Towards Computer-Aided Diagnosis of Fibromuscular Dysplasia

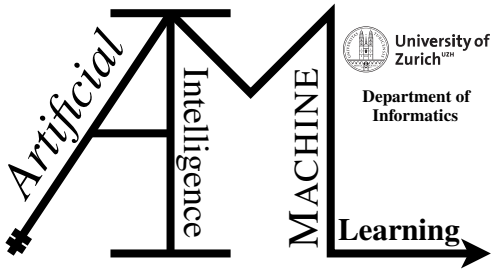
Abnormality Detection in Renal Artery using different Deep Learning Approaches

Master Thesis

Maximilian Achakri
18-700-385

Submitted on
December 2, 2025

Thesis Supervisor
Prof. Dr. Manuel Günther
Prof. Dr. André Anjos, Dr. Oscar Jimenez



Declaration of Independence for Written Work

I hereby declare that I have **composed** this work independently and without the use of any aids other than those declared (including generative AI such as ChatGPT) – the use of generative AI to **improve** my composed work was permitted by the thesis supervisor. I am aware that I take full responsibility for the scientific character of the submitted text myself, even if AI aids were used. All passages taken verbatim or in sense from published or unpublished writings are identified as such. The work has not yet been submitted in the same or similar form or in excerpts as part of another examination.

Zürich, 02.12.2025

Place, Date

Maximilian Achakri

Master Thesis

Author: Maximilian Achakri, maximilian.achakri@uzh.ch

Project period: 02.06.2025 - 02.12.2025

Artificial Intelligence and Machine Learning Group
Department of Informatics, University of Zurich

Acknowledgements

I would like to express my gratitude to Prof. Dr. Manuel Günther for his guidance and expertise throughout the creating of this thesis. I was able to learn a lot from our discussions, and his feedback greatly improved the quality of this work.

I would also like to thank Prof. Dr. André Anjos for his valuable insights and expertise on the project. His knowledge of medical imaging and deep learning was a significant factor in the success of this work. Furthermore, I would like to thank Dr. Oscar Jiménez for his continuous support and constructive feedback during the development of this thesis. He took the time to discuss various aspects of the thesis with me, helping me refine my ideas and improve its overall quality.

I would also like to thank the IDIAP Research Institute for providing access to their computational resources, without which this work would not have been possible.

As this thesis marks the end of my studies, I would also like to thank all my friends for a wonderful time at the University of Zurich. I would also like to thank my girlfriend for her support and patience during the final months of writing this thesis. Finally, I would like to thank my parents for their endless support and encouragement throughout my studies and life in general. Without all of you, this would not have been possible.

Abstract

Fibromuscular dysplasia (FMD) manifests through subtle arterial abnormalities, which renders the automated detection of FMD with deep learning and Computed Tomography Angiography (CTA) particularly challenging. The present thesis investigates deep learning strategies for abnormality detection in a cohort of 126 contrast-enhanced CTA volumes focused on the renal arteries. Three complementary approaches are evaluated: (i) A baseline 3D Convolutional Neural Network (CNN) that was trained from scratch for binary abnormality classification, (ii) a 3D CNN that takes presegmented vessel mask as input, (iii) a feature-extraction framework that used VesselFM, a foundation model that was pretrained on large-scale heterogeneous imaging data.

The experimental results demonstrate the reliability of CNN-based architectures in detecting arterial abnormalities, even in scenarios with limited data. The utilisation of presegmenting vessels has been demonstrated to enhance performance by orienting the model's focus towards clinically pertinent anatomy, thereby mitigating the risk of shortcut learning. Using VesselFm as a feature extractor yields comparable outcomes to the baseline model, yet it demonstrates a lower performance in comparison to a model that is trained from the beginning with presegmented vessels.

The results of the thesis indicate a potentially beneficial application of Computer-Aided Diagnosis (CAD) systems in the domain of vascular imaging.

Contents

1 Introduction **1**
1.1 Research Questions . . . 2

2 Clinical Background **5**
2.1 Fibromuscular Dysplasia (FMD) . . . 5
2.1.1 Definition and Epidemiology of FMD . . . 5
2.1.2 Symptoms of FMD . . . 5
2.1.3 Diagnosis of FMD . . . 5
2.1.4 Arterial abnormalities associated with FMD . . . 6

3 Related work **9**
3.1 CT imaging . . . 9
3.1.1 The history of Computed Tomography (CT) . . . 9
3.1.2 CTA imaging . . . 10
3.2 Deep Learning for Diagnosis with CT Imaging . . . 10
3.3 Deep learning for CAD of Arterial Diseases . . . 11
3.4 Shortcut learning and Inductive biases . . . 13

4 Technical Background **15**
4.1 Network architectures - Foundational concepts . . . 15
4.1.1 Convolutional Neural Networks, its layers and components . . . 15
4.1.2 U-Net . . . 19
4.1.3 NNU-Net . . . 20
4.1.4 Mednet and its 3D CNN architecture . . . 21
4.2 Foundation models . . . 22
4.2.1 The power of pretrained models . . . 22
4.2.2 VesselFM . . . 23
4.2.3 TotalSegmentator . . . 25
4.3 Evaluation metrics . . . 25
4.3.1 True Positive Rate and False Positive Rate . . . 25
4.3.2 Threshold-dependent metrics . . . 25
4.3.3 Threshold-independent metrics . . . 26
4.3.4 Wilcoxon Signed-Rank Test . . . 27

5 Data **29**
5.1 Data origin . . . 29
5.2 Data Characteristics before preprocessing . . . 29
5.3 Preprocessing . . . 31
5.3.1 Scan Filtering . . . 31
5.3.2 Data preperation . . . 32
5.3.3 Image processing . . . 34
5.4 Data Characteristics after preprocessing . . . 34

6 Approach **37**
6.1 Baseline: 3D CNN for Abnormality Detection . . . 37
6.2 Classification with Segmentation-Based Inductive Bias . . . 38
6.3 Leveraging Knowledge from a Pretrained Vascular Foundation Model . . . 38
6.4 Hypotheses . . . 39

7 Experiments **41**
7.1 Experimental Setup . . . 41
7.1.1 Dataset and Splits . . . 41
7.1.2 Hardware and HPC environment . . . 41
7.1.3 Training Parameters . . . 42
7.1.4 Evaluation . . . 42
7.2 Arterial abnormality detection - RQ1 . . . 43
7.3 Presegmentation of vessels - RQ2 . . . 44
7.4 Foundation model as feature extractor - RQ3 . . . 44
7.5 Models . . . 45
7.5.1 Mednet 3D CNN . . . 45
7.5.2 VesselFM as feature extractor . . . 45

8 Results **47**
8.1 RQ1 - Benchmarking Arterial Abnormality Detection . . . 47
8.2 RQ2 - Arterial Abnormality Detection with presegmented Vessels . . . 48
8.3 RQ3 - Arterial Abnormality Detection with foundation model backbone . . . 49
8.3.1 RQ3.1 - Finding the best model configuration . . . 49
8.3.2 RQ3.2 - Comparison to Benchmark . . . 50
8.3.3 RQ3.3 - Comparison to Presegmentation model . . . 51

9 Discussion **53**
9.1 Feasibility of CNN-based abnormality detection . . . 53
9.2 Segmentation as a means to enforce inductive bias . . . 54
9.3 Foundation Model as Feature Extractor . . . 55
9.3.1 The search for the best configuration . . . 55
9.3.2 Comparison to training convolutional models from scratch . . . 56
9.4 Dataset-related limitations . . . 57
9.5 Future research directions . . . 58

10 Conclusion **59**

Chapter 1

Introduction

Fibromuscular Dysplasia (FMD) is a non-atherosclerotic, non-inflammatory arteriopathy that predominantly affects renal and cervicocephalic vessels, and often presents with subtle or nonspecific clinical findings. The condition is frequently underdiagnosed, as it is rare and poorly understood by many healthcare providers (Olin and Sealove, 2011). Many patients remain asymptomatic, and a substantial fraction of cases are detected incidentally during imaging for unrelated indications (Persu et al., 2021). As a result of these challenges, the number of cases diagnosed is likely only a fraction of the true prevalence of FMD in the general population (Olin and Sealove, 2011). This is shown by Cragg et al. (1989) that found that, in a study group of kidney donors, 6.6% had evidence of FMD. At the same time, FMD is associated with a spectrum of arterial abnormalities, such as multifocal stenoses producing the classic string-of-beads appearance, focal stenosis, aneurysm, dissection, ectasia, tortuosity, and fistulae. All of these carry clinically meaningful risks for patients (Olin and Sealove, 2011; Varennes et al., 2015). These lesions can precipitate severe outcomes, including renovascular hypertension, transient ischemic attacks, stroke due to cervical artery dissection, or haemorrhage from aneurysm rupture (Persu et al., 2021; Gornik et al., 2019; Plouin et al., 2007).

FMD is primarily diagnosed through imaging modalities such as Computed Tomography Angiography (CTA), Magnetic Resonance Angiography (MRA) Olin and Sealove (2011). CTA is often the preferred imaging modality for diagnosing FMD due to its high spatial resolution (Gornik et al., 2019). A rotating X-ray tube captures multiple images from different angles, which are then reconstructed into detailed cross-sectional images of the body (Kumamaru et al., 2010). The resulting images are made up of voxels with a specific greyscale intensity, which is called a Hounsfield Unit. Depending on the amount of radiation going through the scanned structure, the Hounsfield Unit varies (Güçük and Uyetürk, 2014). This greyscale representation makes it possible to identify different tissues and structures within the body, including blood vessels. Today these images are manually reviewed by radiologists to identify arterial abnormalities associated with FMD. The goal of this work is to design a Computer-Aided Diagnosis (CAD) system that can assist radiologists in detecting arterial abnormalities associated with FMD in CTA images.

The present thesis concentrates on deep learning-based convolution-based architectures, with the goal of identifying different approaches to improve the detection of arterial abnormalities associated with FMD. As CTA images are three-dimensional volumes, they contain a substantial amount of abundant information that can be utilised for the detection of arterial abnormalities (Kim et al., 2019). This is of significance, as it seeks to address several challenges encountered in the field of medical imaging research. It is important to note that medical problems are unique in nature, and as a result, datasets are frequently constrained in size. To effectively train deep learning models, a substantial amount of annotated data is required (Razzak et al., 2018). Pretrained foundation models try to address this challenge by leveraging large-scale datasets to learn generalizable features that can be fine-tuned for specific tasks with limited data (Yuan, 2023).

In this work we try to reduce the complexity of the task by cropping the input images to a defined region of interest around the kidneys, so that we can focus on the prevelance in the renal arteries. In order to further refine the region of interest, an analysis of the potential advantages of segmenting the vascular structures within that same region is conducted.

A further challenge is the heterogeneity that is present within the datasets (Wang, 2017). Consequently the identification of a dataset that meets the criteria of the research and is large enough to train deep learning models can be challenging. After filtering the available dataset for this thesis the dataset contains only 126 CTA images. We try to address this by reducing noise and inducing biases, by presegmenting the vessels and in another approach try to use a pretrained model as a feature extractor.

In summary, this thesis evaluates the performance of a baseline convolution-based model for arterial abnormality detection, investigates how presegmenting vascular structures influences detection accuracy, and examines the use of a foundation model trained for vessel segmentation as a feature extractor to further improve detection of arterial abnormalities.

1.1 Research Questions

Within this thesis, we aim to answer the following research questions (RQs):

RQ1: Are current CNN-based models able to detect arterial abnormalities? The primary objective of RQ1 is to evaluate the performance of a convolutional architecture in detecting arterial abnormalities associated with FMD within a defined region of interest around the renal arteries in CTA images. To investigate this, this study trains a convolutional model with the architecture proposed by Güler et al. (2024), from scratch to classify the presence or absence of arterial abnormalities within the region of interest.

RQ2: Does presegmentation of vessels improve arterial abnormality detection? RQ2 aims to investigate whether presegmenting vessels in CTA images enhances the detection of arterial abnormalities associated with FMD. To address this question, the model used in RQ1 serves as the baseline for comparison. While the model architecture remains the same, the input data is modified. To presegment the vessels, the foundation model proposed and trained by Wittmann et al. (2025) is used to segment the vascular structures within the region of interest around the renal arteries.

RQ3: Can a foundation model trained for vessel segmentation, when used as a feature extractor, beat the performance of 3D CNN's trained from scratch in the task of arterial abnormality detection? The aim of RQ3 is to explore whether using a foundation model trained for vessel segmentation as a feature extractor can further improve the detection of arterial abnormalities associated with FMD. This pipeline uses the same input images as RQ1, but rather than undergoing full training, the objective is to utilise the capabilities of a pretrained foundation model. The VesselFM model by Wittmann et al. (2025), which was employed in RQ2 for vessel segmentation, serves as the feature extractor within this research question. The plan is to use the encoder component of the VesselFM model to extract relevant vessel features from the CTA images. These features are then fed into a classifier head to detect arterial abnormalities.

Three sub-research questions, that are investigated to provide the most conclusive response to this research question.

- RQ 3.1: Which architecture of the adapted VesselFM model performs best? This includes the number of encoding layers, the type of head (linear or nonlinaer) and the pooling strategies.

- RQ 3.2: Is the performance of arterial abnormality detection better when a foundation model trained for vessel segmentation is used as a feature extractor compared to when a simple 3D convolutional network that is trained from scratch with the same input volumes?
- RQ 3.3: Is the performance of arterial abnormality detection better when a foundation model trained for vessel segmentation is used as a feature extractor compared to when a simple 3D convolutional network that is trained from scratch with presegmented vessels?

The subsequent sub-research questions are designed to provide clarification regarding the foundation model's potential and ability in the context of arterial abnormality detection. To achieve the best possible outcomes, a robust architecture must be provided for the adapted VesselFM model in response to RQ 3.1. The adapted VesselFM model is then compared with the baseline model from RQ 1 to answer RQ 3.2, and with the model from RQ 2 to answer RQ 3.3.

Chapter 2

Clinical Background

This chapter introduces the clinical background to the thesis. First, Fibromuscular Dysplasia (FMD) is explained, and then the arterial abnormalities associated with FMD that are included in the dataset are presented.

2.1 Fibromuscular Dysplasia (FMD)

2.1.1 Definition and Epidemiology of FMD

As stated by Leadbetter and Burkland (1938), FMD was first described in the context of the connection between hypertension and renal disease (Olin and Sealove, 2011). FMD is defined as a nonatherosclerotic and non-inflammatory artery disease. The arteries primarily affected are the renal arteries and the carotid arteries (Plouin et al., 2007). The illness manifests predominantly in middle-aged women. The underlying cause of the disease remains uncertain. However, there have been several potential factors postulated. Genetic factors may have a role to play, given the higher prevalence of FMD observed in specific family lines. A further potential etiology is hormonal, as evidenced by the higher incidence in women (Begelman and Olin, 2000). It has been hypothesised that smoking could be a contributing factor, but this claim is not universally accepted, as evidenced by the existence of studies which contradict this hypothesis (Persu et al., 2021).

2.1.2 Symptoms of FMD

Gornik et al. (2019) presented a list of symptoms associated with FMD, which are outlined below: The manifestation of symptoms is contingent upon the affected artery. The clinical signs and symptoms associated with FMD encompass a wide spectrum of manifestations, including hypertension of varying degrees and intensity, renal artery stenosis, and abdominal bruit. The presence of FMD in the carotid arteries has been demonstrated to be a contributing factor to the development of various health complications, including migraine headaches, tinnitus, strokes and dizziness.

2.1.3 Diagnosis of FMD

The precise prevalence of FMD remains undetermined. As stated by Persu et al. (2021), the underdiagnosis of FMD is frequently observed, with a significant proportion of cases remaining asymptomatic and being identified incidentally. As Olin and Sealove (2011) pointed out, Cragg

et al. (1989) conducted a study that investigated the presence of FMD in potential kidney donors. The presented study revealed that 6.6% of potential donors exhibited indications of FMD on angiography. It is evident that the prevalence of FMD is significantly higher than the number of confirmed cases, a conclusion that is substantiated by the fact that the disease is often detected incidentally. Advancements in non-invasive imaging techniques, such as CTA and MRA, have resulted in an increase in incidental findings of FMD (Olin and Sealove, 2011). Catheter-based angiography is conventionally employed in addition to other imaging techniques for diagnostic purposes. This approach is necessitated by the requirement to exclude other potential vascular diseases, including atherosclerosis, monogenic and inflammatory arterial diseases (der Niepen et al., 2021).

2.1.4 Arterial abnormalities associated with FMD

It is evident that FMD can result in a variety of arterial abnormalities. Varennes et al. (2015) detailed in their work, there are various presentations in medical imaging caused by FMD. The most prevalent manifestation is characterised by a string-of-beads appearance. It is important to note that less common findings may include vascular loops, fusiform vascular ectasia, arterial dissection, stenosis and aneurysms. As Varennes et al. (2015) also stated, aneurysms can lead to further complications, including ruptures and arterovenous fistulas. As stated by Persu et al. (2021), tortuosity has been identified as a contributing factor to arterial lesions associated with FMD.

Varennes et al. (2015) also demonstrated, subarachnoid haemorrhage and hypoperfusion are possible complications associated with FMD.

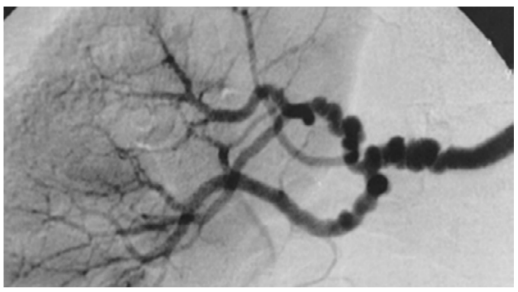
Stenosis

Solan (2024) defined stenosis as a narrowing or constriction of a blood vessel that reduces blood flow. Stenoses can manifest in a variety of forms, each of which exhibits a distinct correlation with FMD. The form typical of FMD, which is associated with the string-of-beads appearance, is multifocal stenosis (Plouin et al., 2007). Focal stenosis in its tubular form is also an important sign for FMD diagnosis (Persu et al., 2021). The symptoms associated with stenoses are contingent on the reduced blood flow and its location. For instance, aortic stenosis results in complications due to the diminished blood flow to the aorta and, consequently, to the body. The resulting symptoms include shortness of breath, fatigue and syncope (Solan, 2024).

String of beads

The most common observed sign for FMD is the string-of-beads appearance (Plouin et al., 2007). The string-of-beads sign is a subject which has been described in a number of different ways. As Plouin et al. (2007) described, the phenomenon is characterised by multiple stenoses in succession. Conversely, Begelman and Olin (2000) offered a different characterisation, describing the phenomenon as beads of a larger diameter than the vessel at the proximal point. In Figure 2.1a the string-of-beads appearance can be seen as an example.

The definition employed in this work is that of Leckie et al. (2021), who offered a characterisation analogous to that of Plouin et al. (2007). Leckie et al. (2021) propose a definition of the string-of-beads appearance as a result of multifocal stenoses and dilatations. String-of-beads are predominantly located in the distal two-thirds renal artery (Begelman and Olin, 2000).



(a) String of Beads (Plouin et al., 2007)

Figure 2.1: ARTERIAL ABNORMALITIES ASSOCIATED WITH FMD. *In (a) the string-of-beads appearance is shown.*

Aneurysm

An arterial aneurysm is defined as a persistent dilatation and widening of a vessel. Aneurysms are generally asymptomatic, but complications can arise from ruptures (Sakalihasan et al., 2005).

Dissection

As defined by Bax et al. (2022), arterial dissections are defined as sudden tears in the arterial wall, resulting in the accumulation of blood within the arterial wall. Dissections are known to carry a significant risk of severe complications. For instance, cervical artery dissections have been demonstrated to result in the occurrence of strokes (Plouin et al., 2007).

