoTICI. However, for three of the four observers, the time required for scoring did increase. This increase could be attributed to the need to process and incorporate more information, especially when the model produced incorrect predictions, requiring observers to evaluate whether to rely on the model's output. The scoring time for eTICI grade 1 showed the most significant decrease, with a reduction of 26% (Table 3.4), likely because this grade is particularly challenging to determine due to conceptual confusion between reperfusion and recanalization, and autoTICI feedback may have provided additional confidence in these cases.

The inter-observer agreements observed in this study were consistent with those reported in the literature [47, 50, 51]. Table 3.5 provides a summary of similar studies investigating inter-observer agreement in visual eTICI scoring. Volny et al. reported fair to moderate agreement, with Krippendorff's alpha values ranging from 0.36 to 0.56, while Heiferman et al. found a fair to moderate agreement with weighted kappa values between 0.35 and 0.57, both of which are lower than the substantial agreement found in our study [47, 51]. Notably, these studies were conducted in clinical settings, where operators tend to overestimate scores by 33% compared to core-laboratory members, as shown by Fahed et al. [50]. When Heiferman et al. repeated their analysis in a core-laboratory setting, they observed moderate to substantial agreement (κ = 0.42-0.70), which more closely aligns with our findings [51]. Similar to our study, Nielsen et al. developed a deep learning-based method for automated and objective TICI scoring to improve inter-observer agreement [20]. Their method achieved a substantial agreement (κ = 0.61) when combined with expert TICI ratings, which is comparable to the agreement levels observed in our study [20].

Table 3.5: Summary of studies investigating the inter-observer agreement of eTICI scoring, compared to our method.

Authors (year of publication)	Observers	eTICI grading system	Measure	Results
AutoTICI (2024)	Core-laboratory members (n = 4)	eTICI, eTICI + autoTICI	Cκ	0.65, 0.67
Nielsen et al. (2021)	Core-laboratory members (n = 2)	mTICI + automated approach	Cκ	0.65
Heiferman et al. (2020)	Core-laboratory members (n = 4)	mTICI	Cκ	0.42–0.70
	Clinical operators (n = 4)	mTICI	Cκ	0.35–0.57
Volny et al. (2017)	Clinical operators (n = 3)	eTICI	Kα	0.36–0.56

Cκ = Cohen's kappa; Kα = Krippendorff's alpha

The results of this study should be interpreted with several limitations in mind. The first limitation regards the sample size calculation. The anticipated difference between the scoring sessions, with the

observer agreement to increase with one class, was overestimated. This overestimation, along with the extra exclusion of four patients, may have led to an insufficient sample size to detect a small difference. However, given that the actual difference between the two groups was only 0.02, a larger sample size would likely still have resulted in a non-significant finding.

The second limitation pertains to the measurement of scoring efficiency, which was conducted automatically by the local executable. As observers performed the scoring independently on their local PC's, recorded times may have been affected by external factors, such as forgetting to close the executable during breaks or when attending to other tasks. These factors likely contributed to the large variation in scoring times across observers, suggesting that the results may not accurately reflect the true scoring times. Therefore, these findings should be interpreted with caution. Conducting the scoring in a controlled setting would have yielded more reliable measurements.

The third limitation is the absence of an intra-observer analysis with and without autoTICI feedback. Intra-observer variability in the literature ranges from 0.45 to 0.70 [51], indicating that the differences observed between the first and second scoring sessions could partly result from variability within individual observers. Incorporating an intra-observer analysis would have provided valuable insights into this influence and clarified whether autoTICI had a significant effect in reducing such variability.

3.5. Conclusions

This core-laboratory study evaluated whether the implementation of autoTICI could enhance inter-observer agreement in eTICI scoring. We found similar agreement levels for eTICI scoring without and with autoTICI, with weighted kappa values of 0.65 and 0.67, respectively. These substantial agreement levels align with findings from previous studies in the literature. Despite issues identified in autoTICI, the overall eTICI scores and observer agreements remained unaffected. Furthermore, the addition of autoTICI feedback did not significantly impact the overall observer efficiency, indicating that the supplementary information did not prolong the time required for eTICI scoring — both encouraging results for the potential clinical application of autoTICI.

Future research should focus on addressing the identified issues in autoTICI. Repeating the second scoring session with an improved version of autoTICI could provide a more accurate assessment of its impact on inter-observer agreement. Additionally, including clinicians with varying levels of experience would help evaluate the tool's potential effectiveness in a clinical context.

4

Segmentation of vascular territories on cerebral DSA using deep learning

4.1. Introduction

4.1.1. Clinical background

The standard treatment for acute ischemic strokes (AIS) with a large vessel occlusion (LVO) is an endovascular thrombectomy (EVT), a minimally invasive mechanical procedure for removing blood clots or thrombi [10, 11]. During EVT, digital subtraction angiography (DSA) is the imaging modality used to visualize cerebral blood vessels and perfusion. The effectiveness of an EVT can be assessed by estimating the reperfusion status on DSA images using the Thrombolysis In Cerebral Infarction (TICI) grading system [14]. However, TICI scoring suffers from moderate to substantial inter-observer variability, presenting challenges for structural outcome comparison [20, 23–26]. To address these issues, our research group developed and evaluated 'autoTICI', a fully automated and objective method for assessing reperfusion after EVT treatment, detailed in the studies published by Su et al. and Van der Sluijs et al. [27, 28] and in section 1.1.4[1].

While autoTICI has shown potential in predicting clinical outcomes with accuracy comparable to traditional eTICI scoring, its robustness still requires improvement. Our previous studies have shown that challenges in the autoTICI pipeline often arise from unsuccessful vascular territory segmentation [27, 28], a critical step in the autoTICI pipeline where the occluded cerebral artery's vascular area is mapped to the post-EVT minimum intensity projection (MinIP) using atlas registration. The segmentation defines the measurement area for subsequent steps in the autoTICI pipeline, allowing for comparison of reperfused areas before and after EVT treatment.

4.1.2. Atlas registration

Since cerebral DSA images only display vessel structures and lack clear anatomical boundaries, direct segmentation of vascular territories is challenging, especially in cases of proximal occlusions. This issue can be addressed by bringing prior information to the segmentation task, for which atlas registration is a widely used method, which involves aligning a standardized anatomical 'atlas' with a patient-specific image [43]. An atlas refers to a reference image that depicts a typical anatomy under normal conditions and includes predefined segmentations of anatomical regions. Registering the atlas to the patient's image yields a transformation which allows the atlas segmentation to be transformed and treated as a segmentation estimate for the patient image [55]. This is particularly useful for delineating specific tissues or regions that are difficult to identify [42].

In autoTICI, atlas registration is employed to map predefined vascular territories, specifically of the internal carotid artery (ICA) and middle cerebral artery (MCA) — commonly affected in ischemic strokes — onto the post-EVT MinIP [3–5]. To account for anatomical variation and topological differences, 21

[1] This is one chapter in a larger thesis. See chapter 1 for a full background.

atlases are used, selected from a non-stroke cerebral DSA dataset and segmented by an experienced interventional radiologist (SC). These atlases are spatially aligned to the post-EVT MinIP using affine registration, with the final atlas selected based on the highest Mattes Mutual Information (MI).

However, in the context of autoTICI, atlas registration presents particular challenges due to the need for inter-patient alignment, which is difficult with DSA images, because of variations in vascular anatomy and the lack of anatomical structures. This complexity is further compounded by stroke-related perfusion changes caused by the occlusion [42–44]. In a recent study, Van der Sluijs et al. showed that atlas registration was only successful in 79% of anteroposterior (AP) and 70% of lateral cases, leading to incorrect determination of the measurement area and autoTICI results [28]. Given that the model requires both AP and lateral views for a prediction, autoTICI was considered unreliable in 30% of cases.

Before autoTICI can be considered ready for clinical implementation, substantial improvements to its vascular segmentation task are essential. A previously conducted literature review on state-of-the-art registration techniques for cerebral DSA found that inter-patient registration techniques, such as atlas registration, were not addressed in the literature [44]. This research gap presents challenges for further improvement and validation of the atlas registration method. Another challenge is the lengthy processing time required by atlas registration (2.4.3). For autoTICI to be effectively used as a decision-support tool during EVT procedures, this processing time needs to be significantly reduced. Artificial intelligence (AI) offers a promising solution to both challenges by potentially enabling direct prediction of segmentation areas on cerebral DSA images, thereby eliminating the need for atlas registration.

4.2. Objectives and overview

This study aims to improve the robustness of autoTICI by developing an automated approach for direct vascular territory segmentation on cerebral DSA. Initially, a deep learning-based approach for automatic segmentation of vascular territories was developed and validated on an internal test set. The effectiveness of the proposed method was assessed by comparing the automated segmentations with those derived from atlas registration.
Section 4.3 describes the technical details of the proposed methodology, section 4.4 describes the tests performed and section 4.5 describes their results.

4.3. Methods

Figure 4.1 outlines the proposed methodology. First, we manually created reference standard segmentations (4.3.3 and 4.3.4), which were used to train (4.3.5) a deep learning framework (4.3.2) on 90% of the imaging dataset (4.3.1). After validating this model (4.3.7), we tested both the model and the atlas registration method on the remaining 10% and compared their performance (4.4.4). In addition, we conducted three experiments (4.4.1 - 4.4.3) to assess the impact of training variations on model performance.

4.3.1. Data inclusion

Imaging data was included from a subset of the MR CLEAN Registry (parts 1-3) [4, 15], which comprised 365 stroke patients and 1336 DSA sequences. This subset, previously used in a study on arterial input functions (AIF) in cerebral DSA, was selected based on the availability of ICA and MCA vascular territory segmentations for all post-EVT DSA sequences (n = 730). Patients were included in this subset if their DSAs showed an ICA, M1, or M2 occlusion and an eTICI grade higher than 2A, indicating ‘high’ perfusion levels, which were necessary for AIF calculations. However, this selection may have caused a selection bias towards higher perfusion outcomes. To mitigate this and enhance dataset diversity, pre-EVT sequences (n = 606), indicative of ‘low’ perfusion levels, were also included in this study.
DSA sequences were excluded from this study if their MinIP (1) showed significant motion artefacts, (2) failed to capture the largest part of the MCA or ICA within the field of view (FoV), (3) had poor image quality, or (4) were oblique acquisitions (in between AP and lateral), resulting from patient head rotation during the procedure. This exclusion criterion was established to focus the model’s predictive ability primarily on AP or lateral views, in line with autoTICI’s output.

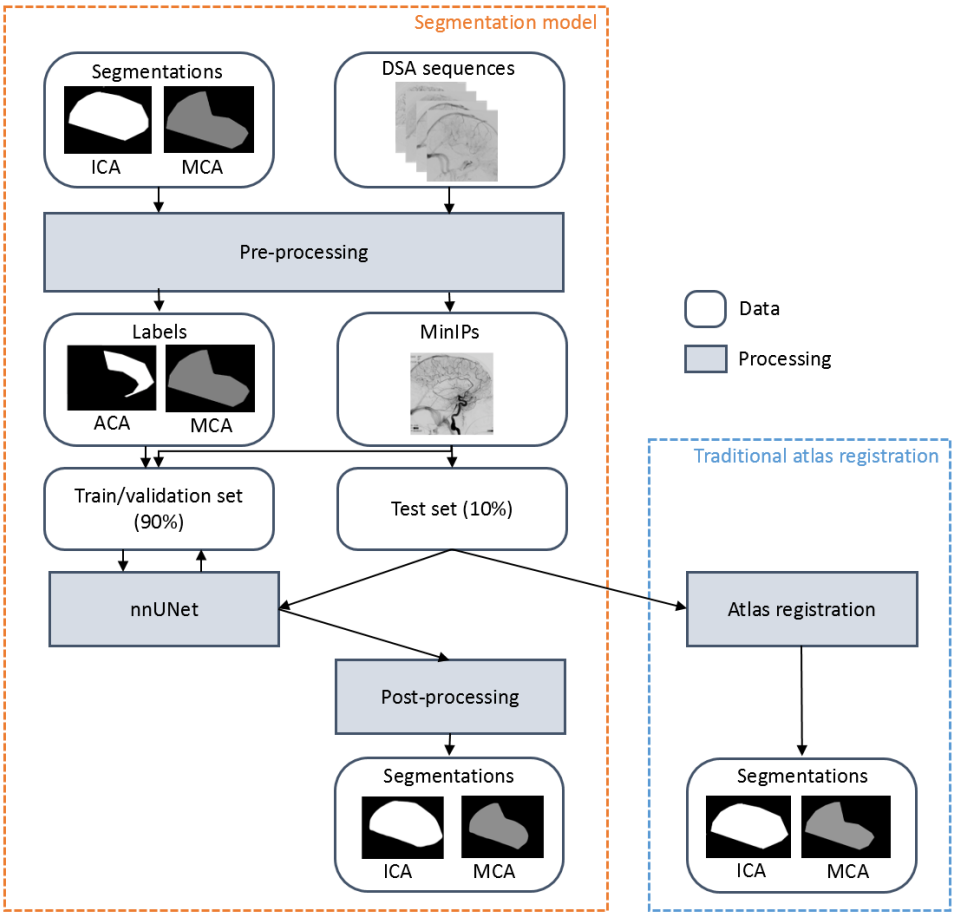


Figure 4.1: Overview of the steps involved in training and testing of the proposed segmentation model, including pre-processing, the division of included data into the internal training, validation, and test sets, and post-processing. The traditional atlas registration approach is performed on the same internal test set images to facilitate segmentation comparisons.

4.3.2. nnUNet

For our deep learning framework, we chose no-new-UNet (nnUNet), which is widely recognized as the state of the art for medical image segmentation [56, 57]. NnUNet is robust and versatile across various imaging modalities and features a self-configurable framework capable of tuning the entire pipeline based on the provided dataset, including pre-processing, network architecture selection, model training, and post-processing. Figure 4.2 illustrates the 'standard' architecture of the nnUNet framework used for 2D modalities [58].

4.3.3. Reference standard segmentations

Deep learning frameworks need labelled reference input to 'learn' the task for which they are created. In our case, we needed to provide nnUNet with reference standards for both the pre-and post-EVT DSA images. For the post-EVT DSAs, the existing segmentations served as reference standards, which were created for the post-EVT MinIPs using the autoTICI atlases. This was done manually by Matthijs van der Sluijs (MS), a PhD candidate with expertise in stroke imaging and a contributor to autoTICI. Reference standards for the pre-EVT DSAs were created by LS by manually registering the post-EVT segmentations to the corresponding pre-EVT MinIPs and adjusting the masks for size, rotation and scaling. All annotations were performed in MeVisLab 3.0.2, using an in-house tool.

4.3.4. Pre-processing

The initial pre-processing step involved generating MinIPs (2D) from the included DSA sequences (2D + t). To meet the input specifications required by nnUNet, the MinIPs and segmentations were resized

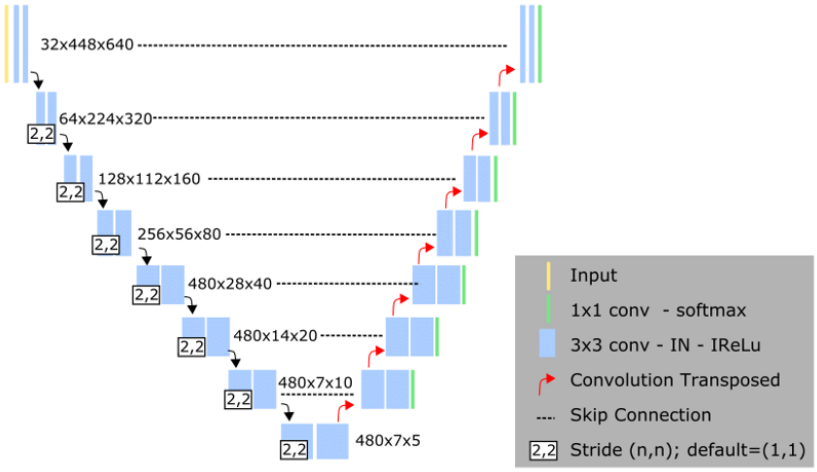


Figure 4.2: 'Standard' architecture of the 2D nnUNet framework [58]

to 1024x1024 pixels, ensuring uniform size and pixel spacing [57]. Three labels were assigned to the background, MCA and ACA segmentations. The ACA mask was created by subtracting the MCA from the ICA, as the MCA is fully enclosed within the ICA. Figure 4.3 shows the 2D MinIPs alongside the corresponding reference standard segmentations.

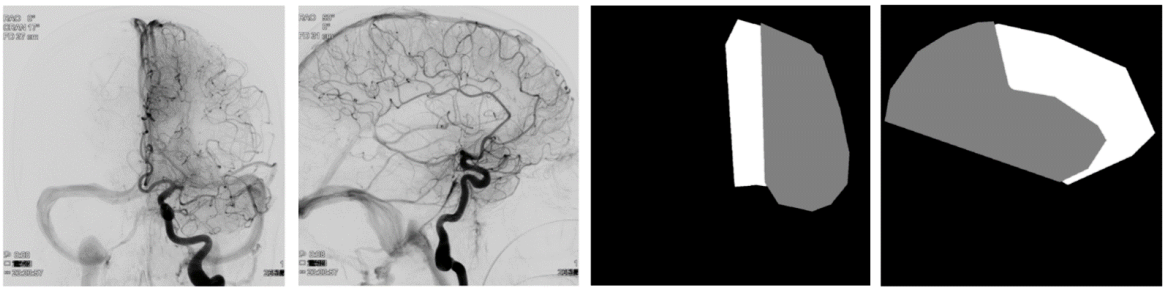


Figure 4.3: Examples of the 2D MinIP input images are displayed in AP and lateral views. On the right, the corresponding reference standard segmentations are shown. The labels were assigned as follows: the background was labelled '0' (black), the MCA mask was labelled '1' (grey), and the ACA mask (ICA - MCA) was labelled '2' (white).

During the annotation of the reference standards, scaling caused the larger ICA mask to scale more than the MCA mask, leaving residual lines at the ICA boundaries that were not fully covered by the MCA mask and, therefore, not subtracted. Residual lines were removed from the ACA mask using a series of morphological operations: erosion to separate the residual ICA lines, connected component analysis to isolate and retain the largest mask, and dilation to restore the mask to its original size. Figure 4.4 shows the result of the morphological operations on the subtracted ACA mask.

4.3.5. Model training

We employed the 2D model of nnUNet (version 2) for training and testing [56]. We divided the input data, the MinIPs and corresponding labels, into 10% for testing and 90% for training. The training data was further divided into training and validation sets using an 80/20 split. A patient-based randomized stratification was employed to ensure that images from the same patient remained within the same split. Besides that, stratified sampling was performed to ensure class balance across the underlying occlusion locations (ICA, M1, M2).
Model training was conducted on the GPU cluster of the Erasmus MC, using NVIDIA GeForce RTX 2080 Ti 11GB and NVIDIA A40 48GB GPUs, and on a local PC (Ubuntu 24.04 LTS, Linux 6.8.0) containing an NVIDIA GeForce RTX GPU. During training, five-fold cross-validation was performed with 1000 epochs. Model ensembling was performed by averaging the weights of the five models trained on each fold. No

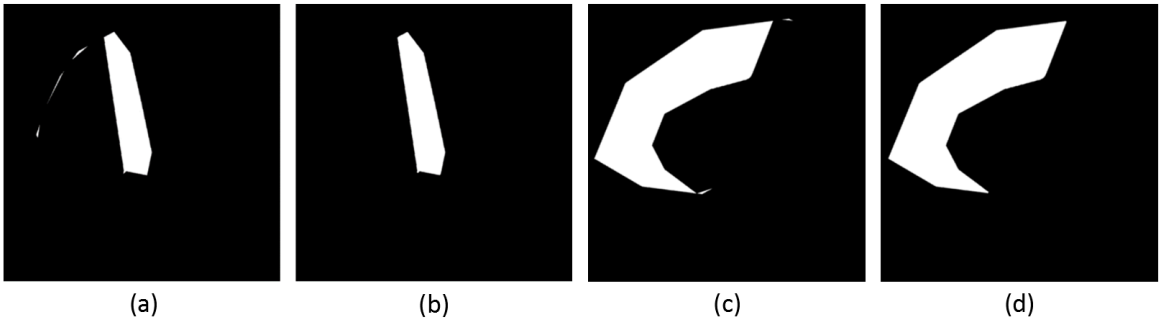


Figure 4.4: Visualization of the morphological operations — erosion, connected component analysis, and dilation — on the ACA mask (ICA - MCA): (a) ACA mask (AP), with residual ICA mask lines; (b) ACA mask (AP) after preprocessing; (c) ACA mask (lateral) with residual ICA mask lines; (d) ACA mask (lateral) after preprocessing.

adjustments were made beyond the default settings provided by nnUNetv2. Initially, separate models were trained for the AP and lateral views. Subsequently, a combined model capable of distinguishing between the AP and lateral views was developed by combining all data.

4.3.6. Post-processing

The predicted segmentations underwent the same morphological operations — erosion, connected components analysis, and dilation — as were applied during pre-processing to ensure consistency. Additionally, the ICA mask was reconstructed by combining the ACA and MCA labels. All pre-and post-processing tasks were conducted in Visual Studio Code, using Python version 3.11.8.

4.3.7. Performance metrics

The model performance was evaluated by calculating the Dice Similarity Coefficient (DSC) and Average Surface Distance (ASD) for the predicted and reference standard segmentations. The DSC was computed using the following equation:

$$DSC = \frac{2 \times |X \cap Y|}{|X| + |Y|}, \tag{4.1}$$

where X and Y represent the set of pixel locations in the predicted and reference standard segmentations, and $|X \cap Y|$ is the number of elements in both sets. The DSC quantifies the global overlap between the two segmentations, with its value ranging from 0 to 1. A DSC of 1 indicates perfect agreement, whereas a DSC of 0 signifies no overlap at all.

To complement the global assessment provided by the DSC, ASD was employed as a metric to measure local distances between the predicted and reference standard segmentations. The ASD was calculated as follows:

$$ASD = \frac{1}{|S_X| + |S_Y|} \left(\sum_{x \in S_X} \min_{y \in S_Y} d(x, y) + \sum_{y \in S_Y} \min_{x \in S_X} d(y, x) \right), \tag{4.2}$$

where $|S_X|$ and $|S_Y|$ are the sets of all points on the surface of segmentations X and Y. $d(x, y)$ calculates the Euclidean distance between point x on the predicted segmentation and point y on the reference standard segmentation. The sums compute the minimum distances for all points on both surfaces, and the overall expression averages these minimum distances. A lower ASD value indicates closer alignment of the segmentation contours, with a value of 0 representing perfect conformity. For both outcome measures, either the means and 95% Confidence Intervals (CIs) or medians and Inter Quartile Ranges (IQRs) were calculated, based on the normality of the data.

4.4. Experiments and benchmarking

By varying the characteristics of the training dataset, we were able to test the impact of training variations on model performance. We tested three variations: combined versus separate AP and lateral models (4.4.1), full-sized versus half-sized training datasets (4.4.2), and full-phase versus single-phase datasets (4.4.3). We also compared the proposed model against the existing atlas registration method, evaluating both overall performance and performance across different occlusion locations and EVT stages (4.4.4).

4.4.1. Comparison of combined and separate AP/lateral models

In an initial experiment, we assessed the performance of the segmentation models trained separately for AP and lateral views and compared these to the combined model. This analysis aimed to determine whether such a combined model, which was preferred due to its simple implementation, could achieve similar results or even outperform the separate models for each view. The decision to proceed with either the separate models or the combined model for further experiments was based on their respective performance outcomes (DSC and ASD). The models were statistically compared using either a two-sided paired Student's t-test or a Mann-Whitney U-test, depending on the normality of the results.

4.4.2. Evaluation of training size

A second experiment was conducted using two different dataset sizes to assess the impact of training data volume on the model performance. The initial model, described in section 4.3.5, was trained and tested using all input MinIPs. A second dataset was introduced, containing only half of its original size, which was selected using patient-based randomized stratification to maintain a balanced representation of the underlying occlusion locations. Both models were evaluated with the full-sized test set. The models' performances were statistically compared using a two-sided paired Student's t-test or a Mann-Whitney U-test, depending on the normality of the data.

4.4.3. Phase-specific analysis

The segmentation model was trained using MinIPs that included all vascular phases—arterial, parenchymal, venous, and non-contrast frames. To discern which phases influenced the model predictions, we compared the performance of segmentations predicted on the full-phase MinIP with those predicted on MinIPs created for each phase. Phase-specific MinIPs were created from the original DSA sequences using the deep-learning phase selection model implemented in the autoTICI pipeline [27]. The non-contrast phase was included to assess whether the model uses the skull outline present in frames with motion artefacts. Statistical testing was conducted using either a Mann-Whitney U-test or a two-sided paired Student's t-test, depending on the normality of the data.

4.4.4. Comparison against traditional method

We benchmarked the performance of the proposed segmentation model against the traditional atlas registration method. For this comparison, segmentations were generated for the test set MinIPs using the atlas registration method in autoTICI. The segmentations produced by both methods were evaluated using the performance metrics (4.3.7), along with additional outcomes robustness and computational time, to assess their suitability for implementation in autoTICI.

Computational time was defined as the time required by the model to generate a segmentation for one test set MinIP, excluding time spent on pre- or post-processing. Due to the potential hardware influence on computational time, both methods were executed on the same PC (DELL Latitude 5510) using a CPU under controlled conditions to reduce the potential impact from active programs. The computation time was reported for both methods as means and 95% CIs for normally distributed data, or medians and IQRs for non-normally distributed data.

Robustness was defined as the percentage of clinically successful segmentations, which was determined using a Likert scale ranging from 0 to 3 to indicate anatomical correctness. A value of 0 indicated a 'failure', where the segmentation was located in the incorrect hemisphere and/or was outside the boundaries of the expected anatomical area; 1 was 'marginal' indicating that the segmentation was in the correct hemisphere, but substantially outside the boundaries of the anatomical area; 2 indicated 'acceptable', meaning the segmentation was largely inside the expected anatomical area, and 3 de-

noted a perfect segmentation that matched anatomical expectations. For segmentations rated from 0 to 2, inaccurate mask boundaries (superior, inferior, left, right) were identified, and the reasons for the errors were noted. Errors were categorized based on incorrect scaling (either too large, encompassing the territory from another vessel or the skull, or too small, missing part of the intended vascular territory), improper rotation (incorrect angle of the mask), or incorrect location of the mask on the input image. The success rate was defined as the percentage of annotations that received a Likert score of 2 and 3. This classification was based on the operational requirements of autoTICI, where exact mask placement is not critical, and minor deviations are acceptable for satisfactory reperfusion calculation. For both methods, the segmentations were independently annotated by two observers, LS and MS. In cases of discrepancies, a consensus was reached to determine a final Likert score for each segmentation.

Furthermore, we compared the performance and robustness of both model's segmentations across different underlying occlusion locations (ICA, M1, M2) and EVT stages (pre- and post-procedure).

4.5. Results

4.5.1. Data

Data inclusion

Figure 4.5 presents a flowchart illustrating the data inclusion process, with numbers provided for the patients and corresponding pre-EVT, post-EVT, AP, and lateral DSA sequences. Table 4.1 details the underlying distributions of EVT-stages, views and occlusion locations. A total of 119 DSA sequences were excluded, with the majority from the pre-EVT group (n = 104) compared to the post-EVT group (n = 15). This was mainly due to the lower quality of pre-EVT acquisitions, primarily intended for verifying catheter positioning rather than for detailed inspection of reperfusion status. The AP and lateral sequences had a similar number of exclusions, 59 and 60 respectively. As a result, 1217 eligible DSA sequences from 361 stroke patients were included in the analysis, distributed across a training/validation set (n = 1097) and a test set (n = 120). The dataset contained slightly more post-EVT than pre-EVT images, with proportions of 41% and 59% and 46% and 54%, in the train and test set, respectively (Table 4.1). The distributions of occlusion locations aligned with the MR CLEAN Registry (parts 1-3) where ICA, M1, and M2 occlusions are reported at 19%, 55% and 25% [4].

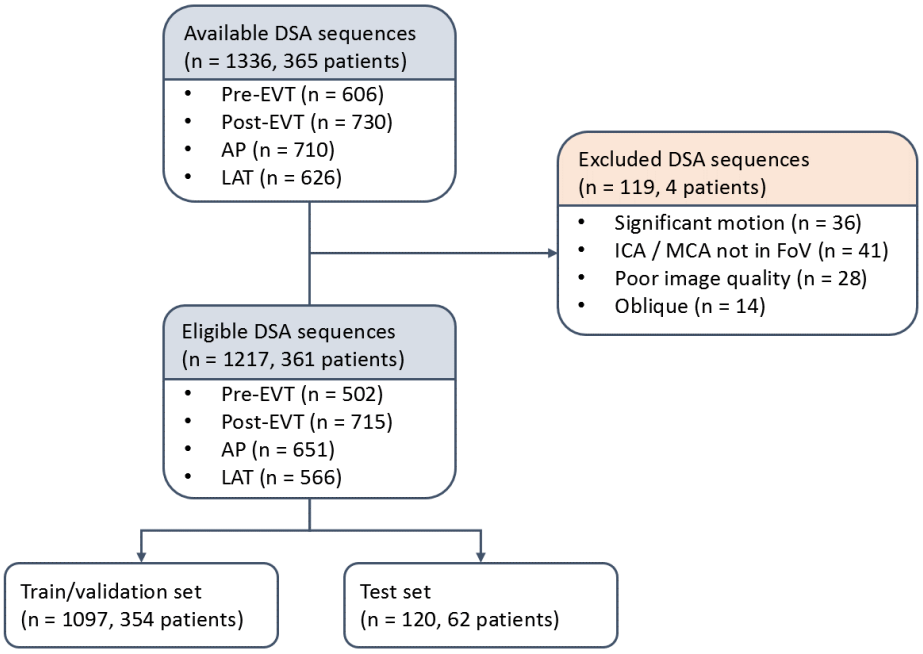


Figure 4.5: Flowchart of the process of data inclusion.

Table 4.1: Distribution of EVT-stages, views and occlusion locations of included DSA sequences, across the training/validation and test sets.

Variables	Internal training/validation set	Internal test set
No. patients	354	62
No. DSA sequences		
- Total	1097	120
- Pre-EVT (%)	447 (41)	55 (46)
- Post-EVT (%)	650 (59)	65 (54)
- AP (%)	588 (54)	63 (53)
- Lateral (%)	509 (46)	57 (47)
No. occlusion locations		
- ICA (%)	159 (15)	23 (19)
- M1 (%)	587 (53)	67 (56)
- M2 (%)	351 (32)	30 (25)

AP = anteroposterior; ICA = Internal Carotid Artery; M1 = first segment of Middle Cerebral Artery; M2 = second segment of Middle Cerebral Artery; EVT = endovascular thrombectomy

4.5.2. Comparison of combined and separate AP/lateral models

All models were trained using the nnUNet framework with a plain convolutional U-Net architecture, detailed in Table F.1 (Appendix F). The AP, lateral, and combined models demonstrated comparable DSC and ASD scores, with no statistically significant differences as determined by the Mann-Whitney U-test (Table 4.2). The p-values were 0.88 for the DSC in both segmentations and 0.99 and 0.82 for the ASD in the ICA and MCA segmentations, respectively. The combined model was selected for further experiments due to its simpler implementation, with median DSC and ASD scores for the ICA of 0.96 (IQR: 0.94–0.97) and 13.8 (IQR: 10.1-20.5), respectively. The model performance for the MCA was lower, with median DSC and ASD scores of 0.94 (95% CI: 0.92–0.96) and 41.6 (IQR: 35.0–47.2).