Fistula

An arteriovenous fistula is defined as a direct connection between an artery and a vein without an intervening capillary bed (Leckie et al., 2021). In younger individuals, the presence of fistulas often goes unnoticed, as the associated symptoms are typically mild. However, as individuals age, the symptoms tend to intensify. Coronary arterial fistulas have been demonstrated to exert a significant physiological burden on the heart, potentially resulting in complications such as heart attacks (Mangukia, 2012).

Tortuosity

Arterial tortuosity is defined as the presence of abnormal twists and turns in arteries. The condition has been demonstrated to be associated with hypertension and other cardiovascular pathologies (Ciurică et al., 2019).

Ectasia

Ectasia is more frequently described in the context of coronary arteries (Doi et al., 2017; Mavrogeni, 2010). Coronary artery ectasia is defined as a dilatation of an arterial segment to at least 1.5 times that of the adjacent normal coronary artery. The distinguishing factor between an arterial aneurysm and an ectasia is that the latter involves the entire vessel. Impaired blood circulation can result in the onset of angina pectoris or myocardial infarction (Mavrogeni, 2010).

Rupture

As articulated by Leckie et al. (2021) and Plouin et al. (2007), arterial rupture may emerge as a potential consequence of diverse arterial malformations. The clinical impact of the rupture site is a key consideration, and can range from localised bleeding to severe complications. For instance, aneurysms have been observed to rupture, precipitating haemorrhage (Plouin et al., 2007). This, in turn, can result in life-threatening conditions, including, but not limited to, haemorrhagic stroke (Howard E. LeWine, MD, 2023).

Chapter 3

Related work

This chapter provides an overview of recent research relevant to this work, focusing on the application of deep learning in medical imaging. Starting with a brief history and explanation of Computed Tomography (CT) imaging, the chapter then delves into various deep learning approaches for diagnosis using CT imaging and Computer-Aided Diagnosis (CAD) for arterial diseases. A brief discussion on inductive biases and shortcut learning is presented at the end of the chapter. These are important concepts for understanding the limitations and challenges of deep learning models in medical imaging, which this thesis attempts to address.

3.1 CT imaging

3.1.1 The history of Computed Tomography (CT)

The first medical image to be created without the use of invasive procedures was developed by Röntgen in the late 19th century. The discovery of the X-ray was accidental, while working with cathode rays. The first image was then taken of his wife's hand, famously while wearing the wedding ring. Following Röntgen's discovery, the capacity to observe the interior of the human body without the necessity of dissection was realised (Robb, 2006). Fluoroscopy, real-time X-ray imaging, was further developed for the purpose of visualising dynamic processes within the human body. The administration of iodine-based contrast media is frequently employed to enhance the visibility of blood vessels and organs during diagnostic imaging procedures. The utilisation of angiographic imaging facilitates the identification of abnormalities within arteries by a physician. The most significant impediment associated with X-rays is the fact that the resulting images are two-dimensional projections. In the period spanning 1967 and 1968, Godfrey Hounsfield undertook a significant re-evaluation of the field of medical three-dimensional imaging. The concept of a three-dimensional object was defined by the author as a stack of two-dimensional slices. It was on the basis of this theory that Hounsfield and his team proceeded to develop prototypes and the first working machines. Subsequent developments in the field of medical imaging, as evidenced by the advent of CT scans, have led to numerous advancements and contributions, including the introduction of PET scans. This advanced imaging modality has the capacity to reveal both anatomical and metabolic activity within the body, thus providing valuable insights into various physiological processes (Graves, 2024).

3.1.2 CTA imaging

The advent of CT Angiography (CTA) in the early 1990s has led to its widespread utilisation as a non-invasive angiographic modality. The tool is employed in a variety of clinical scenarios, encompassing arterial malfunctions such as aneurysms, dissections, and the other abnormalities delineated in Section 2.1.4. The emergence of CTA can be attributed to significant advancements in technology, particularly those aimed at overcoming the most significant constraints. Longitudinal anatomical coverage is defined as the extent of the body or organ system imaged along the patient's long axis (Chow and Rubin, 2002).

CTA is ordinarily carried out using helical CT acquisition, in which the X-ray tube rotates continuously while the patient table advances, thus enabling volumetric coverage of the vasculature within a single breath hold (Kumamaru et al., 2010; Hoffmann et al., 2006). The acquisition of optimal imaging results necessitates the synchronisation of the CT acquisition with the arrival of the iodinated contrast bolus, thereby ensuring uniform vascular enhancement. Two primary strategies are delineated in the following discourse: the test bolus technique, which involves the administration of a minute contrast injection to ascertain circulation time, and automatic bolus tracking, a process in which recurrent low-dose monitoring scans detect contrast arrival and initiate the full acquisition process (Kumamaru et al., 2010; Hoffmann et al., 2006). It is acknowledged that vascular motion, particularly that originating from the beating heart, has the potential to compromise the quality of the image. Consequently, ECG gating is frequently employed as a technique to mitigate this effect. In the context of retrospective ECG gating, continuous data acquisition occurs throughout the cardiac cycle, with images reconstructed at the desired phase. However, it should be noted that this approach increases radiation exposure. In contrast, prospective gating initiates X-ray exposure exclusively during predefined diastolic intervals, thereby reducing the dose while preserving diagnostic image quality (Kumamaru et al., 2010; Hoffmann et al., 2006).

3.2 Deep Learning for Diagnosis with CT Imaging

Several approaches and network architectures have been proposed to automate image interpretation, reduce radiologists' workload, and improve diagnostic accuracy.

Faghani et al. (2024) analysed deep learning algorithms with the goal of automating gout detection, as eventhough colour coding the crystals automatically manual work is still necessary. Two different architectures were tested, a convolutional neural network (CNN)-based SegResNet and a transformer-based SwinUNETR. The CNN-based SegResNet model outperformed the transformer-based SwinUNETR model in thier work.

Chen et al. (2025) developed a segmentation model for the kidneys and kidney stones from CT scans. In order to train the model, three distinct annotated CT scan datasets were used. The three datasets under consideration were non-contrast computed tomography (NCT), corticomedullary CT (CTC) and enhanced CT of excretory (CTE). The kidney stone itself was detected most effectively with the ResU-Net. In the case of the remaining two image types, the ResU-Net and the 3D U-Net demonstrated analogous results and exhibited superior performance in comparison to the SwinUNETR. Following the segmentation process, the three image types are then combined with a proposed registration method. The registration method employed is of particular interest, as it was designed to align images of differing types, given the variability in the positioning of kidneys across different images and image types. The registration method has been adopted from the ITK (Lowekamp et al., 2013). The registration method is comprised of six primary components: the fixed image, the moving image, the metric, the interpolator, the optimizer, and the transform. The programme utilises the segmented images as an input, and calculates the extent to which the fixed and moving images differ. The registration process was instrumental to the authors, as the

alignment of the kidneys was instrumental for the alignment of the inner structures of the kidney and therefore important to detect kidney diseases.

The focus of the work of Zhang et al. (2024) were kidney diseases, and the recent progress and future prospects of these diseases were summarised. In the context of segmentation, the authors once more made reference to a variety of models, which are, once again, based on convolution and U-Net. It is noteworthy that Zhang et al. (2024) hypothesises that deep learning models will enhance diagnostic capabilities in the future. However, a number of challenges must be addressed. A significant challenge in this regard pertains to the substantial volume of data and its distribution that is necessary for the training of these models. Another challenge in deep learning is the variability in patients, image modalities and image acquisition protocols.

Overall, the reviewed works demonstrated that deep learning can significantly enhance the accuracy and efficiency of CT-based diagnostics. However, while promising, these methods also face challenges related to data availability, generalizability, and variability in imaging protocols. Building upon these findings, the next section focuses on CAD approaches specifically targeting arterial diseases.

3.3 Deep learning for CAD of Arterial Diseases

Computer-aided systems and deep learning techniques for the detection and analysis of arterial diseases aim to improve the automated identification of vascular abnormalities such as stenosis, plaque, and aneurysms using CTA data.

Fu et al. (2023) developed a convolutional neural network (CNN) model for the automated detection and classification of stenoses and plaque in neck and head CTA's. The radiologist was assisted by artificial intelligence in the final analysis of the CTAs, with the objective of ensuring the quality of the training data. The subjects of the study had access to a dataset of considerable size, specifically 3,266. This can be regarded as a substantial dataset within this particular domain. The model utilised in their work was divided into two distinct components. The model incorporates both a two-dimensional ResU-Net and a three-dimensional ResU-Net. The 2D model was utilised for segmentation and stenosis detection, while the 3D model was employed for plaque classification. A centre-line extraction algorithm to enhance the lumen segmentation outcomes. The trained algorithm demonstrated a high degree of agreement with the expert on stenosis detection, plaque classification and diameter stenosis.

Building on these advancements, subsequent research has extended deep learning applications to different aspects of arterial disease characterization. In a related study, Li et al. (2022) also worked on deep learning models for arterial abnormalities. In their work, the researchers concentrated on the quantification of plaque and stenosis in the coronary arteries. In their work, Li et al. (2022) also consider a model with convolutional layers. The model incorporated a ConvLSTM with two branches. Each individual branch of the network was equipped with its own feature extractor and a segmentation head. In one such branch, a ConvLSTM was employed to extract features from the current cross-section and five adjacent sections. In the other branch, a DenseNet block was utilised for the extraction of features from the current vessel cross-section. In this study, long short-term memory (LSTM) layers were utilised as a substitute for 3D convolutions. Due to the recurrent structure of LSTMs, the model is capable of capturing information from adjacent slices while maintaining its focus on the current slice. In this manner, three-dimensional context is captured without the elevated computational expense associated with three-dimensional convolutions. The outputs of both branches were then combined using an attention head. The dataset is of a smaller size than the one utilised by Fu et al. (2023), yet it is nevertheless substantial, with a sample size of 921 patients. The findings were consistent with those of experts in the field, thereby underscoring the efficacy of deep learning models in supporting medical professionals in

the diagnosis and risk assessment of arterial abnormalities.

Spinella et al. (2023) aimed to construct a segmentation model for abdominal aortic aneurysms (AAA) with a comparatively limited dataset of 73 abdominal CTAs and utilised a 2.5D CNN as their architecture. Following the segmentation of the aorta into the lumen and thrombus, the diameter of the aorta was measured in order to detect aneurysms. The methodology for measuring AAA diameters was improved by building upon the research group's previous work, which included the description of the segmentation architecture (Fantazzini et al., 2020) and the measurement of aortic diameter (Brutti et al., 2022). The model under consideration combines three single-view U-Nets, for the axial, sagittal and coronal view, with a final multi-view integration phase. The integration phase combines the three segmentation outputs through the implementation of a simple averaging approach. In comparison with a two-dimensional approach, the two-and-a-half-dimensional convolutional neural network (2.5 CNN) incorporates greater spatial context. However, it requires less training data than a fully three-dimensional CNN, thereby enabling the utilisation of more complex architectures. The aorta is segmented and categorised into regions, with the aim of adapting the aneurysm diagnosis depending on the region. The cropped aorta is subjected to a process of automatic extraction of a centreline from the lumen. It is important to note that, with all of the information available, the diameter can be calculated. Should the diameter be higher than the threshold diameter, then the patient is defined as aneurysmal. The lumen segmentation process was successful, with a DICE score of 0.89 being attained. Similarly, the thrombus segmentation process yielded a notable result, achieving a DICE score of 0.93. Cases with larger diameters ($\geq 30mm$) were all detected, which demonstrates that the method can achieve high sensitivity for clinically relevant AAAs, even with a limited dataset. The diameter measurement was satisfactory, however, the authors posited that its enhancement could be achieved through integration with thrombus volume data. This, however, necessitates the acquisition of a more substantial dataset.

La Barbera et al. (2023) presented a series of methodologies for the segmentation of tubular structures in adults, and subsequently applied these methodologies to a paediatric dataset. The emphasis of the study was on the analysis of abdominal visceral Contrast-Enhanced CT (CECT) images in cases of renal tumours. In their analysis, the vesselness filters were identified as a potential method for this purpose. In ceCT images, vessels are characterised by a specific geometric structure (tubular) and exhibit high levels of brightness, attributable to the use of contrast injection. The aforementioned features are utilised to compute a Hessian matrix. The Hessian matrix is a statistical term denoting the change in image intensity surrounding a voxel. The 3×3 matrix consists of the second derivatives of the image intensity $I(x, y, z)$.

Each term in the Hessian matrix signifies the manner in which the brightness undergoes a transformation in two distinct directions, operating in a concurrent fashion. For a vessel-like structure, La Barbera et al. (2023) asserts that, simultaneously, one eigenvalue is negligible, signifying minimal alterations in intensity along the vessel's axis. Concurrently, the remaining two eigenvalues are substantial and negative, indicating pronounced intensity variations across the vessel's walls. In order to reduce noise, it is possible to employ a Gaussian filter prior to the use of the Hessian matrix. Another method presented in their work is that of rule-based methods. For the purpose of illustration, the authors cite the work of Lesage et al. (2009) as a case study of a rule-based method. Lesage et al. (2009) utilised the observation that vessels are distinguished in ceCTs to segment the vessels. It has been demonstrated that this method is not universally applicable, due to the fact that vessels do not exhibit equivalent levels of contrast during all phases of the vessel analysis. La Barbera et al. (2023) proceeded to analyse deep learning methods for vessel segmentation. The efficacy of these deep learning methods in vessel segmentation is well-documented. It has been posited by a model that a notable model for this work is the nnU-Net (Isensee et al., 2021), as elucidated in Section 4.1.3. It is evident that the majority of the aforementioned approaches utilise a convolutional neural network (CNN)-based architecture, with a

particular emphasis on U-Net-based (see Section 4.1.2) architectures.

Zohranyan et al. (2024) extended the segmentation task to include the detection of anomalies. The proposed end-to-end framework, designated Dr-SAM, is intended to facilitate the segmentation, diameter estimation, and anomaly detection of angiography images. The Segment Anything Model (SAM) is utilised for the purpose of segmentation (Kirillov et al., 2023). This model is a highly effective tool for the execution of image segmentation tasks. The subsequent stage in their pipeline involves the estimation of vessel diameter, which is then followed by the implementation of histogram-driven anomaly detection. As the segmentation of vessels has been comprehensively described, the diameter estimation was the primary focus of this study. As with other documented works, the diameter is taken as the basis for anomaly detection. In addition to aneurysms (i.e. dilatation of vessels), Zohranyan et al. (2024) also attempts to detect stenoses (i.e. narrowing of vessels). In order to enhance the efficacy of detection, clustering is employed to reduce potential noise and focus the detection process at a local level. The diameter is measured and then used as the basis for the decision regarding the anomaly, with the aid of a previously extracted centreline.

In summary, recent research demonstrates substantial progress in the application of deep learning and CAD for vascular analysis. Despite notable improvements in segmentation accuracy and anomaly detection, there remain challenges related to data standardization, model interpretability, and clinical integration. In the context of detecting arterial abnormalities, the predominant approach involves the measurement of vessel diameter as a primary metric for identifying such abnormalities. However, it is important to note that not all abnormalities can be detected through diameter measurements alone.

3.4 Shortcut learning and Inductive biases

There are some important concepts to consider when it comes to the distributions and features that deep learning models learn, namely shortcut learning and inductive biases. Geirhos et al. (2020) identified shortcut learning as one of the underlying issues affecting the performance of many deep learning models. Wu et al. (2023) explained in their work that deep neural networks tend to learn easier features instead of semantic features. Shortcuts are decision rules that perform well on training data but fail to generalise to more challenging scenarios, such as those encountered in the real world (Geirhos et al., 2020). This behaviour must be kept in mind when developing deep learning models, particularly when working with CT scans, where the point of interest may be very small in comparison to the total amount of information included in the volume.

Baxter (2000) emphasised the central role of inductive bias in machine learning, highlighting that learners must select a hypothesis space large enough to encompass good solutions yet small enough to ensure reliable generalisation from limited data. Traditionally, such biases are introduced manually through expert knowledge about the task. Goyal and Bengio (2022) recently argued that, given a finite dataset, many solutions may perform equally well on the training distribution. Therefore, successful generalization requires additional assumptions or preferences to be built into the learning system. These inductive biases, implemented through model architecture or training design, guide the learner to prioritise certain solutions over others, potentially leading to better generalisation beyond the observed data distribution.

Chapter 4

Technical Background

4.1 Network architectures - Foundational concepts

In this section we will introduce important architectural concepts that lay the foundation for this work. We will first introduce Convolutional Neural Networks (CNNs), its layers and components in Section 4.1.1. Thereafter, we will present U-Net in Section 4.1.2 and NNU-Net in Section 4.1.3. In this section the focus is on the underlying components of the final architectures used in this work. These concepts are essential to understand the final architectures used. The final adapted architectures used in this work are presented in the methods Chapter 6.

4.1.1 Convolutional Neural Networks, its layers and components

Deep learning methods have enhanced the complexity and precision of image tasks, including classification, detection, and segmentation (Affonso et al., 2017). CNNs represent a class of Deep Learning (DL) networks of significant importance to the field of computer vision (Li et al., 2022). In the standard configuration, CNNs are typically configured to accept a 3D matrix as input for a 2D image. The three dimension configures as rows, columns and three channels (RGB colour channels). CNNs underwent further development to enable the handling of higher-order tensors, which in turn facilitated the management of three-dimensional images (Wu, 2017). CNNs remain feed-forward networks, however they have benefitted from enhancements that are integral to their success in the field of computer vision. Local connections, in which neurons do not connect to all subsequent neurons, have been shown to facilitate the preservation of localised information in images. In addition to the emphasis on local information, the utilisation of down-sampling and dimensionality reductions serves to incorporate global information into the deeper layers (Li et al., 2022).

Input

The first part of CNNs is the input layer, which stores the voxels of the 3D images (O'Shea and Nash, 2015).

Convolutional layer

The convolutional layer represents the most substantial component of CNNs (Wu, 2017). The following concepts are of particular importance in the convolutional layer: the kernel, stride, and

padding. As this work utilises volumetric (three-dimensional) data, the ensuing exposition of these concepts is provided in the context of three-dimensional images. The subsequent definition is based upon the explanations provided in the work of Wu (2017).

It is proposed that the input to the l-th layer shall be defined as a fourth-order tensor.

$$\mathbf{X}^{(l)} \in \mathbb{R}^{C^{(l)} \times D^{(l)} \times H^{(l)} \times W^{(l)}} \tag{4.1}$$

where $C^{(l)}$ denotes the number of input channels, and $D^{(l)}$, $H^{(l)}$, and $W^{(l)}$ represent the spatial depth, height, and width of the input volume, respectively. A convolutional kernel, if consisting of one filter is likewise a fourth-order tensor. However, a convolutional layer contains $C^{(l+1)}$ such filters, one for each output channel, so the full convolutional kernel is a fifth-order tensor:

$$\mathbf{K} \in \mathbb{R}^{C^{(l+1)} \times C^{(l)} \times k_D \times k_H \times k_W}. \tag{4.2}$$

When the kernel is overlaid on the input tensor at spatial position (d, h, w), the convolution output at that location for one output channel is obtained by computing the sum of elementwise products across all input channels and spatial positions. Strictly speaking, this operation corresponds to cross-correlation, since the kernel is not flipped along any dimension. Although deep learning libraries typically refer to this operation as "convolution," its implementation follows the definition of cross-correlation rather than true mathematical convolution (Chen et al., 2020).

$$Y_{d,h,w} = \sum_{c=1}^{C^{(l)}} \sum_{i=0}^{k_D-1} \sum_{j=0}^{k_H-1} \sum_{k=0}^{k_W-1} X^{(l)}_{c,\, d+i,\, h+j,\, w+k}\, K_{c,\, i,\, j,\, k}. \tag{4.3}$$

This operation is repeated as the kernel moves across the spatial dimensions (D, H, W) of the input volume with a defined stride s, thereby generating a three-dimensional feature map. The stride s specifies how many voxels the kernel shifts at each step along the spatial dimensions.