Table 4.2: DSC and ASD for ICA and MCA segmentations predicted by the trained models. The atlas registration method is tabulated for comparison. Values in the table represent medians and Inter Quartile Ranges [IQRs].

	ICA territory		MCA territory	
	DSC	**ASD (px)**	**DSC**	**ASD (px)**
AP	0.95 [0.94, 0.96]	13.1 [9.35, 16.8]	0.94 [0.93, 0.95]	37.2 [32.0, 43.0]
Lateral	0.96 [0.94, 0.97]	16.8 [11.6, 22.1]	0.94 [0.92, 0.96]	44.4 [41.3, 54.7]
Combined	**0.96 [0.94, 0.97]**	**13.8 [10.1, 20.5]**	**0.94 [0.92, 0.96]**	**41.6 [35.0, 47.2]**
Combined half size	0.95 [0.93, 0.97]	14.3 [11.0, 21.9]	0.94 [0.92, 0.96]	41.2 [36.1, 47.4]
Atlas	0.82 [0.76, 0.89]***	47.3 [30.4, 66.5]***	0.78 [0.69, 0.85]***	68.6 [50.1, 88.4]***

Statistical differences between the trained models/atlas method and the combined model are indicated as follows: $p \leq 0.05$ (*), $p \leq 0.01$ (**), $p \leq 0.001$ (***), $p \leq 0.0001$ (****).

ICA = Internal Carotid Artery; MCA = Middle Cerebral Artery; DSC = Dice Similarity Coefficient; ASD = Average Surface Distance; px = pixels; AP = separate model trained for the AP view; Lateral = separate model trained for the lateral view; Combined = model trained for AP and lateral views; Combined half size = combined model trained with half of the training data.

4.5.3. Evaluation of training size

The performance of the combined model trained on the full training set and the combined model trained on the half-size dataset did not show statistically significant differences (p = 0.50 and p = 0.63 for DSC; p = 0.55 and p = 0.84 for ASD, for the ICA and MCA, respectively) (Table 4.2). Figure 4.6 presents the training and validation loss curves for the combined model trained on the full and half training data. Both models showed similar loss curves, with comparable gaps between the training and validation losses, suggesting that the reduction in training data did not lead to increased overfitting.

4.5.4. Phase-specific analysis

The segmentations predicted on the arterial and parenchymal phases showed no significant differences in DSC and ASD scores compared to the combined phases (p = 0.07 for the DSC and p = 0.11 for the ASD of ICA, and p = 0.96 for the ASD of MCA), except for the DSC of the MCA in the arterial phase (p = 0.04) (Figure 4.7). This is also illustrated in two examples in Figure 4.8, where phase-specific segmentations are visualized in yellow alongside the combined model's segmentations in blue. The

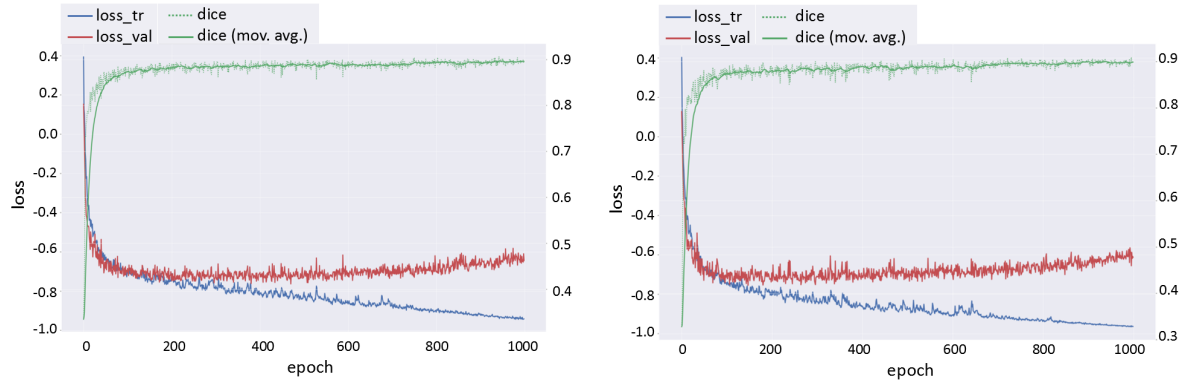


Figure 4.6: Train and validation loss of the combined model trained with (left) 1097 and (right) 549 images. Train and validation losses are depicted in blue and red, respectively, and the corresponding Dice Similarity Coefficient (DSC) is visualized in green.

arterial and parenchymal phase predictions displayed good accuracy in both views, with only minor deviations in mask delineation compared to the combined model.
In contrast, the model performance for the venous phase MinIPs was significantly lower than the combined model, except for the ASD score for the MCA (p = 0.0001 and p = 0.05 for the DSC, and p = 0.01 and p = 0.07 for the ASD, for ICA and MCA, respectively). In the example in the AP view shown in Figure 4.8, the segmentation was positioned in the wrong hemisphere, resulting in a DSC of 0.0. In the lateral view, the MCA mask had an incorrect orientation with an uneven contour, while the ICA mask remained relatively accurate, especially at the superior border.
The non-contrast phase had the worst DSC and ASD scores, with notably large IQRs, as shown in Figure 4.7, with all results significantly worse than the combined phases ($p \leq 0.001$). In this phase, the model produced inaccurate and random segmentations in the example in the AP view, whereas the lateral view, which had motion artefacts, yielded slightly more accurate results, particularly at the superior border (Figure 4.8).

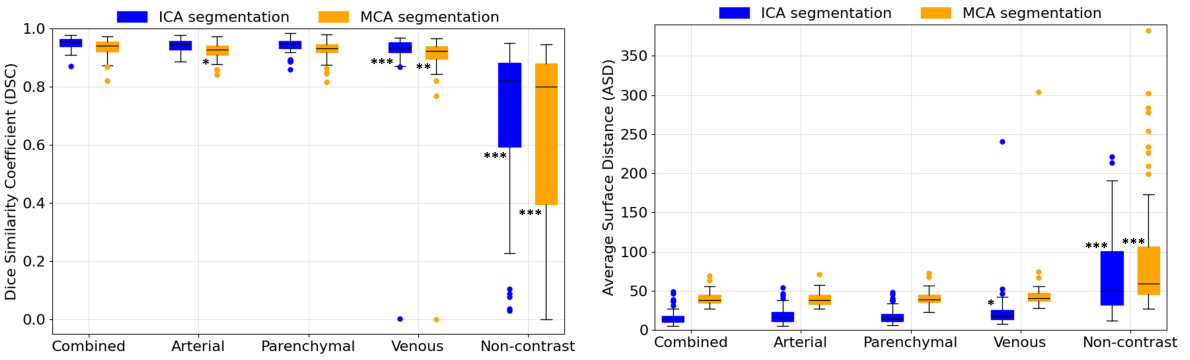


Figure 4.7: Boxplots visualizing the DSC and ASD for ICA and MCA segmentations across the arterial, parenchymal, venous and non-contrast phases. The left graph represents DSC values. The right graph represents ASD values. ICA and MCA segmentations are depicted in blue and orange, respectively. Statistical differences between the performance of the individual vascular versus the combined phases are indicated as follows: $p \leq 0.05$ (*), $p \leq 0.01$ (**), $p \leq 0.001$ (***), $p \leq 0.0001$ (****).
ICA = Internal Carotid Artery; MCA = Middle Cerebral Artery; DSC = Dice Similarity Coefficient; ASD = Average Surface Distance

4.5.5. Comparison against traditional method

The combined model significantly outperformed the atlas registration method in both DSC and ASD values ($p \leq 0.001$)(Table 4.2). This finding is further supported by the visualizations of both methods' segmentations illustrated in Figure 4.9. The segmentations generated by the combined model closely approximate the reference standards in both views. Additionally, the combined model effectively captured variations in the angle of the head and accurately predicted segmentations for zoomed-in images

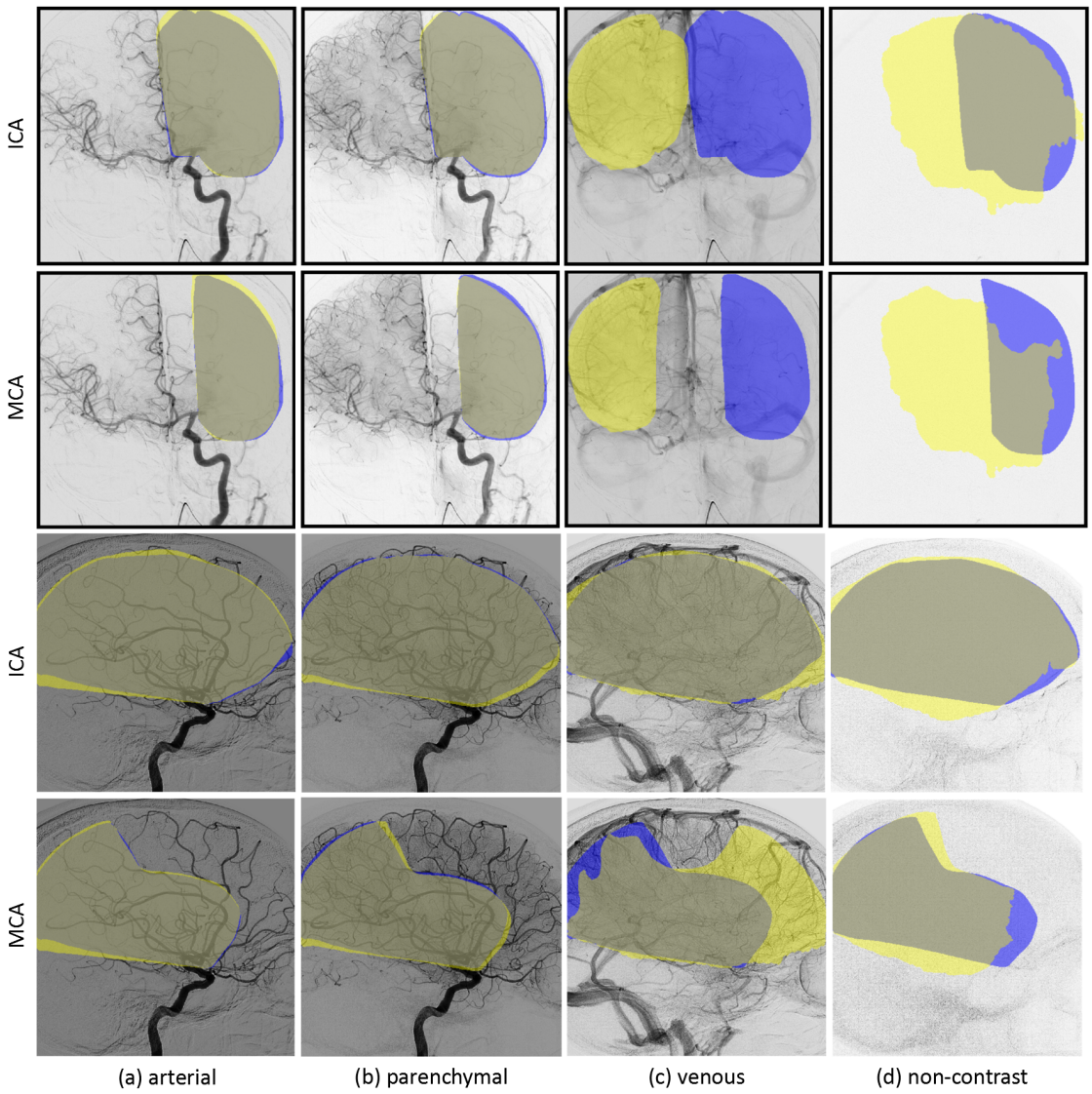


Figure 4.8: Visualization of the ICA and MCA segmentations predicted for the phase-specific MinIPs of example test set images in the AP and lateral view, depicted in yellow. Blue and yellow masks represent the segmentations of the combined model. *ICA = Internal Carotid Artery; MCA = Middle Cerebral Artery; AP = Anteroposterior*

or images that contained text or black borders. In contrast, the atlas registration segmentations are too small in both views and misaligned in the AP view, extending beyond the skull. It can also be observed that the nnUNet segmentations were more rounded compared to the reference standard segmentations and those produced by the atlas method.

In terms of computational time, the atlas registration method was significantly slower, taking 141 seconds compared to 4.19 seconds for the segmentation model (Table 4.3) ($p \leq 0.0001$). Regarding robustness, Figure 4.10 shows that the combined model predominantly received scores of 3 for both ICA and MCA segmentations, with a few instances of scores of 2, indicating a 100% success rate. In contrast, the atlas registration method primarily yielded scores of 1 and 2, with fewer instances of 0 and 3. This resulted in a success rate of 52.5% for ICA segmentations and 53.3% for MCA segmentations. Examples of segmentations corresponding to each Likert score are provided in Appendix G. Of the eight segmentations predicted by the combined model that received a Likert score of 2, six were lateral, all with an oversized inferior border. In the AP view, the segmentations were slightly undersized. For the atlas registration method, among the 108 segmentations that received scores between 0 and 2,

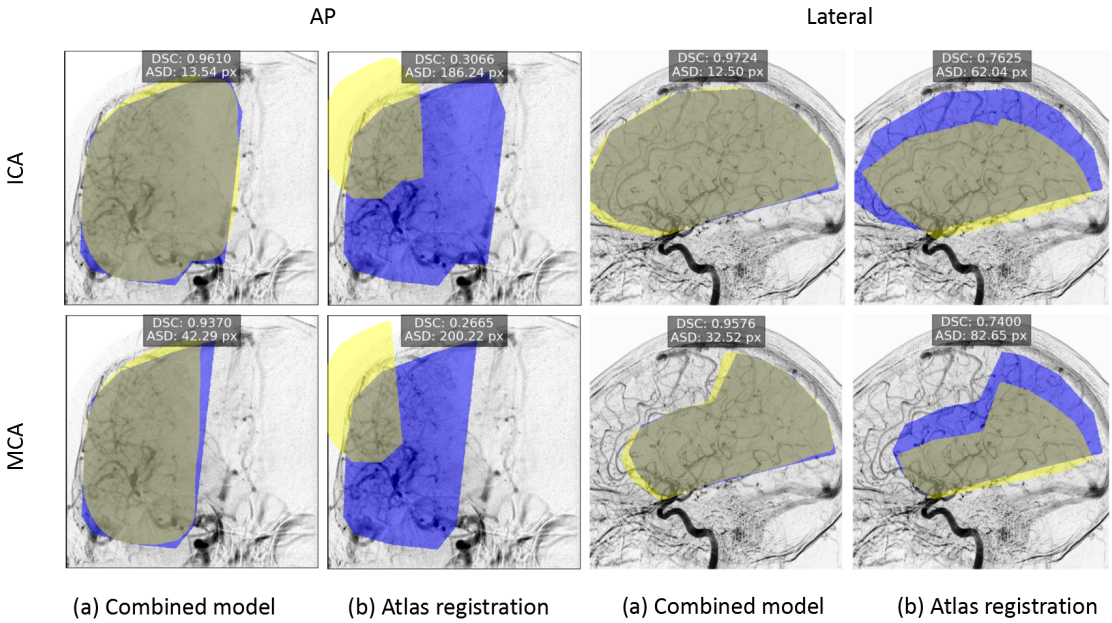


Figure 4.9: Visualization of the ICA and MCA segmentations of the combined segmentation model and the atlas registration method of example test set images in the AP and lateral views. Blue and yellow masks represent the reference standard and predicted segmentations, respectively. *ICA = Internal Carotid Artery; MCA = Middle Cerebral Artery; AP = Anteroposterior; DSC = Dice Similarity Coefficient; ASD = Average Surface Distance*

the majority (n = 84) exhibited errors across multiple borders, with the superior border being the most commonly affected (n = 39). The most frequent issues were scaling errors (n = 74), typically resulting in undersized masks (n = 66). Additional errors included incorrect positioning (n = 19) and rotation inaccuracies (n = 11).

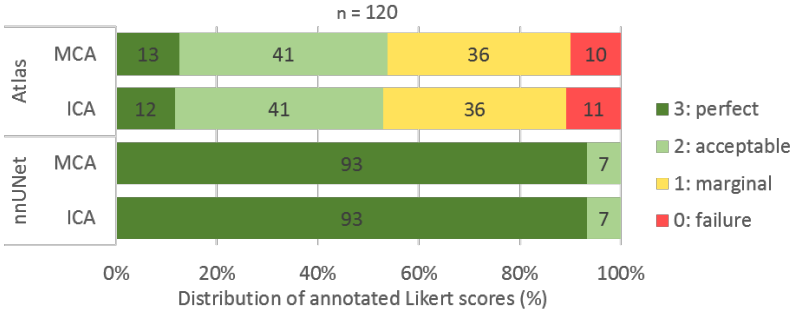


Figure 4.10: Distributions (%) of annotated Likert scores for the proposed combined segmentation model and the traditional atlas registration method, for the ICA and MCA segmentations. *ICA = Internal Carotid Artery; MCA = Middle Cerebral Artery*

Table 4.3: Results of the success rate (%) and computational time (median and Inter Quartile Ranges [IQRs]) for the ICA and MCA segmentations produced by the combined model and the atlas registration method.

	Success rate (%)		
	ICA territory (AP, lateral)	**MCA territory (AP, lateral)**	**Computational time (s)**
Combined model	**100 (100, 100)**	**100 (100, 100)**	**4.19 [3.88, 4.46]**
Atlas	52.5 (55.6, 49.1)	53.3 (57.1, 42.9)	141 [87.6 , 223]****

Statistical differences between the computational times of the atlas method and combined model are indicated as follows: $p \leq 0.05$ (*), $p \leq 0.01$ (**), $p \leq 0.001$ (***), $p \leq 0.0001$ (****).
ICA = Internal Carotid Artery; MCA = Middle Cerebral Artery; CI = Confidence Interval

4.5.6. Effect of occlusion location

Both the atlas and the proposed segmentation method exhibited similar patterns in model performance across different occlusion locations, with the overall performance of the atlas registration method being inferior to that of the segmentation model (Figure 4.11). Overall, no clear differences were observed in model performance across the ICA, M1, and M2 occlusion locations, though a slight increase in ASD was observed for the MCA segmentations for M1 occlusions created by the atlas registration method. Besides that, a slight difference in model performance was observed between the EVT stages for both the DSC and ASD outcome measures; pre-EVT segmentations were notably worse than those predicted for the post-EVT MinIPs in both the ICA and MCA segmentations across both methods.

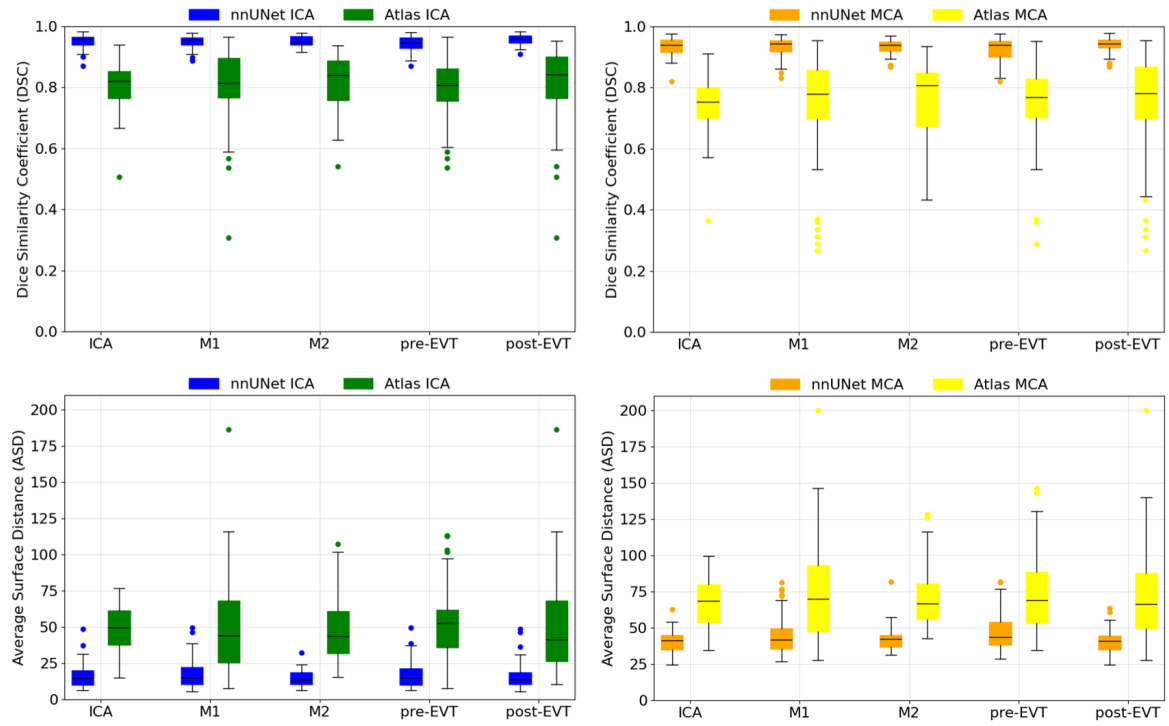


Figure 4.11: Boxplots displaying the DSC (top) and ASD (bottom) for the ICA (left) and MCA (right) segmentations created by the proposed segmentation model and the traditional atlas registration method, across different occlusion locations (ICA, M1, M2) and EVT stages (pre and post). *ICA = Internal Carotid Artery; MCA = Middle Cerebral Artery; M1 = first segment MCA; M2 = second segment MCA; EVT = Endovascular thrombectomy*

Regarding robustness, Figure 4.12 shows that the combined model maintains consistent Likert score distributions across all occlusion locations and for both EVT stages. In contrast, for the atlas registration method, there is a noticeable decrease in the proportion of scores of 0 and 1 for the underlying occlusion locations ICA, M1 and M2 respectively, accompanied by an increase in scores of 2 and 3. This trend is reflected in the corresponding increasing success rates of 39.1% for ICA, 55.2% for M1, and 62.5% for M2. Additionally, there is a difference between the EVT stages, with pre-EVT segmentations showing more scores of 0 and 1 compared to post-EVT, resulting in success rates of 45.3% for pre-EVT and 58.2% for post-EVT.

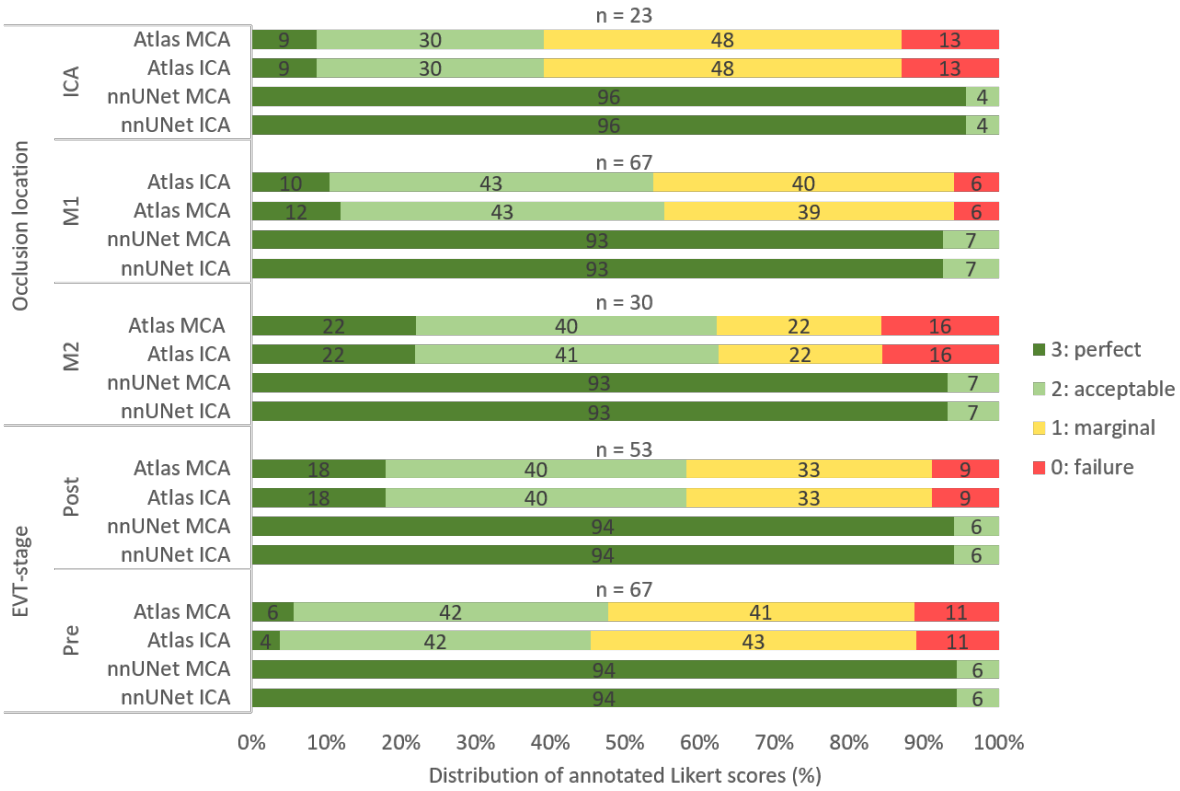


Figure 4.12: Visualization of the distributions (%) of annotated Likert scores for the proposed segmentation model and the traditional atlas registration method, presented for the segmentations across different occlusion locations (ICA, M1, M2) and for the EVT stages (pre and post). Numbers in the stacked bars represent percentages, noted alongside the sample sizes. *ICA = Internal Carotid Artery; MCA = Middle Cerebral Artery; M1 = first segment MCA; M2 = second segment MCA; EVT = Endovascular thrombectomy*

4.6. Discussion

In a previous study, autoTICI results proved unreliable in 70% of lateral and 79% of AP cases due to unsuccessful vascular territory segmentation using atlas registration [28]. Therefore, significant improvements in robustness are necessary before autoTICI can be considered viable for clinical implementation. To address this issue, we proposed an automatic deep-learning approach for direct segmentation of ICA and MCA vascular territories on cerebral DSA and benchmarked it against the atlas registration method. The proposed model achieved excellent results on the held-out test set, with a 100% success rate in both views. This indicated that our approach is well-suited for implementation in the autoTICI pipeline.
For the proposed model, 93% of segmentations were rated as anatomically perfect, and 7% showed slight deviations due to scaling errors at the inferior border in the lateral view. These deviations may be due to difficulties in manually delineating the lateral inferior border due to PCA overprojection, leading to variations in the reference standards and, subsequently, in the predicted segmentations. In contrast, the atlas registration method yielded lower success rates: 55.6% in the AP view and 49.1% in the lateral view, and 57.1% in the AP view and 42.9% in the lateral view for ICA and MCA, respectively (Table 4.3, Figure 4.9). Most errors were due to undersized masks (55%), indicating that the atlas registration method struggled primarily with correct scaling (Figure 4.9 and G.3, Appendix G). The success rates for the atlas registration method found in this study were lower than those reported by van der Sluijs and Su et al. [27, 28], which could be caused by differences in the included imaging data, and the smaller validation dataset of 120 images (62 patients) in our study compared to 1684 images (421 patients) in theirs. Additionally, van der Sluijs and Su et al. pre-processed DSA sequences by removing motion frames, text and black borders, which was not performed in our analysis.
It was observed that the nnUNet segmentations were more rounded compared to the reference standard segmentations and those produced by the atlas method. This is due to the model predicting a probability distribution for each pixel's label, resulting in rounded masks without straight boundaries.

In terms of model performance, the proposed approach significantly outperformed the atlas registration method. It was observed that ICA segmentations consistently achieved higher performance than MCA segmentations, as reflected by a small drop in DSC and a relatively larger increase in ASD (Table 4.2). This difference is likely due to the model's difficulty in accurately delineating the MCA-ACA border, which is located in the middle of the brain and challenging to define, whereas ICA segmentations have clearer borders aligned with the skull in the lateral view and the sagittal midline in the AP view (Figures 4.8 and 4.9). Additionally, the larger mask area of the ICA results in greater overlap, leading to higher DSC and lower ASD values. The nature of these metrics can explain the observed relative differences between the DSC and ASD. DSC calculates the overlap between masks, and since the masks cover a large portion of the MinIP, the DSC remains high due to the substantial overlap. Conversely, ASD does not consider the total pixel count of the mask, leading to more pronounced differences. This highlights a limitation of the ASD metric, as it does not account for variations in mask size. Future research could incorporate a local metric better suited to capturing differences in segmentation scaling.

Initially, we included all eligible images due to the high variability that exists in DSA acquisitions. However, our experiment using only half of the training data demonstrated that the combined model could still achieve similar performance.

The phase split analysis (4.5.4) revealed that the combined model mainly relied on contrast information from the arterial and parenchymal phases for creating accurate predictions in the correct hemisphere and lateral orientation. This is evidenced by errors occurring in the venous and non-contrast phases (Figure 4.8) and the non-significant differences of the parenchymal phase with the combined model, which experiences overlap from the arterial phase due to residual contrast agent in the arteries when it first reaches the parenchyma. This finding is advantageous for clinical applicability since the arterial and parenchymal phases are almost always present in DSAs, whereas venous frames may be absent due to early acquisition termination. Nonetheless, we recommend including all available phases, as the venous and non-contrast phases can assist in delineating the superior mask borders in cases of low perfusion, particularly when there is slight motion making the skull outline visible. This analysis was conducted primarily to gain insight into which phases the combined model relied on. To directly compare model performance across different phases, separate models should be trained for each phase individually. For such analyses, it is important to consider phase overlap (e.g., arterial contrast

persisting into the parenchymal phase) for which the use of a Frangi filter could be explored for removing potential arterial and venous vessels from the parenchymal phase [59].

The experiment on occlusion location (4.5.6) demonstrated that the proposed model maintained consistent performance and robustness across different occlusion locations, despite varying distributions of input data. This indicates that the approach of randomized stratification was effective. A minor difference in performance was observed between EVT stages, potentially suggesting better model performance on highly perfused examples; however, the difference was small and had no impact on robustness (Figure 4.12). In contrast, the atlas registration approach showed increased robustness from ICA to M2, and pre- to post-EVT (Figure 4.12). This could be explained by the fact that post-EVT MinIPs of patients with distal occlusions are more similar to the non-stroke atlases. Further investigation into the relationship between perfusion levels and performance metrics for the atlas registration method would be valuable.

The proposed model was significantly faster than the atlas registration method (33 times faster). However, these times could have been influenced by the local CPU of the PC, which could have affected the results due to unknown background tasks. Besides that, the computational time for the atlas method depends on the number of atlases and could have been optimized. Nevertheless, the substantial difference in speed between the methods, combined with the fact that nnUNet can be tuned for GPU acceleration, underscores the potential of the proposed model for clinical application in the autoTICI pipeline.

To our knowledge, this was the first study to investigate an automated deep learning-based approach for segmenting vascular territories on cerebral DSA. Other studies using deep-learning methods for DSA segmentation focused exclusively on vessels [60–66] and intracranial aneurysms [61, 66, 67], as shown in Table 4.4. All studies used convolutional neural networks (CNNs), with most employing the U-Net architecture [60, 62–65, 67], achieving mean DSC scores between 0.8 and 0.9, which is lower than the performance of our model. However, direct comparison is limited, as these studies focused on vessel segmentation, whereas our approach targeted vascular territories. Patel et al. (2020) compared DeepMedic, a 3D convolutional neural network with multiple input channels, to a U-Net model for intracranial vessel segmentation, finding DSC scores of 0.94 ± 0.02 for DeepMedic and 0.92 ± 0.02 for U-Net, both similar to our results [66]. In 2023, Patel et al. applied DeepMedic to a different dataset and reported a DSC of 0.87 ± 0.008 [61]. This difference in outcome may be attributed to the use of different ground truths in the latter study, derived from CTA and coregistered to DSA through 3D-2D registration [61]. We were the only study focused on segmenting vascular territories without clear anatomical boundaries, highlighting a research gap in this area. Given the high performance and robustness of our method, extending this approach to other vascular territories, such as the PCA, or other brain areas critical for stroke outcomes, like the Wernicke and Broca areas, could offer substantial clinical value [68].