It is generally accepted that multiple kernels are applied in parallel, thus producing several output channels. These channels then combine to form the output tensor of the layer.

The 3D convolution operation facilitates the extraction of volumetric features, capturing spatial dependencies not only in height and width but also along the depth dimension. It is noteworthy that such operations are particularly well-suited to medical imaging tasks, where the input data frequently consists of volumetric structures such as magnetic resonance or computed tomography scans represented as three-dimensional tensors (Yamashita et al., 2018).

Pooling layer

Typically positioned between two convolutional layers, the pooling layer is designed to achieve shift-invariance. A reduction in the resolution of the feature maps enables a transition from local information in the early layers of the CNN to the extraction of more abstract features in the deeper layers (Gu et al., 2018). Consequently, the pooling layers perform spatial down-sampling of the input feature maps. There are two well-known pooling operations: Max Pooling and Average Pooling (Li et al., 2022). Following, we will again adapt the explanation of Wu (2017) to three-dimensional data.

Let the input to the l-th pooling layer be a fourth-order tensor

$$\mathbf{X}^{(l)} \in \mathbb{R}^{C^{(l)} \times D^{(l)} \times H^{(l)} \times W^{(l)}} \tag{4.4}$$

where $C^{(l)}$ denotes the number of channels, and $D^{(l)}$, $H^{(l)}$, and $W^{(l)}$ are the spatial dimensions. Typically, the pooling layer divides each channel into non-overlapping subregions of size $p_D \times p_H \times p_W$, optionally separated by a stride (s_D, s_H, s_W). The output tensor

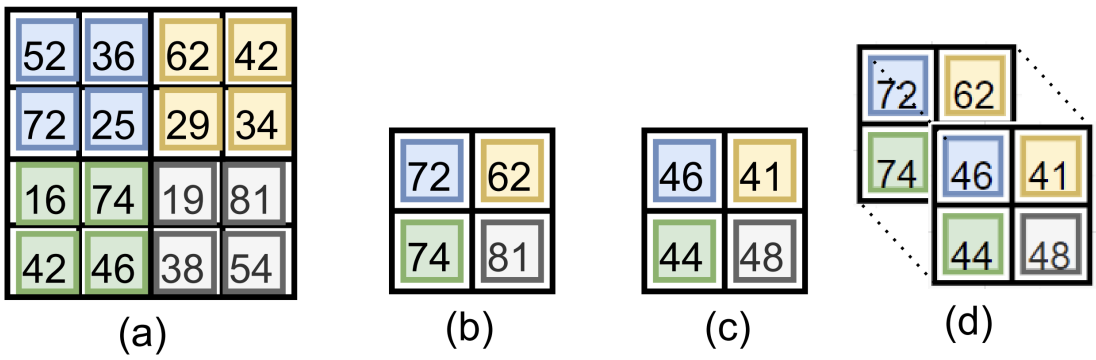


Figure 4.1: COMPARISON OF POOLING METHODS. *Comparison of different pooling methods tested in RQ 3.1 by Doğan (2023). (b) is Max Pooling, (c) is Average Pooling and (d) is combining both Max and Average Pooling. (a) shows the original feature map.*

$$\mathbf{Y}^{(l+1)} \in \mathbb{R}^{C^{(l)} \times D^{(l+1)} \times H^{(l+1)} \times W^{(l+1)}} \tag{4.5}$$

Within each subregion, the pooling operation maps all elements to a single output value according to a specified aggregation function. Defined here max and average pooling the two most common pooling operations:

$$\text{Max pooling: } Y^{(l+1)}_{c,d,h,w} = \max_{0 \le i < p_D,\, 0 \le j < p_H,\, 0 \le k < p_W} X^{(l)}_{c,\, s_D d+i,\, s_H h+j,\, s_W w+k}, \tag{4.6}$$

$$\text{Average pooling: } Y^{(l+1)}_{c,d,h,w} = \frac{1}{p_D p_H p_W} \sum_{i=0}^{p_D-1} \sum_{j=0}^{p_H-1} \sum_{k=0}^{p_W-1} X^{(l)}_{c,\, s_D d+i,\, s_H h+j,\, s_W w+k}. \tag{4.7}$$

Max pooling selects the maximum value within each local subregion of the input, preserving the most dominant activation and thereby emphasizing strong local features. The process of average pooling involves the calculation of the mean value within each subregion, resulting in a representation of the input features that is both smoother and more generalised. In this manner, each pooling operation summarises a local spatial neighbourhood into a single representative value. As Max pooling selects the most prominent features, information about less dominant features is lost. Average pooling, on the other hand, retains more information about the overall feature distribution, but may dilute the impact of strong features. Therefore Doğan (2023) investigated the performance of both pooling methods concatenated (Concat(Avg,Max)). In Figure 4.1 the different pooling methods are illustrated. The results indicated that depending on application scenario one method may outperform the other.

Fully connected layer

It is a common practice to place fully connected layers at the end of CNNs. The neurons of a fully connected layer are interconnected with all neurons of the preceding layer (O'Shea and Nash, 2015).

The adapted definition (Wu, 2017) presents as follows: The output of the layer is computed as

$$\mathbf{y}^{(l+1)} = \mathbf{W}^{(l)} \mathbf{x}^{(l)}, \tag{4.8}$$

where $\mathbf{W}^{(l)} \in \mathbb{R}^{N^{(l+1)} \times N^{(l)}}$ denotes the weight matrix. Each output element $y_i^{(l+1)}$ is therefore influenced by all input elements of $\mathbf{x}^{(l)}$.

In this sense, each output neuron aggregates information from all spatial positions and channels of the input feature maps.

Fractionally-Strided Transposed Convolution for Upsampling Layers

Transposed convolution, often referred to as fractionally-strided convolution, is an upsampling operation that increases the spatial resolution of a feature map by performing the mathematical transpose of a standard convolution. Zeiler et al. (2010) explained this operation in the context of "deconvolutional networks", where it is used to reconstruct higher-resolution representations from lower-dimensional feature maps. In contrast to standard convolution, which aggregates information from a local neighborhood to produce a single output value, the transposed convolution distributes each input value across a larger output region according to learned kernel weights (Zeiler et al., 2010). This viewpoint allows the operation to be interpreted as reversing the forward convolutional mapping, making it well-suited for decoder architectures, generative models, and semantic segmentation.

Dropout Layer

As large models tend to overfit with a small amount of input data Hinton et al. (2012) proposed dropout. Dropout works by randomly removing neurons during training so that co-adaptation between units is prevented and overfitting can be mitigated.

Normalisation

As distributions of each layers inputs change during training, as the parameters of the previous layers change, training a Deep Neural Network (DNN) can be difficult. To mitigate these challenges Ioffe and Szegedy (2015) proposed to integrate normalisation into model architectures. Normalisation stabilizes distributions of activations, improves gradient flow and enables faster training (Ioffe and Szegedy, 2015). Within this thesis Batch normalisation (BN) and Layer normalisation(LN) are used.

Batch Normalisation simply normalizes layer inputs using each training mini-batch to fix the means and variances of layer inputs so that training becomes faster. This also enables strong regularization effect, but it depends on batch statistics (Ioffe and Szegedy, 2015).

Layer Normalisation normalizes all of the summed inputs to the neurons in a layer on a single training case so that the normalisation does not introduce any new dependencies between training cases and works the same in training and testing (Ba et al., 2016).

Activation Functions

Activation functions introduce non-lineatity in neural networks (Dubey et al., 2022). Activation functions used in this thesis include Rectified Linear Units (ReLU), Leaky ReLU (LReLU) and Gaussian Error Linear Units (GeLU).

The ReLU activation funciton is mathematically given by Nair and Hinton (2010):

$$h^{(i)} = \max\left(w^{(i)T}x,\, 0\right) = \begin{cases} w^{(i)T}x, & w^{(i)T}x > 0, \\ 0, & \text{else.} \end{cases} \tag{4.9}$$

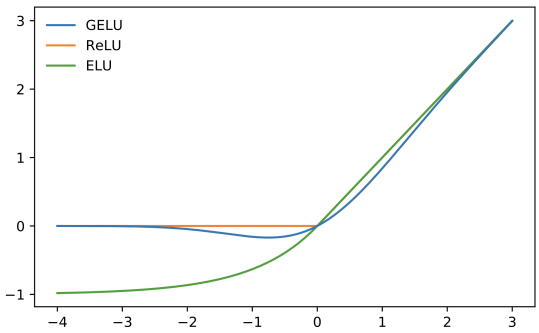


Figure 4.2: ACTIVATION FUNCTIONS. *Visual representation of the activation functions used in this thesis.*

Nair and Hinton (2010) next defined an alteration LReLU like this:

$$h^{(i)} = \max\left(w^{(i)T}x,\, 0\right) = \begin{cases} w^{(i)T}x, & w^{(i)T}x > 0, \\ 0.01\, w^{(i)T}x, & \text{else.} \end{cases} \tag{4.10}$$

In both equations, x denotes the input vector, $w^{(i)}$ is the weight vector of neuron i, $w^{(i)T}x$ is the pre-activation (the weighted sum of inputs), and $h^{(i)}$ is the resulting activation Nair and Hinton (2010). The difference between ReLU and LReLU is that ReLU outputs zero for all negative inputs, while LReLU allows a small non-zero slope for negative inputs so that the unit can still learn even when the pre-activation is negative.

Hendrycks and Gimpel (2023) introduced a further activation function. They defined it like this:

$$\mathrm{GELU}(x) = x\, P(X \le x) = x\, \Phi(x) = x \cdot \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{x}{\sqrt{2}}\right)\right]. \tag{4.11}$$

The GELU activation weights inputs by their value, rather than gates inputs by their sign as in ReLUs ($x\, \mathbf{1}_{x>0}$), so instead of turning negative inputs off like ReLU or Leaky ReLU, it smoothly scales each input by the probability $\Phi(x)$ that a standard normal variable is less than x, giving the deterministic form $\mathrm{GELU}(x) = x\, \Phi(x)$ (Hendrycks and Gimpel, 2023).

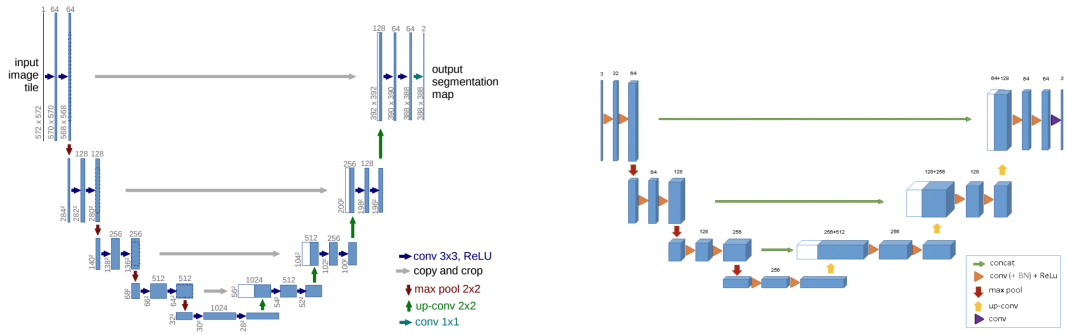
A visual comparison of the three activation functions can be seen in figure Figure 4.2, created by Hendrycks and Gimpel (2023).

Residual Connections

Residual connections, introduced by He et al. (2016) work by taking the original input x and adding it back to the output of a few stacked layers. The stacked layers learn to approximate a function $F(x)$, and the shortcut connection adds the input x to produce the final output $F(x) + x$. This is implemented through shortcut connections that skip one or more layers. Residual connections make very deep networks easier to optimize and allow them to gain accuracy from increased depth, although extremely deep models can still overfit (He et al., 2016).

4.1.2 U-Net

Ronneberger et al. (2015) first introduced this architecture for biomedical image segmentation. As illustrated in Figure Figure 4.3(a), the U-Net architecture is displayed. The architectural design



(a) U-Net architecture (b) 3D U-Net architecture

Figure 4.3: U-NET AND 3D U-NET. *In (a) the U-Net architecture developed by Ronneberger et al. (2015) is shown and in (b) the 3D adaption of Çiçek et al. (2016) U-Net architecture is shown.*

comprises two fundamental components: a downsampling (contracting) element and an upsampling (expansive) element. During the process of downsampling, repeated convolutional layers are employed at each stage, followed by ReLU and max-pooling. In the course of the upsampling process, an up-convolution is employed at each stage. This is then followed by concatenation, two additional convolutions, and ReLU. In the final stage of the process, a 1x1x1 convolution is employed to map each feature vector to the desired number of classes. Between the downsampling and upsampling path there is a bottleneck, which consists of two convolutional layers with ReLU activations. Moreover, skip connections are utilised for the purpose of concatenating feature maps from the downsampling path to the upsampling path (Ronneberger et al., 2015). This architectural design has been demonstrated to yield favourable outcomes with a limited number of annotated images and a reasonable training time (Ronneberger et al., 2015).

Azad et al. (2024) conducted a review of medical image segmentation. To this day, U-Net and its derivatives remain the most widely employed architectures for medical image segmentation (Siddique et al., 2021). The architecture under discussion has achieved a high level of utilisation, and this has resulted in a variety of adaptations of its initial design (Siddique et al., 2021).

The 3D U-Net was first introduced by Çiçek et al. (2016). In Figure Figure 4.3(b), the 3D U-Net architecture is presented. It is evident that the architectural design of the 3D U-Net is analogous to that of the 2D U-Net, with all operations having undergone adaptation to the three dimensions. The primary motivation for this three-dimensional adaptation is that medical images can be volumetric in nature and consequently, a model designed to segment these images should also be capable of handling volumetric data.

4.1.3 NNU-Net

In the biomedical imaging domain imaging modality, image size, voxel spacing and class ratio are different from task to task (Isensee et al., 2021). Creating a model that can adapt to these different tasks was the main motivation for the NNU-Net architecture for Isensee et al. (2021).

Based on the original U-Net architecture by Ronneberger et al. (2015) and on its three dimensional adaption by Çiçek et al. (2016), Isensee et al. (2021) created a framework that can adapt to differences in tasks and can configure itself automatically. They used three parameter groups for this purpose. They used three parameter groups for that purpose. These were fixed parameters that remained constant for all tasks, such as the structure, Dice + Cross Entropy loss, and leaky ReLu. The second group comprised rule-based parameters, which are defined as heuristic rules

to enable instant adaptation to the dataset and other factors.

The second group consists of rule-based parameters. These are defined using heuristic rules so that the model can instantly adapt to dataset characteristics and other influencing factors. For example, batch size and patch size are chosen based on GPU memory limitations or on the image downsampling strategy, which itself depends on voxel spacing.

The third group comprises empirical parameters, which are learned during training. These include elements such as post-processing strategies.

The result is a model that configures itself based on the dataset and task at hand and can eliminate the trial and error configuration of models. This includes the architecture, preprocessing, training and postprocessing and can therefore adapt to different tasks within the biomedical domain (Isensee et al., 2021).

4.1.4 Mednet and its 3D CNN architecture

Mednet is a deep-learning framework developed by Güler et al. (2024) and Laibacher and Anjos (2019). The development of this technology was initiated with the objective of facilitating the creation of medical imaging models in 2D or 3D, utilising a variety of network architectures. The package is openly available.

In this study, the 3D CNN architecture from the mentioned package is employed. A full schematic representation of the architecture is provided in Figure 4.4. The architectural design is delineated as follows:

The network has been designed for the purpose of processing volumetric medical data, and it follows the general architectural principles of convolutional neural networks as introduced in Section 4.1.1. The 3D CNN architecture is composed of four consecutive convolutional blocks, each of which is followed by BN, ReLU activations, and pooling operations (see Section 4.1.1). Normalisation and ReLU, among other activation functions, are advances in DNNs to mitigate the problem of vanishing and exploding gradients (Ebski et al., 2018). The first three convolutional blocks are equal and the fourth block differs slightly in the the absence of a subsequent pooling layer. Both types of convolutional blocks are shown in Figure 4.5. Each block successively increases the number of feature channels, thereby enabling the network to extract increasingly abstract and high-level representations of the volumetric input data. The initial three blocks are succeeded by Max Pooling layers. Each convolutional block incorporates residual connections. Following the fourth convolutional block Average Pooling is used as the pooling method of choice. This process results in the generation of a single feature vector for each sample. The resulting 64-dimensional feature representation is processed through two fully connected layers with an intermediate dropout layer ($p = 0.3$) in order to mitigate the risk of overfitting. The initial fully connected layer serves to reduce the feature dimensionality from 64 to 32, followed by a ReLU activation. The final layer is responsible for mapping the features to the number of output classes and producing linear logits, which are subsequently passed to a sigmoid activation during inference.

The Mednet 3D CNN architecture offers a compact yet expressive 3D architecture that effectively captures both local and global spatial dependencies in volumetric data, rendering it suitable for detecting arterial abnormalities from three-dimensional medical images.

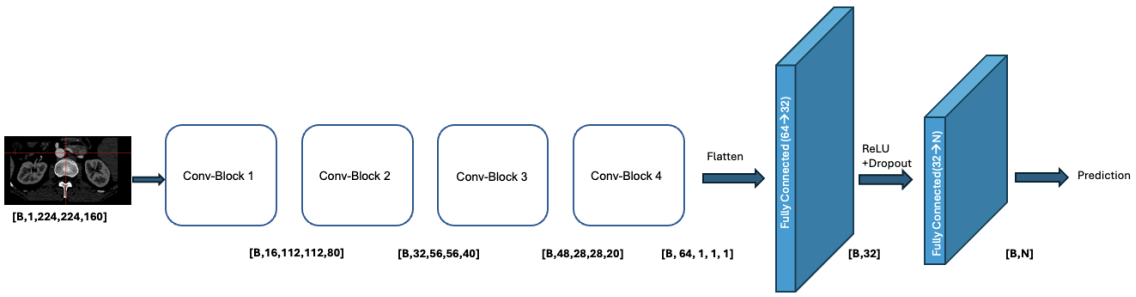


Figure 4.4: MEDNET 3D CNN ARCHITECTURE. *Schematic representation of the 3D CNN architecture employed in this study designed by (Güler et al., 2024). The architecture consists of four convolutional blocks with residual connections, followed by global average pooling and two fully connected layers for classification.*

4.2 Foundation models

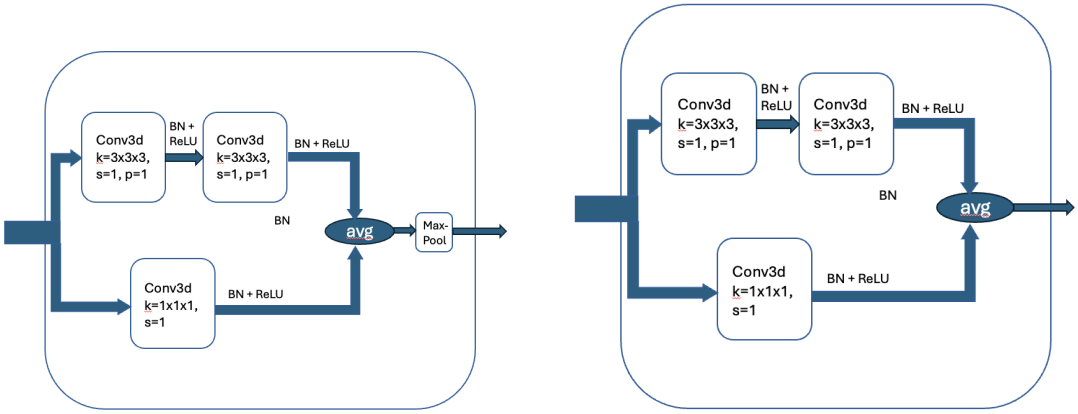
4.2.1 The power of pretrained models

Foundation models are machine learning models that have been trained on large datasets and are capable of performing a range of tasks (Wornow et al., 2023). Foundation models have the potential to improve performance in a variety of tasks within the field of medical imaging. They can facilitate the development of accurate and robust models, reduce the need for large amounts of labelled data and protect the privacy and confidentiality of patient data because knowledge can be transferred without sharing patient information Zhang and Metaxas (2024).

One of todays most popular foundation models is Open AIs GPT series (Wornow et al., 2023). The development of foundation models in the field of medical imaging presents a considerable challenge, given the wide range of domains, tasks and imaging modalities involved. In the medical domain, foundation models are characterised by a high degree of categorisation, as opposed to a more generalised approach, as can be seen with Large Language Models (LLMs) like the GPT series. The categorisation of tasks can be conducted on the basis of region, organ, or task family (Zhang and Metaxas, 2024).

Yuan (2023) provided a theoretical explanation for the versatility of foundation models. In accordance with the principles outlined in their Theorem 2, a foundation model that has undergone pretraining on a sufficiently informative pretext task is capable of capturing the complete structural information of its domain. This capacity enables the model to solve a wide range of downstream tasks through the process of fine-tuning or transfer learning. This theorem is employed in the context of this study as a fundamental principle to refine a foundation model that has been pretrained with a substantial amount of data, subsequently utilised for a downstream task.

How much of the foundation model should be used depends on the task. Yue-Hei Ng et al. (2015) showed that features extracted from earlier convolutional layers of deep networks capture more local patterns of objects. In contrast, deeper layers produce more abstract and semantically global representations, illustrating a fundamental trade-off when choosing which layer to use as a feature extractor.



(a) Mednet Convolutional Blocks 1-3 (Güler et al., 2024)

(b) Mednet Convolutional Block 4 (Güler et al., 2024)

Figure 4.5: MEDNET 3D CNN CONVOLUTIONAL BLOCKS. *Both types of convolutional blocks used in the Mednet 3D CNN architecture are shown. In (a) the first three convolutional blocks are illustrated, each followed by a MaxPool3D layer. In (b) the fourth convolutional block is shown, which does not include a subsequent pooling layer.*

4.2.2 VesselFM

The foundation model used in this work is VesselFM, a model specifically designed for universal three-dimensional (3D) blood vessel segmentation. Wittmann et al. (2025) developed VesselFM with the objective of creating a robust, generalizable foundation model capable of accurately segmenting vascular structures across diverse imaging modalities and anatomical regions. Unlike earlier segmentation foundation models such as SAM-Med3D by Wang et al. (2024) or MedSAM-2 by Ma et al. (2025), which focused on general-purpose or modality-specific segmentation tasks, VesselFM is explicitly tailored to the vascular domain and demonstrates superior performance in zero-shot, one-shot, and few-shot scenarios. The strentgh of this model stems from the dataset accumulation, where important steps were done to increase the size of the training set.

Dataset accumulation, Training and Relevance

VesselFM was trained using a supervised approach involving three heterogeneous data sources, combining real, synthetic and generative datasets to achieve robust cross-domain generalisation. The model was trained using three different data sources. D_{real}, is a large curated dataset containing over 115,000 annotated 3D image patches from 23 different sources. To the knowledge of Wittmann et al. (2025) this dataset is the largest annotated dataset for blood vessel segmentation. The annotations are on a voxel-level. To improve generalisation and zero-shot capabilities, it covers a wide variety of imaging modalities and anatomical regions of the body. D_{drand} is the first of two synthetic datasources used to train VesselFM. It is a domain-randomised dataset that was generated using a domain randomisation strategy that was tailored for this model. The strategy combined randomised vascular foreground geometries with generated background textures and artefacts to produce realistic image-mask pairs. And finally D_{flow} is a dataset generated using a generative model that has been trained on real vessel segmentations. This approach uses a U-Net model proposed by Dorjsembe et al. (2024) called Med-DDPM, which was extended by Wittmann

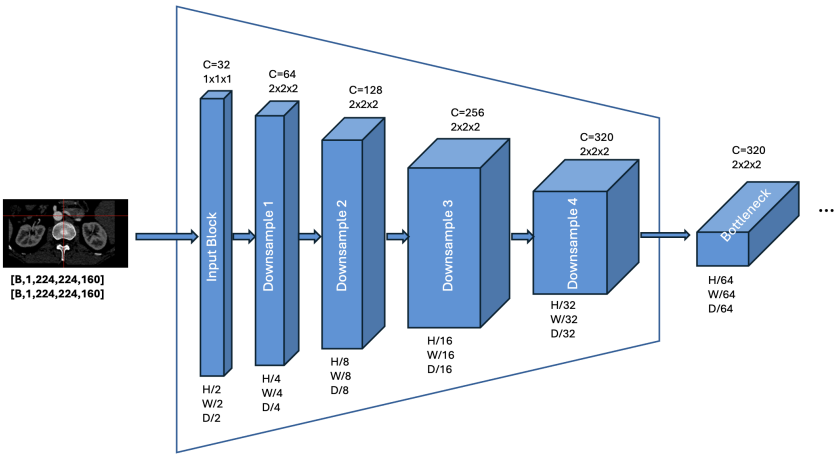


Figure 4.6: VESSELFM ENCODER ARCHITECTURE. *Encoder and Bottleneck of MONAI's NN-Unet implementation (Isensee et al., 2021). The residual connections were left out to simplify the figure.*

et al. (2025) with flow-matching capabilities.

The final training set combines all three data sources, weighted approximately as 70% D_{drand}, 20% D_{real}, and 10% D_{flow}. This composition ensures both large-scale data diversity and anatomical realism. VesselFM was trained on 3D image patches of size 128^3 using a UNet-based segmentation backbone, optimised to preserve the connectivity and topology of vascular structures.

VesselFM introduces several innovations that make it a particularly well-suited foundation model for vascular imaging tasks. The model achieves excellent performance across different benchmarks in vessel segmentation tasks thanks to its data diversity and robustness, outperforming other state-of-the-art models in zero-, one- and few-shot scenarios.

Architecture

VesselFM's architecture is based on the NN-Unet architecture proposed by Isensee et al. (2021), as implemented in MONAI. Further elaboration on NNU-Nets can be found in Section 4.1.3.

VesselFM employs a 3D NN-Unet backbone with one input and one output channel (in_channels=1, out_channels=1), that operates in three spatial dimensions. t is characterised by a symmetric encoder-decoder design with skip connections between corresponding layers. This configuration ensures that fine-grained spatial details from early layers are preserved during upsampling. The encoder pathway is comprised of six resolution stages, with convolutional strides configured as `[[1,1,1], [2,2,2], [2,2,2], [2,2,2], [2,2,2], [2,2,2]]`. This is analogous to progressive spatial downsampling by a factor of two at each level. Figure 4.6 illustrates the encoder and bottleneck of the architecture

As outlined in Section 4.1.1, each stage implements convolutional kernels of size $3 \times 3 \times 3$ with BN and LReLU. It is important to note that the number of feature channels increases with depth, in accordance with the filter configuration `[32, 64, 128, 256, 320, 320]`. After the Input Block 4 Downsampling blocks do follow. Four downsampling blocks follow the input block. These are named Downsample 1, Downsample 2, Downsample 3 and Downsample 4. The decoder is structurally analogous to the encoder, with upsampling layers configured with a kernel size of `[2,2,2]` at each stage. Skip connections between encoder and decoder layers enable the preservation of spatial information. The final output layer maps the highest-resolution feature maps to a single-channel probability mask representing the vessel foreground

4.2.3 TotalSegmentator

TotalSegmentator

TotalSegmentator is a publicly available deep learning model for comprehensive multi-organ segmentation in computed tomography (CT) images. It was introduced by Wasserthal et al. (2023) with the goal of providing a robust and ready-to-use tool that can segment most clinically relevant anatomic structures without additional training. The model segments a total of 104 structures, including organs, bones, muscles, and major vessels, and was trained on a heterogeneous dataset of 1204 clinical CT examinations acquired from routine imaging. This diversity enables the model to generalize well across different scanners, contrast phases, anatomical coverage, and pathological variations.

Technically, TotalSegmentator is based on the nnU-Net framework, which described in Section 4.1.3 Wasserthal et al. (2023) demonstrated reliable segmentation performance even in the presence of abnormalities such as organ deformations, fractures, or surgical modifications.

4.3 Evaluation metrics

4.3.1 True Positive Rate and False Positive Rate

To calculate the metrics used in this thesis we need to first define the True Positive Rate (TPR) and the False Positive Rate (FPR). We will use the definition from Provost and Fawcett (1997). Given a decision threshold τ, the TPR is computed as

$$TPR(\tau) = \frac{\mathrm{TP}(\tau)}{\mathrm{TP}(\tau) + \mathrm{FN}(\tau)}, \tag{4.12}$$

where $\mathrm{TP}(\tau)$ and $\mathrm{FN}(\tau)$ denote the number of true positives and false negatives obtained when predictions greater than τ are classified as positive. Similarly, the FPR is:

$$FPR(\tau) = \frac{\mathrm{FP}(\tau)}{\mathrm{FP}(\tau) + \mathrm{TN}(\tau)}. \tag{4.13}$$

4.3.2 Threshold-dependent metrics

Accuracy

Accuracy is considered to be one of the most frequently employed performance metrics in the domain of binary classification. The term denotes the proportion of instances that have been correctly classified out of all those evaluated. Given a decision threshold τ, Accuracy is computed as (Nellore, 2015):

$$\mathrm{Accuracy}(\tau) = \frac{TP(\tau) + TN(\tau)}{TP(\tau) + FP(\tau) + FN(\tau) + TN(\tau)} \tag{4.14}$$

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively. Accuracy provides an overall estimate of a classifier's effectiveness but does not account for class imbalance, as it treats both classes equally regardless of their distribution.

Recall

Recall measures how effectively a system finds all relevant items in the dataset (Buckland and Gey, 1994) and is exactly equal the TPR. Recall is computed as:

$$\text{Recall}(\tau) = TPR(\tau) \tag{4.15}$$

Recall measures a system's ability to include relevant information in its results, representing retrieval completeness. While high recall is desirable, achieving it often lowers Precision, since retrieving more items increases the likelihood of including nonrelevant ones (Buckland and Gey, 1994).

Precision

Precision measures how many of the items returned by the system are actually relevant (Buckland and Gey, 1994). Precision is computed as:

$$\text{Precision}(\tau) = \frac{TP(\tau)}{TP(\tau) + FP(\tau)} \tag{4.16}$$

High precision implies that most retrieved instances are relevant, meaning the system effectively excludes nonrelevant items from its output.

F1-score

The F1 score represents the harmonic mean of precision and recall, combining both aspects of a classifier's performance into a single metric (Chicco and Jurman, 2020). It is defined as:

$$\text{F1}(\tau) = \frac{2 \cdot TP(\tau)}{2 \cdot TP(\tau) + FP(\tau) + FN(\tau)} = \frac{2 \cdot \text{Precision}(\tau) \cdot \text{Recall}(\tau)}{\text{Precision}(\tau) + \text{Recall}(\tau)} \tag{4.17}$$

The F1 score ranges from 0 (worst) to 1 (best) and reaches its maximum when both precision and recall are 1. As noted by Chicco and Jurman (2020), F1 is independent of the number of true negatives and therefore may provide overly optimistic estimates on imbalanced datasets.

4.3.3 Threshold-independent metrics

The evaluation of binary classification tasks can be undertaken using a variety of metrics. The selection of the metric(s) to be utilised is a pivotal factor in determining the outcome of the model. This phenomenon is particularly evident in the context of threshold-dependent metrics, where the type of metric employed, as well as the specific threshold utilised, play a pivotal role in determining the results. In the binary classification task in this thesis, the model outputs a value between 0 and 1, which can be interpreted as the probability of the input belonging to the positive class. The quality of the numerical results is contingent upon the decision threshold and therefore superior or inferior results can be obtained (Yuan et al., 2015). It is evident that the employment of both threshold-independent and threshold-dependent metrics is justified in this work, with the emphasis placed on the two threshold-independent metrics outlined in this section.

Area under the Curve (AUC)

In this work, AUC refers specifically to the Area Under the Receiver Operating Characteristic Curve (AUROC). The Receiver Operating Characteristic (ROC) curve is a graphical representation

of the relationship between the TPR and the FPR as the decision threshold is varied. This provides a comprehensive assessment of the classifier's performance. The AUC offers a single, threshold-independent summary of performance (Bradley, 1997).

The AUC is commonly estimated by trapezoidal integration,

$$\mathrm{AUC} = \sum_{i=1}^{N-1} (FPR(i+1) - FPR(i)) \frac{TPR(i+1) + TPR(i)}{2}, \tag{4.18}$$

which approximates the integral of the ROC curve without assumptions about the underlying score distributions. The index i enumerates the N sampled ROC points sorted by increasing FPR, so that $i = 1, \ldots, N-1$. Despite its slightly conservative nature, this method demonstrates robust performance when multiple thresholds are sampled. The trapezoidal rule is a numerical method that approximates the area under the ROC curve by summing the areas of trapezoids formed between successive points. This provides a simple and non-parametric estimate of the integral when the curve is defined by discrete thresholds (Yeh, 2002).

As Bradley (1997) demonstrated, the AUC is a metric that quantifies the probability that a positive instance will be ranked higher than a negative one when both are chosen at random,

$$\mathrm{AUC} = P(x_p > x_n) \tag{4.19}$$

Consequently, AUC provides a probabilistic interpretation of a classifier's discriminative ability.

It is evident that these properties contribute to the robustness of the metric, thereby facilitating its widespread adoption in the evaluation of binary classification models (Bradley, 1997).

Average Precision (AP)

Average Precision (AP) summarizes the area under the Precision–Recall (PR) curve by integrating precision over recall. Following the formal definition of Zhu (2004), the average precision of a scoring model is given by the integral of the precision–recall function,

$$AP = \int_0^1 p(r)\, dr, \tag{4.20}$$

where $p(r)$ denotes precision as a function of recall r. Since the PR curve is evaluated at a finite number of operating points in practice, this integral is approximated by a discrete sum (Zhu, 2004):

$$AP = \sum_{n=1}^{N} Precision(n)\, \Delta Recall(n), \tag{4.21}$$

where Precision(n) is the precision at the n-th threshold, and $\Delta Recall(n) = Recall(n) - Recall(n-1)$ denotes the increase in recall. This formulation corresponds to computing a weighted average of precisions, where the weights reflect changes in recall.

As discussed by Davis and Goadrich (2006), AP is particularly informative in imbalanced classification settings, because precision and recall depend only on the distribution of positive examples and are unaffected by the large number of true negatives. Consequently, AP measures how well a model ranks positive instances ahead of negative ones across all possible decision thresholds.

4.3.4 Wilcoxon Signed-Rank Test

The Wilcoxon signed-rank test Wilcoxon (1945) is a non-parametric paired test used to determine whether two paired measurements differ in a systematic way.

Given paired observations (X_i, Y_i) for $i = 1, \ldots, n$, define the differences

$$d_i = X_i - Y_i. \tag{4.22}$$

Differences equal to zero are removed. The absolute differences $|d_i|$ are ranked from smallest to largest, and R_i denotes the rank of $|d_i|$. Let $\operatorname{sgn}(d_i)$ be the sign of the difference. The Wilcoxon signed-rank statistic is then

$$W^+ = \sum_{i:d_i>0} R_i, \qquad W^- = \sum_{i:d_i<0} R_i, \tag{4.23}$$

where W^+ and W^- are the sums of ranks corresponding to positive and negative differences, respectively. A common test statistic is

$$W = \min(W^+, W^-). \tag{4.24}$$

Under the null hypothesis that the distribution of paired differences is symmetric around zero, the sampling distribution of W is known exactly for small sample sizes, and well approximated by a normal distribution for larger n. The resulting p-value quantifies the statistical significance of the performance difference between the two models.

Because it is based on ranks rather than distributional assumptions, the Wilcoxon signed-rank test is robust to outliers and non-normality.

Chapter 5

Data

In this chapter the dataset used in this work is introduced. First the origin and the characteristics of the original dataset are described. Then the preprocessing steps are summarised. After the preprocessing, we will take a closer look into the input data of the models.

5.1 Data origin

The dataset used in this work is provided to IDIAP and obtained from the Service d'Angiologie - Département cœur-vaisseaux, at the Centre Hospitalier Universitaire Vaudois (CHUV). It consists of a lot of different types of CTA scans of patients and corresponding annotations. The dataset is filtered and preprocessed down to a subset, which is then used for model training. These steps are described in more detail in Section 5.3.

5.2 Data Characteristics before preprocessing

As Fibromuscular Dysplasia (FMD) is more prevalent in middle-aged women, the age and gender distribution of the dataset is analysed (Persu et al., 2021). The dataset consists of two groups: an FMD group comprising confirmed cases of FMD, and a control group comprising individuals without FMD. The FMD group consists of 42 participants, and the control group consists of 72 participants. Table 5.1 shows the gender distribution in the FMD dataset and in the control group, which matches the expected distribution with a majority of female patients. The gender distribution in the control group differs by only 3%, making it comparable.

Table 5.1: **Gender distribution in FMD and control groups.** Percentages of gender within both groups are shown

Gender	FMD (%)	Control (%)
Male	33.3	36.1
Female	66.7	63.8

As can be seen in Table 5.2, the age distribution of the FMD group also matches the distribution described in Persu et al. (2021). It can be observed that the control group has a more balanced

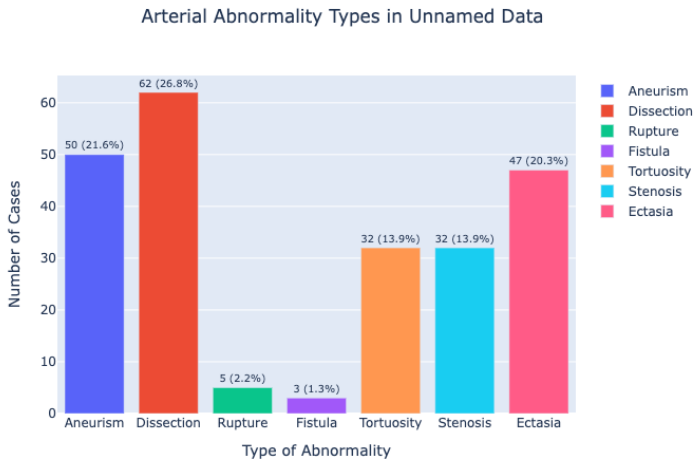


Figure 5.1: ABNORMALITY DISTRIBUTION AFTER PREPROCESSING. *Distribution of the abnormality types within this dataset*

distribution, including a much higher percentage of young participants (13.9% and 19.3%), but a lower percentage (16.7%) in the 50–59 age group than the FMD group.

Both distributions should be kept in mind when analysing the model results later on, as these distributions could lead to shortcut learning. This is especially true for the gender distribution, as the higher prevelance of the female gender may lead to biased decisions by the model.

Table 5.2: **Comparison of Age distributions in FMD and control groups.** Percentages of age group is shown

Age group	FMD (%)	Control (%)
<18	0	0
18-29	2.4	13.9
30-39	9.5	19.4
40-49	31.0	31.9
50-59	38.1	16.7
60-69	9.5	6.9
>70	9.5	11.1

As the goal of this work is to detect arterial abnormalities, both groups are combined in the preprocessing and modelling steps, as both groups contain cases of arterial abnormalities. The preprocessing steps are explained in more detail in Section 5.3.

Figure 5.1 illustrates the distribution of arterial abnormality types present in the dataset. The majority of cases are aneurysmal or dissecting abnormalities, indicating that the dataset is dominated by conditions associated with vessel wall weakening or tear formation. Ectasia and stenosis also occur frequently, suggesting a relevant share of dilation- and narrowing-related pathologies. In contrast, tortuosity, rupture and fistula are comparatively rare, contributing only a small proportion of the overall cases.

Table 5.3: **Comparison of Arterial Abnormality Type Distributions.** Percentages of total abnormalities across the three datasets are shown.