The findings of this study should be interpreted whilst taking several limitations into account.
First, the manual annotation of reference segmentations could have introduced bias, particularly given the difficulty in visually delineating borders due to overprojection. Since these annotations were used for both model training and calculation of the performance metrics, the results may not fully reflect the true vascular territories and should therefore be interpreted with caution. Future research should reduce this bias by using more objective and accurate reference standard segmentations delineated on CTA and co-registered with DSA [61].
A second limitation is that the reference standard segmentations for post-EVT MinIPs were performed by MS and those for pre-EVT MinIPs by LS, potentially introducing bias due to inter-observer variability. However, since the pre-EVT segmentations were identical to the post-EVT ones, with adjustments made for motion between pre- and post-EVT MinIPs, differences are expected to be minimal.
A third limitation is that the method was not validated on a larger external test set to assess generalizability. Future research could address this by validating the model on a different subset of the MR CLEAN Registry or patients from the ‘autoTICI to clinic’ study.

A strength of this research is that the proposed segmentation method addressed two key issues identified in chapter 2: improving the robustness and reducing the computational time of autoTICI, bringing autoTICI closer to clinical implementation.

Table 4.4: Comparison of our method to other DL methods for segmentation of cerebral structures on DSA. The DSC scores of the reported studies (and in one study the F1-score) are displayed as mean ± standard deviations or [95% Confidence Intervals]. The DSC scores of our method are reported as median and [Inter Quartile Ranges]

Authors (year of publication)	Segmentation application	Method	Model performance (DSC or F1*)
Our method (2024)	ICA and MCA vascular territories	nnUNet	0.96 [0.94, 0.97] for ICA, 0.94 [0.92, 0.96] for MCA
Liu (2024) [64]	Intracranial arteries	VSS-Net: U-Net capturing spatiotemporal features	0.76
Patel (2023) [61]	Intracranial vessels + IA	DeepMedic	0.87 ± 0.008
Vepa (2022) [63]	Cerebral vessels	Weakly-Supervised CNN	0.84 ± 0.005
Meng (2020) [62]	Cerebral vessels	U-Net with skip connections	0.88*
Jin (2020) [67]	IA	U-Net	0.53
Patel (2020) [66]	Brain vessels with IA	DeepMedic and U-Net	0.94 ± 0.02, 0.92 ± 0.02
Zhang (2020) [60]	Intracranial vessels	U-Net	0.83 ± 0.052
Neumann (2018) [65]	Cerebral arteries	U-Net	0.88

IA = Intracranial aneurysm; CNN = Convolutional Neural Network; DSC = Dice Similarity Coefficient; ICA = Internal Carotid Artery; MCA = Middle Cerebral Artery

4.7. Conclusions

In conclusion, this study proposed a deep learning method for the segmentation of vascular territories on cerebral DSA, demonstrating excellent results on an internal test set. The method significantly outperformed the current atlas registration approach in robustness, accuracy, and computational time, showing promise for clinical application. Future research should focus on retraining the model with a more objective reference standard co-registered from CTA, followed by validation on a larger external dataset. Ultimately, the model should be integrated into the autoTICI pipeline to enhance its robustness, enabling more objective and reliable reperfusion quantification during EVT.

5

General discussion and conclusions

This master thesis aimed to improve and evaluate autoTICI, with the ultimate goal of implementing it in clinical practice to establish a more uniform and reliable method for structurally reporting outcomes of endovascular thrombectomy (EVT) procedures.

Our pilot study showed that the current performance of autoTICI is insufficient for clinical implementation, mainly due to unreliable atlas registration and inaccurate perfusion segmentation. Despite these challenges, autoTICI demonstrated excellent usability within the clinical workflow, highlighting its potential for future clinical use. Besides, multiple interventional radiologists acknowledged the added value of autoTICI as a decision-support tool during EVT, particularly in challenging situations such as night shifts, high-stress conditions, or complex cases. We showed that autoTICI could be effectively integrated into the interventional radiology workflow through the use of the Open Innovation Platform (OIP) (Philips Medical Systems, Best, the Netherlands), which facilitated reliable streaming and processing of digital subtraction angiography (DSA) images in the angiography suite. The current autoTICI processing time on the OIP was 283 ± 1.90 seconds, which is too long for practical clinical use. Since the majority of this time is consumed by atlas registration (66%), this step was identified as the primary target for speed improvement.

The core-laboratory study showed similar inter-observer agreement levels for extended Thrombolysis In Cerebral Infarction (eTICI) scoring without and with autoTICI, with weighted kappa values of 0.65 and 0.67, respectively. Despite the issues identified in autoTICI, the overall eTICI scores and observer agreement remained unaffected. Furthermore, the addition of autoTICI feedback did not significantly impact the overall observer efficiency, indicating that the supplementary information did not prolong the time required for eTICI scoring — both encouraging results for the potential clinical application of autoTICI.

Future research should focus on improving the performance of autoTICI. This can be achieved by implementing the automated vascular territory segmentation method developed in this thesis, which demonstrated excellent results regarding robustness, segmentation accuracy, and computational efficiency on an internal test set. Future research on this segmentation approach should evaluate its robustness on a larger external test set and explore using a more objective reference standard co-registered from CTA. Improving the accuracy of the perfusion segmentation step is another priority for enhancing autoTICI performance. Future research could potentially investigate the use of different Otsu segmentation thresholds for different regions with contrast intensity differences.
Further automation of autoTICI is also an essential step necessary for clinical implementation. This could be achieved by developing a classification model to automatically detect the view, hemisphere, head orientation, and occlusion location, thereby eliminating the need for manual inputs. Since the proposed segmentation model already accurately identified the correct view, hemisphere, and orientation, training a classification model for these tasks seems feasible.
After implementing these improvements, it is recommended to continue the ‘autoTICI to clinic’ study to gather more data, using the enhanced autoTICI method to allow for a more accurate assessment of its

clinical applicability.

Following these advancements, autoTICI is ready for the next step in the AI funnel, which involves conducting a clinical trial to evaluate autoTICI as a decision-support tool in the angiography suite. Additionally, an inter-observer analysis with the improved version of autoTICI should be performed, involving clinicians with varying levels of experience. Conducting this study in a clinical setting rather than a core-laboratory environment would provide a better evaluation of autoTICI's potential effectiveness in a real-world clinical context. The results of these studies can support the approval of autoTICI as a medical device under the Medical Device Regulations (MDR).

In this master thesis, autoTICI was successfully integrated into the OIP's technical infrastructure, enabling its potential clinical use within the workflow of the interventional radiology. The pilot study and inter-observer analysis identified autoTICI's intended clinical role, its added value, future opportunities and key challenges, including technical issues like unreliable atlas registration. By developing a deep-learning approach for vascular territory segmentation, both the robustness and computational time of autoTICI were significantly improved. Although autoTICI is not yet ready for clinical implementation, with further performance optimization needed regarding perfusion segmentation, this thesis has made substantial progress in advancing autoTICI towards its goal of improving the comparability of stroke treatment outcomes across clinical trials.

References

[1] Shavonne Williams, Kimberly Glaser, and Bappaditya Ray. “Strokes and predictors of outcomes”. In: *Critical Care Nursing Clinics* 35.1 (2023), pp. 1–15.

[2] Rebbeca A Grysiewicz, Kurian Thomas, and Dilip K Pandey. “Epidemiology of ischemic and hemorrhagic stroke: incidence, prevalence, mortality, and risk factors”. In: *Neurologic clinics* 26.4 (2008), pp. 871–895.

[3] Peter Sommer et al. “Is functional outcome different in posterior and anterior circulation stroke?” In: *Stroke* 49.11 (2018), pp. 2728–2732.

[4] Ivo GH Jansen, Maxim JHL Mulder, and Robert-Jan B Goldhoorn. “Endovascular treatment for acute ischaemic stroke in routine clinical practice: prospective, observational cohort study (MR CLEAN Registry)”. In: *bmj* 360 (2018).

[5] Kars CJ Compagne et al. “Improvements in endovascular treatment for acute ischemic stroke: a longitudinal study in the MR CLEAN Registry”. In: *Stroke* 53.6 (2022), pp. 1863–1872.

[6] Marie Louise E Bernsen et al. “Importance of occlusion site for thrombectomy technique in stroke: comparison between aspiration and stent retriever”. In: *Stroke* 52.1 (2021), pp. 80–90.

[7] Mayank Goyal et al. “What constitutes the M1 segment of the middle cerebral artery?” In: *Journal of neurointerventional surgery* 8.12 (2016), pp. 1273–1277.

[8] F Mounier-Vehier, D Leys, and JP Pruvo. “Stroke patterns in unilateral atherothrombotic occlusion of the internal carotid artery”. In: *Stroke* 26.3 (1995), pp. 422–425.

[9] Zhang Shi et al. “Time-of-flight intracranial MRA at 3 T versus 5 T versus 7 T: visualization of distal small cerebral arteries”. In: *Radiology* 306.1 (2023), pp. 207–217.

[10] M. Goyal. “et al. Endovascular thrombectomy after large-vessel ischaemic stroke: a meta-analysis of individual patient data from five randomised trials”. In: *Lancet* 387 (2016), pp. 1723–1731.

[11] O.A. Berkhemer. “et al (Mr CLEAN investigators). A Randomized Trial of Intraarterial Treatment for Acute Ischemic Stroke.” In: *N Engl J Med.* 372.1 (2015), pp. 11–20.

[12] Tasneem F Hasan et al. “Endovascular thrombectomy for acute ischemic stroke”. In: *Current Cardiology Reports* 21 (2019), pp. 1–12.

[13] Osama O Zaidat et al. “Recommendations on angiographic revascularization grading standards for acute ischemic stroke: a consensus statement”. In: *Stroke* 44.9 (2013), pp. 2650–2663.

[14] Randall T Higashida and Anthony J Furlan. “Trial design and reporting standards for intra-arterial cerebral thrombolysis for acute ischemic stroke”. In: *stroke* 34.8 (2003), e109–e137.

[15] Johannes Kaesmacher et al. “Cross-Sectional Imaging Modalities in Correlation to the Thrombolysis in Cerebral Infarction Score: The Next Frontier in Adjunctive Endovascular Stroke Therapy”. In: *Stroke: Vascular and Interventional Neurology* (2023), e001063.

[16] IIT Investigators. “The interventional management of stroke (IMS) II study”. In: *Stroke* 38.7 (2007), pp. 2127–2135.

[17] Mohammed A Almekhlafi et al. “Not all “successful” angiographic reperfusion patients are an equal validation of a modified TICI scoring system”. In: *Interventional Neuroradiology* 20.1 (2014), pp. 21–27.

[18] David S Liebeskind et al. “eTICI reperfusion: defining success in endovascular stroke therapy”. In: *Journal of neurointerventional surgery* 11.5 (2019), pp. 433–438.

[19] Elizabeth A Noser et al. “Aggressive mechanical clot disruption: a safe adjunct to thrombolytic therapy in acute stroke?” In: *Stroke* 36.2 (2005), pp. 292–296.

[20] Maximilian Nielsen et al. “Deep learning–based automated thrombolysis in cerebral infarction scoring: a timely proof-of-principle study”. In: *Stroke* 52.11 (2021), pp. 3497–3504.

[21] Joung-Ho Rha and Jeffrey L Saver. “The impact of recanalization on ischemic stroke outcome: a meta-analysis”. In: *stroke* 38.3 (2007), pp. 967–973.

[22] Bruno P Soares, Jeffrey D Chien, and Max Wintermark. “MR and CT monitoring of recanalization, reperfusion, and penumbra salvage: everything that recanalizes does not necessarily reperfuse!” In: *Stroke* 40.3_suppl_1 (2009), S24–S27.

[23] Pooja Khatri et al. “Revascularization end points in stroke interventional trials: recanalization versus reperfusion in IMS-I”. In: *Stroke* 36.11 (2005), pp. 2400–2403.

[24] Albert J Yoo et al. “Refining angiographic biomarkers of revascularization: improving outcome prediction after intra-arterial therapy”. In: *Stroke* 44.9 (2013), pp. 2509–2512.

[25] Maarten G Lansberg et al. “MRI profile and response to endovascular reperfusion after stroke (DEFUSE 2): a prospective cohort study”. In: *The Lancet Neurology* 11.10 (2012), pp. 860–867.

[26] T Tomsick et al. “Revascularization results in the Interventional Management of Stroke II trial”. In: *American Journal of Neuroradiology* 29.3 (2008), pp. 582–587.

[27] Ruisheng Su et al. “autoTICI: automatic brain tissue reperfusion scoring on 2D DSA images of acute ischemic stroke patients”. In: *IEEE transactions on medical imaging* 40.9 (2021), pp. 2380–2391.

[28] P. M. van der Sluijs et al. “Assessment of automated TICI scoring during endovascular treatment in patients with an ischemic stroke”. In: *J. Neurointerv. Surg.* (2024).

[29] Sander Greenland. “Avoiding power loss associated with categorization and ordinal scores in dose-response and trend analysis”. In: *Epidemiology* 6.4 (1995), pp. 450–454.

[30] Lea Strohm et al. “Implementation of artificial intelligence (AI) applications in radiology: hindering and facilitating factors”. In: *European radiology* 30 (2020), pp. 5525–5532.

[31] Katharina Wenderott et al. “Prospective effects of an artificial intelligence-based computer-aided detection system for prostate imaging on routine workflow and radiologists’ outcomes”. In: *European Journal of Radiology* 170 (2024), p. 111252.

[32] Filippo Pesapane, Marina Codari, and Francesco Sardanelli. “Artificial intelligence in medical imaging: threat or opportunity? Radiologists again at the forefront of innovation in medicine”. In: *European radiology experimental* 2 (2018), pp. 1–10.

[33] Dakai Jin et al. “Artificial intelligence in radiology”. In: *Artificial Intelligence in Medicine*. Elsevier, 2021, pp. 265–289.

[34] Rebecca Steketee. “Een proces voor de ontwikkeling, validatie, en het gebruik van AI: kwaliteit, compliance, efficiëntie”. In: (2023).

[35] Mayank Goyal et al. “Standardized reporting of workflow metrics in acute ischemic stroke treatment: why and how?” In: *Stroke: Vascular and Interventional Neurology* 1.1 (2021), e000177.

[36] Zarina Abdul Assis, Bijoy K Menon, and Mayank Goyal. “Imaging department organization in a stroke center and workflow processes in acute stroke”. In: *European journal of radiology* 96 (2017), pp. 120–124.

[37] Prithvi Santana Baskar et al. “In-hospital acute stroke workflow in acute stroke–Systems-based approaches”. In: *Acta Neurologica Scandinavica* 143.2 (2021), pp. 111–120.

[38] Willemijn J Maas et al. “Expediting workflow in the acute stroke pathway for endovascular thrombectomy in the northern Netherlands: a simulation model”. In: *BMJ open* 12.4 (2022), e056415.

[39] CONTRAST Consortium. *DIVINE - The Dutch registry for endovascular therapy in ischaemic stroke with imaging evaluation*. `https://www.contrast-consortium.nl/ongoing-trials/divine/`. Retrieved August 1, 2024.

[40] John Brooke et al. “SUS-A quick and dirty usability scale”. In: *Usability evaluation in industry* 189.194 (1996), pp. 4–7.

[41] Aaron Bangor, Philip Kortum, and James Miller. “Determining what individual SUS scores mean: Adding an adjective rating scale”. In: *Journal of usability studies* 4.3 (2009), pp. 114–123.