Abnormality Type	FMD (%)	Control (%)	Combined (%)
Aneurysm	21.5	21.8	21.6
Dissection	29.0	25.0	26.8
Rupture	0.9	3.2	2.2
Fistula	0.9	1.6	1.3
Tortuosity	15.0	12.9	13.9
Stenosis	15.9	12.1	13.9
Ectasia	16.8	23.4	20.3

This imbalance highlights that most available samples represent clinically common abnormality types, while rarer lesions are under-represented. As a result, the dataset composition may bias the learning process toward frequently occurring vessel pathologies, potentially limiting the model's ability to recognise less prevalent abnormalities.

Furthermore, an investigation was conducted into whether the cohort origin exerts an influence on the distribution of abnormality types, prior to the implementation of preprocessing on the dataset. This phenomenon is further elaborated in Table 5.3. The percentages of total abnormalities across the FMD group and the control group vary only slightly. It is important to note that there are two key differences that should be considered. The FMD group exhibits a 4.0% prevalence of dissections, while the control group demonstrated a 6.6% prevalence of ectasia. Nevertheless, these disparities are not substantial enough to justify the separation of the two groups in the subsequent analysis.

5.3 Preprocessing

The objective of preprocessing was to generate two types of input images. One type consists of the region of interest (ROI) CTA scans, while the other type comprises ROI vessel segmented scans. To reach that state several preprocessing steps were necessary. To begin with, the dataset was filtered to ensure high quality volumes. The next step was to extract the region of interest from the CTA scans and segment the vessels. After that, a series of image processing steps were applied to the ROI CTA Scans to prepare the volumes as inputs for training.

5.3.1 Scan Filtering

The initial step involves the conversion of CTA scans from DICOM folders to NIFTI files, utilising the dicom2nifti library[1]. It is important to note that not all scans in the dataset are suitable for the detection of arterial abnormalities in the defined region of interest. Test scans, scans without contrast agent and scans that did not include the ROI, namely the kidneys, were filtered out of the dataset. The filtration of the scans ensured the acquisition of sufficient quality data for the region of interest.

[1] https://github.com/icometrix/dicom2nifti

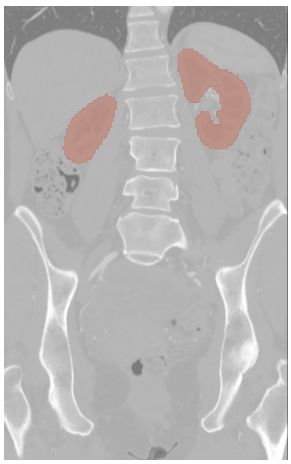


Figure 5.2: KIDNEY SEGMENTATION. *Example of kidney segmentation using the TotalSegmentator model by Wasserthal et al. (2023). The segmented kidneys are highlighted in the CTA scan.*

5.3.2 Data preperation

ROI extraction

The extraction of the region of interest was facilitated by the utilisation of the TotalSegmentator, a model designed by Wasserthal et al. (2023). This foundation model is explained in more detail in Section 4.2.3. The CTA scans were processed with TotalSegmentator in order to extract both kidneys. Such a segmentation is shown in Figure 5.2. A bounding box was created around the segmented kidneys on the full CTA scan, as well as on the segmented vessel image. This was done in order to crop the full image down to the region of interest. The cropping step is of significant importance in order to reduce the amount of irrelevant information in the images and to focus the model on the area where the abnormalities are located. In consideration of the fact that the present work is concerned with arterial abnormalities and focuses on renal arteries, it was deemed essential to incorporate the renal arteries, which extend from the abdominal aorta to the kidneys. Consequently, the bounding box was centred on the body. The extraction of the ROI is a necessary data preparation step to restrict the dataset to the renal artery region. It is not optimised during training and does not represent a modelling choice.

Vessel segmentation

The vessel segmentation process was facilitated by the utilisation of VesselFM. The model is described in great detail in Section 4.2.2. A complete CTA scan is entered into VesselFM, resulting in a binary mask of vessel segmentations (see Figure 5.3). Following the segmentation of the vessels, the images were cropped to the ROI by reusing the coordinates of the ROI cropped CTA scans. The rationale behind this choice is that, despite the fact that VesselFM possesses robust zero-shot capabilities, the input volumes for the model are maintained in a state as unprocessed as possible, thereby preserving the integrity of the original CTA. This approach ensures that the outcomes remain uninfluenced by such preprocessing steps as cropping prior to segmenting.

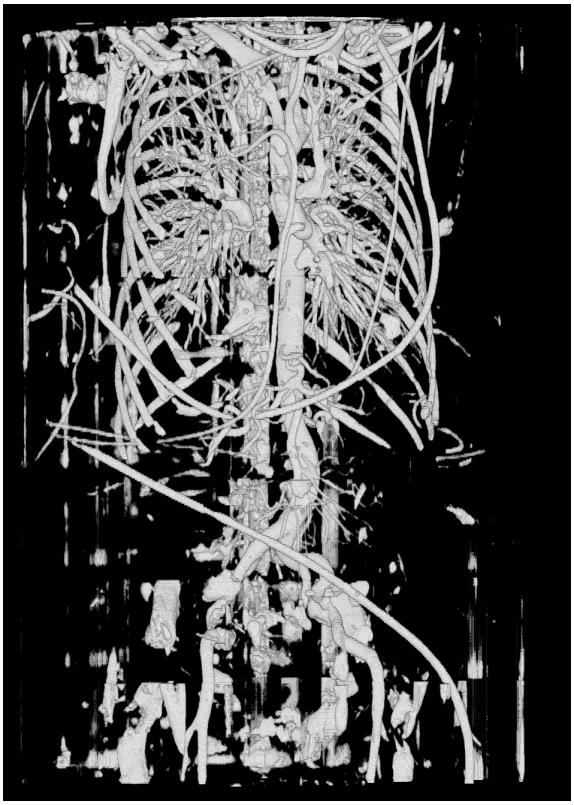


Figure 5.3: VESSEL SEGMENTATION. *Example of vessel segmentation using VesselFM. The output of the model is a binary mask*

5.3.3 Image processing

Up to this point all images persist within the same affine space as the original CTA scan. Consequently, the spatial relationships are transferable between the vessel segmentation images and the CTA scans.

Following the extraction of the ROI and the segmentation of the vessel, the images are resampled to target spacing using the SimpleITK library (Lowekamp et al., 2013). This step is pivotal in standardising the physical resolution of the data, ensuring that each voxel represents an equivalent size in millimetres across all images. This reduction in noisy input for the model is a crucial aspect of the process. The target spacing has been set to (0.9, 0.9, 1.2). The process of interpolation of each image onto the grid is then achieved by the application of SimpleITK's ResampleImageFilter. Linear interpolation is applied to CTA images in order to maintain smooth transitions of Hounsfield values. Conversely, nearest-neighbour interpolation is used for binary label or vessel mask images with the objective of maintaining the mask binary. The original image origin and orientation are preserved, thereby ensuring that all resampled images remain within the same affine coordinate space as the original scans.

Subsequent to resampling to a uniform voxel spacing, each volume is additionally resized to a predefined voxel count using trilinear interpolation. The predefined target size is set to (224, 224, 160). This ensures that all images share the same dimensions required by the neural network input. In the event of a scan being smaller than the target size, zero-padding is applied symmetrically in order to match the desired dimensions, whilst preserving the image content in the centre of the volume.

5.4 Data Characteristics after preprocessing

The dataset is slimmed down after preprocessing. Following preprocessing, the dataset comprises 126 images. This section shows how the abnormalities in the region of interest are distributed. Furthermore, we will examine the images in more detail and their characteristics.

Figure 5.4 illustrates the distribution of arterial abnormality types after restricting the dataset to the renal arteries through preprocessing. Compared with the overall abnormality distribution in the full cohort (Figure 5.1), notable shifts occur. While aneurysms and dissections remain among the most frequent categories, their dominance becomes less pronounced once the analysis is limited to the renal vessels. Instead, stenosis emerges as the most prominent abnormality within the renal arteries.

Furthermore, categories such as rupture and fistula, which are already scarce in the full dataset, disappear completely after preprocessing. It is considered beneficial that ruptures and fistulas, which were exceedingly rare prior to preprocessing, have now been eradicated, thus ensuring that their detection during the training phase of the model will not be hindered. Tortuosity remains present, but constitutes only a minor portion of the renal-specific dataset.

In general, the restriction of the dataset to the renal artery region results in a modification of the distribution of abnormality types. These shifts suggest that the preprocessing step reduces the diversity of abnormality types and increases the imbalance between them. This has implications for the composition of the training data, which may result in the model's exposure to rare or underrepresented categories being limited.

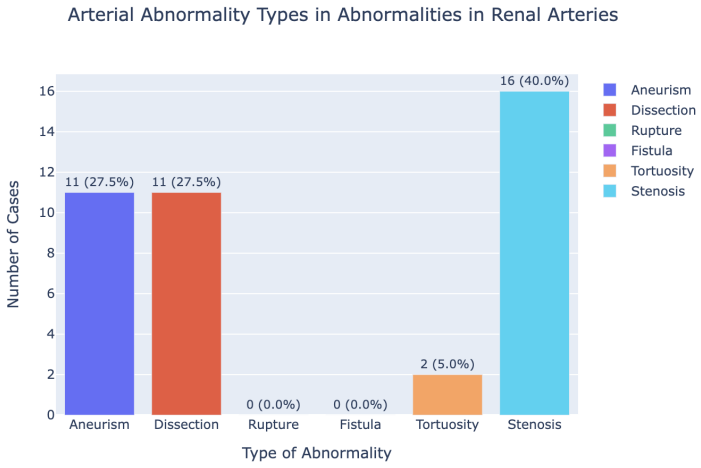


Figure 5.4: ABNORMALITY DISTRIBUTION AFTER PREPROCESSING. *Distribution of arterial abnormalities after preprocessing. All abnormalities that are not in the renal arteries, and therefore not used in this thesis, are removed.*

Chapter 6

Approach

In this chapter, we present the conceptual motivation and methodological strategies underlying the abnormality detection pipelines developed in this work. The decision to create three distinct approaches stems from the idea of reducing noise and focusing on the relevant information within the volumes. Arterial abnormalities associated with Fibromuscular Dysplasia (FMD) are vessel-specific pathologies that manifest as structural alterations of the arterial system, such as stenosis, dissection or aneurysm formation (Varennes et al., 2015). Since these lesions are confined to the vascular lumen and wall, providing models with vessel-focused representations may reduce irrelevant information from surrounding tissues and encourage learning of vascular morphology instead. The input data for all approaches consists of 3D CTA volumes cropped around the renal arteries, as described in Chapter 5. The decision to focus on the renal arteries is motivated by their clinical relevance in FMD and the sufficient number of abnormal cases available for model training. The Region of Interest (ROI) cropping also serves to limit the input size and computational requirements for 3D model training. Instead of inputting the full volumes, which in most cases contain the full upper body, the models can focus on the anatomically relevant sections containing the renal arteries. Convolutional Neural Networks (CNNs) are well suited for processing such data due to their ability to learn spatial hierarchies in three dimensions through convolution and pooling operations (Wu, 2017). In vascular imaging, these properties have been shown to support the extraction of tubular, elongated structures and morphology descriptors (O'Shea and Nash, 2015). However, classification from raw CTA volumes also poses the risk of shortcut learning, where models exploit unintended cues such as scanner-dependent artefacts or background intensity patterns (Zech et al., 2018). These risks and challenges motivate the exploration of vessel-specific inductive biases and pretrained vascular representations in the following approaches.

To start with, a baseline 3D CNN is trained directly on the cropped CTA volumes around the renal arteries. This baseline serves to evaluate whether abnormality detection is generally feasable (RQ1). In a second approach, vessel presegmentation is employed as a preprocessing step to enforce vessel-specific inductive biases (RQ2). Finally, a third approach leverages the pretrained VesselFM foundation model as a feature extractor to reuse vascular representations learned from large-scale data (RQ3). As at the time of model training only binary labels (abnormality present and no abnormality present) the training task at hand for the entire thesis is a binary classification task.

6.1 Baseline: 3D CNN for Abnormality Detection

CTA volumes represent vessels as three-dimensional structures, and abnormalities associated with Fibromuscular Dysplasia (FMD) manifest as local geometric alterations within the vessel lu-

men (Persu et al., 2021). Convolutional neural networks (CNNs) are particularly suitable for such volumetric data because they learn spatial hierarchies through convolution and pooling, enabling the extraction of morphology-relevant patterns across three dimensions (O'Shea and Nash, 2015; Yamashita et al., 2018). In vascular imaging, CNN-based methods have demonstrated strong performance in detecting stenosis and other arterial diseases from CTA data, further supporting their relevance to this task (Fu et al., 2023; Li et al., 2022).

In this work, a compact 3D CNN architecture following the MedNet design is used as a binary classifier (Güler et al., 2024). MedNet emphasises efficient 3D feature extraction through residual convolutional blocks while maintaining low parameter count, which is advantageous when training on relatively small medical datasets such as the 126 CTA scans available for this study. The model operates directly on ROIs cropped around the kidneys, ensuring that the training signal is focused on arterial areas where abnormalities occur. This baseline serves as a reference system to assess whether CTA-based abnormality detection is feasible without enforcing vessel-specific prior knowledge, forming the foundation for evaluating the benefits of segmentation-based inductive bias and pretrained vascular representations in later sections.

6.2 Classification with Segmentation-Based Inductive Bias

Although the baseline model directly processes arterial ROIs, the input still contains non-vascular structures such as surrounding tissues, contrast variations, and scanner-specific signals. These factors introduce variability that is unrelated to vascular pathology and may lead to shortcut learning, where a model relies on spurious correlations instead of pathology-relevant features (Zech et al., 2018). Since abnormalities are confined to the vessel lumen and wall, restricting the input exclusively to segmented vasculature provides a structural inductive bias toward morphology-driven decision making.

To enforce this bias, vessel segmentation masks are generated using the vascular foundation model VesselFM (Wittmann et al., 2025), and the same 3D CNN architecture as in the baseline is trained on masked CTA volumes. The segmentation step removes irrelevant background and reduces input variability, limiting the hypothesis space to tubular geometries. This principle is aligned with established Computer Aided Diagnostics (CAD) pipelines, where segmentation commonly precedes classification or quantitative morphology estimation (Spinella et al., 2023). Furthermore, several vessel analysis strategies explicitly rely on segmented representations to measure lumen diameter, tortuosity, or aneurysm extent (La Barbera et al., 2023; Zohranyan et al., 2024). By integrating segmentation as a preprocessing step, the classifier is encouraged to learn morphological cues that are consistent with how these pathologies are clinically assessed, while still operating on the same ROI-focused classification task as the baseline.

6.3 Leveraging Knowledge from a Pretrained Vascular Foundation Model

While segmentation provides an explicit structural constraint, it does not guarantee that feature extraction is optimal for classification. Foundation models offer an alternative source of prior knowledge by learning general, domain-specific feature representations that can be reused for downstream tasks. VesselFM is trained on diverse real and synthetic vascular data and is designed to produce modality-invariant and topology-preserving vessel features (Wittmann et al.,

2025). Consequently, it should, in theory, be capable of encoding the relevant features of the arterial structures. Secondly, the model was constructed with the objective of attaining optimal zero-shot performance on CT scans (Wittmann et al., 2025). Instead of using it only to produce vessel masks, this research question repurposes its encoder as a feature extractor. A lightweight classification head is placed on top of frozen VesselFM features, enabling abnormality detection without training a full segmentation-classification pipeline from scratch.

Using a foundation model as a feature extractor and adding a classification head can be done in various different ways and therefore also achieve different performances. This is why in this thesis different configurative combinations are tested, with the goal to find the best performing configuration for this task. Starting of different pooling methods, namely Max Pooling, Average Pooling and the combination of both, are compared to each other. The two pooling methods have shown to have different effects on the performance of a model in different scenarios and the combination of both pooling methods has been shown to perform well in such tasks Doğan (2023). The extraction layer of the foundation model features are analysed next, because of the existing locality vs globality trade-off (Yue-Hei Ng et al., 2015). We need to ensure to keep local information encoded, but not discard too much information model from the foundation model. The type of classification head is also analysed. We are comparing a linear head to a nonlinear head, because introducing non-linearity and complexity can improve model performance.

The use of pretrained models is motivated by the limited size and class imbalance of the dataset in this work. Foundation models have been shown to require fewer labelled samples and to generalise more robustly across data acquisitions and medical tasks (Wornow et al., 2023; Zhang and Metaxas, 2024). Since VesselFM learns vascular geometry and lumen continuity directly from large-scale vessel data, its encoder provides a domain-matched starting point for detecting abnormalities without requiring explicit medical annotations of lesion boundaries. For training of the classification head, the same ROI-cropped CTA volumes are used as in the baseline approach. All of this enables a direct comparison between three sources of inductive bias: no prior knowledge (baseline), explicit geometric bias (segmentation), and implicit pretrained vascular representations (foundation model).

6.4 Hypotheses

Based on the design considerations described above, the following hypotheses are investigated in this thesis:

- **H1:** A 3D CNN trained on CTA volumes cropped around the renal arteries can detect arterial abnormalities associated with FMD.
- **H2:** Presegmenting vessels before classification improves abnormality detection performance by enforcing a vessel-specific inductive bias and reducing shortcut learning.
- **H3:** Using VesselFM as a feature extractor outperforms training a 3D CNN from scratch by leveraging pretrained vascular representations learned from large-scale heterogeneous vessel data.

Chapter 7

Experiments

In this chapter, the experiments conducted to address the research questions outlined in Section 1.1 are described in detail. The chapter begins with a description of the experimental setup, including the dataset, hardware environment, training parameters, and evaluation metrics. Subsequently, each experiment is presented in a dedicated section, detailing the objectives, methodologies, and configurations employed to answer the respective research questions. Finally, the architectures of the models utilised in the experiments are described.

7.1 Experimental Setup

7.1.1 Dataset and Splits

The dataset and preprocessing is detailed in Chapter 5. Each remaining case is associated with a binary label indicating the presence or absence of renal abnormalities and with the path to the CTA volume (or mask).

The cleaned dataset is split into a development set (80% of the data, containing train(60%) and validation set(20%)) and an independent test set (20%). This split is performed with stratified sampling so that the proportion of abnormal and normal cases was preserved in both subsets. The test set is fixed, kept seperately and is not used for any fold generation.

On the development set, we pre-compute a 5-fold partition that is reused for all experiments. The folds are created with scikit-learn's StratifiedKFold[1]. Stratification ensures that each fold contains approximately the same class distribution as the full development set. The resulting folds are stored once in JSON format and the same fixed splits are reused across all model training runs. Each set contains the same distribution of abnormalities, namely somewhere between 35% and 37% of images with abnormalities.

7.1.2 Hardware and HPC environment

The provision of the infrastructure is undertaken by IDIAP, with the objective of facilitating the training of the models and the execution of the experiments. Experiments were conducted on a GPU cluster that is managed by SLURM. Each job utilised a single node, equipped with one NVIDIA H100 GPU and 32 GB of system memory.

[1] hiips://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedKFold.html

7.1.3 Training Parameters

Each experiment is trained with a batch-size of 8. For the MedNet configuration, each experiment is trained for up to 75 epochs and trained using the binary cross-entropy loss with logits (BCEWithLogitsLoss). An Adam optimizer by Kingma and Ba (2017) is employed for parameter updates, with optional learning rate scheduling defined within the model configuration. The training is executed utilising the mednet train command-line interface, with multiprocessing for data loading disabled. The classification head of the adapted VesselFM model is trained with a maximum of 50 epochs, utilising the PyTorch Lightning framework with mixed precision (16-mixed) and the AdamW optimiser. For both models gradient accumulation is disabled. The results and checkpoints were automatically saved in a specified output directory. After model training the model with the lowest validation loss during any given epoch is selected for testing.