[42] M Bach Cuadra, Valérie Duay, and J-Ph Thiran. “Atlas-based segmentation”. In: *Handbook of Biomedical Imaging: Methodologies and Clinical Research* (2015), pp. 221–244.

[43] Minjie Wu et al. “Optimum template selection for atlas-based segmentation”. In: *NeuroImage* 34.4 (2007), pp. 1612–1618.

[44] Lotte Strong. “Image registration techniques for cerebral digital subtraction angiography - a review of the state-of-the-art”. In: (2024).

[45] Waleed Brinjikji et al. “e-ASPECTS software improves interobserver agreement and accuracy of interpretation of aspects score”. In: *Interventional Neuroradiology* 27.6 (2021), pp. 781–787.

[46] Kevin A Hallgren. “Computing inter-rater reliability for observational data: an overview and tutorial”. In: *Tutorials in quantitative methods for psychology* 8.1 (2012), p. 23.

[47] Ondrej Volny, Petra Cimflova, and Viktor Szeder. “Inter-rater reliability for thrombolysis in cerebral infarction with TICI 2c category”. In: *Journal of Stroke and Cerebrovascular Diseases* 26.5 (2017), pp. 992–994.

[48] David S Liebeskind et al. “eTICI reperfusion: defining success in endovascular stroke therapy”. In: *Journal of neurointerventional surgery* 11.5 (2019), pp. 433–438.

[49] Guang Zhang et al. “Operator versus core lab adjudication of reperfusion after endovascular treatment of acute ischemic stroke”. In: *Stroke* 49.10 (2018), pp. 2376–2382.

[50] Robert Fahed et al. “Agreement between core laboratory and study investigators for imaging scores in a thrombectomy trial”. In: *Journal of neurointerventional surgery* 10.12 (2018), e30–e30.

[51] DM Heiferman et al. “Reliability of the modified TICI score among endovascular neurosurgeons”. In: *American Journal of Neuroradiology* 41.8 (2020), pp. 1441–1446.

[52] J Richard Landis and Gary G Koch. “The measurement of observer agreement for categorical data”. In: *biometrics* (1977), pp. 159–174.

[53] Terry K Koo and Mae Y Li. “A guideline of selecting and reporting intraclass correlation coefficients for reliability research”. In: *Journal of chiropractic medicine* 15.2 (2016), pp. 155–163.

[54] Julius Sim and Chris C Wright. “The kappa statistic in reliability studies: use, interpretation, and sample size requirements”. In: *Physical therapy* 85.3 (2005), pp. 257–268.

[55] Paul Aljabar et al. “Multi-atlas based segmentation of brain images: atlas selection and its effect on accuracy”. In: *Neuroimage* 46.3 (2009), pp. 726–738.

[56] Fabian 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.

[57] Fabian Isensee et al. “Automated design of deep learning methods for biomedical image segmentation”. In: *arXiv preprint arXiv:1904.08128* (2019).

[58] Esther Puyol-Antón et al. “AI-Enabled Assessment of Cardiac Systolic and Diastolic Function from Echocardiography”. In: *International Workshop on Advances in Simplifying Medical Ultrasound*. Springer. 2022, pp. 75–85.

[59] Alejandro F Frangi et al. “Multiscale vessel enhancement filtering”. In: *Medical Image Computing and Computer-Assisted Intervention—MICCAI’98: First International Conference Cambridge, MA, USA, October 11–13, 1998 Proceedings 1*. Springer. 1998, pp. 130–137.

[60] Min Zhang et al. “A neural network approach to segment brain blood vessels in digital subtraction angiography”. In: *Computer methods and programs in biomedicine* 185 (2020), p. 105159.

[61] Tatsat R Patel et al. “Evaluating a 3D deep learning pipeline for cerebral vessel and intracranial aneurysm segmentation from computed tomography angiography–digital subtraction angiography image pairs”. In: *Neurosurgical Focus* 54.6 (2023), E13.

[62] Cai Meng et al. “Multiscale dense convolutional neural network for DSA cerebrovascular segmentation”. In: *Neurocomputing* 373 (2020), pp. 123–134.

[63] Arvind Vepa et al. “Weakly-supervised convolutional neural networks for vessel segmentation in cerebral angiography”. In: *Proceedings of the IEEE/CVF Winter Conference on applications of computer vision*. 2022, pp. 585–594.

[64] Wentao Liu et al. “DIAS: a dataset and benchmark for intracranial artery segmentation in DSA sequences”. In: *Medical Image Analysis* (2024), p. 103247.

[65] Christian Neumann, Klaus-Dietz Tönnies, and R AngioUnet Pohle-Fröhlich. “a convolutional neural network for vessel segmentation in cerebral DSA series”. In: *Proc VISIGRAPP* (2018).

[66] Tatsat R Patel et al. “Multi-resolution CNN for brain vessel segmentation from cerebrovascular images of intracranial aneurysm: a comparison of U-Net and DeepMedic”. In: *Medical Imaging 2020: Computer-Aided Diagnosis*. Vol. 11314. SPIE. 2020, pp. 677–685.

[67] Hailan Jin et al. “Fully automated intracranial aneurysm detection and segmentation from digital subtraction angiography series using an end-to-end spatiotemporal deep neural network”. In: *Journal of NeuroInterventional Surgery* 12.10 (2020), pp. 1023–1027.

[68] JC Brust et al. “Aphasia in acute stroke.” In: *Stroke* 7.2 (1976), pp. 167–174.

A

Open Innovation Platform

A.0.1. Technical infrastructure

To integrate autoTICI into the intended clinical workflow of the interventional radiology department, autoTICI was implemented in an infrastructure hosted by Philips Medical Systems (Best, the Netherlands), called the Open Innovation Platform (OIP). The OIP is a rapid prototyping platform designed to facilitate the development of cloud-based applications that integrate with interventional procedures. The OIP functions by streaming data from the imaging system via 'CloudCast' hardware to a storage called 'IGT cloud'. From there, imaging data are uploaded to the cloud, where necessary computations are performed for the execution of applications, and results are sent back to the CloudCast. These results can then be connected to the touchscreen interface and display of the imaging system. A connection to the cloud can also be established via a 'Websocket' connection, enabling real-time streaming of imaging data directly to the OIP, bypassing the IGT cloud. However, this streaming option has not yet been implemented for autoTICI to clinic' due to technical issues and the requirement to obtain deferred consent before sending imaging data to the OIP. Figure A.1 shows a schematic overview of the OIP infrastructure.

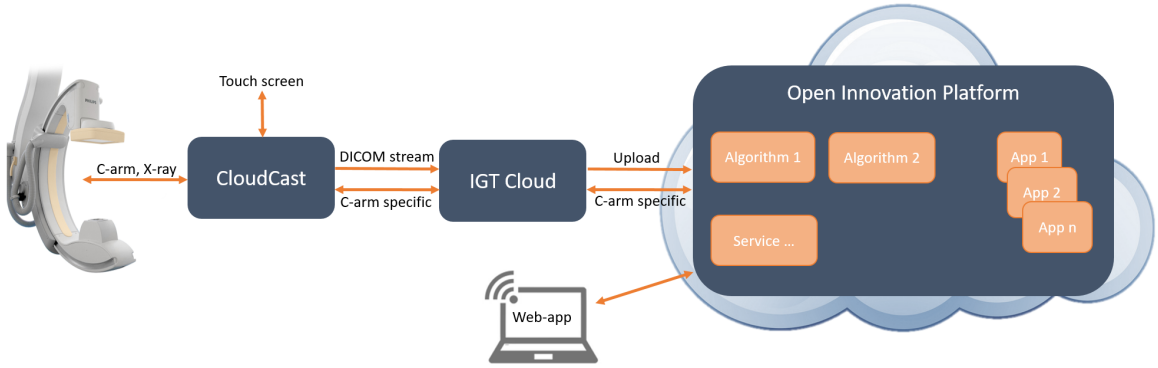


Figure A.1: Schematic overview of the Open Innovation Platform (OIP) infrastructure. Images get streamed from the imaging system to the 'IGT cloud' storage via the 'CloudCast' hardware. From the 'IGT cloud', images are uploaded to the OIP, where they are processed, after which the results are sent back to the CloudCast. These results can then be displayed on the imaging system's touchscreen interface. A connection to the cloud can also be established via a 'Websocket' connection, enabling real-time streaming of imaging data directly to the OIP, bypassing the IGT cloud.

A.0.2. AutoTICI implementation

To implement autoTICI on the OIP, the autoTICI algorithm, along with its pipeline and necessary Python packages, was configured within the OIP development environment, which operates on a Jupyter Hub platform using Python 3.11.

A CE-approved workstation was installed in the angiography system cabinet at Erasmus MC to run the CloudCast software, facilitating secure data streaming to the cloud for which an internet connection with an open port was established. Measures were taken to protect patient privacy, including pseudonymizing DICOM data before streaming and ensuring that no data was shared with Philips. Additionally, multi-factor authentication was required for CloudCast hardware connections. In preparation for clinical use, preliminary testing was conducted on a similar angiography system (Xper FD20/10) at Philips Medical Systems (Best, the Netherlands). This testing verified that the imaging data was correctly pseudonymized and complete.

The CloudCast hardware streamed all imaging data acquired during EVT procedures, temporarily storing the data on the IGT cloud until patient consent was obtained. Once consent was secured, the data was uploaded to the OIP for further processing and analysis. The imaging data acquired during EVT procedures comprised both fluoroscopy images and DSA sequences. To optimize storage, a script was implemented to filter out unnecessary fluoroscopy images, focusing solely on DSA sequences required for autoTICI. Further processing was performed to differentiate between unsubtracted and subtracted frames of the streamed DSA sequences, ensuring that only subtracted images were retained for further analysis. This differentiation was based on the average intensity values derived from their histograms (Figure A.2).

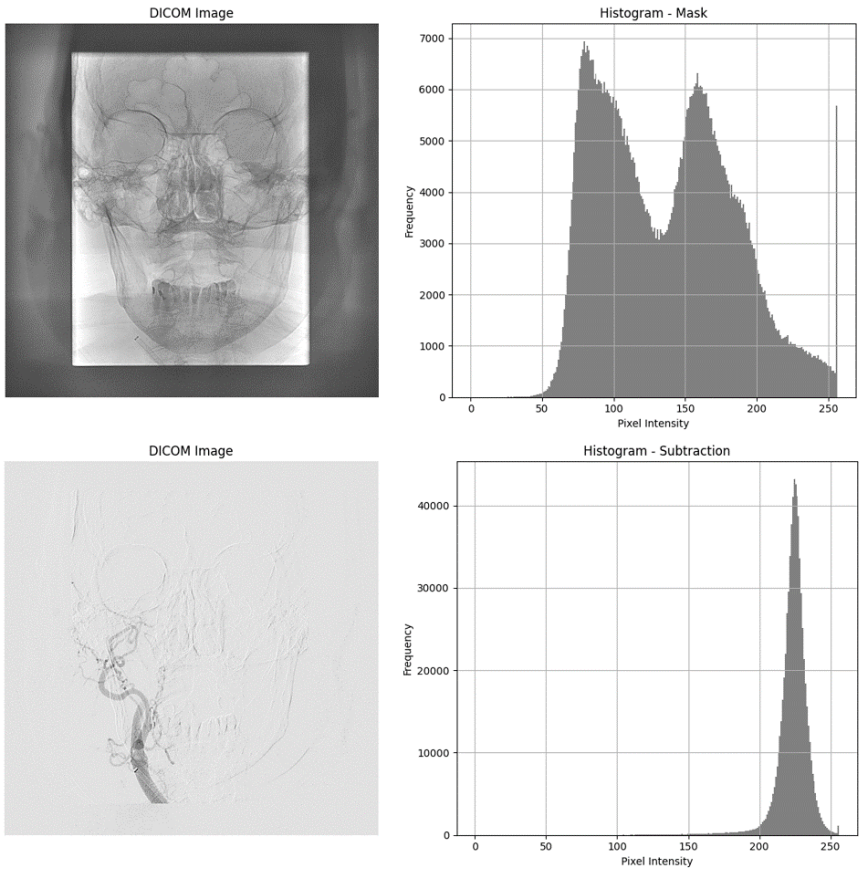


Figure A.2: Visualization of the histograms for unsubtracted and subtracted images from the DSA sequence. The distinction between unsubtracted and subtracted images was made using the average intensity values obtained from their histograms.

Finally, the DSA frames, streamed as single-frame DICOM files in a random order, were organized into multi-frame DICOM files, ensuring that the frames in each sequence were ordered correctly according to the chronological acquisition. These processed sequences were then stored in a designated cloud folder, ready for further processing and analysis by the autoTICI algorithm.

B

User-interface designs

B.0.1. List of requirements

To ensure a smooth adoption of autoTICI amongst interventional radiologists, a user-interface (UI) design was created for visualizing the results of autoTICI in the OIP in a user-friendly and efficient manner. Design requirements were constructed in consultation with both technical and medical experts in the field of stroke imaging and treatment.

We determined that the outcomes of autoTICI should be visually represented through mask overlays on the 2D minimum intensity projections (MINIPs), depicting the calculated reperfused and non-perfused areas. The masks should have a semi-transparent filling ensuring visibility of the structures in the underlying image. For clarity, easily distinguishable colours were chosen: red for non-perfused tissue and green for reperfused tissue, with the corresponding percentages colour-coded accordingly. Although no interventional radiologists involved were colour-blind, the colours can be adjusted to blue and yellow if needed. To enhance the functionality of autoTICI, the UI design should incorporate features that enable visualization of intermediate steps where errors might occur, such as brain area registration and the pre- to post-EVT DSA registration steps. This functionality allows users to assess the accuracy and reliability of the autoTICI predictions and understand any discrepancies. Additionally, incorporating the original input DSA sequences is essential, as they provide insight into the images used by the algorithm and facilitate verification of the visual eTICI score in case of incorrect autoTICI predictions. Additionally, the UI design should be intuitive and quick, as clinicians prefer straightforward features. The design should be created within the OIP environment to ensure compatibility and ease of implementation.

B.0.2. Worked out designs

Based on the list of requirements, three potential UI designs were developed in the OIP development environment using Python 3.11 and open-source packages openCV and ipywidgets: the 1) basic, 2) transparency, and 3) interactive design. Examples of each design can be found in Appendix B.

The basic design included the minimally required information and features as specified in the list of requirements (Figure B.1), including the post-EVT MINIP with the overlayed autoTICI mask displayed next to the four input DSA sequences. The autoTICI mask showed reperfused areas in green and non-perfused areas in red with a fully opaque outline and a 15% opaque filling. The legend displayed the colour-coded percentages of the non-perfused and reperfused areas. The white contour indicated the outline of the registered brain mask, allowing the user to determine if the registration was successful. In the transparency design, the opacity of the mask fill could be adjusted by a slider (Figure B.2). The interactive design included additional interactive widgets (Figure B.3). This allowed the user to display the overlay results on both the pre-and post-EVT MINIPs (Figure B.4). Another feature in the design displayed the original brain mask which was registered to the patient, providing more insight into the brain mask registration process.

The three designs were shown to two experienced neuro-interventional radiologists for comparison in terms of usability and intuitiveness. Both favoured the basic design because it provided sufficient insight

into the accuracy and reliability of the autoTICI predictions while maintaining the most straightforward and simple interface. Therefore, the basic design was selected for displaying the results of autoTICI during the pilot study.

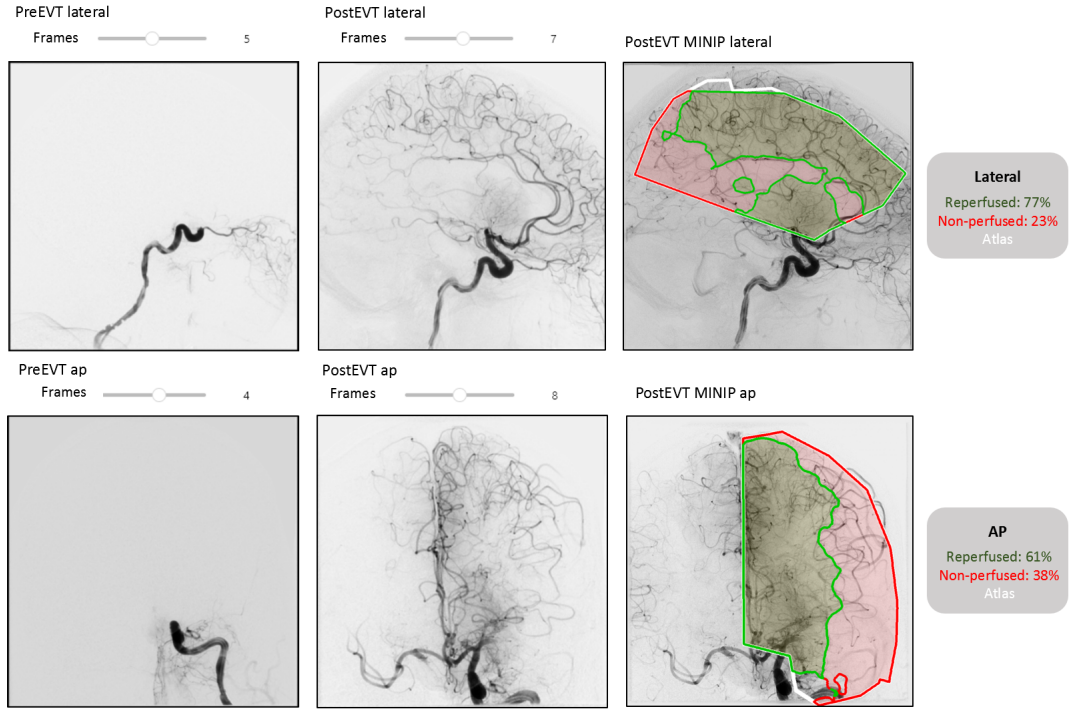


Figure B.1: Basic design: the four input DSA sequences are shown on the left, alongside the post-EVT MINIP with the autoTICI mask overlays. The legend shows the colour-coded percentages.

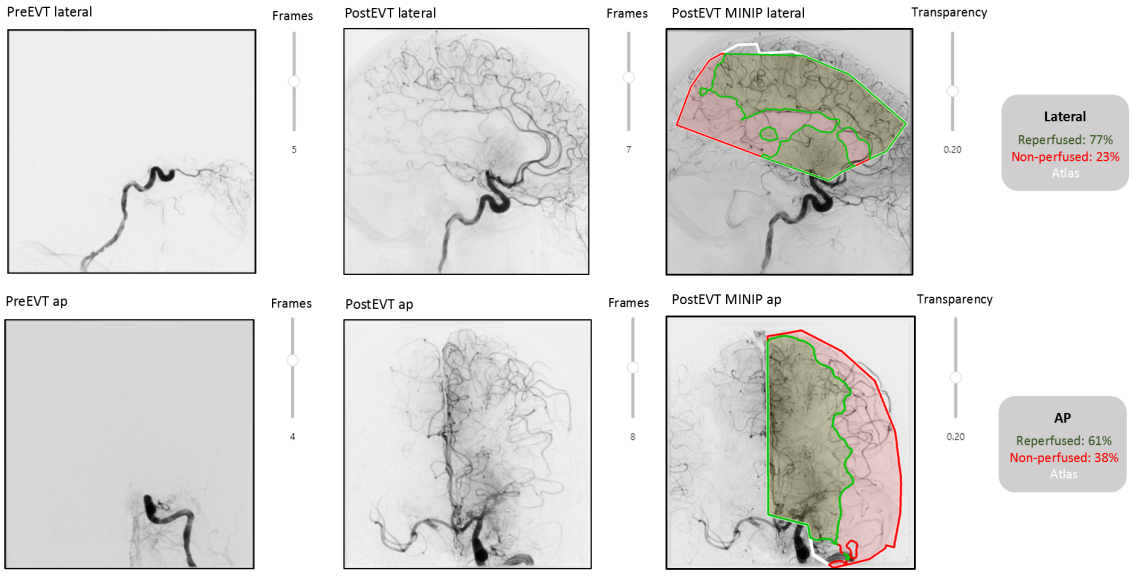


Figure B.2: Transparency design: the four input DSA sequences are shown on the left, alongside the post-EVT MINIP with the autoTICI mask overlays. The transparency of the mask fill can be adjusted with a slider. The legend shows the colour-coded percentages.

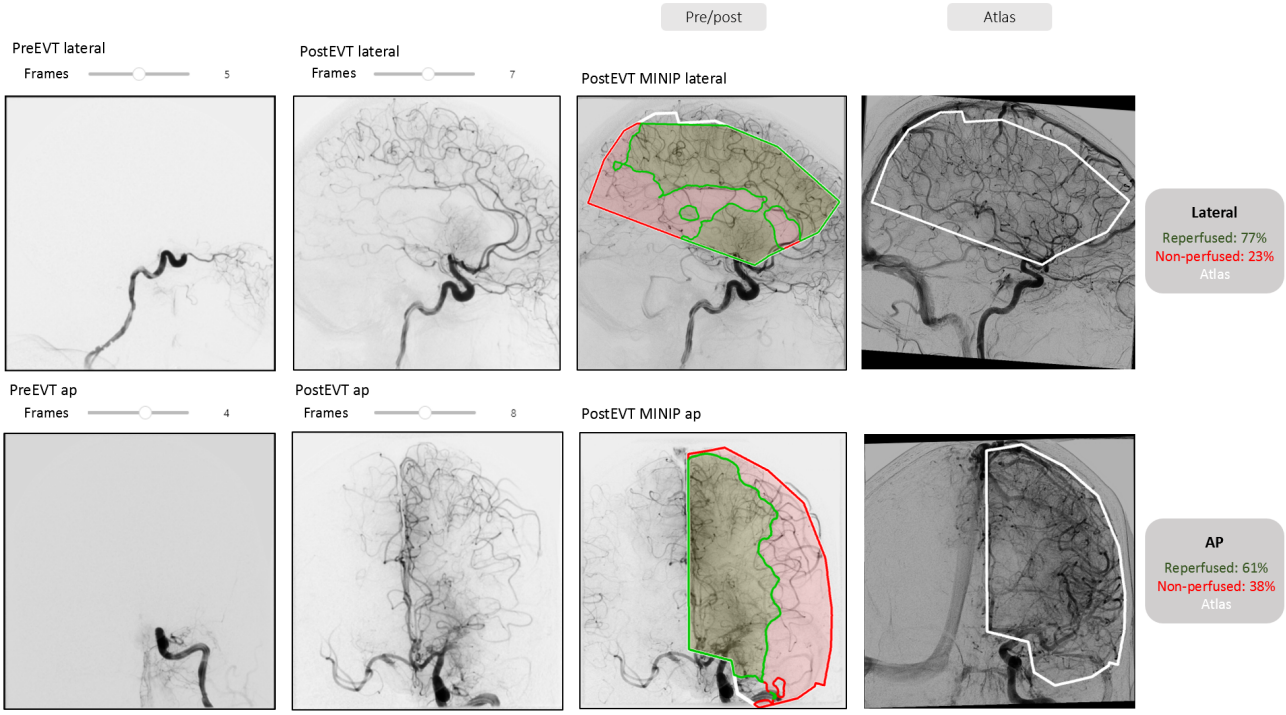


Figure B.3: Interactive design: the four input DSA sequences are shown on the left, alongside the post-EVT MINIP with the autoTICI mask overlays. The original brain atlas is visualized when clicking the 'Atlas' button. The legend shows the colour-coded percentages.

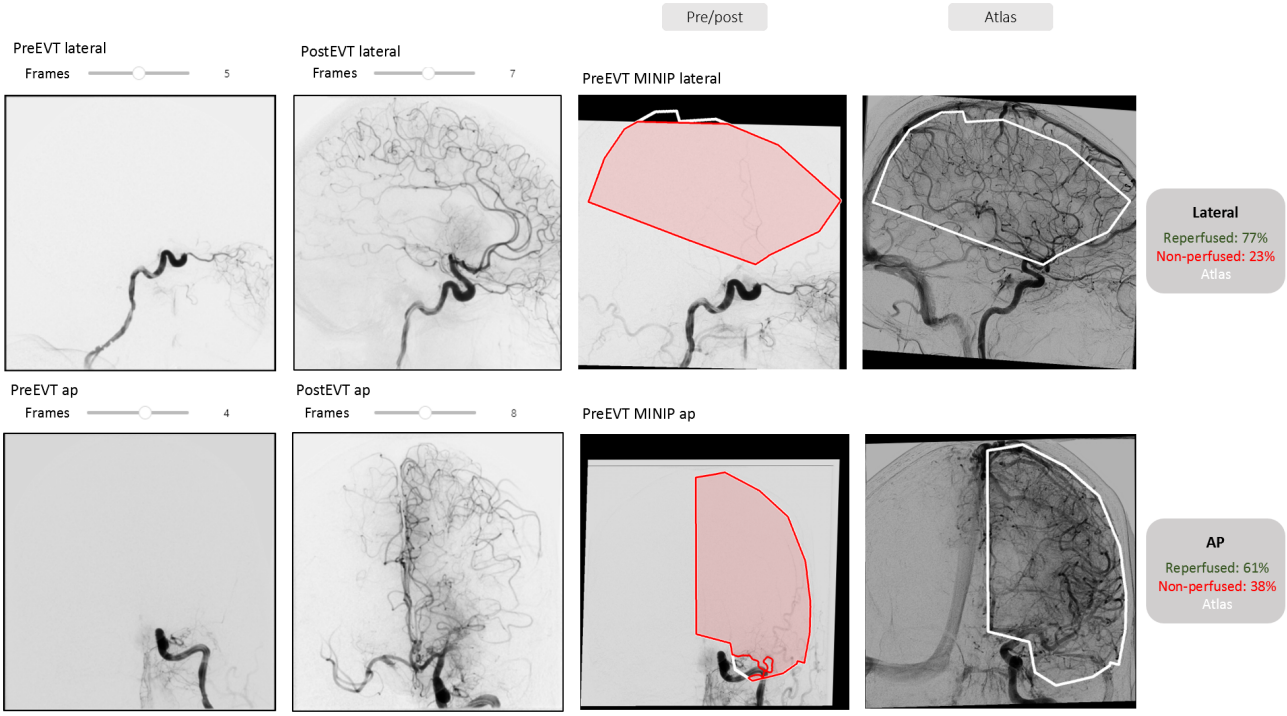


Figure B.4: Interactive design: the four input DSA sequences are shown on the left, alongside the pre-EVT MINIP with the autoTICI mask overlays. By clicking the 'Pre/post' button, the user can switch between the pre-EVT and post-EVT MINIP and corresponding overlays. The original brain atlas is visualized when clicking the 'Atlas' button. The legend shows the colour-coded percentages.

C

AutoTICI to clinic: interview questions

Name:

Role (speciality):

- ☐ Interventional radiologist (neuro)
- ☐ Interventional radiologist (body)
- ☐ Interventional radiologist fellow
- ☐ Interventional radiologist trainee

Number of years of experience in independently performing EVT's:

The following is a questionnaire on the **current** method of visual eTICI scoring:

1. I find the scoring method easy to use
2. I find this scoring method inconsistent
3. I think most interventional radiologists can learn this scoring method very quickly
4. I feel very confident using this scoring method

Please rate each statement from 1 (Strongly Disagree) to 5 (Strongly Agree)

Comments:

What eTICI score did you give during EVT?

Do you have doubts in defining the eTICI score?

☐ No
☐ Yes, why?

What percentage would you give for the amount of brain reperfusion?

The following is a questionnaire on the **automatic** method of eTICI scoring with autoTICI feedback:

1. I would like to use this scoring method frequently
2. I find this scoring method unnecessarily complicated
3. I find the scoring method easy to use
4. I need the support of a (technical) expert to use this scoring method
5. I find the various functions in this scoring method well integrated
6. I find this scoring method inconsistent
7. I think most interventional radiologists can learn this scoring method very quickly
8. I find the scoring method very cumbersome to use
9. I feel very confident using this scoring method
10. I needed to learn a lot of things before I could get going with this scoring method

Rate each statement from 1 (Strongly Disagree) to 5 (Strongly Agree)

Comments:

Do you have doubts about if the vasculatory territory is correctly drawn by the model?

☐ No
☐ Yes, why?

Do you have doubts regarding the reperfusion percentage calculated by the autoTICI model?

☐ No
☐ Yes, why?

Now that you see the reperfusion percentage of the autoTICI model, would you change your earlier given eTICI score?

☐ No
☐ Yes, to which eTICI score would you change it and why?

Would you want to make a change in the EVT procedure based on autoTICI? Would you make another attempt?

☐ No

☐ Yes, why?

Did autoTICI give you new insights? What value does autoTICI have for you?

☐ No

☐ Yes, why?

D

AutoTICI to clinic: clinical implementability interview

Table D.1: Characteristics of included clinicians and summary of their answered questions during the clinical implementability interview.

No. interview	1	2	3	4
Patient	ATC001	ATC001	ATC002	ATC004
Clinician (specialty)*	Obs 1 (body)	Obs 2 (body)	Obs 3 (neuro)	Obs 3 (neuro)
Experience in performing EVT (years)	1	2	9	9
eTICI				
Doubts in eTICI scoring	eTICI scoring is generally straightforward but can be challenging in stressful situations or exceptional cases.	Not for this case. However, eTICI scoring can be inconsistent due to conceptual confusion.	A small borderline thrombus at the edge between 2B and 2C made it difficult to distinguish between these scores.	I was uncertain because, although thrombolysis resolved the occlusion, I noticed a small perfusion defect.
Given TICI score (estimated reperfusion %)	2C (99%)	2C (99%)	2C (90%)**	3 (95%)
autoTICI				
Calculated reperfusion (AP, lateral) (%)	96 (93, 99)	96 (93, 99)	57.5 (59, 56)	0 (0, 0)***
Doubt in brain mask	No	No	Yes, mask is too small, missing part of the MCA vascular territory	Yes, mask partly extends into the skull and does not capture the whole MCA area.
Doubt in perfusion percentage	Yes, reperfusion is too low in the AP view due to segmentation error from collateral overestimation.	Yes, the percentage is too low as segmentation did not account for collaterals, but this is not clinically relevant.	Yes, due to undersized mask, the percentage is underesti-mated.	No, but that is since the reperfusion was already 100% pre-EVT.
SUS score	85	87.5	82.5	70

Comments SUS	Hard to say something about technical support as autoTICI is not yet integrated in the workflow.	If autoTICI is automatically integrated, then it would be clinically suitable. Technical support required if manual input is needed.	Hard to say something about consistency, I have not seen autoTICI for multiple patients.	The consistency is low for autoTICI as it gave the wrong prediction twice.
Added value	AutoTICI is valuable as a decision support tool in stressful or exceptional cases.	AutoTICI has added value in the clinic as a decision support tool.	Yes, autoTICI has added value in both clinical and research settings. It could be very nice in cases of doubt.	In this patient, the added clinical value would be high. However, autoTICI could not support me due to unsuccessful mask registration.
Change in eTICI score	No	No	No	No
Another attempt	No	No	No	No

* All clinicians were interventional radiologists, however had different specialties (neuro, body)
** Obs 3 estimated a eTICI 2C during the interview, however in the EHR a eTICI score of 3 was reported.
*** As no post-EVT DSA was performed, the pre-EVT DSA was also used as the post-EVT DSA for autoTICI. Resulting in a reperfusion percentage of 0%, since the eTICI was already 100% at the time of the pre-EVT DSA acquisition.

Table D.2: Likert scales for each usability aspect of the System Usability Scale (SUS) assigned to autoTICI during each interview.