7.1.4 Evaluation

In this section first all evalutation metrics used to analyse model performance during all experiments are described. Subsequently, the comparative analysis method used to compare the results from different experiments is explained. Since this is a binary classification task, the model can output one of two possible predictions. For each sample, the predicted class (abnormality present or no abnormality present) is compared with the true label. The prediction is then recorded as either correct or incorrect depending on whether it matches the ground truth.

Evaluation Metrics

The objective of this research is to develop a model that can assist physicians in the diagnosis of FMD by detecting arterial abnormalities. Consequently the metrics used to evaluate the models should in combination with each other provide a comprehensive overview of the model's performance in this task. The model will be evaluated on its ability to classify whether an arterial abnormality is present within the input image or not. The evaluation metrics used in this thesis can be split up into threshold dependent and threshold-independent metrics. For the threshold-dependent metrics, a classification threshold is required to convert the model's predicted probabilities into binary class labels. The threshold chosen in this work is 0.5, but can in future work be optimised based on clinical requirements. As threshold dependent metrics, accuracy, precision, recall and F1-score are used. The two threshold independent metrics utilised for the evaluation of the models were Area Under the Curve (AUC) and Average Precision.

After training the models on all five folds, each of the five resulting models is evaluated on the same independent test set. Once all evaluations are completed, the mean and standard deviation of the test metrics across the five models are reported. This approach reduces the influence of unusually good or bad results caused by specific training-set compositions and provides more reliable and meaningful performance estimates. How the evaluation metrics are calculated is further explained in Section 4.3.

Comparative Analysis

Following the calculation of the evaluation metrics for each experiment, a comparative analysis is conducted to determine whether there are any differences in performance between models. Here depending on the experiment there are some differences in the comparative analysis method.

For all comparative experiments, the mean and standard deviation of each evaluation metric across all folds is reported and compared to each other. With this comparison an assesement

of which model performs best on which metric can be made and the research questions can be answered.

In certain cases it is possible that the mean performance of different metrics differ when comparing two models. For example one model might achieve a higher mean AUC, while another model achieves a higher mean Average Precision. For this case different prioritization rules are defined. Within this thesis the threshold independent metrics are prioritised over the threshold dependent metrics. The reason for that is that threshold independent metrics provide a more holistic overview of the model performance across all possible classification thresholds. A second reason is that this work focuses on developing a model with a good overall performance. Setting an optimal classification threshold should be done with experts and while taking clinical requirements into account. If the mentioned usecase arises that one model performs better on AUC, while another model performs better on Average Precision, the model with the higher AUC is prioritised. Even though AP is better suited for imbalance datasets, the imbalance in this dataset is not too severe (approx. 35% positive samples). Additionally the size of the test set is relatively small (approx. 25 samples per fold) and therefore one or two correct/incorrect classifications can have a large impact on the AP.

Although the mean and standard deviation provide an overview of model performance, statistical testing can help determine whether differences in performance are statistically significant. A paired test is conducted over all folds for each metric in a second step, namely a Wilcoxon signed-rank test, so that the results from the two models being compared can be compared. In this study, a one-sided Wilcoxon signed-rank test was used to determine whether one model performed significantly better than the other, with statistical significance being reached when the resulting p-value fell below 0.05. The results of the statistical relevance test are reported in Chapter 9 to evaluate the impact of the different approaches and results. The Wilcoxon test is described in more detail in Section 4.3.4.

7.2 Arterial abnormality detection - RQ1

For this purpose, a 3D CNN is trained, the details of which are described later in this chapter in Section 7.5.1. The input data for this experiment consist of ROI-extracted and preprocessed CTA scans, as described in Section 5.3. The establishment of a baseline is imperative for RQ1, as it will serve as the foundation for the subsequent research questions. The results will be interpreted against a random baseline to determine the model's capacity to learn a distribution that can classify arterial abnormalities better than a randomized model.

The random baseline is defined in table 7.1 and was calculated based on the class distribution of the dataset.

Table 7.1: Random baseline calculated based on the class distribution of the test set

Metric	Random (Mean $\pm$ Std)
AUC	0.500 ± 0.000
Average Precision	0.346 ± 0.000
Accuracy	0.500 ± 0.000
Precision	0.346 ± 0.000
Recall	0.645 ± 0.000
F1-score	0.409 ± 0.000

7.3 Presegmentation of vessels - RQ2

In this experiment, the same 3D CNN architecture employed in RQ1 will be utilised. The input data differs from that used in the baseline model created in RQ1. In the course of this experiment, the vessels were also subjected to segmentation from the CTA scans. The objective of this approach is to minimise input data and noise, thereby enabling the model to concentrate on the pertinent sections of the CTA scans (elucidated in section subsec:vessel-segmentation). The results will then be compared to the results from the baseline model in order to answer RQ2.

7.4 Foundation model as feature extractor - RQ3

In order to achieve to answer RQ 3, it is necessary to provide responses to the three sub-research questions. In the initial sub-research question (RQ 3.1), the optimal architecture for the adapted VesselFM model is determined. In RQ 3.2, a comparison is made between the adapted VesselFM model and the baseline model from RQ 1. In RQ 3.3, a comparison is made between the adapted VesselFM model and the model from RQ 2.

In this experiment, an adaptation of the VesselFM model, as described in Section 7.5.2, will be employed.

Optimal architecture - RQ 3.1

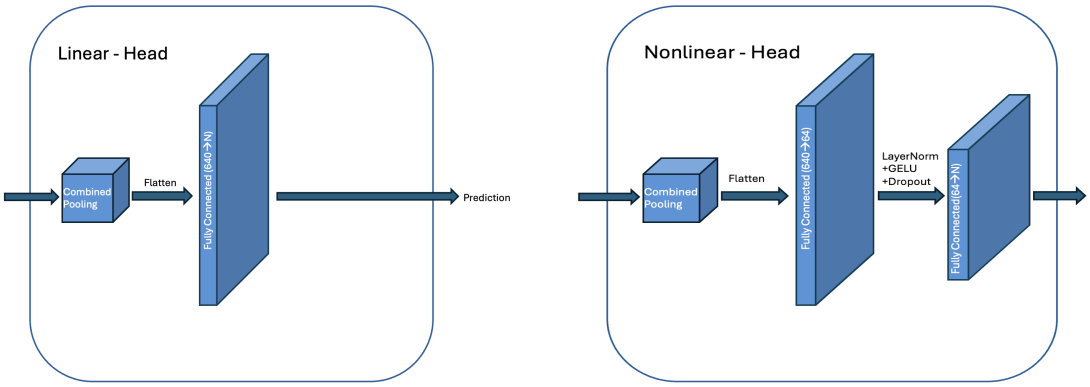
To ensure best possible performance of the adapted VesselFM model, various configurations of the model will be tested in RQ 3.1. Three different configuration categories are tested. First the optimal pooling method for this usecase is determined. Next the extraction layer of the VesselFM encoder is varied, to test various different depths. In this thesis the Bottleneck and Downsample 4 are compared to each other. Finally, two different classification heads are tested, a linear head and a nonlinaer head. In Figure 7.1 the two different classification heads are illustrated. The different configurations are evaluated based on their performance on the test set using the thresholde dependent evaluation metrics described in Section 7.1.4. If different configurations achieve varying ranks on different metrics the AUC is prioritised.

Comparison to baseline model - RQ 3.2

The results of the best possible configuration from RQ 3.1 are evaluated and compared to the results from RQ 1. This comparison is intended to evaluate the hypothesis that utilising a foundation model that has been trained for vessel segmentation as a feature extractor will demonstrate superior performance to that of a 3D CNN that has been trained from scratch with the same input images in terms of arterial abnormality detection. The comparison metrics and method is detailed in Section 7.1.4.

Comparison to 3D CNN trained with presegmented vessels - RQ 3.3

In RQ 3.3, a comparison will be made between the model results and those from RQ 2. Given that the model employed to generate the vessel segmentations utilised in RQ2 is VesselFM, it is worthwhile to investigate whether the employment of VesselFM as a feature extractor can surpass the performance of arterial abnormality detection in comparison to a 3D CNN trained from the beginning with presegmented vessels. The comparison metrics and method is detailed in Section 7.1.4.



(a) Mednet Convolutional Blocks 1-3 (Güler et al., 2024)

(b) Mednet Convolutional Block 4 (Güler et al., 2024)

Figure 7.1: LINEAR AND NONLINEAR CLASSIFICATION HEAD. *In this figure the two different classification heads tested in RQ 3.1 are illustrated. In (a) the linear head is shown, which consists of a global average pooling layer followed by a single fully connected layer. In (b) the nonlinaer head is illustrated, which consists of a global average pooling layer followed by two fully connected layers with a ReLU activation function in between. The final output layer uses a sigmoid activation function to output probabilities for the binary classification task.*

7.5 Models

7.5.1 Mednet 3D CNN

In this study, the 3D CNN architecture proposed by Mednet is employed as one of the two architectures utilised (Güler et al., 2024). The architecture is described in great detail in Section 4.1.4. A key aspect of the MedNet 3D CNN configuration in this thesis is adapting the final layer to output a single neuron, since this is a binary classification task. In this work, no modifications were made to the architecture of the Mednet 3D CNN.

The model has been trained from scratch. This model architecture is employed on two occasions in the present thesis. Initially, the model will be trained to serve as benchmark for the entire thesis and to answer RQ 1. Subsequently, the same architectural framework will be employed, but with presegmented vessel volumes, for RQ 2.

7.5.2 VesselFM as feature extractor

The second architecture employed in this work is an adaptation of VesselFM (Wittmann et al., 2025). The original model is described and shown in Section 4.2.2. In this study, the pretrained weights of the encoder from VesselFM were utilised as a feature extractor and were kept Frozen during the training process. On top of the frozen encoder, a classification head is added to facilitate the binary classification task of arterial abnormality detection. Different configurations of the classification head were tested, as described in Section 7.4. The final architecture of the adapted VesselFM model is determined based on the results from RQ 3.1 and therefore shown in the results chapter in Section 8.3.1.

Chapter 8

Results

8.1 RQ1 - Benchmarking Arterial Abnormality Detection

Table 8.1 presents the results of the benchmark model trained from scratch using the 3D CNN architecture by Güler et al. (2024). Across all five folds, the model demonstrates consistent performance on the validation and test sets with no large deviations between folds, indicating stable generalisation. For threshold-independent metrics, the Area Under the Curve (AUC) on the test sets ranges between 0.824 and 0.856, while the Average Precision (AP) varies between 0.743 and 0.791. These results confirm that the model is capable of learning meaningful class-discriminative features for arterial abnormality detection.

The benchmark model outperforms a random baseline with statistically significant margins across all metrics (e.g. ROC AUC difference of 0.349 ± 0.018, $p = 0.031$). Therefore, the research question can be clearly answered with:

RQ1: Yes, convolution-based CNN models are capable of detecting arterial abnormalities and perform significantly better than random prediction.

Table 8.1: Arterial abnormality detection: Train/Val/Test results for five folds for the baseline model

	Fold 1			Fold 2			Fold 3			Fold 4			Fold 5		
Metric	Train	Val	Test	Train	Val	Test	Train	Val	Test	Train	Val	Test	Train	Val	Test
AUC	0.960	0.979	0.856	0.962	0.910	0.850	0.961	1.000	0.876	0.956	0.903	0.824	0.962	0.931	0.843
Average Precision	0.945	0.972	0.813	0.955	0.934	0.770	0.938	1.000	0.791	0.932	0.878	0.743	0.951	0.932	0.754
Accuracy	0.893	0.960	0.769	0.867	0.920	0.808	0.880	0.920	0.692	0.800	0.720	0.692	0.893	0.880	0.769
Precision	0.857	1.000	0.667	0.781	0.889	0.750	0.828	1.000	0.571	0.659	0.583	0.533	0.885	0.875	0.714
Recall	0.857	0.889	0.667	0.893	0.889	0.667	0.857	0.778	0.444	0.964	0.778	0.667	0.821	0.778	0.556
F1	0.857	0.941	0.667	0.833	0.889	0.706	0.842	0.875	0.500	0.783	0.889	0.667	0.852	0.824	0.625

Table 8.2: Comparison between a random model and benchmark. The results of the training with 5 different folds are averaged.

Metric	Random (Mean ± Std)	Benchmark (Mean ± Std)	Diff (Mean ± Std)
AUC	0.500 ± 0.000	0.849 ± 0.018	0.349 ± 0.018
Average Precision	0.346 ± 0.000	0.774 ± 0.028	0.428 ± 0.028
Accuracy	0.500 ± 0.000	0.746 ± 0.052	0.246 ± 0.052
Precision	0.346 ± 0.000	0.647 ± 0.093	0.301 ± 0.093
Recall	0.645 ± 0.000	0.600 ± 0.100	0.044 ± 0.100
F1-score	0.409 ± 0.000	0.633 ± 0.080	0.224 ± 0.080

8.2 RQ2 - Arterial Abnormality Detection with presegmented Vessels

In RQ2, the same benchmark classification model is applied, but vessel masks generated by the VesselFM foundation model are used as presegmented inputs. Classification results are summarised in Table 8.3. The presegmentation approach yields higher test performance across all five folds compared to the benchmark without segmentation. Specifically, ROC AUC ranges between **0.824 and 0.974**, and AP ranges between **0.776 and 0.958**, demonstrating that vessel-focused image representations provide a clearer and more discriminative signal for classification.

As with RQ1, the results exhibit a slight overfitting effect, reflected by nearly perfect validation scores in multiple folds. Furthermore, in contrast to the benchmark model, the performance variability between folds is higher. This indicates that segmentation amplifies individual fold differences, likely due to variations in segmentation quality, local vessel morphology, or class imbalance across folds.

Therefore, the research question can be answered as follows:

RQ2: Presegmenting vessels from CTA's before training a model does improve model performance.

Table 8.3: Arterial abnormality detection: Train/Val/Test results for five folds for a 3D CNN trained with presegmented vessel masks as input.

	Fold 1			Fold 2			Fold 3			Fold 4			Fold 5		
Metric	Train	Val	Test	Train	Val	Test	Train	Val	Test	Train	Val	Test	Train	Val	Test
AUC	0.955	1.000	0.824	0.997	0.979	0.850	0.995	1.000	0.974	0.967	0.972	0.915	0.991	1.000	0.935
Average Precision	0.862	1.000	0.776	0.995	0.967	0.840	0.993	1.000	0.958	0.960	0.958	0.869	0.985	1.000	0.932
Accuracy	0.893	0.920	0.808	0.960	0.920	0.769	0.947	1.000	0.885	0.920	0.920	0.808	0.933	0.960	0.885
Precision	0.812	0.818	0.700	0.963	0.889	0.800	0.929	1.000	0.875	0.923	1.000	0.833	0.897	1.000	1.000
Recall	0.929	1.000	0.778	0.929	0.889	0.444	0.929	1.000	0.778	0.857	0.778	0.556	0.929	0.889	0.667
F1	0.867	0.900	0.737	0.945	0.889	0.571	0.929	1.000	0.824	0.889	0.875	0.667	0.912	0.941	0.800

Table 8.4: Comparison between benchmark and presegmented mask models. The results of the training with 5 different folds are averaged.

Metric	Benchmark (Mean ± Std)	Masks (Mean ± Std)	Diff (Mean ± Std)
AUC	0.849 ± 0.018	0.900 ± 0.062	0.050 ± 0.062
Average Precision	0.774 ± 0.028	0.875 ± 0.073	0.101 ± 0.088
Accuracy	0.746 ± 0.052	0.831 ± 0.052	0.085 ± 0.088
Precision	0.647 ± 0.093	0.842 ± 0.110	0.195 ± 0.140
Recall	0.600 ± 0.100	0.645 ± 0.145	0.044 ± 0.217
F1-score	0.633 ± 0.080	0.720 ± 0.103	0.087 ± 0.174

8.3 RQ3 - Arterial Abnormality Detection with foundation model backbone

8.3.1 RQ3.1 - Finding the best model configuration

To identify the optimal architecture, the performance of three design variations is compared. The first analysis focuses on the pooling operation used in the classification head. Table 8.5 summarises the results obtained using maximum pooling, average pooling, and a combined pooling strategy, enabling a direct comparison of the effect of each approach on the overall model performance.

Table 8.5: Comparison of pooling methods for classification head

Metric	Avg Pooling	Max Pooling	Combined Pooling
AUC	0.733 ± 0.048	0.681 ± 0.099	0.779 ± 0.049
Average Precision	0.719 ± 0.061	0.681 ± 0.118	0.728 ± 0.055

The combined pooling strategy produces the highest AUC and the best AP. This pooling method is therefore used to evaluate all subsequent configurations. The next design aspect examined is the choice of extraction layer in the foundation model. Comparing the performance of features extracted from the bottleneck layer with those from the preceding downsampling layer (Downsample 4) provides insights into which level of abstraction is most informative for classification. This comparison shows that using features from the downsampling layer improves the AUC but reduces AP (see Table 8.6).

Table 8.6: Comparison of extraction layer

Metric	Bottleneck	Downsample 4
AUC	0.779 ± 0.049	0.851 ± 0.016
Average Precision	0.728 ± 0.055	0.679 ± 0.011

Table 8.7: Comparison of linear and nonlinear head

Metric	linear head	nonlinear head
AUC	0.851 ± 0.016	0.793 ± 0.062
Average Precision	0.679 ± 0.011	0.670 ± 0.068

Since the AUC is prioritised when determining the best-performing model, features extracted from Downsample 4 are selected for the final configuration. The next analysis examines the influence of nonlinear activation on performance (see Table 8.7).

Introducing nonlinearity does not improve the model results. Therefore the final model architecture can be seen in Figure 8.1 at the end of this chapter, combining the VesselFM foundation model as feature extractor with a classification head using combined pooling and a linear layer.

RQ3.1: The best performing model architecture includes a combined pooling method, a linear head and takes the features from the deepest downsampling layer.

8.3.2 RQ3.2 - Comparison to Benchmark

Following the performance of the foundation model as a feature extractor is compared with that of the benchmark model. The results are presented in Table 8.8.

Table 8.8: Comparison between foundation model as feature extractor and benchmark. The results of the training with 5 different folds are averaged.

Metric	Benchmark (Mean $\pm$ Std)	Foundation model (Mean $\pm$ Std)	Diff (Mean $\pm$ Std)
AUC	0.849 ± 0.018	0.851 ± 0.018	0.002 ± 0.018
Average Precision	0.774 ± 0.028	0.679 ± 0.012	-0.095 ± 0.021
Accuracy	0.746 ± 0.052	0.752 ± 0.043	0.006 ± 0.085
Precision	0.647 ± 0.093	0.580 ± 0.152	-0.067 ± 0.202
Recall	0.600 ± 0.100	0.694 ± 0.079	0.094 ± 0.090
F1-score	0.633 ± 0.080	0.533 ± 0.214	-0.100 ± 0.261

The results suggest that using a foundation model trained for vessel segmentation does not improve performance compared to training a model from scratch. In certain metrics, the approach that was evaluated in this study demonstrated a marginally improved performance in comparison to the established benchmark. In all instances where the efficacy of the tested approach is demonstrated, the standard deviation of the results may indicate a potential for outcomes that fall short of the established benchmark. When compared with a model trained on the same input images, most metrics remain similar (Table 8.8). Therefore, we cannot confirm the hypothesis.

RQ3.2: No, using a foundation model as a feature extractor does not improve model performance in abnormality detection tasks compared to a benchmark 3D CNN trained from scratch.

8.3.3 RQ3.3 - Comparison to Presegmentation model

Finally, the performance of the foundation model as a feature extractor is compared with that of the benchmark model and the presegmentation approach. The results are presented in Table 8.9.

Table 8.9: Comparison between foundation model as feature extractor and a model trained with presegmented vessels. The results of the training with 5 different folds are averaged.

Metric	Presegmented Vessels (Mean ± Std)	Masks (Mean ± Std)	Diff (Mean ± Std)
AUC	0.900 ± 0.062	0.851 ± 0.018	-0.049 ± 0.060
Average Precision	0.875 ± 0.073	0.679 ± 0.012	-0.196 ± 0.076
Accuracy	0.831 ± 0.052	0.752 ± 0.043	-0.079 ± 0.070
Precision	0.842 ± 0.110	0.580 ± 0.152	-0.262 ± 0.147
Recall	0.645 ± 0.145	0.694 ± 0.079	-0.049 ± 0.192
F1-score	0.720 ± 0.103	0.533 ± 0.214	-0.186 ± 0.246

The results show that a model trained on presegmented vessels performs better across all the metrics that were measured and is therefore performing worse. Therefore, not only can we not confirm the hypotheses, we also need to say that the opposite might be the case. The results show that the model that achieved the best results was a simple 3D CNN that was trained from scratch but with a presegmentation bias induced.

RQ3.3: No, a foundation model as a feature extractor does not achieve better performance than a model trained from scratch with presegmented vessels.

After answering all three sub research questions we conclude.

RQ3: No, a foundation model as a feature extractor does not beat the performance of 3D CNNs trained from scratch in the task of arterial abnormality detection.

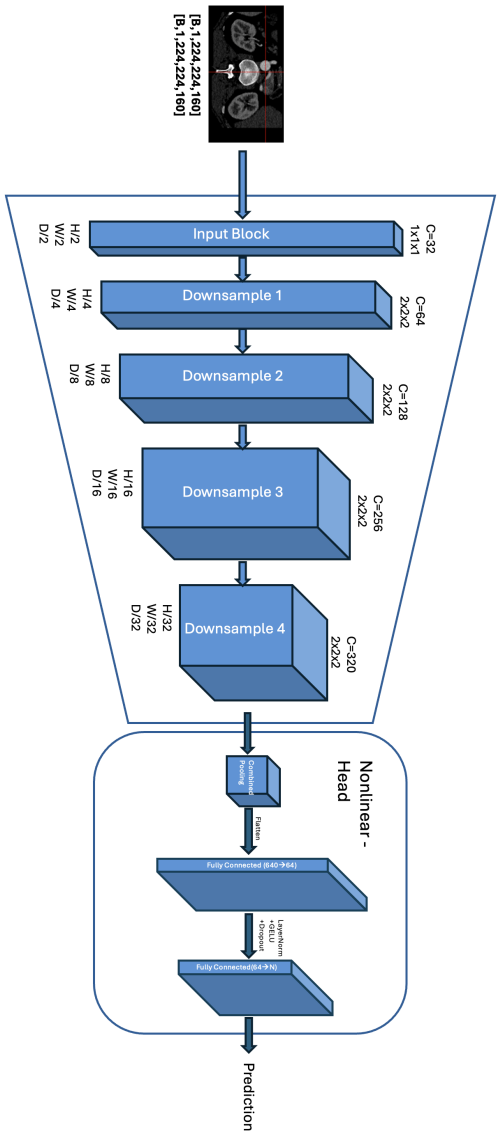


Figure 8.1: FINAL MODEL ARCHITECTURE. *Final model architecture combining the VesselFM foundation model as feature extractor with a classification head using combined pooling and a linear layer. The residual connections were left out to simplify the figure.*

Chapter 9

Discussion

The main objective of this thesis is to detect arterial abnormalities using CTA scans. The impact of presegmentation and the use of foundation models as feature extractors for this use case is also investigated. Detecting small arterial abnormalities within noisy CTA scans is challenging, as the model must learn to distinguish subtle morphological variations between healthy and diseased vessels. First and foremost, this thesis demonstrates that convolutional neural networks can learn a distribution that can detect arterial abnormalities substantially better than by chance. It also demonstrates that presegmenting vessels prior to classification provides a valuable inductive bias that further improves model performance. Finally, the use of a foundation model as a feature extractor is explored; however, the results show that this approach does not outperform training a 3D Convolutional Neural Network (CNN) from scratch.

While the results are promising, limitations in the dataset and approach restrict the interpretability and reliability of the findings. In the following section, we will discuss the results obtained in this thesis, talk about the challenges and limitations, and propose future work.

9.1 Feasibility of CNN-based abnormality detection

In RQ1, a random classifier is compared with a benchmark model that has been trained from scratch. The benchmark model is a simple 3D CNN. The results show that the trained model learns a distribution that can detect arterial abnormalities more effectively than a random model. At the same time, the gap between training and validation/test metrics, particularly visible in validation AP and recall values approaching 1.0 for multiple folds, suggests a degree of slight overfitting to the training data. However, this effect remains moderate, and performance does not collapse on unseen data. In order to make a definitive statistical statement on the ability to detect arterial abnormalities, a Wilcoxon signed-rank test is performed to compare the random model with the trained benchmark model. The Wilcoxon test compares all five folds pairwise. The Wilcoxon method is explained in more detail in Section 4.3.4

Table 9.1 shows that the results are statistically significant across all five folds, which is true for all metrics. It is important to note here that, due to the limited dataset available for this thesis, a statistical test with n = 5 has to be treated with caution as the sample size is very small.

To complement the results, the models from research questions 2 and 3 (presegmented vessels and foundation model as feature extractors, respectively) were also compared against a random model using the Wilcoxon test. The model trained with presegmented vessels as input also shows statistically significant results across all metrics, also always scoring the maximum possible Wilcoxon score of 15 and p-values across the board of 0.031. For the foundation model as a feature extractor, the results are statistically significant for the Area Under the Curve (AUC), Average Precision (AP) and Accuracy. However, for the metrics Precision, Recall and F1-score, the

Table 9.1: Statistical comparison between a random model and benchmark on the test set with a Wilcoxon signed-rank test.

Metric	Wilcoxon	p-value	Signifacant?
AUC	15	0.031	True
Average Precision	15	0.031	True
Accuracy	15	0.031	True
Precision	15	0.031	True
Recall	15	0.031	True
F1-score	15	0.031	True

results are not statistically significant. The results for these three metrics are shown in Table 9.2. Precision achieves a result close to the required Wilcoxon score of 15, meaning that one of the folds does not perform better than expected by chance. Recall does not achieve good scores in the test. On average, recall scores low over the five folds (avg: 0.694, std=0.079), compared to the random model. In conclusion, arterial abnormalities can be detected using convolution-based models, but the architecture impacts the statistical relevance of the results.

Table 9.2: Statistical comparison between a random model and the foundation model as feature extractor on the test set with a Wilcoxon signed-rank test.

Metric	Wilcoxon	p-value	Signifacant?
Precision	14	0.062	False
Recall	7	0.312	False
F1-score	12	0.156	False

Despite the promising results, slight overfitting is consistently observed, as evidenced by near-perfect validation performance, particularly in the recall and AP metrics. This effect likely results from a combination of factors, including the limited dataset size and class imbalance, as well as the possibility that the models converge towards class-conditional artefacts that are unrelated to the underlying disease. Such behaviour is well documented in medical imaging research, where models learn shortcuts and rely on biases that fail to generalise across clinical settings (Zech et al., 2018). Consequently, performance that is better than would be expected by chance does not automatically imply that the learned decision boundary is clinically meaningful or interpretable.

9.2 Segmentation as a means to enforce inductive bias

Presegmenting vessel structures before classification provides meaningful inductive bias by restricting the model's field-of-view to regions of interest, thereby reducing the risk of shortcut learning. Across all folds, this method shows a consistent increase in performance metrics. One plausible explanation is that segmentation removes contextual confounders, forcing the classifier to focus solely on vessel morphology, which is the clinically relevant feature for FMD diagnosis.

For this research question the results from the model trained with presegemented vessels was

compared over the five folds to the results obtained with the model from RQ1 with the Wilcoxon test. Eventhough the results are better on average for every metric, the Wilcoxon score is not high enough to make the statemement that the results are statistically significantly better. Only Precision can be considered statistically significant better (see Table 9.3) But also here the Wilcoxon test was done with n=5, meaning the very small sample size has a big impact on the final results of a statistical relevance test. The model performance is on average better over all five folds and therefore promising. The statistical irrelevance does not only stem from bad model results, but also from the small dataset at hand.

Table 9.3: Statistical comparison between a benchmark model and model trained with presegmented vessels on the test set (5-fold CV) with Wilcoxon.

Metric	Wilcoxon	p-value	Signifacant?
AUC	9	0.125	False
Average Precision	14	0.062	False
Accuracy	13.5	0.094	False
Precision	15	0.031	True
Recall	9	0.406	False
F1-score	8	0.188	False

Also, the lack of statistical significance implies that segmentation alone does not fully resolve the complexity of abnormality detection. Segmentation masks are binary. They capture vessel geometry but discard attenuation variation and wall-texture cues that might correlate with disease presentation, which are encoded in CTA scans with Hounsfield unit. Adding them to the segmentation might improve the ability to detect the arterial abnormalities. Additionally it is worth mentioning that results are dependent on the quality of segmentation results from the preprocessing step. As can be seen in Figure 5.3 VesselFM does segment vessels well, but is not a pure vessel segmentation tool. VesselFM segments tubular structures within the CTA scan, which results in the rib cage being included in most of the segmentations. Removing these unnecessary tubular structures from the input images might improve the results further.

The present results therefore highlight an important trade-off: Presegmentation enforces anatomical focus, but may simultaneously eliminate subtle cues present in raw imaging, if the binary mask is used as model input.

9.3 Foundation Model as Feature Extractor

The evaluation of VesselFM as a feature extractor yields mixed outcomes. Despite the fact that VesselFM has been trained on a range of vascular datasets and incorporates rich morphological information, the extracted features exhibits suboptimal performance in comparison to those obtained from training a CNN from the beginning.

9.3.1 The search for the best configuration

It is demonstrated that the configuration of the head and the final proposed model have a significant impact on the results of the model. In order to perform a more extensive analysis of the results, it is necessary to consider all possible combinations of the three pooling layers, the extraction layer and the linearity of the head. This approach was adopted in order to facilitate a more

comprehensive understanding of the results. As Table 9.4 illustrates, the AUC and AP scores for all 12 combinations are presented.

Table 9.4: AUC scores of all possible combinations for RQ 3.1

Configuration	AUC (Mean ± Std (Rank))	Average Precision (Mean ± Std (Rank))
linear-BN-avg	0.733 ± 0.048 (9)	0.719 ± 0.061 (5)
linear-BN-max	0.681 ± 0.099 (10)	0.681 ± 0.118 (6)
linear-BN-both	0.779 ± 0.049 (7)	0.728 ± 0.055 (4)
linear-DS-avg	0.595 ± 0.141 (12)	0.539 ± 0.122 (11)
linear-DS-max	0.752 ± 0.037 (8)	0.610 ± 0.042 (10)
linear-DS-both	0.851 ± 0.016 (1)	0.679 ± 0.011 (7)
nonlinear-BN-avg	0.800 ± 0.033 (4)	0.783 ± 0.027 (1)
nonlinear-BN-max	0.812 ± 0.062 (3)	0.733 ± 0.054 (3)
nonlinear-BN-both	0.820 ± 0.041 (2)	0.782 ± 0.037 (2)
nonlinear-DS-avg	0.612 ± 0.094 (11)	0.472 ± 0.083 (12)
nonlinear-DS-max	0.797 ± 0.073 (5)	0.651 ± 0.131 (9)
nonlinear-DS-both	0.794 ± 0.062 (6)	0.670 ± 0.068 (8)

A thorough examination of the results presented in Table 9.4 reveals that the configuration demonstrating the most favourable outcomes in this study and in response to RQ 3.1 may in fact be an outlier. The optimal configuration, as determined by the AUC metric, is reported to be a linear head that extracts features from the deepest downsampling layer (Downsample 4). This configuration combines both pooling methods.

Combining both pooling methods appears advantageous in all configurations. However, utilising both a linear head and extracting features from the deepest downsampling layer is not advantageous in isolation. In this thesis, nonlinear heads have been shown to outperform linear ones, and feature extraction from the bottleneck layer has been demonstrated to exceed the performance of the deepest downsampling layer (see Table 9.4). Furthermore, it should be noted that the configuration with the highest AUC ranking only achieves seventh place according to AP. The configurations achieving the highest ranking according to AUC are positioned second, third and first according to AP, respectively. These configurations all have one thing in common: they have a nonlinear head and extract features from the bottleneck layer. This combination appears to be the most robust according to both metrics. These inconsistent results further highlight the challenges of using foundation models as feature extractors, indicating that an extensive hyperparameter search should be conducted if foundation models are to be used for this purpose.

9.3.2 Comparison to training convolutional models from scratch

This thesis demonstrates that using VesselFM as a feature extractor in combination with training a classifier head does not yield superior performance compared to training a 3D CNN from scratch. A comparative analysis using the same dataset reveals that the foundation model approach generates comparable results to a 3D CNN trained from scratch. However, a comparison with a model trained from scratch using presegmented vessels as input shows that the foundation model approach produces inferior results. This finding may be regarded as unexpected given that vessel segmentation was conducted using the same foundation model (VesselFM).

A multitude of factors may be responsible for such outcomes. Firstly, it is important to note

that VesselFM is optimised for segmentation, not abnormality recognition. It is possible that this discrepancy may serve to limit the relevance of the learned features with regard to the use case under investigation in this thesis (Zhou et al., 2024). Furthermore, the extraction of deeper features is associated with a greater tendency for the globality of features, resulting in a loss of spatial resolution in favour of increased abstraction (Li et al., 2022). In the context of arterial abnormalities, the loss of locality is detrimental due to the fact that such abnormalities frequently span only a few voxels. Although abstract vessel embeddings are produced by deeper layers, the removal of fine-grained morphological signatures may provide a rationale for the superior performance of features derived from a shallower layer in comparison to those derived from bottleneck features. An examination of the approach employed in this instance reveals the potential for enhancement. The approach that is employed in this work utilises a single layer for the extraction of features. Furthermore, the employment of the encoder as a feature extractor disregards the crucial function of the decoder in U-Net architectures. The skip connections and decoder part are crucial components of the architecture that are not utilised in this approach, resulting in significant information loss.

A thorough examination of the outcomes and the previously mentioned factors suggests that employing foundation models as feature extractors does not guarantee success. Nonetheless, the results are encouraging. With modifications to the approach, for instance through fine-tuning of the foundation model weights on the designated task or by leveraging additional information from the foundation model architecture, enhanced results may be attainable.

9.4 Dataset-related limitations

In this section, the limitations of the dataset utilised in this thesis are examined. The enhancement of the dataset, and consequently the underlying conditions for training models, has the potential to yield further improvements in results.

The dataset this thesis uses introduces several significant constraints that affect both the reliability and the clinical interpretability of the results. The most fundamental limitation is the absence of lesion-level annotations. In the absence of explicit labelling indicating the location and type of arterial abnormalities, the models are only able to learn from global labels and might suffer from shortcut learning (Geirhos et al., 2020). Consequently, the models are not necessarily trained towards the pathology itself and may instead rely on unrelated visual patterns or proxy features that correlate with disease status. This inadequate form of supervision has the effect of restricting interpretability and making it difficult to ensure that the model truly learns pathological cues rather than artefacts.

The issue is further aggravated by the relatively limited sample size of 126 cases, which reduces statistical power. In such settings, models have a propensity to overfit to small subsets of patients, rendering performance contingent on the data split. Furthermore, the dataset is characterised by significant imbalance, with the prevalence of healthy cases exceeding that of abnormal cases. This imbalance is shown to encourage classifiers to favour the majority class, potentially leading to biased decision boundaries that prioritise correct identification of healthy vessels while failing to generalise towards rare abnormal morphologies. Collectively, these limitations underscore the significance of larger, balanced, and lesion-level annotated datasets obtained under harmonised imaging conditions to facilitate clinically meaningful and interpretable vessel abnormality detection in future research endeavours.

9.5 Future research directions

It is recommended that further research be conducted on the topic of arterial abnormality detection. It has been demonstrated that these abnormalities have the capacity to exert a detrimental effect on the health of an individual. Furthermore, it is pertinent to consider whether the identification of arterial abnormalities could facilitate the diagnosis of rare diseases, such as Fibromuscular Dysplasia(FMD).

A pivotal step in achieving this objective entails the development of datasets with lesion-level annotations, encompassing both abnormality types and their precise anatomical locations. The integration of these labels into the training of models would facilitate the implementation of lesion-level detection, segmentation, multi-task learning, or weakly supervised localisation methods, thereby enhancing interpretability and mitigating the risk of models relying on proxy features.

Beyond the incorporation of learned features, the integration of rule-based measurements, including vessel diameter profiles, local curvature, and topology, has the potential to combine the strengths of human vascular logic with deep learning approaches. The utilisation of hybrid pipelines holds the potential to leverage explicit geometric descriptors to complement learned representations, particularly in the context of small or subtle lesions.

Another approach that could be done is extending the presegmentation approach, by not training a standard 3D CNN, but rather try and use an architecture that is optimised for sparse data.

Extending the Region Of Interest (ROI) to other parts of the body, such as the neck, could improve the model's results further. The different anatomical structures of other parts of the body can therefore help the model to be more generalisable.

Furthermore, subsequent research should investigate data augmentation strategies, as well as more informed choices regarding feature extraction. It is hypothesised that experiments using deeper or alternative network layers, or more complex architectures, may allow models to exploit richer structural and intensity cues. Finally, hybrid segmentation and classification architectures trained in conjunction have the capacity to preserve detailed geometric priors while concurrently supporting disease classification. This may help to bridge the gap between anatomy-focused vessel models and clinically meaningful pathology prediction.

Chapter 10

Conclusion

This thesis examined deep learning methods for detecting arterial abnormalities associated with Fibromuscular Dysplasia (FMD) in CTA volumes. The subtle morphological expression of FMD-related lesions, such as multifocal stenosis, aneurysm, dissection, and tortuosity, together with the limited availability of annotated datasets, poses significant challenges for developing reliable automated detection systems. Because FMD is difficult to diagnose and often underdetected in clinical practice, a Computer-Aided Diagnosis (CAD) system capable of assisting in arterial abnormality detection may meaningfully support clinicians in identifying this rare disease earlier and more consistently.

To address these challenges, three progressively more informed modelling pipelines were explored: a baseline 3D Convolutional Neural Network (CNN) trained from scratch, a 3D CNN that uses presegmented vessel masks as input, and a feature-extraction approach leveraging the vascular foundation model VesselFM. Furthermore, the analysis was restricted to a focused Region Of Interest (ROI) around the kidneys to ephasise the renal arteries that are known to have a high prevelance of FMD related arterial abnormalities.

The baseline model demonstrated that CNN-based architectures are capable of detecting arterial abnormalities. However, the results showed slight overfitting, reflecting the difficulty of generalising from limited training data. Introducing anatomical inductive bias by presegmenting vessels improved performance, as it directed the model's attention to vessel morphology and reduced reliance on shortcut cues. Employing VesselFM as a frozen feature extractor enabled the model to detect abnormal patterns as well, but its performance remained comparable to the baseline and lower than that of the segmentation-enhanced approach, indicating that directly supplying vessel-specific information was more beneficial than relying solely on pretrained vascular features.

Future research should prioritise the development of larger and more diverse datasets, particularly with lesion-level annotations that specify abnormality types and precise anatomical locations. Integrating rule-based vascular descriptors such as diameter profiles, curvature, and topology may further enhance interpretability and model robustness. Additionally, extending the region of interest to other anatomical areas, exploring improved feature extraction strategies, and designing hybrid segmentation–classification models could support more generalisable and anatomically grounded predictions.

Overall, this thesis provides evidence that deep learning methods, especially those integrating anatomical priors, offer a promising direction for CAD in vascular imaging. Such systems represent an important step toward supporting healthcare professionals in detecting and diagnosing FMD more accurately and efficiently, potentially improving patient care and reducing the number of missed cases.