Interview	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
1	5	1	5	1	3	2	5	3	4	1
2	5	1	5	4	3	1	5	1	5	1
3	5	1	5	1	3	3	5	1	2	1
4	4	1	4	2	3	4	4	2	4	2

E

Core-laboratory study: executable for eTICI annotations

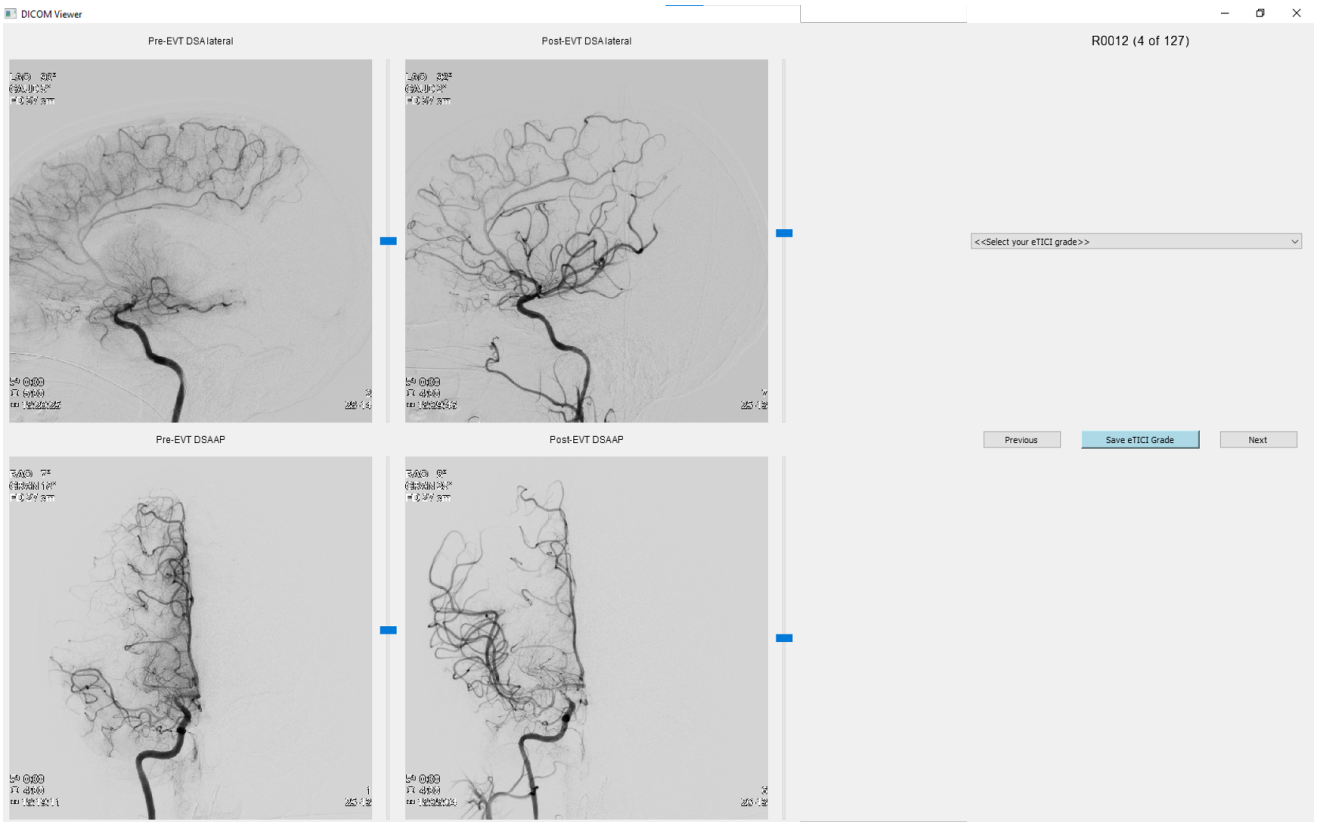


Figure E.1: Executable for annotating eTICI scores, showing the four original DSA sequences and a dropdown menu for selecting the chosen eTICI grade.

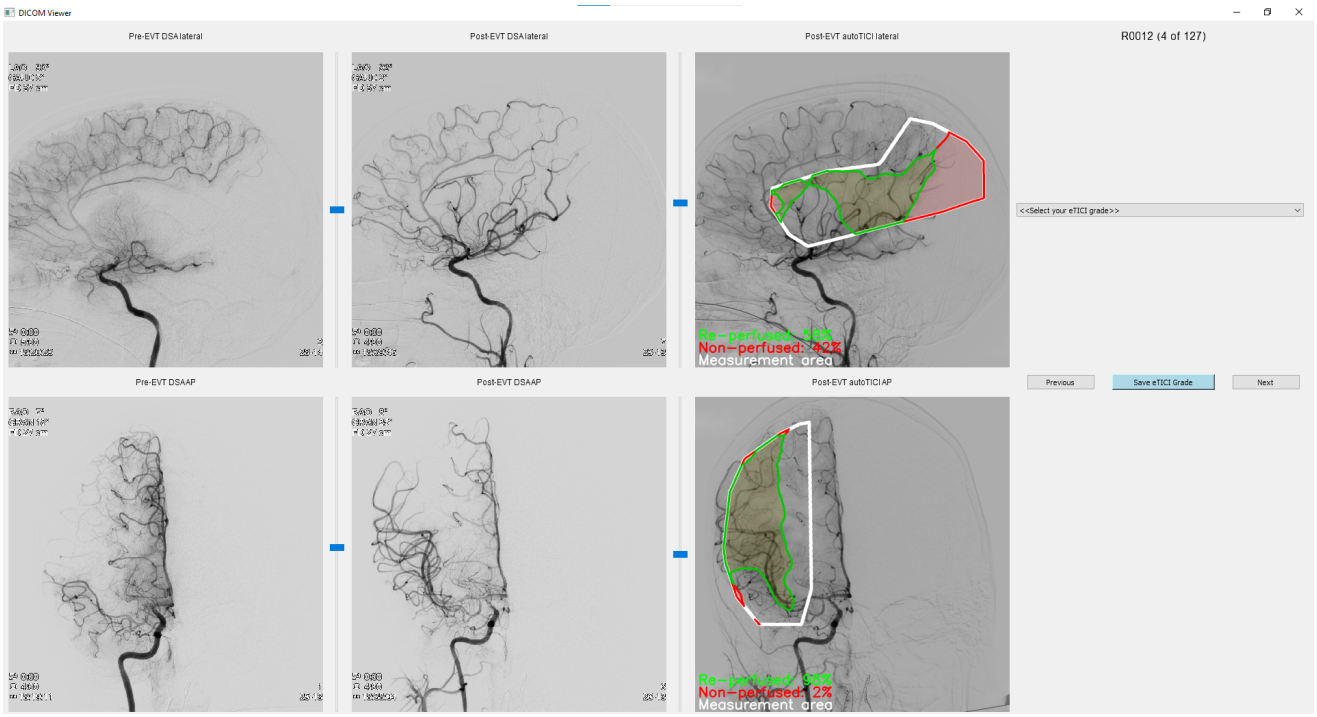


Figure E.2: Executable for annotating eTICI scores, showing the four original DSA sequences alongside the autoTICI result and a dropdown menu for selecting the chosen eTICI grade.

F

Model architecture

F.0.1. Model architecture

The models were trained using a convolutional U-Net architecture featuring a symmetrical contracting (encoder) and expanding path (decoder) interconnected by skip connections. Table F.1 outlines the U-Net layers, including kernel sizes, strides, and output shapes. The final output layer contained three channels corresponding to the background, MCA, and ACA classes. A softmax activation function was applied to convert the output into a probability distribution, and the argmax function was used to determine the class with the highest predicted probability for each pixel.

Table F.1: Configuration of layers for the deep learning model, detailing the layer type, kernel size, stride, and output shape.

Layer type	Kernel size	Stride	Shape
Input	-	-	0, 512, 512
2 conv + leaky ReLU	3x3	2	32, 512, 512
2 conv + leaky ReLU	3x3	2	64, 256, 256
2 conv + leaky ReLU	3x3	2	128, 128, 128
2 conv + leaky ReLU	3x3	2	256, 64, 64
2 conv + leaky ReLU	3x3	2	512, 32, 32
2 conv + leaky ReLU	3x3	2	512, 16, 16
2 conv + leaky ReLU	3x3	2	512, 8, 8
2 conv + leaky ReLU	3x3	2	512, 4, 4
Output	-	-	3, 512, 512

G

Robustness: Likert scale examples

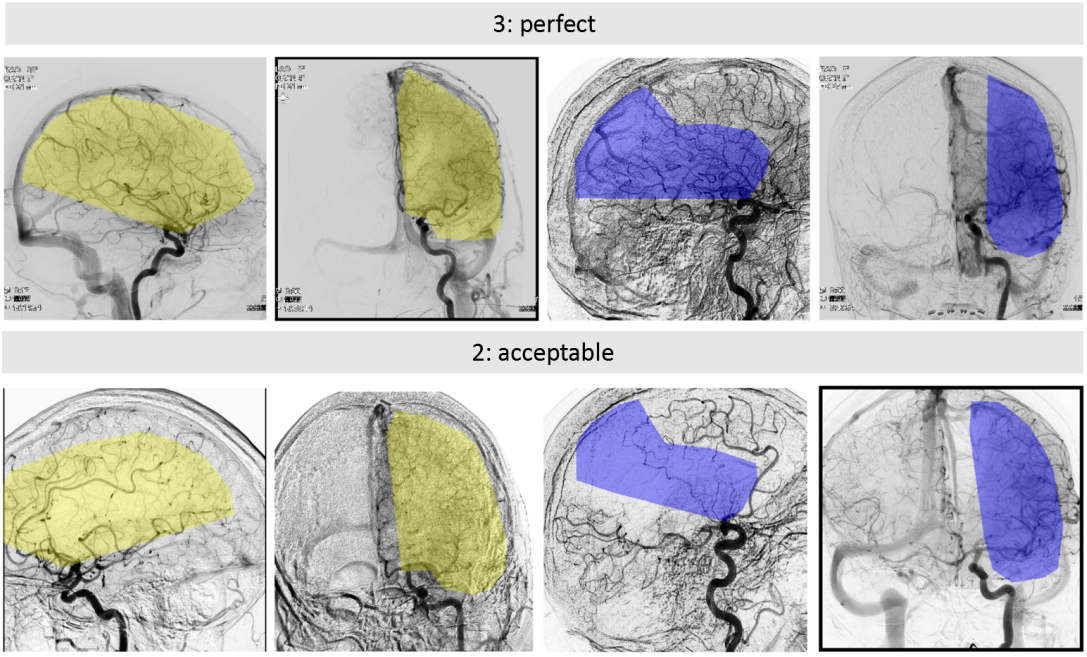


Figure G.1: Examples of Likert scores 2 and 3 for segmentations generated using the atlas registration method. Segmentations of the ICA and MCA are shown in yellow and blue, respectively. *ICA = Internal Carotid Artery; MCA = Middle Cerebral Artery; AP = Anteroposterior*

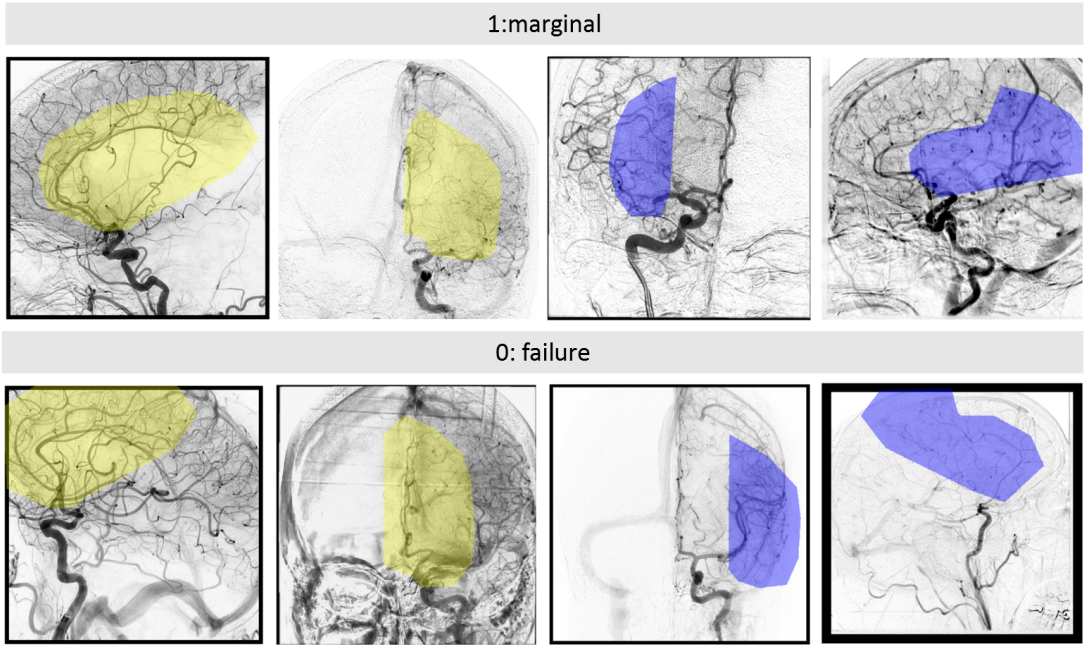


Figure G.2: Examples of Likert scores 0 and 1 for segmentations generated using the atlas registration method. Segmentations of the ICA and MCA are shown in yellow and blue, respectively. *ICA = Internal Carotid Artery; MCA = Middle Cerebral Artery; AP = Anteroposterior*

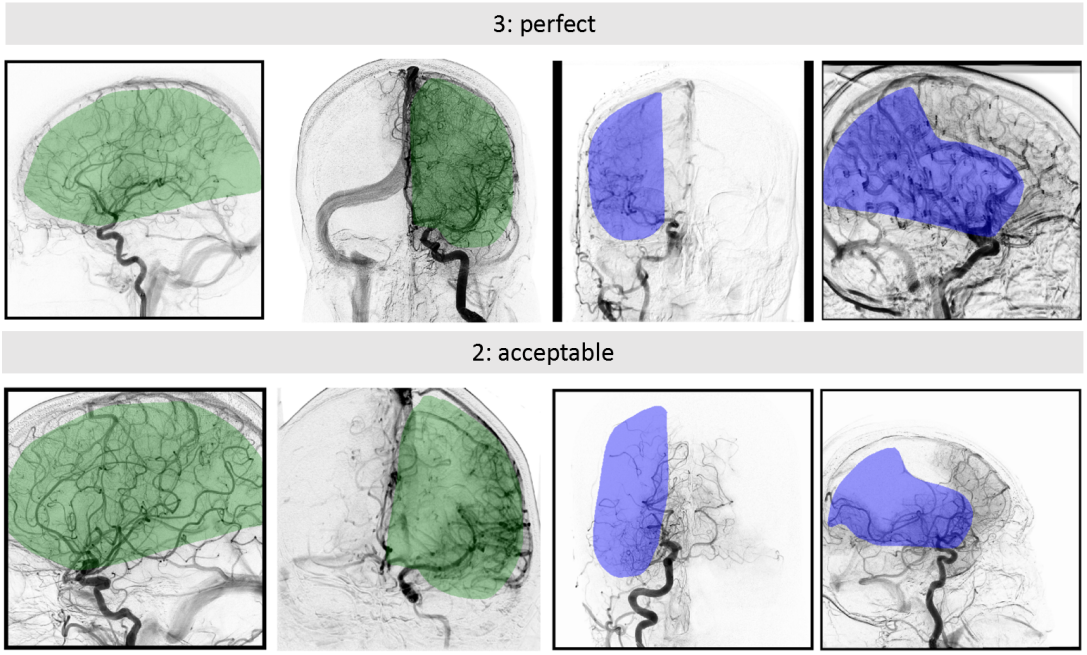


Figure G.3: Examples of Likert scores 2 and 3 for segmentations predicted by the combined segmentation method. Segmentations of the ICA and MCA are shown in yellow and blue, respectively. *ICA = Internal Carotid Artery; MCA = Middle Cerebral Artery; AP = Anteroposterior*