List of Figures

2.1 Arterial abnormalities associated with FMD 7

4.1 Comparison of Pooling Methods 17
4.2 Activation Functions 19
4.3 U-Net and 3D U-Net 20
4.4 Mednet 3D CNN architecture 22
4.5 Mednet 3D CNN Convolutional Blocks 23
4.6 VesselFM Encoder Architecture 24

5.1 Abnormality Distribution after preprocessing 30
5.2 Kidney Segmentation 32
5.3 Vessel Segmentation 33
5.4 Abnormality Distribution after preprocessing 35

7.1 Linear and nonlinear classification head 45

8.1 Final model architecture 52

List of Tables

5.1 Gender distribution 29
5.2 Age distribution 30
5.3 Comparison of Abnormalities 31

7.1 Random baseline calculated based on the class distribution of the test set 43

8.1 Arterial abnormality detection: Train/Val/Test results for five folds for the baseline model 47
8.2 Comparison between a random model and benchmark. The results of the training with 5 different folds are averaged. 48
8.3 Arterial abnormality detection: Train/Val/Test results for five folds for a 3D CNN trained with presegmented vessel masks as input. 48
8.4 Comparison between benchmark and presegmented mask models. The results of the training with 5 different folds are averaged. 49
8.5 Comparison of pooling methods for classification head 49
8.6 Comparison of extraction layer 49
8.7 Comparison of linear and nonlinear head 50
8.8 Comparison between foundation model as feature extractor and benchmark. The results of the training with 5 different folds are averaged. 50
8.9 Comparison between foundation model as feature extractor and a model trained with presegmented vessels. The results of the training with 5 different folds are averaged. 51

9.1 Statistical comparison between a random model and benchmark on the test set with a Wilcoxon signed-rank test. 54
9.2 Statistical comparison between a random model and the foundation model as feature extractor on the test set with a Wilcoxon signed-rank test. 54
9.3 Statistical comparison between a benchmark model and model trained with presegmented vessels on the test set (5-fold CV) with Wilcoxon. 55
9.4 AUC scores of all possible combinations for RQ 3.1 56

Bibliography

Affonso, C., Rossi, A. L. D., Vieira, F. H. A., and de Leon Ferreira de Carvalho, A. C. P. (2017). Deep learning for biological image classification. *Expert Systems with Applications*, 85:114–122.

Azad, R., Aghdam, E. K., Rauland, A., Jia, Y., Avval, A. H., Bozorgpour, A., Karimijafarbigloo, S., Cohen, J. P., Adeli, E., and Merhof, D. (2024). Medical image segmentation review: The success of U-Net. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10076–10095.

Ba, J. L., Kiros, J. R., and Hinton, G. E. (2016). Layer normalization. *arXiv preprint arXiv:1607.06450*.

Bax, M., Romanov, V., Junday, K., Giannoulatou, E., Martinac, B., Kovacic, J. C., Liu, R., Iismaa, S. E., and Graham, R. M. (2022). Arterial dissections: Common features and new perspectives. *Frontiers in Cardiovascular Medicine*, 9.

Baxter, J. (2000). A model of inductive bias learning. *Journal of Artificial Intelligence Research*, 12:149–198.

Begelman, S. M. and Olin, J. W. (2000). Fibromuscular dysplasia. *Current Opinion in Rheumatology*, 12(1).

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159.

Brutti, F., Fantazzini, A., Finotello, A., Müller, L. O., Auricchio, F., Pane, B., Spinella, G., and Conti, M. (2022). Deep learning to automatically segment and analyze abdominal aortic aneurysm from computed tomography angiography. *Cardiovascular Engineering and Technology*, 13(4):535–547.

Buckland, M. and Gey, F. (1994). The relationship between recall and precision. *Journal of the American Society for Information Science*, 45(1):12–19.

Chen, H., Wang, Y., Xu, C., Shi, B., Xu, C., Tian, Q., and Xu, C. (2020). Addernet: Do we really need multiplications in deep learning? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chen, Z., Xiao, C., Liu, Y., Hassan, H., Li, D., Liu, J., Li, H., Xie, W., Zhong, W., and Huang, B. (2025). Comprehensive 3D analysis of the renal system and stones: Segmenting and registering non-contrast and contrast computed tomography images. *Information Systems Frontiers*, 27:97–111. Published online: 16 March 2024.

Chicco, D. and Jurman, G. (2020). The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21(6). Open Access.

Chow, L. C. and Rubin, G. D. (2002). Ct angiography of the arterial system. *Radiologic Clinics*, 40(4):729–749.

Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., and Ronneberger, O. (2016). 3d U-Net: Learning dense volumetric segmentation from sparse annotation. In Ourselin, S., Joskowicz, L., Sabuncu, M. R., Unal, G., and Wells, W., editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, pages 424–432, Cham. Springer International Publishing.

Ciurică, S., Lopez-Sublet, M., Loeys, B. L., Radhouani, I., Natarajan, N., Vikkula, M., Maas, A. H., Adlam, D., and Persu, A. (2019). Arterial tortuosity. *Hypertension*, 73(5):951–960.

Cragg, A. H., Smith, T. P., Thompson, B. H., Maroney, T. P., Stanson, A. W., Shaw, G. T., Hunter, D. W., and Cochran, S. T. (1989). Incidental fibromuscular dysplasia in potential renal donors: long-term clinical follow-up. *Radiology*, 172(1).

Davis, J. and Goadrich, M. (2006). The relationship between precision-recall and ROC curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240.

der Niepen, P. V., Robberechts, T., Devos, H., van Tussenbroek, F., Januszewicz, A., and Persu, A. (2021). Fibromuscular dysplasia: its various phenotypes in everyday practice in 2021. *Polish Heart Journal (Kardiologia Polska)*, 79(7-8):733 – 744.

Doi, T., Kataoka, Y., Noguchi, T., Shibata, T., Nakashima, T., Kawakami, S., Nakao, K., Fujino, M., Nagai, T., Kanaya, T., Tahara, Y., Asaumi, Y., Tsuda, E., Nakai, M., Nishimura, K., Anzai, T., Kusano, K., Shimokawa, H., Goto, Y., and Yasuda, S. (2017). Coronary artery ectasia predicts future cardiac events in patients with acute myocardial infarction. *Arteriosclerosis, Thrombosis, and Vascular Biology*, 37(12):2350–2355.

Dorjsembe, Z., Pao, H.-K., Odonchimed, S., and Xiao, F. (2024). Conditional diffusion models for semantic 3d brain mri synthesis. *IEEE Journal of Biomedical and Health Informatics*, 28(7):4084–4093. Epub 2024 Jul 2.

Doğan, Y. (2023). Which pooling method is better: Max, Avg, or Concat (Max, Avg). *Communications Faculty of Sciences University of Ankara Series A2-A3 Physical Sciences and Engineering*, 66:95–117.

Dubey, S. R., Singh, S. K., and Chaudhuri, B. B. (2022). Activation functions in deep learning: A comprehensive survey and benchmark. *Neurocomputing*, 503:92–108.

Ebski, S., Arpit, D., Ballas, N., Verma, V., Che, T., and Bengio, Y. (2018). Residual connections encourage iterative inference. pages 1–14. International Conference on Learning Representations, ICLR ; Conference date: 30-04-2018 Through 03-05-2018.

Faghani, S., Nicholas, R. G., Patel, S., Baffour, F. I., Moassefi, M., Rouzrokh, P., Khosravi, B., Powell, G. M., Leng, S., Glazebrook, K. N., Erickson, B. J., and Tiegs-Heiden, C. A. (2024). Development of a deep learning model for the automated detection of green pixels indicative of gout on dual energy CT scan. *Research in Diagnostic and Interventional Imaging*, 9:100044.

Fantazzini, A., Esposito, M., Finotello, A., Auricchio, F., Pane, B., Basso, C., Spinella, G., and Conti, M. (2020). 3d automatic segmentation of aortic computed tomography angiography combining Multi-View 2D convolutional neural networks. *Cardiovascular Engineering and Technology*, 11(5):576–586.

Fu, F., Shan, Y., Yang, G., Zheng, C., Zhang, M., Rong, D., Wang, X., and Lu, J. (2023). Deep learning for head and neck CT angiography: Stenosis and plaque classification. *Radiology*, 307(3):e220996. PMID: 36880944.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F. A. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

Gornik, H. L., Persu, A., Adlam, D., Aparicio, L. S., Azizi, M., Boulanger, M., Bruno, R. M., de Leeuw, P., Fendrikova-Mahlay, N., Froehlich, J., Ganesh, S. K., Gray, B. H., Jamison, C., Januszewicz, A., Jeunemaitre, X., Kadian-Dodov, D., Kim, E. S., Kovacic, J. C., Mace, P., Morganti, A., Sharma, A., Southerland, A. M., Touzé, E., van der Niepen, P., Wang, J., Weinberg, I., Wilson, S., Olin, J. W., and Plouin, P.-F. (2019). First international consensus on the diagnosis and management of fibromuscular dysplasia. *Vascular Medicine*, 24(2):164–189.

Goyal, A. and Bengio, Y. (2022). Inductive biases for deep learning of higher-level cognition. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 478(2266):20210068.

Graves, M. J. (2024). Shadows and signals: a brief history of medical imaging. In *Journal of Physics: Conference Series*, volume 2877, page 012113, Oxford, United Kingdom. IOP Publishing. HAPP Centre: 10th Anniversary Commemorative Volume, 3rd June 2024.

Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G., Cai, J., and Chen, T. (2018). Recent advances in convolutional neural networks. *Pattern Recognition*, 77:354–377.

Güçük, A. and Uyetürk, U. (2014). Usefulness of Hounsfield unit and density in the assessment and treatment of urinary stones. *World Journal of Nephrology*, 3(4):282–286.

Güler, O. A., Günther, M., and Anjos, A. (2024). Refining tuberculosis detection in CXR imaging: addressing bias in deep neural networks via interpretability. Published in EUVIP.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Hendrycks, D. and Gimpel, K. (2023). Gaussian error linear units (GELUs). *arXiv preprint arXiv:1606.08415*.

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*.

Hoffmann, U., Ferencik, M., Cury, R. C., and Pena, A. J. (2006). Coronary CT angiography. *Journal of Nuclear Medicine*, 47(5):797–806.

Howard E. LeWine, MD (2023). Hemorrhagic stroke. *Harvard Health*. Reviewed by Howard E. LeWine, MD.

Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*.

Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J., and Maier-Hein, K. H. (2021). nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211.

Kim, M., Yun, J., Cho, Y., Shin, K., Jang, R., jin Bae, H., and Kim, N. (2019). Deep learning in medical imaging. *Neurospine*, 16(4):657–668. Correction published in Neurospine. 2020;17(2):471.

Kingma, D. P. and Ba, J. (2017). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollar, P., and Girshick, R. (2023). Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4015–4026.

Kumamaru, K. K., Hoppel, B. E., Mather, R. T., and Rybicki, F. J. (2010). Ct angiography: current technology and clinical use. *Radiologic Clinics of North America*, 48(2):213–vii.

La Barbera, G., Rouet, L., Boussaid, H., Lubet, A., Kassir, R., Sarnacki, S., Gori, P., and Bloch, I. (2023). Tubular structures segmentation of pediatric abdominal-visceral ceCT images with renal tumors: Assessment, comparison and improvement. *Medical Image Analysis*, 90:102986.

Laibacher, T. and Anjos, A. (2019). On the evaluation and real-world usage scenarios of deep vessel segmentation for retinography.

Leadbetter, W. and Burkland, C. E. (1938). Hypertension in unilateral renal disease. *The Journal of Urology*, 39(5):611–626.

Leckie, A., Tao, M. J., Narayanasamy, S., Khalili, K., Schieda, N., and Krishna, S. (2021). The renal vasculature: What the radiologist needs to know. *RadioGraphics*, 41(5):1531–1548.

Lesage, D., Angelini, E. D., Bloch, I., and Funka-Lea, G. (2009). A review of 3D vessel lumen segmentation techniques: Models, features and extraction schemes. *Medical Image Analysis*, 13(6):819–845. Includes Special Section on Computational Biomechanics for Medicine.

Li, Z., Liu, F., Yang, W., Peng, S., and Zhou, J. (2022). A survey of convolutional neural networks: Analysis, applications, and prospects. *IEEE Transactions on Neural Networks and Learning Systems*, 33(12):6999–7019.

Lowekamp, B. C., Chen, D. T., Ibanez, L., and Blezek, D. (2013). The design of SimpleITK. *Frontiers in Neuroinformatics*, Volume 7 - 2013.

Ma, J., Yang, Z., Kim, S., Chen, B., Baharoon, M., Fallahpour, A., Asakereh, R., Lyu, H., and Wang, B. (2025). Medsam2: Segment anything in 3d medical images and videos. *arXiv preprint arXiv:2504.03600*.

Mangukia, C. V. (2012). Coronary artery fistula. *The Annals of Thoracic Surgery*, 93(6):2084–2092.

Mavrogeni, S. (2010). Coronary artery ectasia: From diagnosis to treatment. *Hellenic Journal of Cardiology*, 51:158–163.

Nair, V. and Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814.

Nellore, S. B. (2015). Various performance measures in binary classification: An overview of ROC study. *International Journal of Innovative Science, Engineering and Technology (IJISET)*, 2(9):—.

Olin, J. W. and Sealove, B. A. (2011). Diagnosis, management, and future developments of fibromuscular dysplasia. *Journal of Vascular Surgery*, 53(3):826–836.e1.

O'Shea, K. and Nash, R. (2015). An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*.

Persu, A., Dobrowolski, P., Gornik, H. L., Olin, J. W., Adlam, D., Azizi, M., Boutouyrie, P., Bruno, R. M., Boulanger, M., Demoulin, J.-B., Ganesh, S. K., J. Guzik, T., Januszewicz, M., Kovacic, J. C., Kruk, M., de Leeuw, P., Loeys, B. L., Pappaccogli, M., Perik, M. H. A. M., Touzé, E., Van der Niepen, P., Van Twist, D. J. L., Warchoł-Celińska, E., Prejbisz, A., and Januszewicz, A. (2021). Current progress in clinical, molecular, and genetic aspects of adult fibromuscular dysplasia. *Cardiovascular Research*, 118(1):65–83.

Plouin, P.-F., Perdu, J., La Batide-Alanore, A., Boutouyrie, P., Gimenez-Roqueplo, A.-P., and Jeunemaitre, X. (2007). Fibromuscular dysplasia. *Orphanet Journal of Rare Diseases*, 2(1):28.

Provost, F. and Fawcett, T. (1997). Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions. In *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining*, pages 43–48.

Razzak, M. I., Naz, S., and Zaib, A. (2018). *Deep Learning for Medical Image Processing: Overview, Challenges and the Future*, pages 323–350. Springer International Publishing.

Robb, R. A. (2006). Biomedical imaging: Past, present and predictions. *Medical Imaging Technology*, 24(1):25.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In Navab, N., Hornegger, J., Wells, W. M., and Frangi, A. F., editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham. Springer International Publishing.

Sakalihasan, N., Limet, R., and Defawe, O. (2005). Abdominal aortic aneurysm. *The Lancet*, 365(9470):1577–1589.

Siddique, N., Paheding, S., Elkin, C. P., and Devabhaktuni, V. (2021). U-net and its variants for medical image segmentation: A review of theory and applications. *IEEE Access*, 9:82031–82057.

Solan, M. (2024). An inside look at aortic stenosis. *Harvard Health Publishing*. Reviewed by Howard E. LeWine, MD.

Spinella, G., Fantazzini, A., Finotello, A., Vincenzi, E., Boschetti, G. A., Brutti, F., Magliocco, M., Pane, B., Basso, C., and Conti, M. (2023). Artificial intelligence application to screen abdominal aortic aneurysm using computed tomography angiography. *Journal of Digital Imaging*, 36(5):2125–2137.

Varennes, L., Tahon, F., Kastler, A., Grand, S., Thony, F., Baguet, J. P., Detante, O., Touzé, E., and Krainik, A. (2015). Fibromuscular dysplasia: what the radiologist should know: a pictorial review. *Insights into Imaging*, 6(3):295–307.

Wang, H., Guo, S., Ye, J., Deng, Z., Cheng, J., Li, T., Chen, J., Su, Y., Huang, Z., Shen, Y., Fu, B., Zhang, S., He, J., and Qiao, Y. (2024). Sam-med3d: Towards general-purpose segmentation models for volumetric medical images.

Wang, L. (2017). Heterogeneous data and big data analytics. *Automatic Control and Information Sciences*, 3(1):8–15.

Wasserthal, J., Breit, H.-C., Meyer, M. T., Pradella, M., Hinck, D., Sauter, A. W., Heye, T., Boll, D. T., Cyriac, J., Yang, S., Bach, M., and Segeroth, M. (2023). Totalsegmentator: Robust segmentation of 104 anatomic structures in ct images. *Radiology: Artificial Intelligence*, 5(5):e230024.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83.

Wittmann, B., Wattenberg, Y., Amiranashvili, T., Shit, S., and Menze, B. (2025). vesselfm: A foundation model for universal 3d blood vessel segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20874–20884.

Wornow, M., Xu, Y., Thapa, R., Patel, B., Steinberg, E., Fleming, S., Pfeffer, M. A., Fries, J., and Shah, N. H. (2023). The shaky foundations of large language models and foundation models for electronic health records. *npj Digital Medicine*, 6(1):135.

Wu, J. (2017). Introduction to convolutional neural networks. *National Key Lab for Novel Software Technology. Nanjing University. China*. LAMDA Group.

Wu, S., Chen, S., Xie, C., and Huang, X. (2023). One-pixel shortcut: on the learning preference of deep neural networks. *arXiv preprint arXiv:2205.12141*.

Yamashita, R., Nishio, M., Do, R. K. G., and Togashi, K. (2018). Convolutional neural networks: an overview and application in radiology. *Insights into Imaging*, 9(4):611–629.

Yeh, S.-T. (2002). Using trapezoidal rule for the area under a curve calculation. In *Proceedings of the 15th Annual NorthEast SAS Users Group Conference (NESUG 15)*, Collegeville, PA, USA. GlaxoSmithKline. Poster presentation.

Yuan, Y. (2023). On the power of foundation models. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J., editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 40519–40530. PMLR.

Yuan, Y., Su, W., and Zhu, M. (2015). Threshold-free measures for assessing the performance of medical screening tests. *Frontiers in Public Health*, 3:57.

Yue-Hei Ng, J., Yang, F., and Davis, L. S. (2015). Exploiting local features from deep networks for image retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., and Oermann, E. K. (2018). Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLOS Medicine*, 15(11):e1002683.

Zeiler, M. D., Krishnan, D., Taylor, G. W., and Fergus, R. (2010). Deconvolutional networks. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2528–2535.

Zhang, M., Ye, Z., Yuan, E., Lv, X., Zhang, Y., Tan, Y., Xia, C., Tang, J., Huang, J., and Li, Z. (2024). Imaging-based deep learning in kidney diseases: recent progress and future prospects. *Insights into Imaging*, 15(1):50.

Zhang, S. and Metaxas, D. (2024). On the challenges and perspectives of foundation models for medical image analysis. *Medical Image Analysis*, 91:102996.

Zhou, Y., Du, S., Li, H., Yao, J., Zhang, Y., and Wang, Y. (2024). Reprogramming distillation for medical foundation models. In Linguraru, M. G., Dou, Q., Feragen, A., Giannarou, S., Glocker, B., Lekadir, K., and Schnabel, J. A., editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, pages 533–543. Springer Nature Switzerland.

Zhu, M. (2004). Recall, precision and average precision. Technical report, Department of Statistics & Actuarial Science, University of Waterloo.

Zohranyan, V., Navasardyan, V., Navasardyan, H., Borggrefe, J., and Navasardyan, S. (2024). Dr-sam: An end-to-end framework for vascular segmentation diameter estimation and anomaly detection on angiography images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 5113–5121.